Neural Network Theory

Philipp Christian Petersen

University of Vienna

April 18, 2022

Contents

1 Introduction

In these notes, we study a mathematical structure called *neural networks*. These objects have recently received much attention and have become a central concept in modern machine learning. Historically, however, they were motivated by the functionality of the human brain. Indeed, the first neural network was devised by McCulloch and Pitts [18] in an attempt to model a biological neuron.

A *McCulloch and Pitts neuron* is a function of the form

$$\mathbb{R}^d \ni x \mapsto \mathbb{1}_{\mathbb{R}^+}\left(\sum_{i=1}^{d} w_i x_i - \theta\right),$$

where $d \in \mathbb{N}$, $\mathbb{1}_{\mathbb{R}^+} : \mathbb{R} \to \mathbb{R}$, with $\mathbb{1}_{\mathbb{R}^+}(x) = 0$ for $x < 0$ and $\mathbb{1}_{\mathbb{R}^+}(x) = 1$ else, and $w_i, \theta \in \mathbb{R}$ for $i = 1, \dots d$. The function $\mathbb{1}_{\mathbb{R}^+}$ is a so-called *activation function*, θ is called a *threshold*, and w_i are *weights*. The McCulloch and Pitts neuron, receives d input signals. If their combined weighted strength exceeds θ, then the neuron *fires*, i.e., returns 1. Otherwise the neuron remains inactive.

A network of neurons can be constructed by linking multiple neurons together in the sense that the output of one neuron forms an input to another. A simple model for such a network is the *multilayer perceptron*[*] as introduced by Rosenblatt [28].

Definition 1.1. *Let $d, L \in \mathbb{N}$, $L \geq 2$ and $\varrho : \mathbb{R} \to \mathbb{R}$. Then a* multilayer perceptron (MLP) with d-dimensional input, L layers, and activation function ϱ *is a function F that can be written as*

$$x \mapsto F(x) \coloneqq T_L\left(\varrho\left(T_{L-1}\left(\dots \varrho\left(T_1\left(x\right)\right)\dots\right)\right)\right), \tag{1.1}$$

where $T_\ell(x) = A_\ell x + b_\ell$, and $(A_\ell)_{\ell=1}^L \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, $b_\ell \in \mathbb{R}^{N_\ell}$ for $N_\ell \in \mathbb{N}$, $N_0 = d$, and $\ell = 1, \dots, L$. Here $\varrho : \mathbb{R} \to \mathbb{R}$ is applied coordinate-wise.

The neurons in the MLP correspond again, to the applications of $\varrho : \mathbb{R} \to \mathbb{R}$ even though, in contrast to the McCulloch and Pitts neuron, we now allow arbitrary ϱ. In Figure 1.1, we visualise a MLP. We should notice that the MLP does not allow arbitrary connections between neurons, but only between those, that are in adjacent layers, and only from lower layers to higher layers.

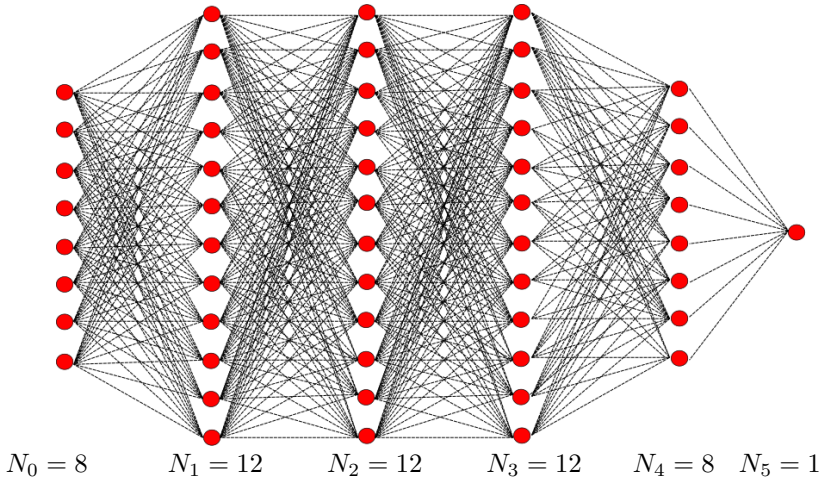

Figure 1.1: Illustration of a multi-layer perceptron with 5 layers. The red dots correspond to the neurons.

While the MLP or variations thereof, are probably the most widely used type of neural network in practice, they are very different from their biological motivation. Connections only between layers and arbitrary

[*] We will later introduce a notion of neural networks, that differs slightly from that of a multilayer perceptron.

activation functions make for an efficient numerical scheme but are not a good representation of the biological reality.

Nowadays, the field of neural network theory draws most of its motivation from the fact that deep neural networks are applied in a technique called *deep learning* [12]. In deep learning, one is concerned with the algorithmic identification of the most suitable deep neural network for a specific application. It is, therefore, reasonable to search for purely mathematical arguments why and under which conditions a MLP is an adequate architecture in practice instead of taking the motivation from the fact that biological neural networks perform well.

In this note, we will study deep neural networks with a very narrow focus. We will exclude all algorithmic aspects of deep learning and concentrate fully on a functional analytical and well-founded framework. One the one hand, following this focussed approach, it must be clear that we will not be able to provide a comprehensive answer to how deep learning methods perform in practice. For an attempt to explain the full pipeline, see [4]. On the other hand, we will see that this restricted focus allows us to make rigorous statements which do provide explanations and intuition as to why certain neural network architectures are preferable over others.

Concretely, we will identify many mathematical properties of sets of MLPs which explain, to some extent, practically observed phenomena in machine learning. For example, we will see explanations of why deep neural networks are, in some sense, superior to shallow neural networks or why the neural network architecture can efficiently reproduce high dimensional functions when most classical approximation schemes cannot.

2 Classical approximation results by neural networks

The very first question that we would naturally ask ourselves is which functions we can express as a MLP. Given that the activation function is fixed, it is conceivable that the set of functions that can be represented or approximated could be quite small.

Example 2.1.
- *For linear activation functions $\varrho(x) = ax$, $a \in \mathbb{R}$ it is clear that every MLP with this activation function is an affine linear map.*
- *More generally, if ϱ is a polynomial of degree $k \in \mathbb{N}$, then every MLP with L layers is a polynomial of degree at most k^{L-1}.*[*]

Example 2.1 demonstrates that under some assumptions on the activation function not every function can be represented and not even approximated by MLPs with fixed depth.

2.1 Universality

One of the most famous results in neural network theory is that, under minor conditions on the activation function, the set of networks is very expressive, meaning that every continuous function on a compact set can be arbitrarily well approximated by a MLP. This theorem was first shown by Hornik [14] and Cybenko [8].

To talk about approximation, we first need to define a topology on a space of functions of interest. We define, for $K \subset \mathbb{R}^d$

$$C(K) \coloneqq \{f : K \to \mathbb{R} \colon f \text{ continuous}\}$$

and we equip $C(K)$ with the uniform norm

$$\|f\|_\infty \coloneqq \sup_{x \in K} |f(x)|.$$

If K is a compact space, then the representation theorem of Riesz [30, Theorem 6.19] tells us that the topological dual space of $C(K)$ is the space

$$\mathcal{M} \coloneqq \{\mu \colon \mu \text{ is a signed Borel measure on } K\}.$$

[*] A diligent student would probably want to verify this.

Having fixed the topology on $C(K)$, we can define the concept of *universality* next.

Definition 2.2. *Let $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous, $d, L \in \mathbb{N}$ and $K \subset \mathbb{R}^d$ be compact. Denote by $\mathrm{MLP}(\varrho, d, L)$ the set of all MLPs with d-dimensional input, L layers, $N_L = 1$, and activation function ϱ.*

We say that $\mathrm{MLP}(\varrho, d, L)$ is universal, if $\mathrm{MLP}(\varrho, d, L)$ is dense in $C(K)$.

Example 2.1 demonstrates that $\mathrm{MLP}(\varrho, d, L)$ is *not* universal for every activation function.

Definition 2.3. *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$, compact. A continuous function $f : \mathbb{R} \to \mathbb{R}$ is called* discriminatory *if the only measure $\mu \in \mathcal{M}$ such that*

$$\int_K f(ax - b)d\mu(x) = 0, \quad \text{for all } a \in \mathbb{R}^d, b \in \mathbb{R}$$

is $\mu = 0$.

Theorem 2.4 (Universal approximation theorem [8])**.** *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, and $\varrho : \mathbb{R} \to \mathbb{R}$ be discriminatory. Then $\mathrm{MLP}(\varrho, d, 2)$ is universal.*

Proof. We start by observing that $\mathrm{MLP}(\varrho, d, 2)$ is a linear subspace of $C(K)$. Assume towards a contradiction, that $\mathrm{MLP}(\varrho, d, 2)$ is not dense in $C(K)$. Then there exists $h \in C(K) \setminus \overline{\mathrm{MLP}(\varrho, d, 2)}$.

By the theorem of Hahn-Banach [30, Theorem 5.19] there exists a functional

$$0 \neq H \in C(K)'$$

so that $H = 0$ on $\mathrm{MLP}(\varrho, d, 2)$. Since, for $a \in \mathbb{R}^d, b \in \mathbb{R}$,

$$x \mapsto \varrho(ax - b) =: \varrho_{a,b} \in \mathrm{MLP}(\varrho, d, 2),$$

we have that $H(\varrho_{a,b}) = 0$ for all $a \in \mathbb{R}^d, b \in \mathbb{R}$. Finally, by the identification $C(K)' = \mathcal{M}$ there exists a non-zero measure μ so that

$$\int_K \varrho_{a,b} d\mu = 0, \quad \text{for all } a \in \mathbb{R}^d, b \in \mathbb{R}.$$

This is a contradiction to the assumption that ϱ is discriminatory. □

At this point, we know that all discriminatory activation functions lead to universal spaces of MLPs. Since the property of being discriminatory seems hard to verify directly, we are now interested in identifying more accessible sufficient conditions guaranteeing this property.

Definition 2.5. *A continuous function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \to 1$ for $x \to \infty$ and $f(x) \to 0$ for $x \to -\infty$ is called* sigmoidal*.*

Proposition 2.6. *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ be compact. Then every sigmoidal function $f : \mathbb{R} \to \mathbb{R}$ is discriminatory.*

Proof. Let f be sigmoidal. Then it is clear from Definition 2.5 that, for $\lambda \to \infty$,

$$f\left(\lambda(ax - b) + \theta\right) \to \begin{cases} 1 & \text{if } ax - b > 0 \\ f(\theta) & \text{if } ax - b = 0 \\ 0 & \text{if } ax - b < 0. \end{cases}$$

As f is bounded and K compact, we conclude by the dominated convergence theorem that, for every $\mu \in \mathcal{M}$,

$$\int_K f(\lambda(a \cdot -b) + \theta)d\mu \to \int_{H_{a,b,>}} 1 d\mu + \int_{H_{a,b,=}} f(\theta)d\mu,$$

where

$$H_{a,b,>} \coloneqq \{x \in K : ax - b > 0\} \text{ and } H_{a,b,=} \coloneqq \{x \in K : ax - b = 0\}.$$

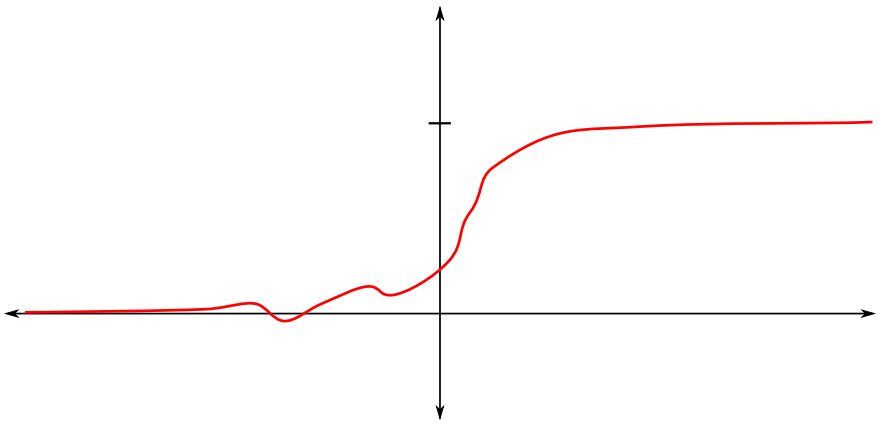

Figure 2.1: A sigmoidal function according to Definition 2.5.

Now assume that

$$\int_K f(\lambda(a \cdot -b) + \theta) d\mu = 0$$

for all $a \in \mathbb{R}^d, b \in \mathbb{R}$. Then

$$\int_{H_{a,b,>}} 1 d\mu + \int_{H_{a,b,=}} f(\theta) d\mu = 0$$

and letting $\theta \to -\infty$, we conclude that $\int_{H_{a,b,>}} 1 d\mu = 0$ for all $a \in \mathbb{R}^d, b \in \mathbb{R}$.

For fixed $a \in \mathbb{R}^d$ and $b_1 < b_2$, we have that

$$0 = \int_{H_{a,b_1,>}} 1 d\mu - \int_{H_{a,b_2,>a}} 1 d\mu = \int_K \mathbb{1}_{[b_1,b_2]}(ax) d\mu(x).$$

By linearity, we conclude that

$$0 = \int_K g(ax) d\mu(x) \tag{2.1}$$

for every step function g. By a density argument and the dominated convergence theorem, we have that (2.1) holds for every bounded continuous function g. Thus (2.1) holds, in particular, for $g = \sin$ and $g = \cos$. We conclude that

$$0 = \int_K \cos(ax) + i \sin(ax) d\mu(x) = \int_K e^{iax} d\mu(x).$$

This implies that the Fourier transform of the measure μ vanishes. This can only happen if $\mu = 0$, [29, p. 176]. □

Remark 2.7. *Universality results can be achieved under significantly weaker assumptions than sigmoidality. For example, in [16] it is shown that Example 2.1 already contains all continuous activation functions that do not generate universal sets of MLPs.*

2.2 Approximation rates

We saw in Theorem 2.4 that MLPs form universal approximators. However, neither the result nor the proof of it give any indication of how ”large” MLPs need to be to achieve a certain approximation accuracy.

Before we can even begin to analyse this question, we need to introduce a precise notion of the size of a MLP. One option could certainly be to count the number of neurons, i.e., $\sum_{\ell=1}^{L} N_\ell$ in (1.1) of Definition 1.1. However, since a MLP was defined as a function, it is by no means clear if there is a unique representation with a unique number of neurons. Hence, the notion of "number of neurons" of a MLP requires some clarification.

Definition 2.8. *Let $d, L \in \mathbb{N}$. A* neural network (NN) with input dimension d and L layers *is a sequence of matrix-vector tuples*

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

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

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

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

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

$$\begin{aligned} &x_0 \coloneqq x, \\ &x_\ell \coloneqq \varrho\left(A_\ell\, x_{\ell-1} + b_\ell\right) \quad \text{for } \ell = 1, \dots, L-1, \\ &x_L \coloneqq A_L\, x_{L-1} + b_L. \end{aligned} \tag{2.2}$$

Here ϱ is understood to act component-wise.

We call $N(\Phi) \coloneqq d + \sum_{j=1}^{L} N_j$ the number of neurons of the NN Φ, *$L(\Phi) \coloneqq L$ the* number of layers *or* depth, *and $M(\Phi) \coloneqq \sum_{j=1}^{L} M_j(\Phi) \coloneqq \sum_{j=1}^{L} \|A_j\|_0 + \|b_j\|_0$ the* number of weights of Φ. *Here $\|.\|_0$ denotes the number of non-zero entries of a matrix or vector.*

According to the notion of Definition 2.8, a MLP is the realisation of a NN.

2.3 Basic operations of networks

Before we analyse how many weights and neurons NNs need to possess so that their realisations approximate certain functions well, we first establish a couple of elementary operations that one can perform with NNs. This formalism was developed first in [24].

To understand the purpose of the following formalism, we start with the following question: Given two realisations of NNs $f_1 \colon \mathbb{R}^d \to \mathbb{R}^d$ and $f_2 \colon \mathbb{R}^d \to \mathbb{R}^d$, is it the case that the function

$$x \mapsto f_2(f_1(x))$$

is the realisation of a NN and how many weights, neurons, and layers does this new function need to have?

Given two functions $f_1 : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $f_2 : \mathbb{R}^{d'} \to \mathbb{R}^{d''}$, where $d, d', d'' \in \mathbb{N}$, we denote by $f_1 \circ f_2$ the composition of these functions, i.e., $f_1 \circ f_2(x) = f_1(f_2(x))$ for $x \in \mathbb{R}^d$. Indeed, a similar concept is possible for NNs.

Definition 2.9. *Let $L_1, L_2 \in \mathbb{N}$ and let $\Phi^1 = ((A_1^1, b_1^1), \dots, (A_{L_1}^1, b_{L_1}^1)), \Phi^2 = ((A_1^2, b_1^2), \dots, (A_{L_2}^2, b_{L_2}^2))$ be two NNs such that the input layer of Φ^1 has the same dimension as the output layer of Φ^2. Then $\Phi^1 \bullet \Phi^2$ denotes the following $L_1 + L_2 - 1$ layer network:*

$$\Phi^1 \bullet \Phi^2 \coloneqq \left(\left(A_1^2, b_1^2\right), \dots, \left(A_{L_2-1}^2, b_{L_2-1}^2\right), \left(A_1^1 A_{L_2}^2, A_1^1 b_{L_2}^2 + b_1^1\right), \left(A_2^1, b_2^1\right), \dots, \left(A_{L_1}^1, b_{L_1}^1\right)\right).$$

We call $\Phi^1 \bullet \Phi^2$ the concatenation of Φ^1 and Φ^2.

It is left as an exercise to show that

$$\mathrm{R}\left(\Phi^1 \bullet \Phi^2\right) = \mathrm{R}\left(\Phi^1\right) \circ \mathrm{R}\left(\Phi^2\right).$$

A second important operation is that of parallelisation.

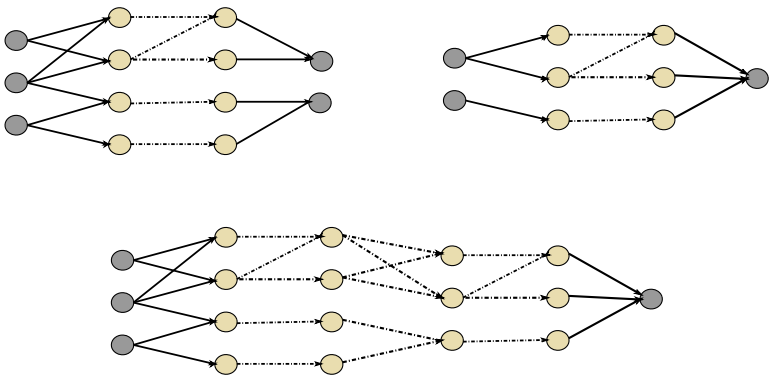

Figure 2.2: **Top:** Two networks. **Bottom:** Concatenation of both networks according to Definition 2.9.

Definition 2.10. *Let $L, d_1, d_2 \in \mathbb{N}$ and let $\Phi^1 = ((A_1^1, b_1^1), \dots, (A_L^1, b_L^1)), \Phi^2 = ((A_1^2, b_1^2), \dots, (A_L^2, b_L^2))$ be two NNs with L layers and with d_1-dimensional and d_2-dimensional input, respectively. We define*

1. $\mathrm{P}\left(\Phi^1, \Phi^2\right) \coloneqq \left(\left(\widehat{A}_1, \widehat{b}_1\right), \left(\tilde{A}_2, \tilde{b}_2\right), \dots, \left(\tilde{A}_L, \tilde{b}_L\right)\right)$, *if* $d_1 = d_2$,
2. $\mathrm{FP}\left(\Phi^1, \Phi^2\right) \coloneqq \left(\left(\tilde{A}_1, \tilde{b}_1\right), \dots, \left(\tilde{A}_L, \tilde{b}_L\right)\right)$, *for arbitrary* $d_1, d_2 \in \mathbb{N}$,

where

$$\widehat{A}_1 \coloneqq \begin{pmatrix} A_1^1 \\ A_1^2 \end{pmatrix}, \quad \widehat{b}_1 \coloneqq \begin{pmatrix} b_1^1 \\ b_1^2 \end{pmatrix}, \quad \text{and} \quad \tilde{A}_\ell \coloneqq \begin{pmatrix} A_\ell^1 & 0 \\ 0 & A_\ell^2 \end{pmatrix}, \quad \tilde{b}_\ell \coloneqq \begin{pmatrix} b_\ell^1 \\ b_\ell^2 \end{pmatrix} \quad \text{for } 1 \leq \ell \leq L.$$

$\mathrm{P}(\Phi^1, \Phi^2)$ *is a NN with d-dimensional input and L layers, called the* parallelisation with shared inputs of Φ^1 and Φ^2. $\mathrm{FP}(\Phi^1, \Phi^2)$ *is a NN with $d_1 + d_2$-dimensional input and L layers, called the* parallelisation without shared inputs of Φ^1 and Φ^2.

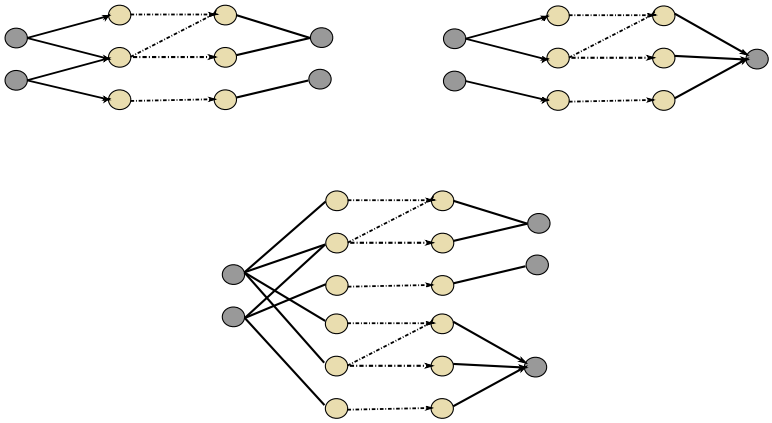

Figure 2.3: **Top:** Two networks. **Bottom:** Parallelisation with shared inputs of both networks according to Definition 2.10.

One readily verifies that $M(\mathrm{P}(\Phi^1, \Phi^2)) = M(\mathrm{FP}(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$, and

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

We depict the parallelisation of two networks in Figure 2.3. Using the concatenation, we can, for example, increase the depth of networks without significantly changing their output if we can build a network that realises the identity function. We demonstrate how to approximate the identity function below. This is our first quantitative approximation result.

Proposition 2.11. *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, and $\varrho : \mathbb{R} \to \mathbb{R}$ be differentiable and not constant on an open set. Then, for every $\epsilon > 0$, there exists a NN $\Phi = ((A_1, b_1), (A_2, b_2))$ such that $A_1, A_2 \in \mathbb{R}^{d\times d}$, $b_1, b_2 \in \mathbb{R}^d$, $M(\Phi) \leq 4d$, and*

$$|\mathrm{R}(\Phi)(x) - x| < \epsilon,$$

for all $x \in K$.

Proof. Assume $d = 1$, the general case of $d \in \mathbb{N}$ then follows immediately by parallelisation without shared inputs.

Let $x^* \in \mathbb{R}$ be such that ϱ is differentiable on a neighbourhood of x^* and $\varrho'(x^*) = \theta \neq 0$. Define, for $\lambda > 0$

$$b_1 \coloneqq x^*, \quad A_1 \coloneqq 1/\lambda, \quad b_2 \coloneqq -\lambda\varrho(x^*)/\theta, \quad A_2 \coloneqq \lambda/\theta.$$

Then we have, for all $x \in K$,

$$|\mathrm{R}(\Phi)(x) - x| = \left|\lambda \frac{\varrho(x/\lambda + x^*) - \varrho(x^*)}{\theta} - x\right|. \tag{2.4}$$

If $x = 0$, then (2.4) shows that $|\mathrm{R}(\Phi)(x) - x| = 0$. Otherwise

$$|\mathrm{R}(\Phi)(x) - x| = \frac{|x|}{|\theta|}\left|\frac{\varrho(x/\lambda + x^*) - \varrho(x^*)}{x/\lambda} - \theta\right|.$$

By the definition of the derivative, we have that $|\mathrm{R}(\Phi)(x) - x| \to 0$ for $\lambda \to \infty$ and all $x \in K$. □

Remark 2.12. *It follows from Proposition 2.11 that under the assumptions of Theorem 2.4 and Proposition 2.11 we have that $\mathrm{MLP}(\varrho, d, L)$ is universal for every $L \in \mathbb{N}$, $L \geq 2$.*

The operations above can be performed for quite general activation functions. If a special activation is chosen, then different operations are possible. In Section 3, we will, for example, introduce an exact emulation of the identity function by realisations of networks with the so-called ReLU activation function.

2.4 Reapproximation of dictionaries

Approximation theory is a well-established field in applied mathematics. This field is concerned with establishing the trade-off between the size of certain sets and their capability of approximately representing a function. Concretely, let $\mathcal{H}$ be a normed space and $(A_N)_{N\in\mathbb{N}}$ be a nested sequence (i.e. $A_N \subset A_{N+1}$ for every $N \in \mathbb{N}$) of subsets of $\mathcal{H}$ and let $\mathcal{C} \subset \mathcal{H}$.

For $N \in \mathbb{N}$, we are interested in the following number

$$\sigma(A_N, \mathcal{C}) \coloneqq \sup_{f\in\mathcal{C}} \inf_{g\in A_N} \|f - g\|_{\mathcal{H}}. \tag{2.5}$$

Here, $\sigma(A_N, \mathcal{C})$ denotes the worst-case error when approximating every element of $\mathcal{C}$ by the closest element in A_N. Quite often, it is not so simple to precisely compute $\sigma(A_N, \mathcal{C})$ but instead we can only establish an *asymptotic approximation rate*. If $h\colon \mathbb{N} \to \mathbb{R}^+$ is such that

$$\sigma(A_N, \mathcal{C}) = \mathcal{O}(h(N)), \quad \text{for } N \to \infty, \tag{2.6}$$

then we say that $(A_N)_{N\in\mathbb{N}}$ *achieves an approximation rate of h for $\mathcal{C}$.*

Definition 2.13. *A typical example of nested spaces of which we want to understand the approximation capabilities are spaces of sparse representations in a basis or more generally in a dictionary. Let $D \coloneqq (f_i)_{i=1}^{\infty} \subset \mathcal{H}$ be a* dictionary*. We define the spaces*

$$A_N \coloneqq \left\{ \sum_{i=1}^{\infty} c_i f_i \colon \|c\|_0 \leq N \right\}. \tag{2.7}$$

Here $\|c\|_0 = \#\{i \in \mathbb{N} \colon c_i \neq 0\}$.

With this notion of A_N, we call $\sigma(A_N, \mathcal{C})$ the best N-term approximation error of $\mathcal{C}$ with respect to D. Moreover, if h satisfies (2.6) *then we say that D* achieves a rate of best N-term approximation error of h for $\mathcal{C}$.

We can introduce a simple procedure to lift approximation theoretical results for N-term approximation to approximation theoretical results of NNs.

Theorem 2.14. *Let $d \in \mathbb{N}$, $\mathcal{H} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ be a normed space, $\varrho : \mathbb{R} \to \mathbb{R}$, and $D \coloneqq (f_i)_{i=1}^{\infty} \subset \mathcal{H}$ be a dictionary.*

Assume that there exist $L, C \in \mathbb{N}$, such that, for every $i \in \mathbb{N}$, and for every $\epsilon > 0$ there exists a NN Φ_i^{ϵ} such that

$$L\left(\Phi_i^{\epsilon}\right) = L, \quad M\left(\Phi_i^{\epsilon}\right) \leq C, \quad \left\|\mathrm{R}\left(\Phi_i^{\epsilon}\right) - f_i\right\|_{\mathcal{H}} \leq \epsilon. \tag{2.8}$$

For every $\mathcal{C} \subset \mathcal{H}$, define A_N as in (2.7) *and*

$$B_N \coloneqq \{\mathrm{R}(\Phi) : \Phi \text{ is a NN with } d\text{-dim input}, L(\Phi) = L, M(\Phi) \leq N\}.$$

Then, for every $\mathcal{C} \subset \mathcal{H}$,

$$\sigma\left(B_{CN}, \mathcal{C}\right) \leq \sigma\left(A_N, \mathcal{C}\right).$$

Proof. We aim to show that there exists $C > 0$ such that every element in A_N can be approximated by a NN with CN weights to arbitrary precision.

Let $a \in A_N$, then $a = \sum_{j=1}^{N} c_{i(j)} f_{i(j)}$. Let $\epsilon > 0$ then, by (2.8), we have that there exist NNs $(\Phi_j)_{j=1}^{N}$ such that

$$L\left(\Phi_j\right) = L, \quad M\left(\Phi_j\right) \leq C, \quad \left\|\mathrm{R}\left(\Phi_j\right) - f_{i(j)}\right\|_{\mathcal{H}} \leq \epsilon / \left(N\|c\|_{\infty}\right). \tag{2.9}$$

We define, $\Phi^c \coloneqq (([c_{i(1)}, c_{i(2)}, \ldots, c_{i(N)}], 0))$ and $\Phi^{a,\epsilon} \coloneqq \Phi^c \bullet \mathrm{P}(\Phi_1, \Phi_2, \cdots, \Phi_N)$. Now it is clear, by the triangle inequality, that

$$\left\|\mathrm{R}\left(\Phi^{a,\epsilon}\right) - a\right\| = \left\| \sum_{j=1}^{N} c_{i(j)} \left(f_{i(j)} - \mathrm{R}\left(\Phi_j\right)\right) \right\| \leq \sum_{j=1}^{N} |c_{i(j)}| \left\| \left(f_{i(j)} - \mathrm{R}\left(\Phi_j\right)\right) \right\| \leq \epsilon.$$

Per Definition 2.9, $L(\Phi^c \bullet \mathrm{P}(\Phi_1, \Phi_2, \cdots, \Phi_N)) = L(\mathrm{P}(\Phi_1, \Phi_2, \cdots, \Phi_N)) = L$ and it is not hard to see that

$$M\left(\Phi^c \bullet \mathrm{P}\left(\Phi_1, \Phi_2, \cdots, \Phi_N\right)\right) \leq M\left(\mathrm{P}\left(\Phi_1, \Phi_2, \cdots, \Phi_N\right)\right) \leq N \max_{j=1,\ldots,N} M\left(\Phi_j\right) \leq NC.$$

□

Remark 2.15. *In words, Theorem 2.14 states that we can transfer a classical N-term approximation result to approximation by realisations of NNs if we can approximate every element from the underlying dictionary arbitrarily well by NNs. It turns out that, under the right assumptions on the activation function, Condition* (2.8) *is quite often satisfied. We will see one instance of such a result in the following subsection and another one in Proposition 3.3 below.*

*We assume here and in the sequel that a dictionary contains only countably many elements. This assumption is not necessary, but simplifies the notation a bit.

2.5 Approximation of smooth functions

We shall proceed by demonstrating that (2.9) holds for the dictionary of multivariate B-splines. This idea, was probably first applied by Mhaskar in [19].

Towards our first concrete approximation result, we therefore start by reviewing some approximation properties of B-splines: The univariate cardinal B-spline on $[0, k]$ of order $k \in \mathbb{N}$ is given by

$$\mathcal{N}_k(x) := \frac{1}{(k-1)!} \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} (x-\ell)_+^{k-1}, \quad \text{for } x \in \mathbb{R}, \tag{2.10}$$

where we adopt the convention that $0^0 = 0$.

For $t \in \mathbb{R}$ and $\ell \in \mathbb{N}$, we define $\mathcal{N}_{\ell,t,k} := \mathcal{N}_k(2^\ell(\cdot - t))$. Additionally, we denote for $d \in \mathbb{N}$, $\ell \in \mathbb{N}$, $t \in \mathbb{R}^d$ *the multivariate B-splines* by

$$\mathcal{N}_{\ell,t,k}^d(x) := \prod_{i=1}^{d} \mathcal{N}_{\ell,t_i,k}(x_i), \quad \text{for } x = (x_1, \dots x_d) \in \mathbb{R}^d.$$

Finally, for $d \in \mathbb{N}$, we define the dictionary of *dyadic B-splines of order* k by

$$\mathcal{B}^k := \left\{ \mathcal{N}_{\ell,t_\ell,k}^d \colon \ell \in \mathbb{N}, t_\ell \in 2^{-\ell}\mathbb{Z}^d \right\}. \tag{2.11}$$

Best N-term approximation by multivariate B-splines is a well studied field. For example, we have the following result by Oswald.

Theorem 2.16 ([22, Theorem 7])**.** *Let $d, k \in \mathbb{N}$, $p \in (0, \infty]$, $0 < s \le k$. Then there exists $C > 0$ such that, for every $f \in C^s([0,1]^d)$, we have that, for every $\delta > 0$, and every $N \in \mathbb{N}$ there exists $c_i \in \mathbb{R}$ with $|c_i| \le C\|f\|_\infty$ and $B_i \in \mathcal{B}^k$ for $i = 1, \dots, N$ such that*

$$\left\| f - \sum_{i=1}^{N} c_i B_i \right\|_{L^p} \lesssim N^{\frac{\delta - s}{d}} \|f\|_{C^s}.$$

In particular, for $\mathcal{C} := \{f \in C^s([0,1]^d) : \|f\|_{C^s} \le 1\}$, we have that $\mathcal{B}^k$ achieves a rate of best N-term approximation error of order $N^{(\delta-s)/d}$ for every $\delta > 0$. [a]

[a] In [22, Theorem 7] this statement is formulated in much more generality. We cite here a simplified version so that we do not have to introduce Besov spaces.

To obtain an approximation result by NN via Theorem 2.14, we now only need to check under which conditions every element of the B-spline dictionary can be represented arbitrarily well by a NN. In this regard, we first fix a class of activation functions.

Definition 2.17. *A function $\varrho : \mathbb{R} \to \mathbb{R}$ is called sigmoidal of order $q \in \mathbb{N}$, if $\varrho \in C^{q-1}(\mathbb{R})$ and*

$$\frac{\varrho(x)}{x^q} \to 0, \text{ for } x \to -\infty, \quad \frac{\varrho(x)}{x^q} \to 1, \text{ for } x \to \infty, \text{ and}$$
$$|\varrho(x)| \lesssim (1 + |x|)^q, \text{ for all } x \in \mathbb{R}.$$

Standard examples of sigmoidal functions of order $k \in \mathbb{N}$ are the functions $x \mapsto \max\{0, x\}^q$. We have the following proposition.

Proposition 2.18. *Let $k, d \in \mathbb{N}$, $K > 0$, and $\varrho : \mathbb{R} \to \mathbb{R}$ be sigmoidal of order $q \ge 2$. There exists a constant $C > 0$ such that for every $f \in \mathcal{B}^k$ and every $\epsilon > 0$ there is a NN Φ^ϵ with $\lceil \log_2(d) \rceil + \lceil \max\{\log_q(k-1), 0\} \rceil + 1$ layers and C weights, such that*

$$\|f - \mathrm{R}_\varrho(\Phi^\epsilon)\|_{L^\infty([-K,K]^d)} \le \epsilon.$$

Proof. We demonstrate how to approximate a cardinal B-spline of order k, i.e., $\mathcal{N}^d_{0,0,k}$, by a NN Φ with activation function ϱ. The general case, i.e., $\mathcal{N}^d_{\ell,t,k}$, follows by observing that shifting and rescaling of the realisation of Φ can be done by manipulating the entries of A_1 and b_1 associated to the first layer of Φ. Towards this goal, we first approximate a univariate B-spline. We observe with (2.10) that we first need to build a network that approximates the function $x \mapsto (x)_+^{k-1}$. The rest follows by taking sums and shifting the function.

It is not hard to see (but probably a good exercise to formally show) that, for every $K' > 0$,

$$\left| a^{-q^T} \underbrace{\varrho \circ \varrho \circ \cdots \circ \varrho(ax)}_{T- \text{ times}} - x_+^{q^T} \right| \to 0 \text{ for } a \to \infty \text{ uniformly for all } x \in [-K', K'].$$

Choosing $T \coloneqq \lceil \max\{\log_q(k-1), 0\} \rceil$ we have that $q^T \geq k-1$. We conclude that, for every $K' > 0$ and $\epsilon > 0$ there exists a NN Φ^*_ϵ with $\lceil \max\{\log_q(k-1), 0\} \rceil + 1$ layers such that

$$\left| \mathrm{R}\left(\Phi^*_\epsilon\right)(x) - x_+^p \right| \leq \epsilon, \tag{2.12}$$

for every $x \in [-K', K']$, where $p \geq k-1$. We observe that, for all $x \in [-K', K']$,

$$\frac{\mathrm{R}\left(\Phi^*_{\delta^2}\right)(x+\delta) - \mathrm{R}\left(\Phi^*_{\delta^2}\right)(x)}{\delta} \to p x_+^{p-1} \text{ for } \delta \to 0. \tag{2.13}$$

One can prove (2.13) directly with the binomial theorem, by observing that

$$\left| \frac{\mathrm{R}\left(\Phi^*_{\delta^2}\right)(x+\delta) - \mathrm{R}\left(\Phi^*_{\delta^2}\right)(x)}{\delta} - p x_+^{p-1} \right| \leq 2\delta + \left| \frac{(x+\delta)_+^p - (x)_+^p}{\delta} - p x_+^{p-1} \right|.$$

Repeating the 'derivative-trick' of (2.13), we can find, for every $K' > 0$ and $\epsilon > 0$ a NN $\Phi^\dagger_\epsilon$ such that, for all $x \in [-K', K']$,

$$\left| \mathrm{R}(\Phi^\dagger_\epsilon)(x) - x_+^{k-1} \right| \leq \epsilon.$$

By (2.10), it is now clear that there exists a NN $\Phi^\vee_\epsilon$ the size of which is independent of ϵ which approximates a univariate cardinal B-spline up to an error of ϵ.

As a second step, we would like to construct a network which multiplies all entries of the d-dimensional output of the realisation of the NN $\mathrm{FP}(\Phi^\vee_\epsilon, \ldots, \Phi^\vee_\epsilon)$. Since ϱ is a sigmoidal function of order larger than 2, we observe by the 'derivative trick' that led to (2.12) that we can also build a fixed size NN with two layers which, for every $K' > 0$ and $\epsilon > 0$, approximates the map $x \mapsto x_+^2$ arbitrarily well for $x \in [-K', K']$.

We have that for every $x = (x_1, x_2) \in \mathbb{R}^2$

$$2x_1x_2 = (x_1+x_2)^2 - x_1^2 - x_2^2 = (x_1+x_2)_+^2 + (-x_1-x_2)_+^2 - (x_1)_+^2 - (-x_1)_+^2 - (x_2)_+^2 - (-x_2)_+^2. \tag{2.14}$$

Hence, we can conclude that, for every $K' > 0$, we can find a fixed size NN $\Phi^{\mathrm{mult}}_\epsilon$ with input dimension 2 which, for every $\epsilon > 0$, approximates the map $(x_1, x_2) \mapsto x_1x_2$ arbitrarily well for $(x_1, x_2) \in [-K', K']^2$.

We assume for simplicity, that $\log_2(d) \in \mathbb{N}$. Then we define

$$\Phi^{\mathrm{mult},d,d/2}_\epsilon \coloneqq \mathrm{FP}(\underbrace{\Phi^{\mathrm{mult}}_\epsilon, \ldots, \Phi^{\mathrm{mult}}_\epsilon}_{d/2-\text{times}}).$$

It is clear that, for all $x \in [-K', K']^d$,

$$\left| \mathrm{R}\left(\Phi^{\mathrm{mult},d,d/2}_\epsilon\right)(x_1, \ldots, x_d) - (x_1x_2, x_3x_4, \ldots, x_{d-1}x_d) \right| \leq \epsilon.$$

Now, we set

$$\Phi^{\mathrm{mult},d,1}_\epsilon \coloneqq \Phi^{\mathrm{mult}}_\epsilon \bullet \Phi^{\mathrm{mult},4,2}_\epsilon \bullet \ldots \bullet \Phi^{\mathrm{mult},d,d/2}_\epsilon. \tag{2.15}$$

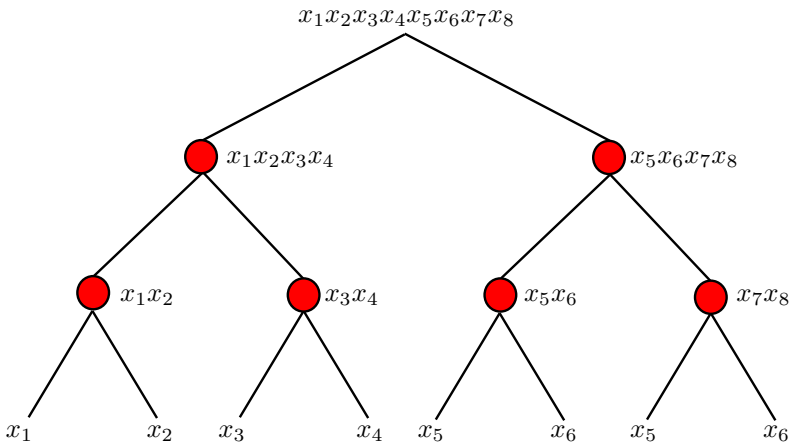

Figure 2.4: Setup of the multiplication network (2.15). Every red dot symbolises a multiplication network $\Phi_\epsilon^{\text{mult}}$ and *not* a regular neuron.

We depict the hierarchical construction of (2.15) in Figure 2.4. Per construction, we have that $\Phi_\epsilon^{\text{mult},d,1}$ has $\log_2(d)+1$ layers and, for every $\epsilon'>0$ and $K'>0$, there exists $\epsilon>0$ such that

$$\left|\Phi_\epsilon^{\text{mult},d,1}(x_1,\dots x_d) - x_1x_2\cdots x_d\right| \leq \epsilon'.$$

Finally, we set

$$\Phi_\epsilon := \Phi_\epsilon^{\text{mult},d,1} \bullet \text{FP}(\underbrace{\Phi_\epsilon^\vee,\dots,\Phi_\epsilon^\vee}_{d-\text{times}}).$$

Per definition of $\bullet$, we have that Φ has $\lceil \max\{\log_q(k-1),0\}\rceil + \log_2(d)+1$ many layers. Moreover, the size of all components of Φ was independent of ϵ. By choosing ϵ sufficiently small it is clear by construction that Φ_ϵ approximates $\mathcal{N}_{0,0,k}^d$ arbitrarily well on $[-K,K]^d$ for sufficiently small ϵ. □

As a simple consequence of Theorem 2.14 and Proposition 2.18 we obtain the following corollary.

Corollary 2.19. *Let $d\in\mathbb{N}$, $s>\delta>0$ and $p\in(0,\infty]$. Moreover let $\varrho:\mathbb{R}\to\mathbb{R}$ be sigmoidal of order $q\geq 2$. Then there exists a constant $C>0$ such that, for every $f\in C^s([0,1]^d)$ with $\|f\|_{C^s}\leq 1$ and every $1/2>\epsilon>0$, there exists a NN Φ such that*

$$\|f-\mathrm{R}(\Phi)\|_{L^p}\leq\epsilon$$

and $M(\Phi)\leq C\epsilon^{-\frac{d}{s-\delta}}$ and $L(\Phi)=\lceil\log_2(d)\rceil+\lceil\max\{\log_q(\lceil s\rceil-1),0\}\rceil+1$.

Remark 2.20. *Corollary 2.19 constitutes the first quantitative approximation result of these notes for a large class of functions. There are a couple of particularly interesting features of this result. First of all, we observe that with increasing smoothness of the functions, we need smaller networks to achieve a certain accuracy. On the other hand, at least in the framework of this theorem, we require more layers if the smoothness s is much higher than the order of sigmoidality of ϱ.*

Finally, the order of approximation deteriorates very quickly with increasing dimension d. Such a behaviour is often called curse of dimension*. We will later analyse to what extent NN approximation can overcome this curse.*

2.6 Fast approximations with Kolmogorov

One observation that we made in the previous subsection is that some activation functions yield better approximation rates than others. In particular, in Theorem 2.19, we see that if the activation function ϱ has a low order of sigmoidality, then we need to use much deeper networks to obtain the same approximation rates than with a sigmoidal function of high order.

Naturally, we can ask ourselves if, by a smart choice of activation function, we could even improve Corollary 2.19 further. The following proposition shows how to achieve an incredible improvement if $d = 1$. The idea for the following proposition and Theorem 2.24 below appeared in [17] first, but is presented in a slightly simplified version here.

Proposition 2.21. *There exists a continuous, piecewise polynomial activation function $\varrho : \mathbb{R} \to \mathbb{R}$ such that for every function $f \in C([0,1])$ and every $\epsilon > 0$ there is a NN $\Phi^{f,\epsilon}$ with $M(\Phi^{f,\epsilon}) \leq 3$, and $L(\Phi^{f,\epsilon}) = 2$ such that*

$$\left\| f - \mathrm{R}\left(\Phi^{f,\epsilon}\right) \right\|_\infty \leq \epsilon. \tag{2.16}$$

Proof. We denote by $\Pi^{\mathbb{Q}}$, the set of univariate polynomials with rational coefficients. It is well-known that this set is countable and dense in $C(K)$ for every compact set K. Hence, we have that $\{\pi_{|[0,1]} \colon \pi \in \Pi^{\mathbb{Q}}\}$ is a countable set and dense in $C([0,1])$. We set $(\pi_i)_{i\in\mathbb{Z}} \coloneqq \{\pi_{|[0,1]} \colon \pi \in \Pi^{\mathbb{Q}}\}$ and define

$$\varrho(x) \coloneqq \begin{cases} \pi_i(x-2i), & \text{if } x \in [2i, 2i+1], \\ \pi_i(1)(2i+2-x) + \pi_{i+1}(0)(x-2i-1), & \text{if } x \in (2i+1, 2i+2). \end{cases}$$

It is clear that ϱ is continuous and piecewise polynomial.

Finally, let us construct the network such that (2.19) holds. For $f \in C([0,1])$ and $\epsilon > 0$ we have by density of $(\pi_i)_{i\in\mathbb{Z}}$ that there exists $i \in \mathbb{Z}$ such that $\|f - \pi_i\|_\infty \leq \epsilon$. Hence,

$$|f(x) - \varrho(x+2i)| = |f(x) - \pi_i(x)| \leq \epsilon. \tag{2.17}$$

The claim follows by defining $\Phi^{f,\epsilon} \coloneqq ((1, 2i), (1, 0))$. □

Remark 2.22. *It is clear that the restriction to functions defined on $[0,1]$ is arbitrary. For every function $f \in C([-K,K])$ for a constant $K > 0$, we have that $f(2K(\cdot - 1/2)) \in C([0,1])$. Therefore, the result of Proposition 2.21 holds by replacing $C([0,1])$ by $C([-K,K])$.*

We will discuss to what extent the activation function ϱ of Proposition 2.21 is sensible a bit further below. Before that, we would like to generalise this result to higher dimensions. This can be done by using Kolmogorov's superposition theorem.

Theorem 2.23 ([15])**.** *For every $d \in \mathbb{N}$, there are $2d^2 + d$ univariate, continuous, and increasing functions $\phi_{p,q}$, $p = 1, \ldots, d$, $q = 1, \ldots, 2d+1$ such that for every $f \in C([0,1]^d)$ we have that, for all $x \in [0,1]^d$,*

$$f(x) = \sum_{q=1}^{2d+1} g_q \left(\sum_{p=1}^{d} \phi_{p,q}(x_p) \right), \tag{2.18}$$

where g_q, $q = 1, \ldots 2d+1$, are univariate continuous functions depending on f.

We can combine Kolmogorov's superposition theorem and Proposition 2.21 to obtain the following approximation theorem for realisations of networks with the special activation function from Proposition 2.21.

Theorem 2.24. *Let $d \in \mathbb{N}$. Then there exists a constant $C(d) > 0$ and a continuous activation function ϱ, such that for every function $f \in C([0,1]^d)$ and every $\epsilon > 0$ there is a NN $\Phi^{f,\epsilon,d}$ with $M(\Phi^{f,\epsilon,d}) \leq C(d)$, and $L(\Phi^{f,\epsilon,d}) = 3$ such that*

$$\left\| f - \mathrm{R}\left(\Phi^{f,\epsilon,d}\right) \right\|_\infty \leq \epsilon. \tag{2.19}$$

Proof. Let $f \in C([0,1]^d)$. Let $\epsilon_0 > 0$ and let $\widetilde{\Phi}^{1,d} \coloneqq (([1, \ldots, 1], 0))$ be a network with d dimensional input and $\widetilde{\Phi}^{1,2d+1} \coloneqq (([1, \ldots, 1], 0))$ be a network with $2d+1$ dimensional input. Let $g_q, \phi_{p,q}$ for $p = 1, \ldots, d$, $q = 1, \ldots, 2d+1$ be as in (2.18).

We have that there exists $C \in \mathbb{R}$ such that

$$\operatorname{ran}(\phi_{p,q}) \subset [-C, C], \text{ for all } p = 1, \dots, d, q = 1, \dots, 2d+1.$$

We define, with Proposition 2.21,

$$\Phi^{q,\epsilon_0} := \widetilde{\Phi}^{1,d} \bullet \mathrm{FP}\left(\Phi^{\phi_{1,q},\epsilon_0}, \Phi^{\phi_{2,q},\epsilon_0}, \dots, \Phi^{\phi_{d,q},\epsilon_0}\right).$$

It is clear that, for $x = (x_1, \dots, x_d) \in [0,1]^d$,

$$\left| \mathrm{R}\left(\Phi^{q,\epsilon_0}\right)(x) - \sum_{p=1}^{d} \phi_{p,q}(x_p) \right| \leq d\epsilon_0 \tag{2.20}$$

and, by construction, $M(\Phi^q) \leq 3d$. Now define, for $\epsilon_1 > 0$,

$$\Phi^f_{\epsilon_0,\epsilon_1} := \widetilde{\Phi}^{1,2d+1} \bullet \mathrm{FP}\left(\Phi^{g_1,\epsilon_1}, \Phi^{g_2,\epsilon_1}, \dots \Phi^{g_{2d+1},\epsilon_1}\right) \bullet \mathrm{P}\left(\Phi^{1,\epsilon_0}, \Phi^{2,\epsilon_0}, \dots, \Phi^{2d+1,\epsilon_0}, \epsilon_0\right), \tag{2.21}$$

where Φ^{g_1,ϵ_1} is according to Remark 2.22 with $K = C + 1$.

Per definition of $\bullet$ it follows that $L(\Phi^f_{\epsilon_0}) \leq 3$ and the size of $\Phi^f_{\epsilon_0}$ is independent of ϵ_0 and ϵ_1. We also have that

$$\mathrm{R}\left(\Phi^f_{\epsilon_0,\epsilon_1}\right) = \sum_{q=1}^{2d+1} \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right) \circ \mathrm{R}\left(\Phi^{q,\epsilon_0}\right).$$

We have by Proposition 2.21 that, for fixed ϵ_1, the map $\mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)$ is uniformly continuous on $[-C-1, C+1]$ for all $q = 1, \dots, 2d+1$ and $\epsilon_0 \leq 1$.

Hence, we have that, for each $\tilde{\epsilon} > 0$, there exists $\delta_{\tilde{\epsilon}} > 0$ such that

$$\left|\mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)(x) - \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)(y)\right| \leq \tilde{\epsilon},$$

for all $x, y \in [-C-1, C+1]$ so that $|x - y| \leq \delta_{\tilde{\epsilon}}$ in particular this statement holds for $\tilde{\epsilon} = \epsilon_1$.

It follows from the triangle inequality, (2.20), and Proposition 2.21 that

$$\begin{aligned}
\left\|\mathrm{R}\left(\Phi^f_{\epsilon_0,\epsilon_1}\right) - f\right\|_\infty &\leq \sum_{q=1}^{2d+1} \left\| \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)\left(\mathrm{R}\left(\Phi^{q,\epsilon_0}\right)\right) - g_q\left(\sum_{p=1}^{d} \phi_{p,q}\right) \right\|_\infty \\
&\leq \sum_{q=1}^{2d+1} \left\| \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)\left(\mathrm{R}\left(\Phi^{q,\epsilon_0}\right)\right) - \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)\left(\sum_{p=1}^{d} \phi_{p,q}\right) \right\|_\infty \\
&\quad + \left\| \mathrm{R}\left(\Phi^{g_q,\epsilon_1}\right)\left(\sum_{p=1}^{d} \phi_{p,q}\right) - g_q\left(\sum_{p=1}^{d} \phi_{p,q}\right) \right\|_\infty \\
&=: \sum_{p=1}^{2d+1} \mathrm{I}_{\epsilon_0,\epsilon_1} + \mathrm{II}_{\epsilon_0,\epsilon_1}.
\end{aligned}$$

Choosing $d\epsilon_0 < \delta_{\epsilon_1}$, we have that $\mathrm{I}_{\epsilon_0,\epsilon_1} \leq \epsilon_1$. Moreover, $\mathrm{II} \leq \epsilon_1$ by construction .

Hence, for every $1/2 > \epsilon > 0$, there exists ϵ_0, ϵ_1 such that $\left\|\mathrm{R}\left(\Phi^f_{\epsilon_0}\right) - f\right\|_\infty \leq (2d+1)\epsilon_1 \leq \epsilon$. We define $\Phi^{f,\epsilon,d} := \Phi^f_{\epsilon_0,\epsilon_1}$ which concludes the proof. □

Without knowing the details of the proof of Theorem 2.24 the statement that any function can be arbitrarily well approximated by a fixed-size network is hardly believable. It seems as if the reason for this result to hold is that we have put an immense amount of information into the activation function. At the very least, we have now established that at least from a certain minimal size on, *there is no aspect of the architecture of a NN that fundamentally limits its approximation power*. We will later develop fundamental lower bounds on approximation capabilities. As a consequence of the theorem above, these lower bounds can only be given for specific activation functions or under further restricting assumptions.

3 ReLU networks

We have already seen a variety of activation functions including sigmoidal and higher-order sigmoidal functions. In practice, a much simpler function is usually used. This function is called *rectified linear unit (ReLU)*. It is defined by

$$x \mapsto \varrho_R(x) \coloneqq (x)_+ = \max\{0, x\} = \begin{cases} x & \text{for } x \geq 0 \\ 0 & \text{else.} \end{cases} \tag{3.1}$$

There are various reasons why this activation function is immensely popular. Most of these reasons are based on its practicality in the algorithms used to train NNs which we do not want to analyse in this note. One thing that we can observe, though, is that the evaluation of $\varrho_R(x)$ can be done much more quickly than that of virtually any non-constant function. Indeed, only a single decision has to be made, whereas, for other activation functions such as, e.g., $\arctan$, the evaluation requires many numerical operations. This function is probably the simplest function that does not belong to the class described in Example 2.1.

One of the first questions that we can ask ourselves is whether the ReLU is discriminatory. We observe the following. For $a \in \mathbb{R}$, $b_1 < b_2$ and every $x \in \mathbb{R}$, we have that

$$H_a(x) \coloneqq \varrho_R(ax - ab_1 + 1) - \varrho_R(ax - ab_1) - \varrho_R(ax - ab_2) + \varrho_R(ax - ab_2 - 1) \to \mathbb{1}_{[b_1, b_2]} \text{ for } a \to \infty.$$

Indeed, for $x < b_1 - 1/a$, we have that $H_a(x) = 0$. If $b_1 - 1/a < x < b_1$, then $H_a(x) = a(x - b_1 + 1/a) \leq 1$. If $b_1 < x < b_2$, then $H_a(x) = \varrho_R(ax - ab_1 + 1) - \varrho_R(ax - ab_1) = 1$. If $b_2 \leq x < b_2 + 1/a$, then $H_a(x) = 1 - \varrho_R(ax - ab_2) = 1 - ax - ab_2 \leq 1$. Finally, if $x \geq b_2 + 1/a$ then $H_a(x) = 0$. We depict H_a in Figure 3.1.

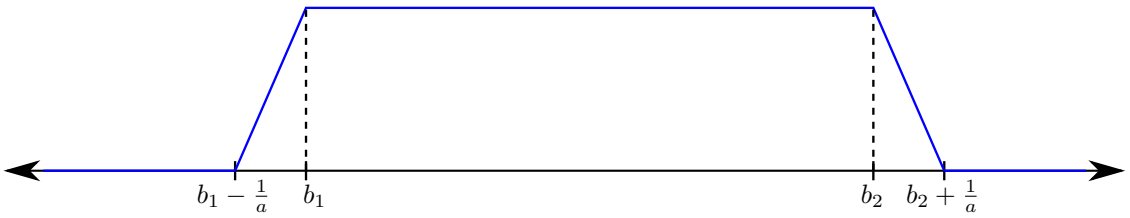

Figure 3.1: Pointwise approximation of a univariate indicator function by sums of ReLU activation functions.

The argument above shows that sums of ReLUs can pointwise approximate arbitrary indicator function. If we had that

$$\int_K \varrho_R(ax + b) d\mu(x) = 0,$$

for a $\mu \in \mathcal{M}$ and all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, then this would imply

$$\int_K \mathbb{1}_{[b_1, b_2]}(ax) d\mu(x) = 0$$

for all $a \in \mathbb{R}^d$ and $b_1 < b_2$. At this point we have the same result as in (2.1). Following the rest of the proof of Proposition 2.6 yields that ϱ_R is discriminatory.

We saw in Proposition 2.18 how higher-order sigmoidal functions can reapproximate B-splines of arbitrary order. The idea there was that, essentially, through powers of x_+^q, we can generate arbitrarily high degrees of polynomials. This approach does not work anymore if $q = 1$. Moreover, the crucial multiplication operation of Equation (2.14) cannot be performed so easily with realisations of networks with the ReLU activation function.

If we want to use the local approximation by polynomials in a similar way as in Corollary 2.19, we have two options: being content with approximation by piecewise linear functions, i.e., polynomials of degree one, or trying to reproduce higher-order monomials by realisations of NNs with the ReLU activation function in a different way than by simple composition.

Let us start with the first approach, which was established in [13].

3.1 Linear finite elements and ReLU networks

We start by recalling some basics on linear finite elements. Below, we will perform a lot of basic operations on sets and therefore it is reasonable to recall and fix some set-theoretical notation first. For a subset A of a topological space, we denote by $\mathrm{co}(A)$ the *convex hull* of A, i.e., the smallest convex set containing A. By $\overline{A}$ we denote the *closure of* A, i.e., the smallest closed set containing A. Furthermore, $\mathrm{int}\ A$ denotes the *interior of* A, which is the largest open subset of A. Finally, the *boundary of* A is denoted by ∂A and $\partial A \coloneqq \overline{A} \setminus \mathrm{int}\ A$.

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$. A set $\mathcal{T} \subset \mathcal{P}(\Omega)$ so that

$$\bigcup \mathcal{T} = \Omega,$$

$\mathcal{T} = (\tau_i)_{i=1}^{M_\mathcal{T}}$, where each τ_i is a d-simplex*, and such that $\tau_i \cap \tau_j \subset \partial\tau_i \cap \partial\tau_j$ is an n-simplex with $n < d$ for every $i \neq j$ is called a *simplicial mesh of* Ω. We call the τ_i the *cells of the mesh* $\mathcal{T}$ and the extremal points of the τ_i, $i = 1 \ldots, M_\mathcal{T}$, the *nodes of the mesh*. We denote the set of nodes by $(\eta_i)_{i=1}^{M_N}$. We will also assume that if $\eta_i \in \tau_k$ for some k and i then η_i is always an extremal point of τ_k to prevent degenerate meshes.

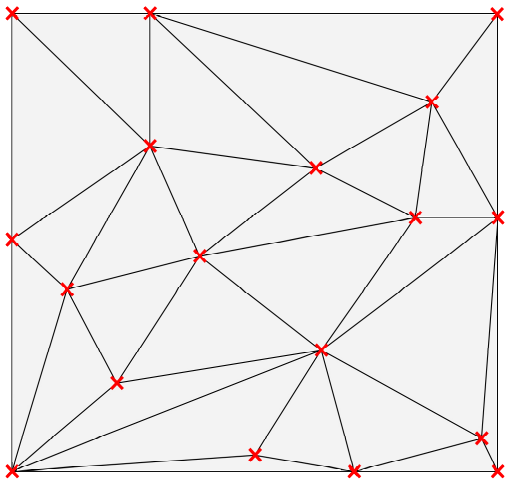

Figure 3.2: A two dimensional simplicial mesh of $[0,1]^2$. The nodes are depicted by red x's.

We say that a mesh $\mathcal{T} = (\tau_i)_{i=1}^{M_\mathcal{T}}$ is *locally convex*, if for every η_i it holds that $\bigcup\{\tau_j : \eta_i \in \tau_j\}$ is convex.

For any mesh $\mathcal{T}$ one defines the linear finite element space

$$V_\mathcal{T} \coloneqq \left\{ f \in C(\Omega) \colon f_{|\tau_i} \text{ affine linear for all } i = 1, \ldots, M_\mathcal{T} \right\}.$$

Since an affine linear function is uniquely defined through its values on $d+1$ linearly independent points, it is clear that every $f \in V_\mathcal{T}$ is uniquely defined through the values $(f(\eta_i))_{i=1}^{M_N}$. By the same token, for every choice of $(y_i)_{i=1}^{M_N}$, there exists a function f in $V_\mathcal{T}$ such that $f(\eta_i) = y_i$ for all $i = 1, \ldots, M_N$.

For $i = 1, \ldots, M_N$ we define the *Courant elements* $\phi_{i,\mathcal{T}} \in V_\mathcal{T}$ to be those functions that satisfy $\phi_{i,\mathcal{T}}(\eta_j) = \delta_{i,j}$. See Figure 3.3 for an illustration.

Proposition 3.1. *Let $d \in \mathbb{N}$ and $\mathcal{T}$ be a simplicial mesh of $\Omega \subset \mathbb{R}^d$, then we have that*

$$f = \sum_{i=1}^{M_N} f(\eta_i)\phi_{i,\mathcal{T}}$$

holds for every $f \in V_\mathcal{T}$.

*A d-simplex is a convex hull of $d+1$ points $v_0, \ldots, v_d$ such that $(v_1 - v_0), (v_2 - v_0), \ldots, (v_d - v_0)$ are linearly independent.

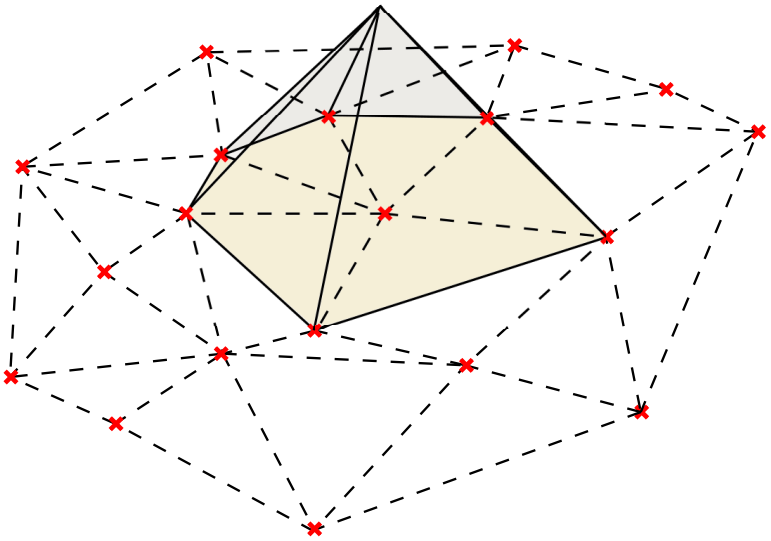

Figure 3.3: Visualisation of a Courant element on a mesh.

As a consequence of Proposition 3.1, we have that we can build every function $f \in V_{\mathcal{T}}$ as the realisation of a NN with ReLU activation function if we can build $\phi_{i,\mathcal{T}}$ for every $i = 1, \dots, M_N$.

We start by making a couple of convenient definitions and then find an alternative representation of $\phi_{i,\mathcal{T}}$. We define, for $i, j = 1, \dots M_N$,

$$F(i) \coloneqq \{j \in \{1, \dots, M_N\} \colon \eta_i \in \tau_j\}, \qquad G(i) \coloneqq \bigcup_{j \in F(i)} \tau_j, \tag{3.2}$$

$$H(j,i) \coloneqq \{\eta_k \in \tau_j, \eta_k \neq \eta_i\}, \qquad I(i) \coloneqq \{\eta_k \in G(i)\}. \tag{3.3}$$

Here $F(i)$ is the set of all indices of cells that contain η_i. Moreover, $G(i)$ is the polyhedron created from taking the union of all these cells.

Proposition 3.2. *Let $d \in \mathbb{N}$ and $\mathcal{T}$ be a locally convex simplicial mesh of $\Omega \subset \mathbb{R}^d$. Then, for every $i = 1, \dots, M_N$, we have that*

$$\phi_{i,\mathcal{T}} = \max\left\{0, \min_{j \in F(i)} g_j\right\}, \tag{3.4}$$

where g_j is the unique affine linear function such that $g_j(\eta_k) = 0$ for all $\eta_k \in H(j,i)$ and $g_j(\eta_i) = 1$.

Proof. Let $i \in \{1, \dots, M_N\}$. By the local convexity assumption we have that $G(i)$ is convex. For simplicity, we assume that $\eta_i \in \operatorname{int} G(i)$.*

Step 1: We show that

$$\partial G(i) = \bigcup_{j \in F(i)} \operatorname{co}(H(j,i)). \tag{3.5}$$

The argument below is visualised in Figure 3.4. We have by convexity that $G(i) = \operatorname{co}(I(i))$. Since η_i lies in the interior of $G(i)$ we have that there exists $\epsilon > 0$ such that $B_\epsilon(\eta_i) \subset G(i)$. By convexity, we have that also the open set $\operatorname{co}(\operatorname{int} \tau_k, B_\epsilon(\eta_i))$ is a subset of $G(i)$. It is not hard to see that $\tau_k \setminus \operatorname{co}(H(k,i)) \subset \operatorname{co}(\operatorname{int} \tau_k, B_\epsilon(\eta_i))$ and

*The case $\eta_i \in \partial G(i)$ needs to be treated slightly differently and is left as an excercise.

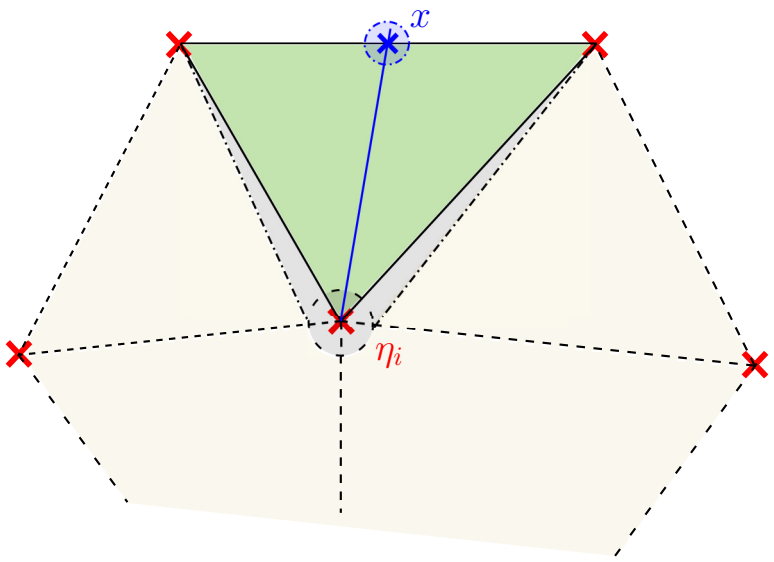

Figure 3.4: Visualisation of the argument in Step 1. The simplex τ_k is coloured green. The grey ball around η_i is $B_\epsilon(\eta_i)$. The blue $\times$ represents x.

hence $\tau_k \setminus \mathrm{co}(H(k,i))$ lies in the interior of $G(i)$. Since we also have that $\partial G(i) \subset \bigcup_{k\in F(i)} \partial\tau_k$, we conclude that

$$\partial G(i) \subset \bigcup_{i\in F(i)} \mathrm{co}(H(j,i)).$$

Now assume that there is j such that $\mathrm{co}(H(j,i)) \not\subset \partial G(i)$. Since $\mathrm{co}(H(j,i)) \subset G(i)$ this would imply that there exist $x \in \mathrm{co}(H(j,i))$ such that x is in the interior of $G(i)$. This implies that there exists $\epsilon' > 0$ such that $B'_\epsilon(x) \subset G(i)$. Hence, the line from η_i to x can be extended for a distance of $\epsilon'/2$ to a point $x^* \in G(i) \setminus \tau_j$. As x^* must belong to a simplex τ_{j^*} that also contains η_i, we conclude that τ_{j^*} intersects the interior of τ_j which cannot be by assumption on the mesh.

Step 2:

For each j, denote by $\mathcal{H}(j,i)$ the hyperplane through $H(j,i)$. The hyperplane $\mathcal{H}(j,i)$ splits $\mathbb{R}^d$ into two subsets, and we denote by $H^{\mathrm{int}}(j,i)$ the set that contains η_i.

We claim that

$$G(i) = \bigcap_{j\in F(i)} H^{\mathrm{int}}(j,i). \tag{3.6}$$

This is intuitively clear and sketched in Figure 3.5.

We first prove the case $G(i) \subset \bigcap_{j\in F(i)} H^{\mathrm{int}}(j,i)$. Assume towards a contradiction that $x' \in G(i)$ is a point in $\mathbb{R}^d \setminus H^{\mathrm{int}}(j,i)$ for a $j \in F(i)$

Since η_i does not lie in the boundary of $G(i)$ there exists $\epsilon > 0$ such that $B_\epsilon(\eta_i) \subset G(i)$ and therefore, by convexity $\mathrm{co}(B_\epsilon(\eta_i), x') \subset G(i)$. Since η_i and x' are on different sides of $\mathcal{H}(j,i)$, we have that there is a point $x'' \in \mathcal{H}(j,i)$ and $\epsilon' > 0$, such that $B_{\epsilon'}(x'') \subset G(i)$. Therefore, $\mathrm{co}(B_{\epsilon'}(x''), \mathrm{int}\ \mathrm{co}(H(j,i))) \subset G(i)$ is open. In particular, $\mathrm{co}(B_{\epsilon'}(x''), \mathrm{int}\ \mathrm{co}(H(j,i))) \cap \partial G(i) = \varnothing$. We conclude that $\mathrm{int}\ \mathrm{co}(H(j,i)) \cap \partial G(i) = \varnothing$. This constitutes a contradiction to (3.5).

Next we prove that $G(i) \supset \bigcap_{j\in F(i)} H^{\mathrm{int}}(j,i)$. Let $x''' \notin G(i)$. Next, we show that x''' lies in $\mathbb{R}^d \setminus H^{\mathrm{int}}(j,i)$ for at least one j. The line between x''' and η_i intersects $G(i)$ and, by Step 1, it intersects $\mathrm{co}(H(j,i))$ for a $j \in F(i)$. It is also clear that x''' is not on the same side as η_i. Hence $x''' \notin H^{\mathrm{int}}(j,i)$.

Step 3: For each $\eta_j \in I(i)$, we have that $g_k(\eta_j) \geq 0$ for all $k \in F(i)$.

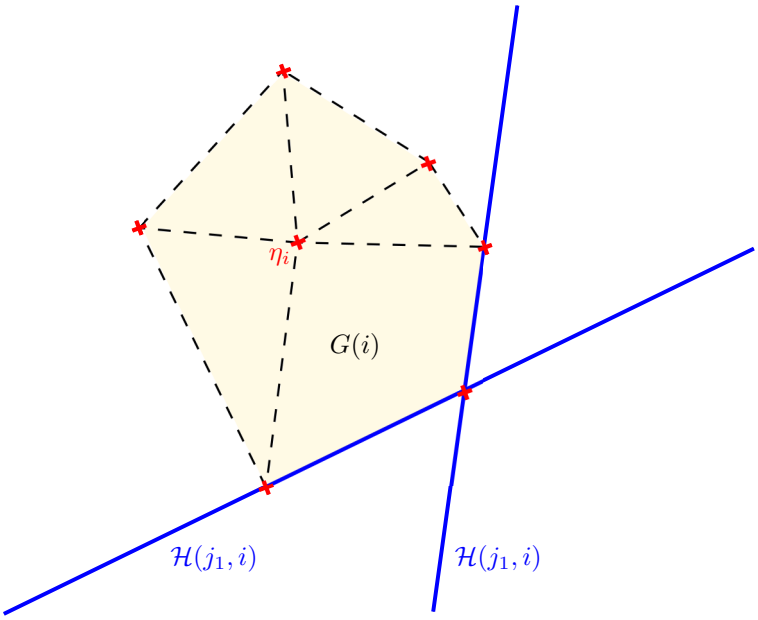

Figure 3.5: The set $G(i)$ and two hyperplanes $\mathcal{H}(j_1,i)$, $\mathcal{H}(j_2,i)$. Since $G(i)$ is convex and $\mathcal{H}(j,i)$ extends its boundary it is intuitively clear that $G(i)$ is only on one side of $\mathcal{H}(j,i)$ and that (3.6) holds.

This is because, by (3.6) $G(i)$ lies fully on one side of each hyperplane $\mathcal{H}(j,i)$, $j \in F(i)$. Since g_k vanishes on $\mathcal{H}(k,i)$ and equals 1 on η_i we conclude that $g_k(\eta_j) \geq 0$ for all $k \in F(i)$

Step 4: For every $k \in F(i)$ we have that $g_k \leq g_j$ on τ_k for all $j \in F(i)$

If for $j \in F(i)$, $g_j(\eta_\ell) \geq g_k(\eta_\ell)$ for all $\eta_\ell \in \tau_k$, then, since $\tau_k = \mathrm{co}(\{\eta_\ell : \eta_\ell \in \tau_k\})$, we conclude that $g_j \geq g_k$. Assume towards a contradiction that $g_j(\eta_\ell) < g_k(\eta_\ell)$ for at least one $\eta_\ell \in I(i)$. Clearly this assumption cannot hold for $\eta_\ell = \eta_i$ since there $g_j(\eta_i) = 1 = g_k(\eta_i)$. If $\eta_\ell \neq \eta_i$, then $g_k(\eta_\ell) = 0$ implying $g_j(\eta_\ell) < 0$. Together with Step 3 this yields a contradiction.

Step 5: For each $z \notin G(i)$, we have that there exists at least one $k \in F(i)$ such that $g_k(z) \leq 0$.

This follows as in Step 3. Indeed, if $z \notin G(i)$ then, by (3.6) we have that there is a hyperplane $\mathcal{H}(k,i)$ so that z does not lie on the same side as η_i. Hence $g_k(z) \leq 0$.

Combining Steps 1-5 yields the claim (3.4). □

Now that we have a formula for the functions $\phi_{i,\mathcal{T}}$, we proceed by building these functions as realisations of NNs.

Proposition 3.3. *Let $d \in \mathbb{N}$ and $\mathcal{T}$ be a locally convex simplicial mesh of $\Omega \subset \mathbb{R}^d$. Let $k_\mathcal{T}$ denote the maximum number of neighbouring cells of the mesh, i.e.,*

$$k_\mathcal{T} \coloneqq \max_{i=1,\dots,M_N} \#\left\{j : \eta_i \in \tau_j\right\}. \tag{3.7}$$

Then, for every $i = 1, \dots, M_N$, there exists a NN Φ_i with

$$L(\Phi_i) = \lceil \log_2(k_\mathcal{T}) \rceil + 2, \text{ and } M(\Phi_i) \leq C \cdot (k_\mathcal{T} + d) k_\mathcal{T} \log_2(k_\mathcal{T})$$

for a universal constant $C > 0$, and

$$\mathrm{R}(\Phi_i) = \phi_{i,\mathcal{T}}, \tag{3.8}$$

where the activation function is the ReLU.

Proof. We now construct the network the realisation of which equals (3.4). The claim (3.8) then follows with Proposition 3.2.

We start by observing that, for $a, b \in \mathbb{R}$,

$$\min\{a,b\} = \frac{a+b}{2} - \frac{|a-b|}{2} = \frac{1}{2}\left(\varrho_R(a+b) - \varrho_R(-a-b) - \varrho_R(a-b) - \varrho_R(b-a)\right).$$

Thus, defining $\Phi^{\min,2} \coloneqq ((A_1,0),(A_2,0))$ with

$$A_1 \coloneqq \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad A_2 \coloneqq \frac{1}{2}[1,-1,-1,-1],$$

yields $\mathrm{R}(\Phi^{\min,2})(a,b) = \min\{a,b\}$, $L(\Phi^{\min,2}) = 2$ and $M(\Phi^{\min,2}) = 12$. Following an idea that we saw earlier for the construction of the multiplication network in (2.15), we construct for $p \in \mathbb{N}$ even, the networks

$$\widetilde{\Phi}^{\min,p} \coloneqq \mathrm{FP}(\underbrace{\Phi^{\min,2},\dots,\Phi^{\min,2}}_{p/2- \text{ times}})$$

and for $p = 2^q$

$$\Phi^{\min,p} = \widetilde{\Phi}^{\min,2} \bullet \widetilde{\Phi}^{\min,4} \dots \bullet \widetilde{\Phi}^{\min,p}.$$

It is clear that the realisation of $\Phi^{\min,p}$ is the minimum operator with p inputs. If p is not a power of two then a small adaptation of the procedure above is necessary. We will omit this discussion here.

We see that $L(\Phi^{\min,p}) = \lceil \log_2(p) \rceil + 1$. To estimate the weights, we first observe that the number of neurons in the first layer of $\widetilde{\Phi}^{\min,p}$ is bounded by $2p$. It follows that each layer of $\Phi^{\min,p}$ has less than $2p$ neurons. Since all affine maps in this construction are linear, we have that

$$\Phi^{\min,p} = ((A_1,b_1),\dots,(A_L,b_L)) = ((A_1,0),\dots,(A_L,0)). \tag{3.9}$$

We have that $g_k = G_k(\cdot) + \theta_k$ for $\theta_k \in \mathbb{R}$ and $G_k \in \mathbb{R}^{1,d}$. Let

$$\Phi^{\mathrm{aff}} \coloneqq \mathrm{P}\left(((G_1,\theta_1)),((G_2,\theta_2)),\dots,\left(\left(G_{\#F(i)},\theta_{\#F(i)}\right)\right)\right).$$

Clearly, Φ^{aff} has one layer, d dimensional input, and $\#F(i)$ many output neurons.

We define, for $p \coloneqq \#F(i)$,

$$\Phi^{i,\mathcal{T}} \coloneqq ((1,0),(1,0)) \bullet \Phi^{\min,p} \bullet \Phi^{\mathrm{aff}}.$$

Per construction and (3.4), we have that $\mathrm{R}(\Phi^{i,\mathcal{T}}) = \phi_{i,\mathcal{T}}$. Moreover, $L(\Phi^{i,\mathcal{T}})) = L(\Phi^{\min,p}) + 1 = \lceil \log_2(p) \rceil + 2$. Also, by construction, the number of neurons in each layer of $\Phi^{i,\mathcal{T}}$ is bounded by $2p$. Since, by (3.9), we have that

$$\Phi^{i,\mathcal{T}} = ((A_1,b_1),(A_2,0),\dots,(A_L,0)),$$

with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_1 \in \mathbb{R}^p$, we conclude that

$$M(\Phi^{i,\mathcal{T}}) \leq p + \sum_{\ell=1}^{L} \|A_\ell\|_0 \leq p + \sum_{\ell=1}^{L} N_{\ell-1}N_\ell \leq p + 2dp + (2p)^2(\lceil \log_2(p) \rceil).$$

Finally, per assumption $p \leq k_{\mathcal{T}}$ which yields the claim. □

As a consequence of Propositions 3.3 and 3.1, we conclude that one can represent every continuous piecewise linear function on a locally compact mesh with N nodes as the realisation of a NN with CN weights where the constant depends on the maximum number of cells neighbouring a vertex $k_{\mathcal{T}}$ and the input dimension d.

Theorem 3.4. *Let $\mathcal{T}$ be a locally convex partition of $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. Let $\mathcal{T}$ have M_N and let $k_\mathcal{T}$ be defined as in* (3.7)*. Then, for every $f \in V_\mathcal{T}$, there exists a NN Φ^f such that*

$$\begin{aligned} &L\left(\Phi^f\right) \leq \lceil \log_2(k_\mathcal{T}) \rceil + 2, \\ &M\left(\Phi_f\right) \leq C M_N \cdot (k_\mathcal{T} + d)\, k_\mathcal{T} \log_2\left(k_\mathcal{T}\right), \\ &\mathrm{R}\left(\Phi^f\right) = f, \end{aligned}$$

for a universal constant $C > 0$.

Remark 3.5. *One way to read Theorem 3.4 is the following: Whatever one can approximate by piecewise affine linear, continuous functions with N degrees of freedom can be approximated to the same accuracy by realisations of NNs with $C \cdot N$ degrees of freedom, for a constant C. If we consider approximation rates, then this implies that realisations of NNs achieve the same approximation rates as linear finite element spaces.*

For example, for $\Omega \coloneqq [0,1]^d$, one has that there exists a sequence of locally convex simplicial meshes $(\mathcal{T}_n)_{n=1}^\infty$ with $M_\mathcal{T}(\mathcal{T}_n) \lesssim n$ such that

$$\inf_{g \in V_{\mathcal{T}_n}} \|f - g\|_{L^2(\Omega)} \lesssim n^{-\frac{2}{d}} \|f\|_{W^{2,2d/(d+2)}(\Omega)},$$

for all $f \in W^{2,2d/(d+2)}(\Omega)$, see, e.g., [13].

3.2 Approximation of the square function

With Theorem 3.4, we are able to reproduce approximation results of piecewise linear functions by realisations of NNs. However, the approximation rates of piecewise affine linear functions when approximating C^s regular functions do not improve for increasing s as soon as $s \geq 1$, see, e.g., Theorem 2.16. To really benefit from higher-order smoothness, one requires piecewise polynomials of higher degree.

Therefore, if we want to approximate smooth functions in the spirit of Corollary 2.19, then we need to be able to efficiently approximate continuous piecewise polynomials of degree higher than 1 by realisations of NNs.

It is clear that this emulation of polynomials cannot be performed as in Corollary 2.19, since the ReLU is piecewise linear. However, if we allow sufficiently deep networks there is, in fact, a surprisingly effective possibility to approximate square functions and thereby polynomials by realisations of NNs with ReLU activation functions.

To see this, we first consider the remarkable construction below.

Efficient construction of saw-tooth functions: Let

$$\Phi^\wedge \coloneqq ((A_1, b_1), (A_2, 0)),$$

where

$$A_1 \coloneqq \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \quad b_1 \coloneqq \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix}, \quad A_2 \coloneqq [1, -2, 1].$$

Then

$$\mathrm{R}\left(\Phi^\wedge\right)(x) = \varrho_R(2x) - 2\varrho_R(2x-1) + \varrho_R(2x-2)$$

and $L(\Phi^\wedge) = 2$, $M(\Phi^\wedge) = 8$, $N_0 = 1, N_1 = 3, N_3 = 1$. It is clear that $\mathrm{R}(\Phi^\wedge)$ is a hat function. We depict it in Figure 3.6.

A quite interesting thing happens if we compose $\mathrm{R}(\Phi^\wedge)$ with itself. We have that

$$\mathrm{R}(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) = \underbrace{\mathrm{R}(\Phi^\wedge) \circ \cdots \circ \mathrm{R}(\Phi^\wedge)}_{n-\text{times}}$$

is a saw-tooth function with 2^{n-1} hats of width 2^{1-n} each. This is depicted in Figure 3.6. Compositions are notoriously hard to picture, hence it is helpful to establish the precise form of $\mathrm{R}(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}})$ formally. We analyse this in the following proposition.

Proposition 3.6. *For $n \in \mathbb{N}$, we have that*

$$F_n = \mathrm{R}(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}})$$

satisfies, for $x \in (0,1)$,

$$F_n(x) \coloneqq \begin{cases} 2^n(x - i2^{-n}) & \text{for } x \in [i2^{-n}, (i+1)2^{-n}], i \text{ even}, \\ 2^n((i+1)2^{-n} - x)) & \text{for } x \in [i2^{-n}, (i+1)2^{-n}], i \text{ odd} \end{cases} \tag{3.10}$$

and $F_n = 0$ for $x \notin (0,1)$. Moreover, $L(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) = n+1$ and $M(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) \leq 12n - 2$.

Proof. The proof follows by induction. We have that, for $x \in [0, 1/2]$,

$$\mathrm{R}(\Phi^\wedge)(x) = \varrho_R(2x) = 2x.$$

Moreover, for $x \in [1/2, 1]$ we conclude

$$\mathrm{R}(\Phi^\wedge)(x) = 2x - 2(2x-1) = 2 - 2x.$$

Finally, if $x \notin (0,1)$, then

$$\varrho_R(2x) - 2\varrho_R(2x-1) + \varrho_R(2x-2) = 0.$$

This completes the case $n = 1$. We assume that we have shown (3.10) for $n \in \mathbb{N}$. Hence, we have that

$$F_{n+1} = F_n \circ \mathrm{R}(\Phi^\wedge), \tag{3.11}$$

where F_n satisfies (3.10). Let $x \in [0, 1/2]$ and $x \in [i2^{-n-1}, (i+1)2^{-n-1}], i$ even. Then $\mathrm{R}(\Phi^\wedge)(x) = 2x \in [i2^{-n}, (i+1)2^{-n}], i$ even. Hence, by (3.11), we have

$$F_{n+1}(x) = 2^n(2x - i2^{-n}) = 2^{n+1}(x - i2^{-n-1}).$$

If $x \in [1/2, 1]$ and $x \in [i2^{-n-1}, (i+1)2^{-n-1}], i$ even, then $\mathrm{R}(\Phi^\wedge)(x) = 2 - 2x \in [2 - (i+1)2^{-n}, 2 - i2^{-n}] = [(2^{n+1} - i - 1)2^{-n}, (2^{n+1} - i)2^{-n}] = [j2^{-n}, (j+1)2^{-n}]$ for $j \coloneqq (2^{n+1} - i - 1)$ odd. We have, by (3.11),

$$\begin{aligned} F_{n+1}(x) &= 2^n(j2^{-n} - (2 - 2x)) = 2^n((2 - 2^{-n}(i+1)) - (2 - 2x)) \\ &= 2^n(2x - 2^{-n}(i+1)) = 2^{n+1}(x - 2^{-n-1}(i+1)). \end{aligned}$$

The cases for i odd follow similarly. If $x \notin (0,1)$, then $\mathrm{R}(\Phi^\wedge)(x) = 0$ and per (3.11) we have that $F_{n+1}(x) = 0$.

It is clear by Definition 3.12 that $L(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) = n+1$. To show that $M(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) \leq 12n - 2$, we observe with

$$\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge =: ((A_1, b_1), \ldots, (A_L, b_L)))$$

that $M(\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge) \leq \sum_{\ell=1}^{n+1} N_{\ell-1}N_\ell + N_\ell \leq (n-1)(3^2+3) + N_0N_1 + N_1 + N_nN_{n+1} + N_{n+1} = 12(n-1) + 3 + 3 + 3 + 1 = 12n - 2$, where we use that $N_\ell = 3$ for all $1 \leq \ell \leq n$ and $N_0 = N_L = 1$. □

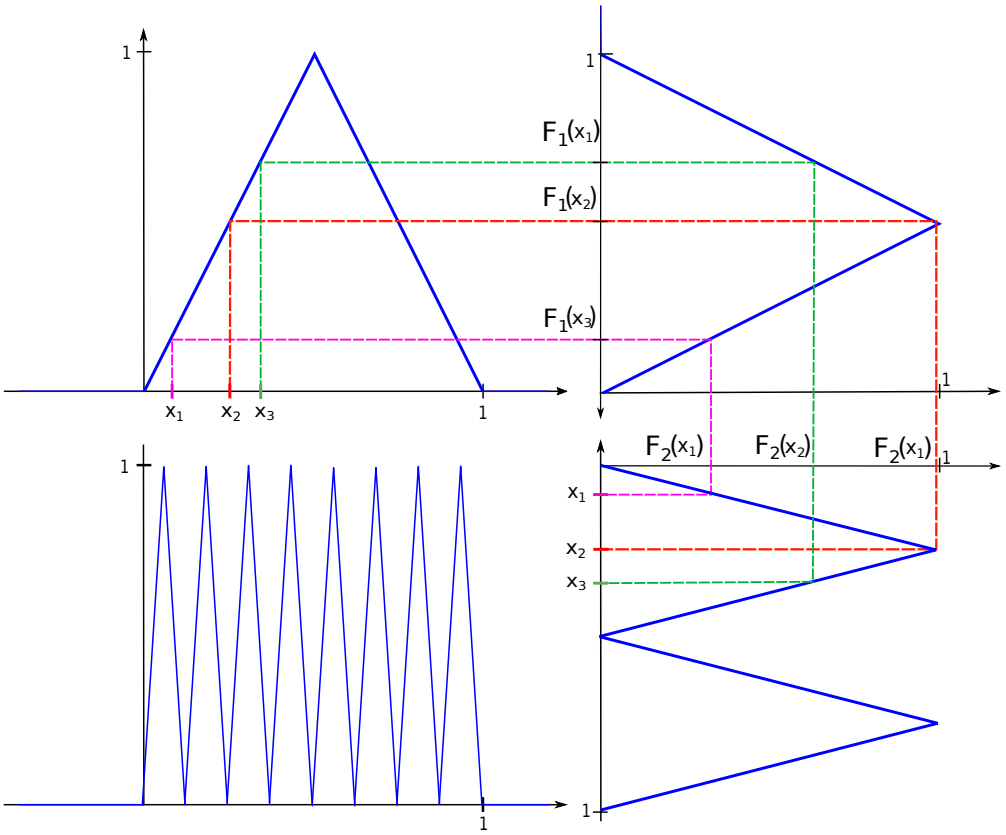

Figure 3.6: **Top Left:** Visualisation of $\mathrm{R}(\Phi^\wedge) = F_1$. **Bottom Right:** Visualisation of $\mathrm{R}(\Phi^\wedge) \circ \mathrm{R}(\Phi^\wedge) = F_2$, **Bottom Left:** F_n **for** $n = 4$**.**

Remark 3.7. *Proposition 3.6 already shows something remarkable. Consider a two layer network Φ with input dimension 1 and N neurons. Then its realisation with ReLU activation function is given by*

$$\mathrm{R}(\Phi) = \sum_{j=1}^{N} c_j \varrho_R(a_j x + b_j) - d,$$

for $c_j, a_j, b_j, d \in \mathbb{R}$. It is clear that $\mathrm{R}(\Phi)$ is piecewise affine linear with at most $N \leq M(\Phi)$ pieces. We see, that with this construction, the resulting networks have not more than $M(\Phi)$ pieces. However, the function F_n from Proposition 3.6 has at least $2^{\frac{M(\Phi)+2}{12}}$ linear pieces.

The function F_n is therefore a function that can be very efficiently represented by deep networks, but not very efficiently by shallow networks. This was first observed in [37].

The surprisingly high number of linear pieces of F_n is not the only remarkable thing about the construction of Proposition 3.6. Yarotsky [40] made the following insightful observation:

Proposition 3.8 ([40])**.** *For every $x \in [0, 1]$ and $N \in \mathbb{N}$, we have that*

$$\left| x^2 - x + \sum_{n=1}^{N} \frac{F_n(x)}{2^{2n}} \right| \leq 2^{-2N-2}. \tag{3.12}$$

Proof. We claim that

$$H_N \coloneqq x - \sum_{n=1}^{N} \frac{F_n}{2^{2n}} \tag{3.13}$$

is a piecewise linear function with breakpoints $k2^{-N}$ where $k = 0, \dots, 2^N$, and $H_N(k2^{-N}) = k^2 2^{-2N}$. We prove this by induction. The result clearly holds for $N = 0$. Assume that the claim holds for $N \in \mathbb{N}$. Then we see that

$$H_N - H_{N+1} = \frac{F_{N+1}}{2^{2N+2}}.$$

Since, by Proposition 3.6, F_{N+1} is piecewise linear with breakpoints $k2^{-N-1}$ where $k = 0, \dots, 2^{N+1}$ and H_N is piecewise linear with breakpoints $\ell 2^{-N-1}$ where $\ell = 0, \dots, 2^{N+1}$ even, we conclude that H_{N+1} is piecewise linear with breakpoints $k2^{-N-1}$ where $k = 0, \dots, 2^{N+1}$. Moreover, by Proposition 3.6, F_{N+1} vanishes for all $k2^{-N-1}$, where k is even. Hence, by the induction hypothesis $H_{N+1}(k2^{-N-1}) = (k2^{-N-1})^2$ for all k even.

To complete the proof, we need to show that

$$\frac{F_{N+1}}{2^{2N+2}}(k2^{-N-1}) = H_N(k2^{-N-1}) - (k2^{-N-1})^2,$$

for all k odd. Since H_N is linear on $[(k-1)2^{-N-1}), (k+1)2^{-N-1})]$, we have that

$$\begin{aligned} H_N(k2^{-N-1}) - (k2^{-N-1})^2 &= \frac{1}{2}\left(((k-1)2^{-N-1})^2 + ((k+1)2^{-N-1})^2\right) - (k2^{-N-1})^2 \\ &= 2^{-2N-2}\left(\frac{1}{2}\left(((k-1))^2 + (k+1)^2\right) - k^2\right) \\ &= 2^{-2(N+1)} = 2^{-2(N+1)} F_{N+1}(k2^{-N-1}), \end{aligned} \tag{3.14}$$

where the last step follows by Proposition 3.6. This shows that $H_{N+1}(k2^{-N-1}) = (k2^{-N-1})^2$ for all $k = 0, \dots, 2^{N+1}$ and completes the induction.

Finally, let $x \in [k2^{-N}, (k+1)2^{-N}]$, $k = 0, \dots, 2^N - 1$, then

$$|H_N(x) - x^2| = H_N - x^2 = (k2^{-N})^2 + \frac{\left((k+1)^2 - k^2\right)2^{-2N}}{2^{-N}}(x - k2^{-N}) - x^2, \tag{3.15}$$

where the first step is because $x \mapsto x^2$ is convex and therefore its graph lies below that of the linear interpolant and the second step follows by representing H_N locally as the linear map that intersects $x \mapsto x^2$ at $k2^{-N}$ and $(k+1)2^{-N}$.

Since (3.15) describes a continuous function on $[k2^{-N}, (k+1)2^{-N}]$ vanishing at the boundary, it assumes its maximum at the critical point

$$x^* \coloneqq \frac{1}{2}\frac{\left((k+1)^2 - k^2\right)2^{-2N}}{2^{-N}} = \frac{1}{2}(2k+1)2^{-N} = (2k+1)2^{-N-1} = \ell 2^{-N-1},$$

for $\ell \in \{1, \dots 2^{N+1}\}$ odd. We have already computed in (3.14) that

$$|H_N(x^*) - (x^*)^2| \leq 2^{-2(N+1)}.$$

This yields the claim. □

Equation 3.12 and Proposition 3.6 make us optimistic that, with sufficiently deep networks, we can approximate the square function very efficiently. Before we can do this properly, we need to enlarge our toolbox slightly and introduce a couple of additional operations on NNs.

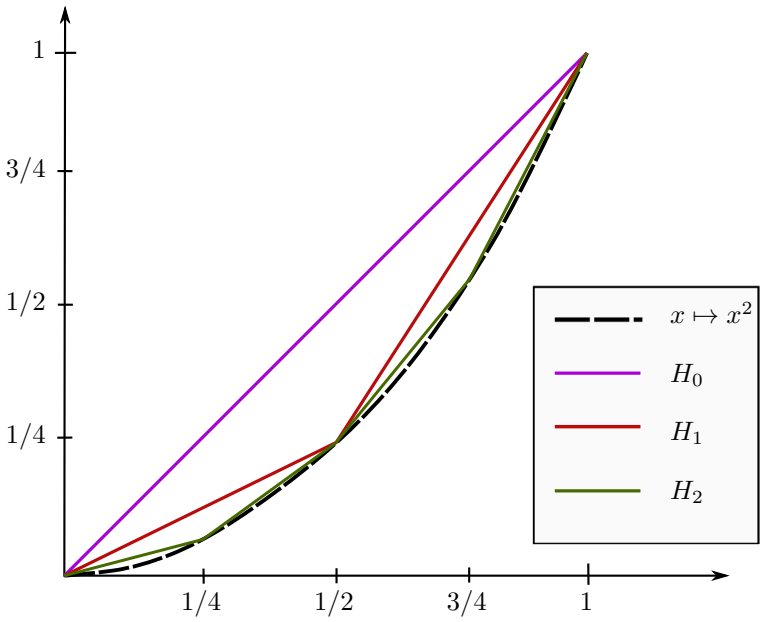

Figure 3.7: Visualisation of the construction of H_N of (3.13).

ReLU specific network operations We saw in Proposition 2.11 that we can approximate the identity function by realisations of NNs for many activation functions. For the ReLU, we can even go one step further and rebuild the identity function exactly.

Lemma 3.9. *Let $d \in \mathbb{N}$, and define*

$$\Phi^{\mathrm{Id}} := ((A_1, b_1), (A_2, b_2))$$

with

$$A_1 := \begin{pmatrix} \mathrm{Id}_{\mathbb{R}^d} \\ -\mathrm{Id}_{\mathbb{R}^d} \end{pmatrix}, \quad b_1 := 0, \quad A_2 := \begin{pmatrix} \mathrm{Id}_{\mathbb{R}^d} & -\mathrm{Id}_{\mathbb{R}^d} \end{pmatrix}, \quad b_2 := 0.$$

Then $\mathrm{R}(\Phi^{\mathrm{Id}}) = \mathrm{Id}_{\mathbb{R}^d}$.

Proof. Clearly, for $x \in \mathbb{R}^d$, $\mathrm{R}(\Phi^{\mathrm{Id}})(x) = \varrho_R(x) - \varrho_R(-x) = x$. □

Remark 3.10. *Lemma 3.9 can be generalised to yield emulations of the identity function with arbitrary numbers of layers. For each $d \in \mathbb{N}$, and each $L \in \mathbb{N}_{\geq 2}$, we define*

$$\Phi^{\mathrm{Id}}_{d,L} := \left(\left(\begin{pmatrix} \mathrm{Id}_{\mathbb{R}^d} \\ -\mathrm{Id}_{\mathbb{R}^d} \end{pmatrix}, 0 \right), \underbrace{(\mathrm{Id}_{\mathbb{R}^{2d}}, 0), \ldots, (\mathrm{Id}_{\mathbb{R}^{2d}}, 0)}_{L-2 \text{ times}}, ([\mathrm{Id}_{\mathbb{R}^d} \mid -\mathrm{Id}_{\mathbb{R}^d}], 0) \right).$$

For $L = 1$, one can simply set $\Phi^{\mathrm{Id}}_{d,1} := ((\mathrm{Id}_{\mathbb{R}^d}, 0))$.

Our first application of the NN of Lemma 3.9 is for a redefinition of the concatentation. Before that, we first convince ourselves that the current notion of concatenation is not adequate if we want to control the number of parameters of the concatenated NN.

Example 3.11. *Let $N \in \mathbb{N}$ and $\Phi = ((A_1, 0), (A_2, 0))$ with $A_1 = [1, \ldots, 1]^T \in \mathbb{R}^{N \times 1}$, $A_2 = [1, \ldots, 1] \in \mathbb{R}^{1 \times N}$. Per definition we have that $M(\Phi) = 2N$.*

Moreover, we have that

$$\Phi \bullet \Phi = ((A_1, 0), (A_1 A_2, 0), (A_2, 0)).$$

It holds that $A_1 A_2 \in \mathbb{R}^{N \times N}$ and every entry of $A_1 A_2$ equals 1. Hence $M(\Phi \bullet \Phi) = N + N^2 + N$.

Example shows that the number of weights of networks behaves quite undesirably under concatenation. Indeed, we would expect that it should be possible to construct a concatenation of networks that implements the composition of the respective realisations and the number of parameters scales *linearly instead of quadratically* in the number of parameters of the individual networks.

Fortunately, Lemma 3.9 enables precisely such a construction, see also Figure 3.8 for an illustration.

Definition 3.12. *Let $L_1, L_2 \in \mathbb{N}$, and let $\Phi^1 = ((A_1^1, b_1^1), \dots, (A_{L_1}^1, b_{L_1}^1))$ and $\Phi^2 = ((A_1^2, b_1^2), \dots, (A_{L_2}^2, b_{L_2}^2))$ be two NNs such that the input layer of Φ^1 has the same dimension d as the output layer of Φ^2. Let Φ^{Id} be as in Lemma 3.9.*

Then the sparse concatenation of Φ^1 and Φ^2 *is defined as*

$$\Phi^1 \odot \Phi^2 := \Phi^1 \bullet \Phi^{\mathrm{Id}} \bullet \Phi^2.$$

Remark 3.13. *It is easy to see that*

$$\Phi^1 \odot \Phi^2 = \left((A_1^2, b_1^2), \dots, (A_{L_2-1}^2, b_{L_2-1}^2), \left(\begin{pmatrix} A_{L_2}^2 \\ -A_{L_2}^2 \end{pmatrix}, \begin{pmatrix} b_{L_2}^2 \\ -b_{L_2}^2 \end{pmatrix} \right), \left(\begin{bmatrix} A_1^1 \,|\, -A_1^1 \end{bmatrix}, b_1^1 \right), (A_2^1, b_2^1), \dots, (A_{L_1}^1, b_{L_1}^1) \right)$$

has $L_1 + L_2$ layers and that $\mathrm{R}(\Phi^1 \odot \Phi^2) = \mathrm{R}(\Phi^1) \circ \mathrm{R}(\Phi^2)$ and $M(\Phi^1 \odot \Phi^2) \leq 2M(\Phi^1) + 2M(\Phi^2)$.

Approximation of the square: We shall now build a NN that approximates the square function on $[0, 1]$. Of course this is based on the estimate (3.12).

Proposition 3.14 ([40, Proposition 2])**.** *Let $1/2 > \epsilon > 0$. There exists a NN $\Phi^{\mathrm{sq},\epsilon}$ such that, for $\epsilon \to 0$,*

$$L(\Phi^{\mathrm{sq},\epsilon}) = \mathcal{O}(\log_2(1/\epsilon)) \tag{3.16}$$

$$M(\Phi^{\mathrm{sq},\epsilon}) = \mathcal{O}(\log_2^2(1/\epsilon)) \tag{3.17}$$

$$\left| \mathrm{R}(\Phi^{\mathrm{sq},\epsilon})(x) - x^2 \right| \leq \epsilon, \tag{3.18}$$

for all $x \in [0, 1]$. In addition, we have that $\mathrm{R}(\Phi^{\mathrm{sq},\epsilon})(0) = 0$.

Proof. By (3.12), we have that, for $N := \lceil -\log_2(\epsilon)/2 \rceil$, it holds that, for all $x \in [0, 1]$,

$$\left| x^2 - x + \sum_{n=1}^{N} \frac{F_n(x)}{2^{2n}} \right| \leq \epsilon. \tag{3.19}$$

We define, for $n = 1, \dots, N$,

$$\Phi_n := \Phi_{1,N-n}^{\mathrm{Id}} \odot (\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}). \tag{3.20}$$

Then we have that $L(\Phi_n) = N - n + L(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) = N + 1$ by Proposition 3.6. Moreover, by Remark 3.13,

$$M(\Phi_n) \leq 2M(\Phi_{1,N-n}^{\mathrm{Id}}) + 2M(\underbrace{\Phi^\wedge \bullet \cdots \bullet \Phi^\wedge}_{n-\text{times}}) \leq 4(N - n) + 2(12n - 2) \leq 24N, \tag{3.21}$$

where the penultimate inequality follows from Remark 3.10 and Proposition 3.6.

Next, we set

$$\Phi^{\mathrm{sq},\epsilon} := \left(\left[1, -1/4, \dots, -2^{-2N} \right], 0 \right) \odot \mathrm{P}\left(\Phi_{1,N+1}^{\mathrm{Id}}, \Phi_1, \dots, \Phi_N \right).$$

Per construction, we have that

$$\mathrm{R}\left(\Phi^{\mathrm{sq},\epsilon} \right)(x) = \mathrm{R}\left(\Phi_{1,N+1}^{\mathrm{Id}} \right)(x) - \sum_{n=1}^{N} 2^{-2n} \mathrm{R}\left(\Phi_j \right)(x) = x - \sum_{n=1}^{N} \frac{F_n(x)}{2^{2n}},$$

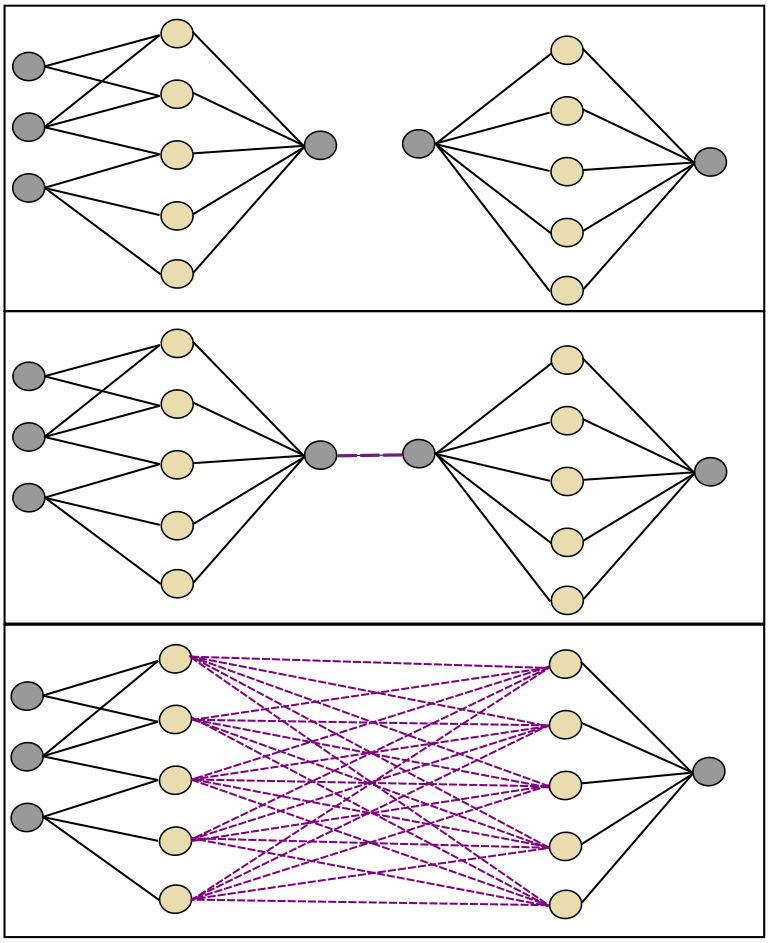

Figure 3.8: **Top:** Two neural networks, **Middle:** Sparse concatenation of the two networks as in Definition 3.12, **Bottom:** Regular concatenation as in Definition 2.9.

and, by (3.19), we conclude (3.18), for all $x \in [0,1]$, and that $\mathrm{R}(\Phi)(0) = 0$. Moreover, we have by Remark 3.13 that

$$L\left(\Phi^{\mathrm{sq},\epsilon}\right) = L\left(\left(\left[1, -1/4, \dots, -2^{-2N}\right], 0\right)\right) + L\left(\mathrm{P}\left(\Phi^{\mathrm{Id}}_{1,N+1}, \Phi_1, \dots, \Phi_N\right)\right) = N + 2 = \lceil -\log_2(\epsilon)/2 \rceil + 2.$$

This shows (3.16). Finally, by Remark 3.13

$$\begin{aligned} M\left(\Phi^{\mathrm{sq},\epsilon}\right) &\leq 2M\left(\left(\left[1, -1/4, \dots, -2^{-2N}\right], 0\right)\right) + 2M\left(\mathrm{P}\left(\Phi^{\mathrm{Id}}_{1,N+1}, \Phi_1, \dots, \Phi_N\right)\right) \\ &= 2(N+1) + 2\left(M\left(\Phi^{\mathrm{Id}}_{1,N+1}\right) + \sum_{n=1}^{N} M\left(\Phi_n\right)\right) \\ &= 2(N+1) + 4(N+1) + 2\sum_{n=1}^{N} M\left(\Phi_n\right) \\ &\leq 6(N+1) + 2\sum_{n=1}^{N} 24N = 6(N+1) + 48N^2, \end{aligned}$$

where we applied (3.21) in the last estimate. Clearly,

$$6(N+1) + 48N^2 = \mathcal{O}\left(N^2\right), \text{ for } N \to \infty,$$

and hence

$$M\left(\Phi^{\mathrm{sq},\epsilon}\right) = \mathcal{O}\left(\log_2^2(1/\epsilon)\right), \text{ for } \epsilon \to 0,$$

which yields (3.17). □

Remark 3.15. *In [31, Theorem 5], a proof of the result above is given that does not require Proposition 3.8, but instead is based on three fascinating ideas:*

- Multiplication can be approximated by finitely many semi-binary multiplications: *For $x \in [0,1]$, we write $x = \sum_{\ell=1}^{\infty} x_\ell 2^{-\ell}$. Then*

$$x \cdot y = \sum_{\ell=1}^{\infty} 2^{-\ell} x_\ell y = \sum_{\ell=1}^{N} 2^{-\ell} x_\ell y + \mathcal{O}(2^{-N}), \text{ for } N \to \infty.$$

- Multiplication on $[0,1]$ by 0 or 1 can be build with a single ReLU: *It holds that*

$$\begin{aligned} \varrho_R(2^{-\ell} y + x_\ell - 1) &= \begin{cases} 2^{-\ell} y & \text{if } x_\ell = 1 \\ 0 & \text{else} \end{cases} \\ &= 2^{-\ell} x_\ell y. \end{aligned}$$

- Extraction of binary representation is efficient: *We have, by Proposition 3.6, that F_ℓ vanishes on all $i2^{-\ell}$ for $i = 0, \dots, 2^\ell$ even and equals 1 on all $i2^{-\ell}$ for $i = 0, \dots, 2^\ell$ odd. Therefore*

$$F_N\left(\sum_{\ell=1}^{N} 2^{-\ell} x_\ell\right) = x_N.$$

  By a short computation this yields that for all $x \in [0,1]$ that $F_N(x - 2^{-N-1}) > 1/2$, if $x_N = 1$ and $F_N(x - 2^{-N-1}) \leq 1/2$, if $x_N = 0$. Hence, by building an approximate Heaviside function $\mathbb{1}_{x \geq 0.5}$ with ReLU realisations of networks, it is clear that one can approximate the map $x \mapsto x_\ell$ for all ℓ.

Building N of the binary multiplications therefore requires N bit extractors and N multipliers by $0/1$. Hence, this requires of the order of N neurons, to achieve an error of 2^{-N}.

3.3 Approximation of smooth functions

With the emulation of the square function on $[0,1]$ we have, in principle, a way of emulating the higher-order sigmoidal function x_+^2 by ReLU networks. As we have seen in Section 2.5, sums and compositions of these functions can be used to approximate smooth functions very efficiently.

Approximation of multiplication: Based on the idea, that we have already seen in the proof of Proposition 2.18, in particular, Equation (2.14), we show how an approximation of a square function yields an approximation of a multiplication operator.

Proposition 3.16. *Let $p \in \mathbb{N}$, $K \in \mathbb{N}$, $\epsilon \in (0, 1/2)$. There exists a NN $\Phi^{\mathrm{mult},p,\epsilon}$ such that for $\epsilon \to 0$*

$$L(\Phi^{\mathrm{mult},p,\epsilon}) = \mathcal{O}(\log_2(K) \cdot \log_2(1/\epsilon)) \tag{3.22}$$

$$M(\Phi^{\mathrm{mult},p,\epsilon}) = \mathcal{O}(\log_2(K) \cdot \log_2^2(1/\epsilon)) \tag{3.23}$$

$$\left| \mathrm{R}(\Phi^{\mathrm{mult},p,\epsilon})(x) - \prod_{\ell=1}^{p} x_\ell \right| \leq \epsilon, \tag{3.24}$$

for all $x = (x_1, x_2, \dots, x_p) \in [-K, K]^p$. Moreover, $\mathrm{R}(\Phi^{\mathrm{mult},p,\epsilon})(x) = 0$ if $x_\ell = 0$ for at least one $\ell \leq p$. Here the implicit constant depends on p only.

Proof. The crucial observation is that, by the parallelogram identity, we have that for $x, y \in [-K, K]$

$$\begin{aligned} x \cdot y &= K^2 \cdot \left(\left(\frac{x+y}{2K} \right)^2 - \left(\frac{x-y}{2K} \right)^2 \right) \\ &= K^2 \left(\left(\frac{\varrho_R(x+y)}{2K} + \frac{\varrho_R(-x-y)}{2K} \right)^2 - \left(\frac{\varrho_R(x-y)}{2K} + \frac{\varrho_R(-x+y)}{2K} \right)^2 \right). \end{aligned}$$

We set

$$\Phi_1 \coloneqq \left(\left(\begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix}, 0 \right), \left(\frac{1}{2K} \cdot \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, 0 \right) \right), \text{ and } \Phi_2 \coloneqq \left(\left(\left[K^2, -K^2 \right], 0 \right) \right).$$

Now we define

$$\Phi^{\mathrm{mult},2,\epsilon} \coloneqq \Phi_2 \odot \mathrm{FP}\left(\Phi^{\mathrm{sq},\epsilon/(2K^2)}, \Phi^{\mathrm{sq},\epsilon/(2K^2)} \right) \odot \Phi_1.$$

It is clear that, for all $x, y \in [-K, K]$ it holds that $|x \pm y|/(2K) \leq 1$ and hence

$$\left| \mathrm{R}\left(\Phi^{\mathrm{mult},2,\epsilon} \right)(x, y) - x \cdot y \right| \leq \epsilon.$$

Moreover, the size of $\Phi^{\mathrm{mult},2,\epsilon}$ is up to a constant that of $\Phi^{\mathrm{sq},\epsilon/K^2}$. Thus (3.23)-(3.24) follow from Proposition 3.14. The construction for $p > 2$ follows by the now well-known strategy of building a binary tree of basic multiplication networks as in Figure 2.4. □

A direct corollary of Proposition 3.16 is the following corollary that we state without proof.

Corollary 3.17. *Let $p \in \mathbb{N}$, $K \in \mathbb{N}$, $\epsilon \in (0, 1/2)$. There exists a NN $\Phi^{\mathrm{pow},p,\epsilon}$ such that, for $\epsilon \to 0$,*

$$\begin{aligned} L(\Phi^{\mathrm{pow},p,\epsilon}) &= \mathcal{O}(\log_2(K) \cdot \log_2(1/\epsilon)) \\ M(\Phi^{\mathrm{pow},p,\epsilon}) &= \mathcal{O}(\log_2(K) \cdot \log_2^2(1/\epsilon)) \\ |\mathrm{R}(\Phi^{\mathrm{pow},p,\epsilon})(x) - x^p| &\leq \epsilon, \end{aligned}$$

for all $x \in [-K, K]$. Moreover, $\mathrm{R}(\Phi^{\mathrm{pow},p,\epsilon})(x) = 0$ if $x = 0$. Here the implicit constant depends on p only.

Approximation of B-splines: Now that we can build a NN the realisation of which is a multiplication of $p \in \mathbb{N}$ scalars, it is not hard to see with (2.10) that we can rebuild cardinal B-splines by ReLU networks.

Proposition 3.18. *Let $d, k, \ell \in \mathbb{N}$, $k \geq 2$, $t \in \mathbb{R}^d$, $1/2 > \epsilon > 0$. There exists a NN $\Phi^d_{\ell,t,k}$ such that for $\epsilon \to 0$*

$$L(d, k) \coloneqq L(\Phi^d_{\ell,t,k}) = \mathcal{O}(\log_2(1/\epsilon)), \tag{3.25}$$

$$M(d, k) \coloneqq M(\Phi^d_{\ell,t,k}) = \mathcal{O}(\log_2^2(1/\epsilon)), \tag{3.26}$$

$$\left| \mathrm{R}(\Phi^d_{\ell,t,k})(x) - \mathcal{N}^d_{\ell,t,k}(x) \right| \leq \epsilon, \tag{3.27}$$

for all $x \in \mathbb{R}^d$.

Proof. Clearly, it is sufficient to show the result for $\ell = 0$ and $t = 0$. We have by (2.10) that

$$\mathcal{N}_k(x) = \frac{1}{(k-1)!} \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} (x - \ell)_+^{k-1}, \text{ for } x \in \mathbb{R}, \tag{3.28}$$

It is well known, see [33], that $\operatorname{supp} \mathcal{N}_k = [0,k]$ and $\|\mathcal{N}_k\|_\infty \leq 1$. Let $\delta > 0$, then we set

$$\Phi_{k,\delta} \coloneqq \left(\left(\frac{1}{(k-1)!} \left[\binom{k}{0}, -\binom{k}{1}, \dots, (-1)^k \binom{k}{k} \right], 0 \right) \right) \odot \mathrm{FP} \left(\underbrace{\Phi^{\mathrm{pow},k-1,\delta}, \dots, \Phi^{\mathrm{pow},k-1,\delta}}_{k+1-\text{times}} \right) \odot ((A_1, b_1), (\mathrm{Id}_{\mathbb{R}^{k+1}}, 0)),$$

where

$$A_1 = [1,1,\dots,1]^T, \qquad b_1 = -[0,1,\dots,k]^T,$$

and $\mathrm{Id}_{\mathbb{R}^{k+1}}$ is the identity matrix on $\mathbb{R}^{k+1}$. Here $K \coloneqq k+1$ in the definition of $\Phi^{\mathrm{pow},k-1,\delta}$ via Corollary 3.17.

It is now clear, that we can find $\delta_\epsilon > 0$ so that

$$|\mathrm{R}(\Phi_{k,\delta_\epsilon})(x) - \mathcal{N}_k(x)| \leq \epsilon/(4d2^{d-1}), \tag{3.29}$$

for $x \in [-k-1, k+1]$. With sufficient care, we see that, we can choose $\delta_\epsilon = \Omega(\epsilon)$, for $\epsilon \to 0$. Hence, we can conclude from Definition 3.12 that $L^{\delta_\epsilon} \coloneqq L(\Phi_{k,\delta_\epsilon}) = \mathcal{O}(L(\Phi^{\mathrm{mult},k+1,\delta_\epsilon})) = \mathcal{O}(\log_2(1/\epsilon))$, and $M(\Phi_{k,\delta_\epsilon}) = \mathcal{O}(\Phi^{\mathrm{mult},k+1,\delta_\epsilon}) \in \mathcal{O}(\log_2^2(1/\epsilon))$, for $\epsilon \to 0$ which yields (3.25) and (3.26). At this point, $\mathrm{R}(\Phi_{k,\delta_\epsilon})$ only accurately approximates $\mathcal{N}_k$ on $[-k-1, k+1]$. To make this approximation global, we multiply $\mathrm{R}(\Phi_{k,\delta_\epsilon})$ with an appropriate indicator function.

Let

$$\Phi^{cut} \coloneqq \left(\left([1,1,1,1]^T, [1,0,-k,-k-1]^T \right), ([1,-1,-1,1], 0) \right).$$

Then $\mathrm{R}(\Phi^{cut})$ is a piecewise linear spline with breakpoints $-1, 0, k, k+1$. Moreover, $\mathrm{R}(\Phi^{cut})$ is equal to 1 on $[0,k]$, vanishes on $[-1,k+1]^c$, and is non-negative and bounded by 1. We define

$$\widetilde{\Phi}_{k,\delta} \coloneqq \Phi^{\mathrm{mult},2,\epsilon/(4d2^{d-1})} \odot \mathrm{P} \left(\Phi_{k,\delta_\epsilon}, \Phi^{\mathrm{Id}}_{1,L^{\delta_\epsilon}-2} \odot \Phi^{cut} \right).$$

Since the realisation of the multiplication is 0 as soon as one of the inputs is zero by Proposition 3.16, we conclude that

$$\left| \mathrm{R} \left(\widetilde{\Phi}_{k,\delta_\epsilon} \right)(x) - \mathcal{N}_k(x) \right| \leq \epsilon/(2d2^{d-1}), \tag{3.30}$$

for all $x \in \mathbb{R}$. Recall that

$$\mathcal{N}^d_{0,0,k}(x) \coloneqq \prod_{j=1}^d \mathcal{N}_k(x_j), \text{ for } x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Now we define

$$\Phi^d_{0,0,k,\epsilon} \coloneqq \Phi^{\mathrm{mult},d,\epsilon/2} \odot \mathrm{FP}(\underbrace{\widetilde{\Phi}_{k,\delta_\epsilon}, \dots, \widetilde{\Phi}_{k,\delta_\epsilon}}_{d-\text{times}}).$$

We have that

$$\left| \mathcal{N}^d_{0,0,k}(x) - \mathrm{R} \left(\Phi^d_{0,0,k,\epsilon} \right)(x) \right| \leq \left| \prod_{j=1}^d \mathcal{N}_k(x_j) - \prod_{j=1}^d \mathrm{R} \left(\widetilde{\Phi}_{k,\delta_\epsilon} \right)(x_j) \right| + \left| \mathrm{R} \left(\Phi^d_{0,0,k,\epsilon} \right)(x) - \prod_{j=1}^d \mathrm{R} \left(\widetilde{\Phi}_{k,\delta_\epsilon} \right)(x_j) \right|.$$

Additionally, we have by (3.30) that

$$\left| \prod_{j=1}^d \mathrm{R} \left(\widetilde{\Phi}_{k,\delta} \right)(x_j) - \mathrm{R} \left(\widetilde{\Phi}^{\mathrm{mult},d,\epsilon/2} \right) \circ \mathrm{R}(\mathrm{FP}(\underbrace{\widetilde{\Phi}_{k,\delta_\epsilon}, \dots, \widetilde{\Phi}_{k,\delta_\epsilon}}_{d-\text{times}}))(x) \right| \leq \epsilon/2,$$

for all $x \in \mathbb{R}^d$. It is clear, by repeated applications of the triangle inequality that for $a_j \in [0,1]$, $b_j \in [-1,1]$, for $j = 1, \dots, d$,

$$\left| \prod_{j=1}^{d} a_j - \prod_{j=1}^{d} (a_j + b_j) \right| \leq d \cdot \left(1 + \max_{j=1,\dots,d} |b_j|\right)^{d-1} \max_{j=1,\dots,d} |b_j| \leq d 2^{d-1} \max_{j=1,\dots,d} |b_j|.$$

Hence,

$$\left| \prod_{j=1}^{d} \mathcal{N}_k(x_j) - \prod_{j=1}^{d} \mathrm{R}\left(\widetilde{\Phi}_{k,\delta_\epsilon}\right)(x_j) \right| \leq \epsilon/2.$$

This yields (3.27). The statement on the size of $\Phi^d_{0,0,k,\epsilon}$ follows from Remark 3.13. □

Approximation of smooth functions: Having established how to approximate arbitrary B-splines with Proposition 3.18, we obtain that we can also approximate all functions that can be written as weighted sums of B-splines with bounded coefficients. Indeed, we can conclude with Theorem 2.16 and with similar arguments as in Theorem 2.14 the following result. Our overall argument to arrive here followed the strategy of [36].

Theorem 3.19. *Let $d \in \mathbb{N}$, $s > \delta > 0$ and $p \in (0, \infty]$. Then there exists a constant $C > 0$ such that, for every $f \in C^s([0,1]^d)$ with $\|f\|_{C^s} \leq 1$ and every $1/2 > \epsilon > 0$, there exists a NN Φ such that*

$$L(\Phi) \leq C \log_2(1/\epsilon), \tag{3.31}$$

$$M(\Phi) \leq C \epsilon^{-\frac{d}{s-\delta}}, \tag{3.32}$$

$$\|f - \mathrm{R}(\Phi)\|_{L^p} \leq \epsilon. \tag{3.33}$$

Here the activation function is the ReLU.

Proof. Let $f \in C^s([0,1]^d)$ with $\|f\|_{C^s} \leq 1$ and let $s > \delta > 0$. By Theorem 2.16 there exist a constant $C > 0$ and, for every $N \in \mathbb{N}$, $c_i \in \mathbb{R}$ with $|c_i| \leq C$ and $B_i \in \mathcal{B}^k$ for $i = 1, \dots, N$ and $k \coloneqq \lceil s \rceil$, such that

$$\left\| f - \sum_{i=1}^{N} c_i B_i \right\|_p \leq C N^{\frac{\delta - s}{d}}.$$

By Proposition 3.18, each of the B_i can be approximated up to an error of $N^{\frac{\delta-s}{d}}/(CN)$ with a NN Φ_i of depth $\mathcal{O}(\log_2(N^{\frac{\delta-s}{d}}/(CN))) = \mathcal{O}(\log_2(N))$ and number of weights $\mathcal{O}(\log_2^2(N^{\frac{\delta-s}{d}}/(CN))) = \mathcal{O}(\log_2^2(N))$ for $N \to \infty$.

We define

$$\Phi_f^N \coloneqq ([c_1, \dots, c_N], 0) \bullet \mathrm{P}(\Phi_1, \dots, \Phi_N).$$

It is not hard to see that, for $N \to \infty$,

$$M(\Phi_f^N) = \mathcal{O}(N \log_2^2(N)) \quad \text{and } L(\Phi_f^N) = \mathcal{O}(\log_2(N)).$$

Additionally, by the triangle inequality

$$\left\| f - \mathrm{R}(\Phi_f^N) \right\|_p \leq 2 N^{\frac{\delta-s}{d}}.$$

To achieve (3.33), we, therefore, need to choose $N = N_\epsilon \coloneqq \lceil (\epsilon/2)^{d/(\delta-s)} \rceil$.

A simple estimate yields that $L(\Phi_f^{N_\epsilon}) = \mathcal{O}(\log_2(1/\epsilon))$ for $\epsilon \to 0$, i.e, (3.31). Moreover, we have that

$$N_\epsilon \log_2^2(N_\epsilon) \leq 4d/(s-\delta)(\epsilon/2)^{d/(\delta-s)} \log_2^2(\epsilon/2) \leq C' \epsilon^{-d/(s-\delta)} \log_2^2(\epsilon),$$

for a constant $C' > 0$. It holds that $\log_2^2(\epsilon) = \mathcal{O}(\epsilon^{-\sigma})$ for every $\sigma > 0$. Hence, for every $\delta' > \delta$ with $s > \delta'$, we have

$$\epsilon^{-d/(s-\delta)} \log_2^2(\epsilon) = \mathcal{O}(\epsilon^{-d/(s-\delta')}), \text{ for } \epsilon \to 0.$$

As a consequence we have that $M(\Phi_f^{N_\epsilon}) = \mathcal{O}(\epsilon^{-d/(s-\delta')})$ for $\epsilon \to 0$. Since δ was arbitrary, this yields (3.32). □

Remark 3.20. • *It was shown in [40] that Theorem 3.19 holds with $\delta = 0$ but with the bound $M(\Phi) \leq C\epsilon^{-d/s} \log_2(1/\epsilon)$. Moreover, it holds for $f \in C^s([-K, K]^d)$ for $K > 0$, but the constant C will then depend on K.*

4 The role of depth

We have seen in the previous results that NNs can efficiently emulate the approximation of classical approximation tools, such as linear finite elements or B-splines. Already in Corollary 2.19, we have seen that deep networks are sometimes more efficient at this task than shallow networks. In Remark 3.7, we found that ReLU-realisations of deep NNs can represent certain saw-tooth functions with N linear pieces using only $\mathcal{O}(\log_2(N))$ many weights, whereas shallow NNs require $\mathcal{O}(N)$ many weights for $N \to \infty$.

In this section, we investigate further examples of representation or approximation tasks that can be performed easily with deep networks but cannot be achieved by small shallow networks or any shallow networks.

4.1 Representation of compactly supported functions

Below we show that compactly supported functions cannot be represented by weighted sums of functions of the form $x \mapsto \varrho_R(\langle a, x\rangle)$, but they can be represented by 3-layer networks. This result is based on [5, Section 3].

Proposition 4.1. *Let $d \in \mathbb{N}, d \geq 2$. The following two statements hold for the activation function ϱ_R:*

- *If $L \geq 3$, then there exists a NN Φ with L layers, such that* $\operatorname{supp} \mathrm{R}(\Phi) = B_{\|.\|_1}(0)$,*
- *If $L \leq 2$, then, for every NN Φ with L layers, such that* $\operatorname{supp} \mathrm{R}(\Phi)$ *is compact, we have that* $\mathrm{R}(\Phi) \equiv 0$.

Proof. It is clear that, for every $x \in \mathbb{R}^d$, we have that

$$\sum_{\ell=1}^{d} (\varrho_R(x_\ell) + \varrho_R(-x_\ell)) = \|x\|_1.$$

Moreover, the function $\varrho_R(1 - \|x\|_1)$ is clearly supported on $B_{\|.\|_1}(0)$. Moreover, we have that $\varrho_R(1 - \|x\|_1)$ can be written as the realisation of a NN with at least 3 layers.

Next we address the second part of the theorem. If $L = 1$, then the set of realisations of NNs contains only affine linear functions. It is clear that the only affine linear function that vanishes on a set of non-empty interior is 0. For $L = 2$, all realisations of NNs have the form

$$x \mapsto \sum_{i=1}^{N} c_i \varrho_R(\langle a_i, x\rangle + b_i) + d, \tag{4.1}$$

for $N \in \mathbb{N}$, $c_i, b_i, d \in \mathbb{R}$ and $a_i \in \mathbb{R}^d$, for $i = 1, \ldots, N$. We assume without loss of generality that all $a_i \neq 0$ otherwise $\varrho_R(\langle a_i, x\rangle + b_i)$ would be constant and one could remove the term from (4.1) by adapting d accordingly.

* Here $\|x\|_p^p := \sum_{k=1}^d |x_k|^p$ for $p \in (0, \infty)$.

We next show that every function of the form (4.1) with compact support vanishes everywhere. For an index i, we have that $\varrho_R(\langle a_i, x\rangle + b_i)$ is not continously differentiable at the hyperplane given by

$$\mathcal{S}_i \coloneqq \left\{ -\frac{ba_i}{\|a_i\|^2} + z \colon z \perp a_i \right\}.$$

Let f be a function of the form (4.1). We define $i \sim j$, if $\mathcal{S}_i = \mathcal{S}_j$. Then we have that, for $J \in \{1, \dots, N\}/\sim$ that $a_i^\perp = a_j^\perp$ for all $i, j \in J$. Hence,

$$\sum_{j \in J} c_j \varrho_R(\langle a_j, x\rangle + b_k),$$

is constant perpendicular to a_j for every $j \in J$. And since the sum is piecewise affine linear, we have that it is either affine linear or not continuously differentiable at every element of $\mathcal{S}_j$. We can write

$$f(x) = \sum_{J \in \{1,\dots,N\}/\sim} \left(\sum_{j \in J} c_j \varrho_R(\langle a_j, x\rangle + b_j) \right) + d.$$

If $i \not\sim j$, then $\mathcal{S}_i$ and $\mathcal{S}_j$ intersect in hyperplanes of dimension $d-2$. Hence, it is clear that, if for at least one $J \in \{1, \dots, N\}/\sim$, $\sum_{j\in J} c_j \varrho_R(\langle a_j, x\rangle + b_j)$ is not linear, then f is not continuously differentiable almost everywhere in $\mathcal{S}_j$ for $j \in J$. Since $\mathcal{S}_j$ is unbounded, this contradicts the compact support assumption on f. On the other hand, if, for all $J \in \{1, \dots, N\}/\sim$, we have that $\sum_{j\in J} c_j \varrho_R(\langle a_j, x\rangle + b_j)$ is affine linear, then f is affine linear. By previous observations we have that this necessitates $f \equiv 0$ to allow compact support of f. □

Remark 4.2. *Proposition 4.1, deals with representability only. However, a similar result is true in the framework of approximation theory. Concretely, two layer networks are inefficient at approximating certain compactly supported functions, that three layer networks can approximate very well, see e.g. [10].*

4.2 Number of pieces

We start by estimating the number of piecewise linear pieces of the realisations of NNs with input and output dimension 1 and L layers. This argument can be found in [37, Lemma 2.1].

Theorem 4.3. *Let $L \in \mathbb{N}$. Let ϱ be piecewise affine linear with p pieces. Then, for every NN Φ with $d = 1, N_L = 1$ and $N_1, \dots, N_{L-1} \leq N$, we have that $\mathrm{R}(\Phi)$ has at most $(pN)^{L-1}$ affine linear pieces.*

Proof. The proof is given via induction over L. For $L = 2$, we have that

$$\mathrm{R}(\Phi) = \sum_{k=1}^{N_1} c_k \varrho(\langle a_k, x\rangle + b_i) + d,$$

where $c_k, a_k, b_i, d \in \mathbb{R}$. It is not hard to see that if f_1, f_2 are piecewise affine linear with n_1, n_2 pieces each, then $f_1 + f_2$ is piecewise affine linear with at most $n_1 + n_2$ pieces. Hence, $\mathrm{R}(\Phi)$ has at most Np many affine linear pieces.

Assume the statement to be proven for $L \in \mathbb{N}$. Let Φ_{L+1} be a NN with $L+1$ layers. We set

$$\Phi_{L+1} =: \left((A_1, b_1), \dots, (A_{L+1}, b_{L+1}) \right).$$

It is clear, that

$$\mathrm{R}(\Phi_{L+1})(x) = A_{L+1}[\varrho(h_1(x)), \dots, \varrho(h_{N_L}(x))]^T + b_{L+1},$$

where for $\ell = 1, \dots, N_L$ each h_ℓ is the realisation of a NN with input and output dimension 1, L layers, and less than N neurons in each layer.

For a piecewise affine linear function f with $\tilde{p}$ pieces, we have that $\varrho \circ f$ has at most $p \cdot \tilde{p}$ pieces. This is because, for each of the $\tilde{p}$ affine linear pieces of f—let us call one of those pieces $A \subset \mathbb{R}$—we have that f is either constant or injective on A and hence $\varrho \circ f$ has at most p linear pieces on A.

By this observation and the induction hypothesis, we conclude that $\varrho \circ h_1$ has at most $p(pN)^{L-1}$ affine linear pieces. Hence,

$$\mathrm{R}(\Phi_{L+1})(x) = \sum_{k=1}^{N_L} (A_{L+1})_k \varrho(h_k(x)) + b_{L+1}$$

has at most $Np(pN)^{L-1} = (pN)^L$ many affine linear pieces. This completes the proof. □

For functions with input dimension more than 1 we have the following corollary.

Corollary 4.4. *Let $L, d \in \mathbb{N}$. Let ϱ be piecewise affine linear with p pieces. Then, for every NN Φ with $N_L = 1$ and $N_1, \dots, N_{L-1} \leq N$, we have that $\mathrm{R}(\Phi)$ has at most $(pN)^{L-1}$ affine linear pieces along every line.*

Proof. Every line in $\mathbb{R}^d$ can be parametrized by $\mathbb{R} \ni t \mapsto x_0 + tv$ for $x_0, v \in \mathbb{R}^d$. For Φ as in the statement of corollary, we have that

$$\mathrm{R}(\Phi)(x_0 + tv) = \mathrm{R}(\Phi \bullet \Phi_0)(t),$$

where $\Phi_0 = ((v, x_0))$, which gives the result via Theorem 4.3. □

4.3 Approximation of non-linear functions

Through the bounds on the number of pieces of a realisation of a NN with an piecewise affine linear activation function, we can deduce a limit on approximability through NNs with bounds on the width and numbers of layers for certain non-linear functions. This is based on the following observation, which can, e.g., be found in [11].

Proposition 4.5. *Let $f \in C^2([a,b])$, for $a < b < \infty$ so that f is not affine linear, then there exists a constant $c = c(f) > 0$ so that, for every $p \in \mathbb{N}$,*

$$\|g - f\|_\infty > cp^{-2},$$

for all g which are piecewise affine linear with at most p pieces.

From this argument, we can now conclude the following lower bound to approximating functions which are not affine linear by realisations of NNs with fixed numbers of layers.

Theorem 4.6. *Let $d, L, N \in \mathbb{N}$, and $f \in C^2([0,1]^d)$, where f is not affine linear. Let $\varrho : \mathbb{R} \to \mathbb{R}$ be piecewise affine linear with p pieces. Then for every NN with L layers and fewer than N neurons in each layer, we have that*

$$\|f - \mathrm{R}(\Phi)\|_\infty \geq c(pN)^{-2(L-1)}.$$

Proof. Let $f \in C^2([0,1]^d)$ and non-linear. Then it is clear that there exists a point x_0 and a vector v so that $t \mapsto f(x_0 + tv)$ is non-linear in $t = 0$.

We have that, for every NN Φ with d-dimensional input, one-dimensional output, L layers, and fewer than N neurons in each layer that

$$\|f - \mathrm{R}(\Phi)\|_\infty \geq \|f(x_0 + \cdot v) - \mathrm{R}(\Phi)(x_0 + \cdot v)\|_\infty \geq c \cdot (pN)^{-2(L-1)},$$

where the last estimate is by Corollary 4.4 and Proposition 4.5. □

Remark 4.7. *Theorem 4.6 shows that Theorem 3.19 would not hold with a fixed, bounded number of layers L as soon as s sufficiently large. In other words, for very smooth functions, shallow networks yield suboptimal approximation rates.*

Moreover, no twice continuously differentiable and non-linear function can be approximated with an error that decays with a super polynomial rate in the number of neurons by NNs with a fixed number of layers. In particular, the approximation rate of Proposition 3.14 is not achievable by sequences of NNs of fixed finite depth.

5 High dimensional approximation

At this point we have seen two things on an abstract level. Deep NNs can approximate functions as well as basically every classical approximation scheme. Shallow NNs do not perform as well as deep NNs in many problems. From these observations we conclude that deep networks are preferable over shallow networks, but we do not see why we should not use a classical tool, such as B-splines in applications instead. What is it that makes deep NNs better than classical tools?

One of the advantages will become clear in this section. As it turns out, deep NNs are quite efficient in approximating high dimensional functions.

5.1 Curse of dimensionality

The *curse of dimensionality* is a term introduced by Bellman [3] which is commonly used to describe an exponentially increasing difficulty of problems with increasing dimension. A typical example is that of function interpolation. We define the following function class, for $d \in \mathbb{N}$,

$$\mathcal{F}_d \coloneqq \left\{ f \in C^\infty([0,1]^d) \colon \sup_{|\alpha|=1} \|D^\alpha f\| \leq 1 \right\}.$$

If one defines $e(n,d)$ as the smallest number such that there exists an algorithm reconstructing every $f \in \mathcal{F}_d$ up to an error of $e(n,d)$ from n point evaluations of f, then

$$e(n,d) = 1$$

for all $n \leq 2^{\lfloor d/2 \rfloor} - 1$, see [21]. As a result, in any statement of the form

$$e(n,d) \leq C_{d,r} n^{-r},$$

we have that the constant $C_{d,r}$ depends exponentially on d.

Another instance of this principle can be observed when approximating non-smooth functions. For example, in Theorem 2.16, we saw that the approximation rate, when approximating functions $f \in C^s([0,1]^d)$ deteriorates exponentially with the dimension d. In fact, the approximation rates of Theorem 2.16 are, up to the δ, optimal under some very reasonable assumptions on the approximation scheme, see [9] and discussions later in the manuscript. Hence, there is a fundamental lower bound on approximation capabilities of any approximation scheme that increases exponentially with the dimension.

Careful inspection of the arguments above show that these arguments also apply to approximation by deep NNs. Hence, whenever we say below, that NNs *overcome the curse of dimensionality* then we mean that under a certain additional assumption on the functions to approximate, we will not see a terrible dependence of the approximation rate on the dimension.

5.2 Hierarchy assumptions

We have seen in Corollary 2.19 and Theorem 3.19 that, to approximate a C^s regular function by a NN with a higher-order sigmoidal function or a ReLU as activation function up to an accuracy $\epsilon > 0$, we need essentially $\mathcal{O}(\epsilon^{-d/s})$ many weights. In contrast to that, a d-dimensional function f so that $f(x) = \sum_{i=1}^d g_i(x_i)$, where all the g_i are one dimensional can be approximated using essentially $d\mathcal{O}(\epsilon^{-1/s})$ many weights, which is asymptotically much less than $\mathcal{O}(\epsilon^{-d/s})$ for $\epsilon \to 0$.

It is, therefore, reasonable to assume that high dimensional functions that are build from lower dimensional functions in a way that can be emulated well with NNs, can be much more efficiently approximated than high dimensional functions without this structure.

This observation was used in [27] to study approximation of so-called compositional functions. The definition of these functions is based on special types of graphs.

Definition 5.1. *Let $d, k, N \in \mathbb{N}$ and let $\mathcal{G}(d, k, N)$ be the set of directed acyclic graphs with N vertices, where the indegree of every vertex is at most k and the outdegree of all but one vertex is at least 1 and the indegree of exactly d vertices is 0.*

For $G \in \mathcal{G}(d, k, N)$, let $(\eta_i)_{i=1}^N$ be a topological ordering of G. In other words, every edge $\eta_i\eta_j$ in G satisfies $i < j$. Moreover, for each $i > d$ we denote

$$T_i := \{j : \eta_j\eta_i \text{ is an edge of } G\},$$

and $d_i = \#T_i \leq k$.

With the necessary graph theoretical framework established, we can now define sets of hierarchical functions.

Definition 5.2. *Let $d, k, N, s \in \mathbb{N}$. Let $G \in \mathcal{G}(d, k, N)$ and let, for $i = d+1, \dots, N$, $f_i \in C^s(\mathbb{R}^{d_i})$ with $\|f_i\|_{C^s(\mathbb{R}^{d_i})} \leq 1$*. For $x \in \mathbb{R}^d$, we define, for $i = 1, \dots, d$, $v_i = x_i$ and, for $i = d+1, \dots, N$, $v_i(x) = f_i(v_{j_1}(x), \dots, v_{j_{d_i}}(x))$, where $j_1, \dots, j_{d_i} \in T_i$ and $j_1 < j_2 < \cdots < j_{d_i}$.*

We call the function

$$f : [0,1]^d \to \mathbb{R}, \qquad x \mapsto v_N(x)$$

a compositional function associated to G with regularity s. *We denote the set of compositional functions associated to any graph in $\mathcal{G}(d, k, N)$ with regularity s by $\mathcal{CF}(d, k, N; s)$.*

We present a visualisation of three types of graphs in Figure 5.1. While we have argued before that it is reasonable to expect that NNs can efficiently approximate these types of functions, it is not entirely clear why this is a relevant function class to study. In [20, 27], it is claimed that these functions are particularly close to the functionality of the human visual cortex. In principle, the visual cortex works by first analysing very localised features of a scene and then combining the resulting responses in more and more abstract levels to yield more and more high-level descriptions of the scene.

If the inputs of a function correspond to spatial locations, e.g., come from several sensors, such as in weather forecasting, then it might make sense to model this function as network of functions that first aggregate information from spatially close inputs before sending the signal to a central processing unit.

Compositional functions can also be compared with Boolean circuits comprised of simple logic gates.

Let us now show how well functions from $\mathcal{CF}(d, k, N; s)$ can be approximated by ReLU NNs. Here we are looking for an approximation rate that increases with s and, hopefully, does not depend too badly on d.

Theorem 5.3. *Let $d, k, N, s \in \mathbb{N}$. Then there exists a constant $C > 0$ such that for every $f \in \mathcal{CF}(d, k, N; s)$ and every $1/2 > \epsilon > 0$ there exists a NN Φ_f with*

$$L(\Phi_f) \leq CN^2 \log_2(k/\epsilon) \tag{5.1}$$

$$M(\Phi_f) \leq CN^4 (2k)^{\frac{kN}{s}} \epsilon^{-\frac{k}{s}} \log_2(k/\epsilon) \tag{5.2}$$

$$\|f - \mathrm{R}(\Phi_f)\|_\infty \leq \epsilon, \tag{5.3}$$

where the activation function is the ReLU.

Proof. Let $f \in \mathcal{CF}(d, k, N; s)$ and let, for $i = d+1, \dots, N$, $f_i \in C^s(\mathbb{R}^{d_i})$ be according to Definition 5.2. By Theorem 3.19 and Remark 3.20, we have that there exists a constant $C > 0$ and NNs Φ_i such that

$$|\mathrm{R}(\Phi_i)(x) - f_i(x)| \leq \frac{\epsilon}{(2k)^N}, \tag{5.4}$$

for all $x \in [-2, 2]^{d_i}$ and $L(\Phi_i) \leq CN \log_2(k/\epsilon)$ and

$$M(\Phi_i) \leq C\epsilon^{-d_i/s} (2k)^{\frac{d_i N}{s}} N \log_2(k/\epsilon) \leq C\epsilon^{-k/s} (2k)^{\frac{kN}{s}} N \log_2(k/\epsilon).$$

*The restriction $\|f_i\|_{C^s(\mathbb{R}^{d_i})} \leq 1$ could be replaced by $\|f_i\|_{C^s(\mathbb{R}^{d_i})} \leq \kappa$ for a $\kappa > 1$, and Theorem 5.3 below would still hold up to some additional constants depending on κ. This would, however, significantly increase the technicalities and obfuscate the main ideas in Theorem 5.3.

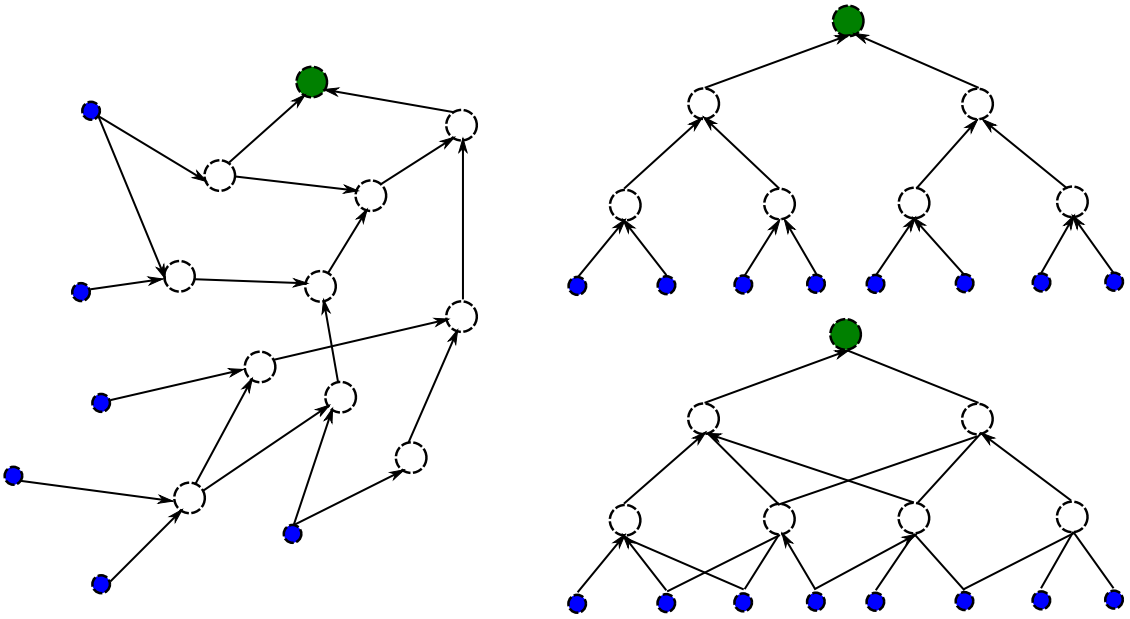

Figure 5.1: Three types of graphs that could be the basis of compositional functions. The associated functions are composed of two or three dimensional functions only.

For $i = d+1, \ldots, N$, let P_i be the orthogonal projection from $\mathbb{R}^{i-1}$ to the components in T_i, i.e, for $T_i =: \{j_1, \ldots, j_{d_i}\}$, where $j_1 < \cdots < j_{d_i}$, we set $P_i((x_k)_{k=1}^{i-1}) = (x_{j_k})_{k=1}^{d_i}$.

Now we define for $j = d+1, \ldots, N-1$,

$$\widetilde{\Phi}^j \coloneqq \mathrm{P}\left(\Phi^{\mathrm{Id}}_{j-1, L(\Phi_j)}, \Phi_j \bullet P_j\right),$$

and

$$\widetilde{\Phi}^N \coloneqq \Phi_N \bullet P_N.$$

Moreover,

$$\Phi_f \coloneqq \widetilde{\Phi}^N \odot \widetilde{\Phi}^{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}.$$

We first analyse the size of Φ_f. It is clear that

$$L(\Phi_f) \leq N \max_{j=d+1}^{N} L\left(\widetilde{\Phi}^j\right) \leq N \max_{j=d+1}^{N} L(\Phi_j) \leq CN^2 \log_2(k/\epsilon),$$

which yields (5.1). Additionally, since

$$\begin{aligned} M\left(\widetilde{\Phi}^N \odot \widetilde{\Phi}^{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right) \leq\ & 2M\left(\widetilde{\Phi}^N \odot \widetilde{\Phi}^{N-1} \odot \ldots \odot \widetilde{\Phi}^{\lceil (N+d+1)/2 \rceil}\right) \\ & + 2M\left(\widetilde{\Phi}^{\lceil (N+d+1)/2 \rceil - 1} \odot \ldots \odot \widetilde{\Phi}^{N+d+1/2}\right), \end{aligned}$$

we have that

$$M(\Phi_f) \lesssim 2^{\lceil \log_2(N) \rceil} N \max_{j=d+1}^{N} M\left(\widetilde{\Phi}^j\right) \lesssim N^2 \max_{j=d+1}^{N} M\left(\widetilde{\Phi}^j\right). \tag{5.5}$$

Furthermore,

$$\begin{aligned}\max_{j=d+1}^{N} M\left(\widetilde{\Phi}^j\right) &\leq \max_{j=d+1}^{N-1} M\left(\Phi^{\mathrm{Id}}_{j-1,L(\Phi_j)}\right) + \max_{j=d+1}^{N} M(\Phi_j)\\ &\leq 2NL\left(\Phi_j\right) + \max_{j=d+1}^{N} M\left(\Phi_j\right)\\ &\leq 2CN^2\log_2(k/\epsilon) + C\epsilon^{-k/s}(2k)^{Nk/s}N\log_2(k/\epsilon),\end{aligned}$$

where the penultimate estimate follows by Remark 3.10. Therefore, by (5.5),

$$M(\Phi_f) \lesssim \epsilon^{-k/s}(2k)^{Nk/s}N^4\log_2(k/\epsilon),$$

which implies (5.2).

Finally, we prove (5.3). We claim that for $N > j > d$ in the notation of Definition 5.2, for $x \in [0,1]^d$,

$$\left|\mathrm{R}\left(\widetilde{\Phi}^j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - [v_1(x), v_2(x), \ldots, v_j(x)]\right| \leq \epsilon/\left(2k\right)^{N-j}. \tag{5.6}$$

We prove (5.6) by induction. Since the realisation of $\Phi^{\mathrm{Id}}_{d,L(\Phi_{d+1})}$ is the identity, we have, by construction that $(\mathrm{R}(\widetilde{\Phi}^{d+1})(x))_k = v_k(x)$ for all $k \leq d$. Moreover, by (5.4), we have that

$$\left|\left(\mathrm{R}\left(\widetilde{\Phi}^{d+1}\right)(x)\right)_{d+1} - v_{d+1}(x)\right| = \left|\left(\mathrm{R}\left(\widetilde{\Phi}^{d+1}\right)(x)\right)_{d+1} - f_{d+1}(x)\right| \leq \epsilon/\left(2k\right)^N.$$

Assume, for the induction step, that (5.6) holds for $N-1 > j > d$.

Again, since the identity is implemented exactly, we have by the induction hypothesis that, for all $k \leq j$,

$$\left|\left(\mathrm{R}\left(\widetilde{\Phi}^{j+1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x)\right)_k - v_k(x)\right| \leq \epsilon/\left(2k\right)^{N-j}.$$

Moreover, we have that $v_{j+1}(x) = f_{j+1}(P_{j+1}([v_1(x), \ldots, v_j(x)]))$. Hence,

$$\begin{aligned}&\left|\left(\mathrm{R}\left(\widetilde{\Phi}^{j+1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x)\right)_{j+1} - v_{j+1}(x)\right|\\ &= \left|\mathrm{R}\left(\Phi_{j+1}\right) \circ P_{j+1} \circ \mathrm{R}\left(\widetilde{\Phi}_j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - v_{j+1}(x)\right|\\ &\leq \left|\mathrm{R}\left(\Phi_{j+1}\right) \circ P_{j+1} \circ \mathrm{R}\left(\widetilde{\Phi}_j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - f_{j+1} \circ P_{j+1} \circ \mathrm{R}\left(\widetilde{\Phi}_j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x)\right|\\ &\quad + \left|f_{j+1} \circ P_{j+1} \circ \mathrm{R}\left(\widetilde{\Phi}_j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - f_{j+1} \circ P_{j+1} \circ [v_1(x), \ldots, v_j(x)]\right| =: \mathrm{I} + \mathrm{II}.\end{aligned}$$

Per (5.4), we have that $\mathrm{I} \leq \epsilon/(2k)^N$ (Note that $P_{j+1} \circ \mathrm{R}\left(\widetilde{\Phi}^j \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) \subset [-2,2]^{d_{j+1}}$ by the induction hypothesis). Moreover, since every partial derivative of f_{j+1} is bounded in absolute value by 1 we have that $\mathrm{II} \leq d_{j+1}\epsilon/\left((2k)^{N-j}\right) \leq \epsilon/\left(2(2k)^{N-j-1}\right)$ by the induction assumption. Hence $\mathrm{I} + \mathrm{II} \leq \epsilon/(2k)^{N-j-1}$

Finally, we compute

$$\begin{aligned}&\left|\mathrm{R}\left(\widetilde{\Phi}^N \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - v_N(x)\right|\\ &= \left|\mathrm{R}\left(\Phi_N\right) \circ P_N \circ \mathrm{R}\left(\widetilde{\Phi}_{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - v_N(x)\right|\\ &\leq \left|\mathrm{R}\left(\Phi_N\right) \circ P_N \circ \mathrm{R}\left(\widetilde{\Phi}_{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - f_N \circ P_N \circ \mathrm{R}\left(\widetilde{\Phi}_{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x)\right|\\ &\quad + \left|f_N \circ P_N \circ \mathrm{R}\left(\widetilde{\Phi}_{N-1} \odot \ldots \odot \widetilde{\Phi}^{d+1}\right)(x) - f_N \circ P_N \circ [v_1(x), \ldots, v_{N-1}(x)]\right| =: \mathrm{III} + \mathrm{IV}.\end{aligned}$$

Using the exact same argument as for estimating I and II above, we conclude that

$$\mathrm{III} + \mathrm{IV} \leq \epsilon,$$

which yields (5.3). □

Remark 5.4. *Theorem 5.3 shows what we had already conjectured earlier. The complexity of approximating a compositional function depends asymptotically not on the input dimension d, but on the maximum indegree of the underlying graph.*

We also see that, while the convergence rate does not depend on d, the constants in (5.2) *are potentially very large. In particular, for fixed s the constants grow superexponentially with k.*

5.3 Manifold assumptions

Realisations of deep NNs are, by definition, always functions on a d dimensional euclidean space. Of course, we may only care about the values that this function takes on subsets of this space. For example, we may only study approximation by NNs on compact subsets of $\mathbb{R}^d$. In this manuscript, we have mostly studied this setup for compact subsets of the form $[A, B]^d$, where $A < B$.

Another approach could be, that we only care about the approximation of functions that live on low dimensional submanifolds $\mathcal{M} \subset \mathbb{R}^d$. In applications, such as image classification, it is conceivable that the input data, only come from the (potentially) low dimensional submanifold of natural images. In that context, it is clear that the approximation properties of NNs are only interesting to us on that submanifold. In other words, we would not care about the behaviour of a NN on inputs that are just unstructured combinations of pixel values.

For a function $f\colon \mathcal{M} \to \mathbb{R}^n$ and $\epsilon > 0$, we now search for a NN Φ with input dimension d and output dimension n, such that

$$|f(x) - \mathrm{R}(\Phi)(x)| \leq \epsilon, \text{ for all } x \in \mathcal{M}.$$

If $\mathcal{M}$ is a d'-dimensional manifold with $d' < d$, and $f \in C^n(\mathcal{M})$, then we would expect to be able to obtain an approximation rate by NNs, that does not depend on d but on d'.

To obtain such a result, we first make a convenient definition of certain types of submanifolds of $\mathbb{R}^d$.

Definition 5.5. *Let $\mathcal{M}$ be a smooth d'-dimensional submanifold of $\mathbb{R}^d$. For $N \in \mathbb{N}, \delta > 0$, we say that $\mathcal{M}$ is (N, δ)-covered, if there exist $x_1, \dots, x_N \in \mathcal{M}$ and such that*

- $\bigcup_{i=1}^N B_{\delta/2}(x_i) \supset \mathcal{M}$
- *the projection*
$$P_i\colon \mathcal{M} \cap B_\delta(x_i) \to T_{x_i}\mathcal{M}$$
is injective and smooth and
$$P_i^{-1}\colon P_i(\mathcal{M} \cap B_\delta(x_i)) \to \mathcal{M}$$
is smooth.

Here $T_{x_i}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at x_i. See Figure 5.2 for a visualisation. We identify $T_{x_i}\mathcal{M}$ with $\mathbb{R}^{d'}$ in the sequel.

Next, we need to define spaces of smooth functions on $\mathcal{M}$. For $k \in \mathbb{N}$, a function f on $\mathcal{M}$ is k-times continuously differentiable if $f \circ \varphi^{-1}$ is k-times continuously differentiable for every coordinate chart φ. If $\mathcal{M}$ is (N, δ) covered, then we can even introduce a convenient C^k- norm on the space of k-times continuously differentiable functions on $\mathcal{M}$ by

$$\|f\|_{C^k,\delta,N} \coloneqq \sup_{i=1,\dots,N} \left\|f \circ P_i^{-1}\right\|_{C^k(P_i(\mathcal{M}\cap B_\delta(x_i)))}.$$

With this definition, we can have the following result which is similar to a number of results in the literature, such as [34, 35, 6, 32].

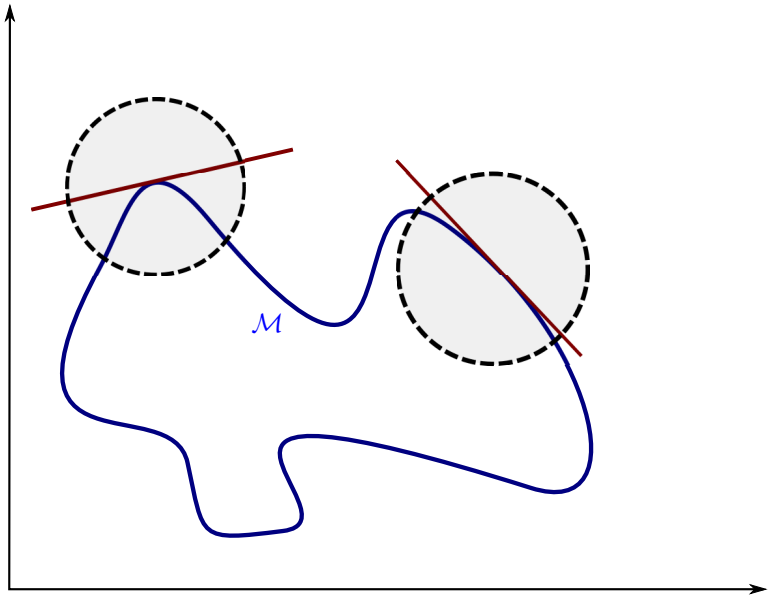

Figure 5.2: One dimensional manifold embedded in 2D. For two points the tangent space is visualised in red. The two circles describe areas where the projection onto the tangent space is invertible and smooth.

Theorem 5.6. *Let $d, k \in \mathbb{N}$, $\mathcal{M} \subset \mathbb{R}^d$ be a (N, δ)-covered d'-dimensional manifold for an $N \in \mathbb{N}$ and $\delta > 0$. Then there exists a constant $c > 0$, such that, for every $\epsilon > 0$, and $f \in C^k(\mathcal{M}, \mathbb{R})$ with $\|f\|_{C^k,\delta,N} \leq 1$, there exists a NN Φ, such that*

$$\begin{aligned} \|f - \mathrm{R}(\Phi)\|_\infty &\leq \epsilon, \\ M(\Phi) &\leq c \cdot \left(\epsilon^{-\frac{d'}{k}} \log_2(1/\epsilon)\right) \\ L(\Phi) &\leq c \cdot (\log_2(1/\epsilon)) \,. \end{aligned}$$

Here the activation function is the ReLU.

Proof. The proof is structured in two parts. First we show a convenient alternative representation of f, then we construct the associated NN.

Step 1: Since $\mathcal{M}$ is (N, δ)-covered, there exists $B > 0$ such that $\mathcal{M} \subset [-B, B]^d$.

Let $\mathcal{T}$ be a simplicial mesh on $[-B, B]^d$ so that for all nodes $\eta_i \in \mathcal{T}$ we have that

$$G(i) \subset B_{\delta/8}(\eta_i).$$

See (3.2) for the definition of $G(i)$ and Figure 5.3 for a visualisation of $\mathcal{T}$.

By Proposition 3.1, we have that

$$1 = \sum_{i=1}^{M_N} \phi_{i,\mathcal{T}}.$$

We denote

$$I_{\mathcal{M}} \coloneqq \{i = 1, \dots, M_N \colon \mathrm{dist}(\eta_i, \mathcal{M}) \leq \delta/8\},$$

where $\mathrm{dist}(a, \mathcal{M}) = \min_{y \in \mathcal{M}} |a - y|$. Per construction, we have that

$$1 = \sum_{i \in I_{\mathcal{M}}} \phi_{i,\mathcal{T}}(x), \quad \text{for all } x \in \mathcal{M}.$$

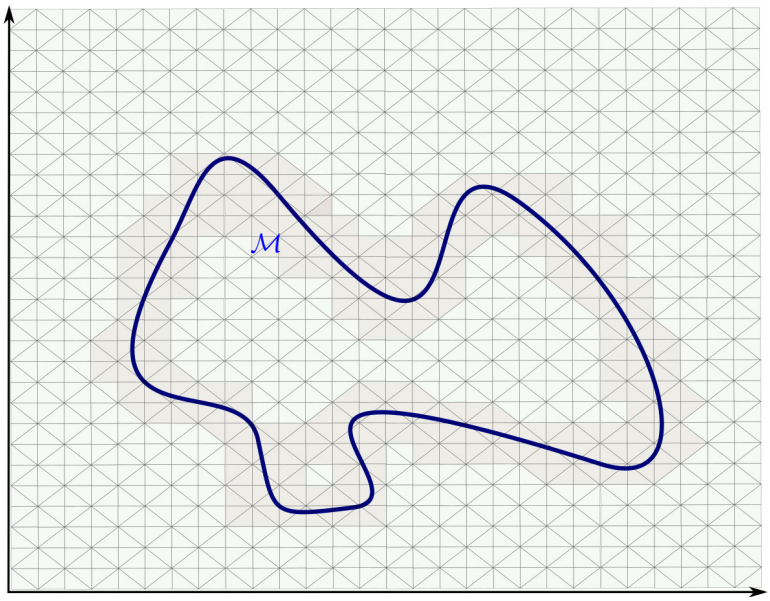

Figure 5.3: Construction of mesh and choice of $I_{\mathcal{M}}$ for a given manifold $\mathcal{M}$

In Figure 5.3, we highlight the cells corresponding to $I_{\mathcal{M}}$.

Moreover, by Definition 5.5, there exist $x_1 \dots x_N \in \mathcal{M}$ such that $\bigcup_{i=1}^N B_{\delta/2}(x_i) \supset \mathcal{M}$. Hence, $\eta_i \in \bigcup_{i=1}^N B_{5\delta/8}(x_i)$ for all $i \in I_{\mathcal{M}}$. Thus, for each η_i there exists $j(i) \in \{1, \dots, N\}$ such that $B_{\delta/8}(\eta_i) \subset B_{3\delta/4}(x_{j(i)})$.

We rewrite f as follows: For $x \in \mathcal{M}$, we have that

$$\begin{aligned} f(x) &= \sum_{i \in I_{\mathcal{M}}} \phi_{i,\mathcal{T}}(x) \cdot f(x) \\ &= \sum_{i \in I_{\mathcal{M}}} \phi_{i,\mathcal{T}}(x) \cdot \left(f \circ P_{j(i)}^{-1} \circ P_{j(i)}(x) \right) \\ &=: \sum_{i \in I_{\mathcal{M}}} \phi_{i,\mathcal{T}}(x) \cdot \left(f_{j(i)} \circ P_{j(i)}(x) \right), \end{aligned} \tag{5.7}$$

where $f_i : P_i(\mathcal{M} \cap B_\delta(x_i)) \to \mathbb{R}$ has C^k norm bounded by 1. We have that

$$P_i(\mathcal{M} \cap B_{3\delta/4}(x_i)) \subset \overline{P_i(\mathcal{M} \cap B_{3\delta/4}(x_i))} \subset P_i(\mathcal{M} \cap B_{7\delta/8}(x_i))$$

and $P_i(\mathcal{M} \cap B_{3\delta/4}(x_i))$, $P_i(\mathcal{M} \cap B_{7\delta/8}(x_i))$ are open. By a C^∞ version of the Urysohn Lemma, there exists a smooth function $\sigma : \mathbb{R}^d \to [0,1]$ such that $\sigma = 1$ on $\overline{P_i(\mathcal{M} \cap B_{3\delta/4}(x_i))}$ and $\sigma = 0$ on $(P_i(\mathcal{M} \cap B_{7\delta/8}(x_i)))^c$.

We define

$$\tilde{f}_i := \begin{cases} \sigma f_i & \text{for } x \in P_i(\mathcal{M} \cap B_\delta(x_i)) \\ 0 & \text{else.} \end{cases}$$

It is not hard to see that $\tilde{f}_i \in C^k(\mathbb{R}^{d'})$ with $\|f\|_{C^k} \leq C_{\mathcal{M}}$, where $C_{\mathcal{M}}$ is a constant depending on $\mathcal{M}$ only and $\tilde{f}_i = f_i$ on $P_i(\mathcal{M} \cap B_{3\delta/4}(x_i))$. Hence, with (5.7), we have that

$$f(x) = \sum_{i \in I_{\mathcal{M}}} \phi_{i,\mathcal{T}}(x) \cdot \left(\tilde{f}_{j(i)} \circ P_{j(i)}(x) \right). \tag{5.8}$$

Step 2: The form of f given by (5.8) suggests a simple way to construct a ReLU approximation of f.

First of all, for every $i \in I_{\mathcal{M}}$, we have that $P_{j(i)}$ is an affine linear map from $[-B,B]^d$ to $\mathbb{R}^{d'}$. We set $\Phi_i^P \coloneqq ((A_1^i, b_1^i))$, where A_1^i, b_1^i are such that $A_1^i x + b_1^i = P_{j(i)}(x)$ for all $x \in \mathbb{R}^d$.

Let $K > 0$ be such that $P_i(\mathcal{M}) \subset [-K,K]^{d'}$ for all $i \in I_{\mathcal{M}}$. For every $i \in I_{\mathcal{M}}$, we have by Theorem 3.19 and Remark 3.20 that for every $\epsilon_1 > 0$ there exists a NN Φ_i^f such that, for all $x \in [-K,K]^{d'}$,

$$\begin{aligned} \left|\tilde{f}_{j(i)}(x) - \mathrm{R}\left(\Phi_i^f\right)(x)\right| &\leq \epsilon_1, \\ M\left(\Phi_i^f\right) &\lesssim \epsilon_1^{-d'/k}\log_2(1/\epsilon_1), \qquad (5.9) \\ L\left(\Phi_i^f\right) &\lesssim \log_2(1/\epsilon_1). \qquad (5.10) \end{aligned}$$

Per Proposition 3.3, there exists, for every $i \in I_{\mathcal{M}}$, a neural network Φ_i^ϕ with

$$\mathrm{R}\left(\Phi_i^\phi\right) = \phi_{i,\mathcal{T}}, \qquad M\left(\Phi_i^\phi\right), L\left(\Phi_i^\phi\right) \lesssim 1, \qquad (5.11)$$

with a constant depending on d.

Now we define, with Proposition 3.16, for $\epsilon_2 > 0$,

$$\Phi_i^{\phi(fP)} \coloneqq \Phi^{\mathrm{mult},2,\epsilon_2} \odot \mathrm{P}\left(\Phi_{1,L^*}^{\mathrm{Id}} \odot \Phi_i^\phi, \Phi_i^f \odot \Phi_i^P\right),$$

where $L^* \coloneqq L(\Phi_i^f \odot \Phi_i^P) - L(\Phi_i^\phi)$. At this point, we assume that $L^* \geq 0$. If $L(\Phi_i^f \odot \Phi_i^P) < L(\Phi_i^\phi)$, then one could instead extend Φ_i^f.

Finally, we define, for $Q \coloneqq |I_{\mathcal{M}}|$,

$$\Phi^{\epsilon_1,\epsilon_2} \coloneqq ((\,[1,\dots,1],0)) \bullet \mathrm{P}\left(\Phi_{i_1}^{\phi(fP)}, \Phi_{i_2}^{\phi(fP)}, \dots, \Phi_{i_Q}^{\phi(fP)}\right).$$

We have, by (5.8) that

$$\begin{aligned} \|f - \mathrm{R}(\Phi^{\epsilon_1,\epsilon_2})\|_\infty &\leq Q \max_{i\in I_{\mathcal{M}}} \left\|\phi_{i,\mathcal{T}} \cdot \left(\tilde{f}_{j(i)} \circ P_{j(i)}\right) - \mathrm{R}\left(\Phi_i^{\phi(fP)}\right)\right\|_\infty \\ &\leq Q \max_{i\in I_{\mathcal{M}}} \left\|\phi_{i,\mathcal{T}} \cdot \left(\tilde{f}_{j(i)} \circ P_{j(i)}\right) - \mathrm{R}\left(\Phi_i^\phi\right) \cdot \mathrm{R}\left(\Phi_i^f \odot \Phi_i^P\right)\right\|_\infty \\ &\quad + Q \max_{i\in I_{\mathcal{M}}} \left\|\mathrm{R}\left(\Phi_i^\phi\right) \cdot \mathrm{R}\left(\Phi_i^f \odot \Phi_i^P\right) - \mathrm{R}\left(\Phi_i^{\phi(fP)}\right)\right\|_\infty =: Q \cdot (\mathrm{I} + \mathrm{II}). \end{aligned}$$

We proceed by estimating I. By (5.11) we have that $\phi_{i,\mathcal{T}} = \mathrm{R}\left(\Phi_i^\phi\right)$ and hence

$$\begin{aligned} \mathrm{I} &= \max_{i\in I_{\mathcal{M}}} \left\|\phi_{i,\mathcal{T}} \cdot \left(\tilde{f}_{j(i)} \circ P_{j(i)}\right) - \phi_{i,\mathcal{T}} \cdot \mathrm{R}\left(\Phi_i^f \odot \Phi_i^P\right)\right\|_\infty \\ &\leq \max_{i\in I_{\mathcal{M}}} \left\|\left(\tilde{f}_{j(i)} \circ P_{j(i)}\right) - \mathrm{R}\left(\Phi_i^f \odot \Phi_i^P\right)\right\|_\infty \leq \epsilon_1. \end{aligned}$$

Moreover, $\mathrm{II} \leq \epsilon_2$ by construction. We have, for $\epsilon > 0$ and $\epsilon_1 \coloneqq \epsilon_2 \coloneqq \epsilon/(2Q)$, that

$$\|f - \mathrm{R}\left(\Phi^{\epsilon_1,\epsilon_2}\right)\|_\infty \leq \epsilon.$$

Finally, we estimate the size of $\Phi^{\epsilon_1,\epsilon_2}$. We have that

$$\begin{aligned} M(\Phi^{\epsilon_1,\epsilon_2}) &\leq Q \max_{i\in I_{\mathcal{M}}} M\left(\Phi_i^{\phi(fP)}\right) \\ &\leq 2Q \cdot \left(M\left(\Phi^{\mathrm{mult},2,\epsilon_2}\right) + M\left(\Phi_{1,L^*}^{\mathrm{Id}} \odot \Phi_i^\phi\right) + M\left(\Phi_i^f \odot \Phi_i^P\right)\right) \\ &\leq 2Q \cdot \left(c_{\mathrm{mult}}\log_2(1/\epsilon) + 4L^* + 2M\left(\Phi_i^\phi\right) + 2M\left(\Phi_i^f\right) + 2M\left(\Phi_i^P\right)\right), \end{aligned}$$

for a constant $c_{\mathrm{mult}} > 0$ by Proposition 3.16 and Remark 3.10. By (5.9), we conclude that

$$M(\Phi^{\epsilon_1,\epsilon_2}) \lesssim \epsilon^{-d'/k} \log_2(1/\epsilon),$$

where the implicit constant depends on $\mathcal{M}$ and d. As the last step, we compute the depth of $\Phi^{\epsilon_1,\epsilon_2}$. We have that

$$\begin{aligned} L(\Phi^{\epsilon_1,\epsilon_2}) &= \max_{i \in I_{\mathcal{M}}} L\left(\Phi_i^{\phi(fP)}\right) \\ &= L\left(\Phi^{\mathrm{mult},2,\epsilon_2}\right) + L\left(\Phi_i^f \odot \Phi_i^P\right), \\ &\lesssim \log_2(1/\epsilon_2) + \log_2(1/\epsilon_1) \lesssim \log_2(1/\epsilon) \end{aligned}$$

by (5.10). □

Remark 5.7. *Theorem 5.6 shows that the approximation rate when approximating C^k regular functions defined on a manifold does not depend badly on the ambient dimension. However, at least in our construction, the constants may still depend on d and even grow rapidly with d. For example, in the estimate in* (5.11) *the implicit constant depends, because of Proposition 3.3 on the maximal number of neighbouring cells of the underlying mesh. For a typical mesh on a grid $\mathbb{Z}^d$ of a d dimensional space, it is not hard to see that this number grows exponentially with the dimension d.*

5.4 Dimension dependent regularity assumption

The last instance of an approximation result without curse of dimension that we shall discuss in this section is arguably the historically first result of this form. In [2], it was shown that, under suitable assumptions on the integrability of the Fourier transform of a function, approximation rates that are (almost) independent of the underlying dimensions are possible.

Here we demonstrate a slightly simplified result compared to that of [2]. Let, for $C > 0$,

$$\Gamma_C \coloneqq \left\{ f \in L^1\left(\mathbb{R}^d\right) : \|\hat{f}\|_1 < \infty, \int_{\mathbb{R}^d} |2\pi\xi| \left|\hat{f}(\xi)\right| d\xi < C \right\},$$

where $\hat{f}$ denotes the Fourier transform of f. By the inverse Fourier transform theorem, the condition $\|\hat{f}\|_1 < \infty$ implies that every element of Γ_C is continuous (in fact it is even continuously differentiable). We also denote the unit ball in $\mathbb{R}^d$ by $B_1^d \coloneqq \{x \in \mathbb{R}^d \colon |x| \leq 1\}$.

We have the following result:

Theorem 5.8 (cf. [2, Theorem 1])**.** *Let $d \in \mathbb{N}$, $f \in \Gamma_C$, $\varrho : \mathbb{R} \to \mathbb{R}$ be sigmoidal and $N \in \mathbb{N}$. Then, for every $c > 4C^2$, there exists a NN Φ with*

$$\begin{aligned} L(\Phi) &= 2, \\ M(\Phi) &\leq N \cdot (d+2) + 1, \\ \frac{1}{|B_1^d|} \int_{B_1^d} |f(x) - \mathrm{R}(\Phi)(x)|^2 \, dx &\leq \frac{c}{N}, \end{aligned}$$

where $|B_1^d|$ denotes the Lebesgue measure of B_1^d.

Remark 5.9. *The approximation rate above cannot be significantly improved. From [25, Proposition 4.6] it follows that the approximation rate cannot be improved beyond $N^{-(2+d)/d}$.*

Before we present the proof of Theorem 5.8, we show the following auxilliary result, which is sometimes called Approximate Caratheodory theorem, [39, Theorem 0.0.2].

Lemma 5.10 ([2, 26]). *Let G be a subset of a Hilbert space and let G be such that the norm of each element of G is bounded by $B > 0$. Let $f \in \overline{\mathrm{co}}(G)$. Then, for every $N \in \mathbb{N}$ and $c' > B^2$ there exist $(g_i)_{i=1}^N \subset G$ and $(c_i)_{i=1}^N \subset [0,1]$ with $\sum_{i=1}^N c_i = 1$ such that*

$$\left\| f - \sum_{i=1}^N c_i g_i \right\|^2 \leq \frac{c'}{N}. \tag{5.12}$$

Proof. Let $f \in \overline{\mathrm{co}}(G)$. For every $\delta > 0$, there exists $f^* \in \mathrm{co}(G)$ so that

$$\|f - f^*\| \leq \delta.$$

Since $f^* \in \mathrm{co}(G)$, there exists $m \in \mathbb{N}$ so that

$$f^* = \sum_{i=1}^m c_i' g_i'$$

for some $(g_i')_{i=1}^m \subset G$, $(c_i')_{i=1}^m \subset [0,1]$, with $\sum_{i=1}^m c_i' = 1$.

At this point, there exists an at most m dimensional linear space L_m such that $(g_i')_{i=1}^m \subset L_m$ which is isometrically isomorphic to $\mathbb{R}^m$. Hence, we can think of g_i', and f^* to be elements of $\mathbb{R}^m$ in the sequel.*

Let σ be a probability distribution on $\{1, \ldots, m\}$ with $\mathbb{P}_\sigma(k) = c_k'$ for $k \in \{1, \ldots, m\}$. Let $(g_j)_{j=1}^\infty$ be i.i.d random variable with $g_j = g_{i_j}'$, where $i_j \sim \sigma$. We have that

$$\mathbb{E}(g_j) = \mathbb{E}(g_1) = \sum_{i=1}^m c_i' g_i' = f^*,$$

and therefore we can find $(X_j)_{j=1}^\infty \coloneqq g_{i_j} - f^*$, for $i_j \sim \sigma$, are i.i.d random variables with $\mathbb{E}(X_j) = 0$. Since the X_j are independent random variables, we have that

$$\begin{aligned} \mathbb{E}\left(\left\| \frac{1}{N} \sum_{j=1}^N X_j \right\|^2 \right) &= \frac{1}{N^2} \sum_{j=1}^N \mathbb{E}\left(\|X_j\|^2 \right) = \frac{1}{N^2} \sum_{j=1}^N \mathbb{E}_{i\sim\sigma}\left(\|g_i\|^2 - 2\langle g_i, f^* \rangle + \|f^*\|^2 \right) \\ &= \frac{1}{N^2} \sum_{j=1}^N \mathbb{E}_{i\sim\sigma}\left(\|g_i\|^2 \right) - \|f^*\|^2 \leq \frac{B^2}{N}. \end{aligned} \tag{5.13}$$

The first identity above follows from Bienaymé's identity whereas the rest of the argument is, of course, commonly known as the weak law of large numbers. Because of (5.13) there exists at least one event ω such that

$$\left\| \frac{1}{N} \sum_{j=1}^N X_j(\omega) \right\|^2 \leq \frac{B^2}{N}$$

and hence

$$\left\| \frac{1}{N} \sum_{j=1}^N g_{i_j(\omega)} - f^* \right\|^2 \leq \frac{B^2}{N}.$$

By the triangle inequality, we have that

$$\left\| \frac{1}{N} \sum_{i=1}^N g_{i_j(\omega)} - f \right\|^2 \leq \frac{B^2}{N} + \delta.$$

Since δ was arbitrary this yields the result. □

*This simplification is not necessary at all, but some people might find it easier to think of real-valued random variables instead of Hilbert-space-valued.

We can have a more intuitive and elementary argument yielding Lemma 5.10 if $G = (\phi_i)_{i=1}^\infty$ is an orthonormal basis. This is based on an argument usually referred to as Stechkin's estimate, see e.g. [7, Lemma 3.6]. Let $f \in \overline{\mathrm{co}}(G)$, then

$$f = \sum_{i=1}^\infty \langle f, \phi_i \rangle \phi_i =: \sum_{i=1}^\infty c_i(f)\phi_i \tag{5.14}$$

with $\|c_i\|_1 = 1$. We have now that if Λ_n corresponds the indices of the n largest of $(|c_i(f)|)_{i=1}^\infty$ in (5.14), then

$$\left\| f - \sum_{j\in\Lambda_n} c_j(f)\phi_j \right\|^2 = \left\| \sum_{j\notin\Lambda_n} c_j(f)\phi_j \right\|^2 = \sum_{j\notin\Lambda_n} |c_j(f)|^2. \tag{5.15}$$

by Parseval's identity. Let $(\tilde{c}_k(f))_{k=1}^\infty$ be a non-increasing rearrangement of $(|c_j(f)|)_{j=1}^\infty$. We have that

$$\sum_{j\notin\Lambda_n} |c_j(f)|^2 = \sum_{k\geq n+1} \tilde{c}_k(f)^2 \leq \tilde{c}_{n+1}(f) \sum_{k\geq n+1} \tilde{c}_k(f) \leq \tilde{c}_{n+1}(f), \tag{5.16}$$

Since $(n+1)\tilde{c}_{n+1} \leq \sum_{j=1}^{n+1} \tilde{c}_j \leq 1$, we have that $\tilde{c}_{n+1} \leq (n+1)^{-1}$ and hence, the estimate

$$\left\| f - \sum_{j\in\Lambda_n} c_j(f)\phi_j \right\|^2 \leq (n+1)^{-1},$$

follows. Therefore, in the case that G is an orthogonal basis, we can explicitly construct the g_i and c_i of Lemma 5.10.

Remark 5.11. *Lemma 5.10 allows a quite powerful procedure. Indeed, to achieve an approximation rate of $1/N$ for a function f by superpositions of N elements of a set G, it suffices to show that* any *convex combination of elements of G approximates f.*

In the language of NNs, we could say that every function that can be represented by an arbitrary wide two-layer NN with bounded activation function and where the weights in the last layer are positive and sum to one can also be approximated with a network with only N neurons in the first layer and an error proportional to $1/N$.

In view of Lemma 5.10, to show Theorem 5.8, we only need to demonstrate that each function in Γ_C is in the convex hull of functions representable by superpositions of sigmoidal NNs with few weights. Before we prove this, we show that each function $f \in \Gamma_C$ is in the convex hull of functions of the set

$$G_C \coloneqq \left\{ B_1^d \ni x \mapsto \gamma \cdot \mathbb{1}_{\mathbb{R}^+}(\langle a, x\rangle + b) \colon a \in \mathbb{R}^d, b \in \mathbb{R}, |\gamma| \leq 2C \right\}.$$

Lemma 5.12. *Let $f \in \Gamma_C$. Then $f_{|B_1^d} - f(0) \in \overline{\mathrm{co}}(G_C)$. Here the closure is taken with respect to the norm $\|\cdot\|_{L^{2,\diamond}(B_1^d)}$, defined by*

$$\|g\|_{L^{2,\diamond}(B_1^d)} \coloneqq \left(\frac{1}{|B_1^d|} \int_{B_1^d} |g(x)|^2 dx \right)^{1/2}.$$

Proof. Since $f \in \Gamma_C$ is continuous and $\hat{f} \in L^1(\mathbb{R}^d)$, we have by the inverse Fourier transform that

$$\begin{aligned} f(x) - f(0) &= \int_{\mathbb{R}^d} \hat{f}(\xi) \left(e^{2\pi i \langle x, \xi\rangle} - 1 \right) d\xi \\ &= \int_{\mathbb{R}^d} \left| \hat{f}(\xi) \right| \left(e^{2\pi i \langle x, \xi\rangle + i\kappa(\xi)} - e^{i\kappa(\xi)} \right) d\xi \\ &= \int_{\mathbb{R}^d} \left| \hat{f}(\xi) \right| \left(\cos(2\pi\langle x, \xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)) \right) d\xi, \end{aligned}$$

where $\kappa(\xi)$ is the phase of $\hat{f}(\xi)$ and the last inequality follows since f is real-valued. Moreover, we have that

$$\int_{\mathbb{R}^d} \left|\hat{f}(\xi)\right| (\cos(2\pi\langle x,\xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)))\, d\xi = \int_{\mathbb{R}^d} \frac{(\cos(2\pi\langle x,\xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)))}{|2\pi\xi|} |2\pi\xi| \left|\hat{f}(\xi)\right| d\xi.$$

Since $f \in \Gamma_C$, we have $\int_{\mathbb{R}^d} |2\pi\xi||\hat{f}(\xi)|d\xi \leq C$, and thus Λ such that

$$d\Lambda(\xi) \coloneqq \frac{1}{C}|2\pi\xi||\hat{f}(\xi)|d\xi$$

is a finite measure with $\Lambda(\mathbb{R}^d) = \int_{\mathbb{R}^d} d\Lambda(\xi) \leq 1$. In this notion, we have

$$f(x) - f(0) = C \int_{\mathbb{R}^d} \frac{(\cos(2\pi\langle x,\xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)))}{|2\pi\xi|} d\Lambda(\xi).$$

Since $(\cos(2\pi\langle x,\xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)))/|2\pi\xi|$ is continuous and bounded by 1 by the Lipschitz continuity of $\cos$, and hence integrable with respect to $d\Lambda(\xi)$ we have by the dominated convergence theorem that, for $n \to \infty$,

$$\left| C \int_{\mathbb{R}^d} \frac{(\cos(2\pi\langle x,\xi\rangle + \kappa(\xi)) - \cos(\kappa(\xi)))}{|2\pi\xi|} d\Lambda(\xi) - C \sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} \frac{(\cos(2\pi\langle x,\theta\rangle + \kappa(\theta)) - \cos(\kappa(\theta)))}{|2\pi\theta|} \cdot \Lambda(I_\theta) \right| \to 0, \tag{5.17}$$

where $I_\theta \coloneqq [0, 1/n]^d + \theta$. Since $f(x) - f(0)$ is continuous and thus bounded on B_1^d and

$$C \left| \sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} \frac{(\cos(2\pi\langle x,\theta\rangle + \kappa(\theta)) - \cos(\kappa(\theta)))}{|2\pi\theta|} \cdot \Lambda(I_\theta) \right| \leq C,$$

we have by the dominated convergence theorem that

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f(x) - f(0) - C \sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} \frac{(\cos(2\pi\langle x,\theta\rangle + \kappa(\theta)) - \cos(\kappa(\theta)))}{|2\pi\theta|} \cdot \Lambda(I_\theta) \right|^2 dx \to 0. \tag{5.18}$$

Since $\sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} \Lambda(I_\theta) = \Lambda(\mathbb{R}^d) \leq 1$, we conclude that $f(x) - f(0)$ is in the $L^{2,\diamond}(B_1^d)$ closure of convex combinations of functions of the form

$$x \mapsto g_\theta(x) \coloneqq \alpha_\theta \frac{\cos(2\pi\langle x,\theta\rangle + \kappa(\theta)) - \cos(\kappa(\theta))}{|2\pi\theta|},$$

for $\theta \in \mathbb{R}^d$ and $0 \leq \alpha_\theta \leq C$. The result follows, if we can show that each of the functions g_θ is in $\overline{\text{co}}(G_C)$. Setting $z = \langle x, \theta/|\theta|\rangle$, it suffices to show that the map

$$[-1,1] \ni z \mapsto \alpha_\theta \frac{\cos(2\pi|\theta|z + \kappa(\theta)) - \cos(\kappa(\theta))}{|2\pi\theta|} =: \tilde{g}_\theta(z),$$

can be approximated arbitrarily well by convex combinations of functions of the form

$$[-1,1] \ni z \mapsto \gamma \cdot \mathbb{1}_{\mathbb{R}^+}(a'z + b'), \tag{5.19}$$

where $a', b' \in \mathbb{R}$ and $|\gamma| \leq 2C$.

Per definition, we have that $\|\tilde{g}'_\theta\| \leq C$. We define, for $T \in \mathbb{N}$,

$$g_{T,+} \coloneqq \sum_{i=1}^{T} \frac{|\tilde{g}_\theta(i/T) - \tilde{g}_\theta((i-1)/T)|}{2C} \cdot (2C \cdot \operatorname{sign}(\tilde{g}_\theta(i/T) - \tilde{g}_\theta((i-1)/T)) \cdot \mathbb{1}_{\mathbb{R}^+}(x - i/T)),$$

$$g_{T,-} \coloneqq \sum_{i=1}^{T} \frac{|\tilde{g}_\theta(-i/T) - \tilde{g}_\theta((1-i)/T)|}{2C} (2C \cdot \operatorname{sign}(\tilde{g}_\theta(-i/T) - \tilde{g}_\theta((1-i)/T)) \cdot \mathbb{1}_{\mathbb{R}^+}(-x + i/T)).$$

Clearly, $(g_{T,-}) + (g_{T,+})$ converges to $\tilde{g}_\theta$ for $T \to \infty$ and since

$$\sum_{i=1}^{T} \frac{|\tilde{g}_\theta(i/T) - \tilde{g}_\theta((i-1)/T)|}{2C} + \sum_{i=1}^{T} \frac{|\tilde{g}_\theta(-i/T) - \tilde{g}_\theta((1-i)/T)|}{2C} \leq 2 \sum_{i=1}^{T} \|\tilde{g}'_\theta\|_\infty/(2CT) \leq 1$$

we have that $\tilde{g}_\theta$ can be arbitrarily well approximated by convex combinations of the form (5.19). Therefore, we have that $g_\theta \in \overline{\operatorname{co}}(G_C)$ and by (5.18) this yields that $f - f(0) \in \overline{\operatorname{co}}(G_C)$. □

Proof of Theorem 5.8. Let $f \in \Gamma_C$, then, by Lemma 5.12, we have that

$$f_{|B_1^d} - f(0) \in \overline{\operatorname{co}}(G_C).$$

Moreover, for every element $g \in G_C$ we have that $\|g\|_{L^{2,\diamond}(B_1^d)} \leq 2C$. Therefore, by Lemma 5.10, applied to the Hilbert space $L^{2,\diamond}(B_1^d)$, we get that for every $N \in \mathbb{N}$, there exist $|\gamma_i| \leq 2C$, $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, for $i = 1, \ldots, N$, so that

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f_{B_1^d}(x) - f(0) - \sum_{i=1}^{N} \gamma_i \mathbb{1}_{\mathbb{R}^+}(\langle a_i, x\rangle + b_i) \right|^2 dx \leq \frac{4C^2}{N}.$$

Since $\varrho(\lambda x) \to \mathbb{1}_{\mathbb{R}^+}(x)$ for $\lambda \to \infty$ almost everywhere, it is clear that, for every $\delta > 0$, there exist $\tilde{a}_i$, $\tilde{b}_i$, $i = 1, \ldots, N$, so that

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f_{B_1^d}(x) - f(0) - \sum_{i=1}^{N} \gamma_i \varrho\left(\langle \tilde{a}_i, x\rangle + \tilde{b}_i\right) \right|^2 dx \leq \frac{4C^2}{N} + \delta.$$

The result follows by observing that

$$\sum_{i=1}^{N} \gamma_i \varrho\left(\langle \tilde{a}_i, x\rangle + \tilde{b}_i\right) + f(0)$$

is the realisation of a network Φ with $L(\Phi) = 2$ and $M(\Phi) \leq N \cdot (d+3)$. This is clear by setting

$$\Phi \coloneqq (([\gamma_1, \ldots, \gamma_N], f(0))) \bullet \mathrm{P}\left(((\tilde{a}_1, \tilde{b}_1)), \ldots, ((\tilde{a}_N, \tilde{b}_N))\right).$$

□

Remark 5.13. *The fact, that the approximation rate of Theorem 5.8 is independent from the dimension is quite surprising at first. However, the following observation might render Theorem 5.8 more plausible. The assumption of having a finite Fourier moment is comparable to a certain* dimension dependent regularity assumption*. In other words, the condition of having a finite Fourier moment becomes more restrictive in higher dimensions, meaning that the complexity of the function class does not, or only mildly grow with the dimension. Indeed, while this type of regularity is not directly expressible in terms of classical orders of smoothness, Barron notes that a necessary condition for $f \in \Gamma_C$, for some $C > 0$, is that f has bounded first-order derivatives. A sufficient condition is that all derivatives of order up to $\lfloor d/2 \rfloor + 2$ are square-integrable, [2][Section II]. The sufficient condition amounts to $f \in W^{\lfloor d/2 \rfloor + 2, 2}(\mathbb{R}^d)$ which would also imply an approximation rate of N^{-1} in the squared L^2 norm by sums of at most N B-splines, see [22, 9].*

Example 5.14. *A natural question, especially in view of Remark 5.13, is which well known and relevant functions are contained in Γ_C. In [2, Section IX], a long list with properties of this set and elements thereof is presented. Among others, we have that*

1. *If $g \in \Gamma_C$, then*
$$a^{-d} g\left(a(\cdot - b)\right) \in \Gamma_C,$$
for every $a \in \mathbb{R}^+$, $b \in \mathbb{R}^d$.

2. *For $g_i \in \Gamma_C$, $i = 1, \dots, m$ and $c = (c_i)_{i=1}^m$ it holds that*
$$\sum_{i=1}^m c_i g_i \in \Gamma_{\|c\|_1 C}.$$

3. *The Gaussian function: $x \mapsto e^{-|x|^2/2}$ is in Γ_C for $C = \mathcal{O}(d^{1/2})$.*

4. *Functions of high smoothness. If the first $\lceil d/2 \rceil + 2$ derivatives of a function g are square integrable on $\mathbb{R}^d$, then $g \in \Gamma_C$, where the constant C depends linearly on $\|g\|_{W^{\lfloor d/2 \rfloor + 2, 2}}$.*

The last three examples show quite nicely how the assumption $g \in \Gamma_C$ includes an indirect dependence on the dimension.

6 Complexity of sets of networks

Until this point, we have mostly tried understanding the capabilities of NNs through the lens of approximation theory. This analysis is based on two pillars: First, we are interested in asymptotic performance, i.e., we are aiming to understand the behaviour of NNs for increasing sizes. Second, we measure our success over a continuum by studying L^p norms for $p \in [1, \infty]$.

This point of view is certainly not the only possible, and different applications require a different analysis of the capabilities of NNs. Consider, for example, a binary classification task, i.e., a process, where values $(x_i)_{i=1}^N$ should be classified as either 0 or 1. In this scenario, it is interesting to establish if for every possible classification of the values $(x_i)_{i=1}^N$ as 0 or 1 there exists a NN the realisation of which is a function performing this classification.

This question, in contrast to the point of view of approximation theory, is non-asymptotic and only studies the success of networks on a finite, discrete set of samples. Nonetheless, we will later see, that the complexity measures that we will introduce below are also closely related to some questions in approximation theory and can be used to establish lower bounds on approximation rates.

The following sections are strongly inspired by [1, Sections 3-8].

6.1 The growth function and the VC dimension

We now introduce two notions of the capability of a class of functions to perform binary classification of points: Let X be a space, $H \subset \{h \colon X \to \{0, 1\}\}$ and $S \subset X$ be finite. We define by
$$H_S \coloneqq \{h_{|S} \colon h \in H\},$$
the *restriction* of H to S. We define, the *growth function of H* by
$$\mathcal{G}_H(m) \coloneqq \max\left\{|H_S| \colon S \subset X, |S| = m\right\}, \quad \text{for } m \in \mathbb{N}.$$
The growth function counts the number of functions that result from restricting H to the best possible set S with m elements. Intuitively, in the framework of binary classification, the growth function tells us in how many ways we can classify the elements of the best possible sets S of any cardinality by functions in H.

It is clear that for every set S with $|S| = m$, we have that $|H_S| \leq 2^m$ and hence $\mathcal{G}_H(m) \leq 2^m$. We say that a set S with $|S| = m$ for which $|H_S| = 2^m$ is *shattered by H*.

A second, more compressed notion of complexity in the context of binary classification is that of the *Vapnik–Chervonenkis dimension* (VC Dimension), [38]. We define $\mathrm{VCdim}(H)$ to be the largest integer m such that there exists $S \subset X$ with $|S| = m$ that is shattered by H. In other words,

$$\mathrm{VCdim}(H) \coloneqq \max\{m \in \mathbb{N} \colon \mathcal{G}_H(m) = 2^m\}.$$

Example 6.1. *Let $X = \mathbb{R}^2$.*

1. *Let $H \coloneqq \{0, 1\}$, then $\mathcal{G}_H(m) = 2$ for all $m \geq 1$. Hence, $\mathrm{VCdim}(H) = 1$.*
2. *Let $H \coloneqq \{0, \chi_\Omega, \chi_{\Omega^c}, 1\}$ for some fixed non-empty set $\Omega \subsetneq \mathbb{R}^2$. Then, choosing $S = (x_1, x_2)$ with $x_1 \in \Omega, x_2 \in \Omega^c$, we have $\mathcal{G}_H(2) = 4$ for all $m \geq 2$. Hence, $\mathrm{VCdim}(H) = 2$.*
3. *Let $h \coloneqq \chi_{\mathbb{R}^+}$ and*

$$H \coloneqq \left\{ h_{\theta,t} \coloneqq h\left(\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}^T \cdot -t \right) \;:\; \theta \in [-\pi, \pi],\ t \in \mathbb{R}^2 \right\}.$$

Then H is the set of all linear classifiers. It is not hard to see, that if S contains 3 points in general position, then $|H_{|S}| = 8$, see Figure 6.1. Hence, these sets S are shattered by H. We will later see that H does not shatter any set of points with at least 4 elements. Hence $\mathrm{VCdim}(H) = 3$. This is intuitively clear when considering Figure 6.2.

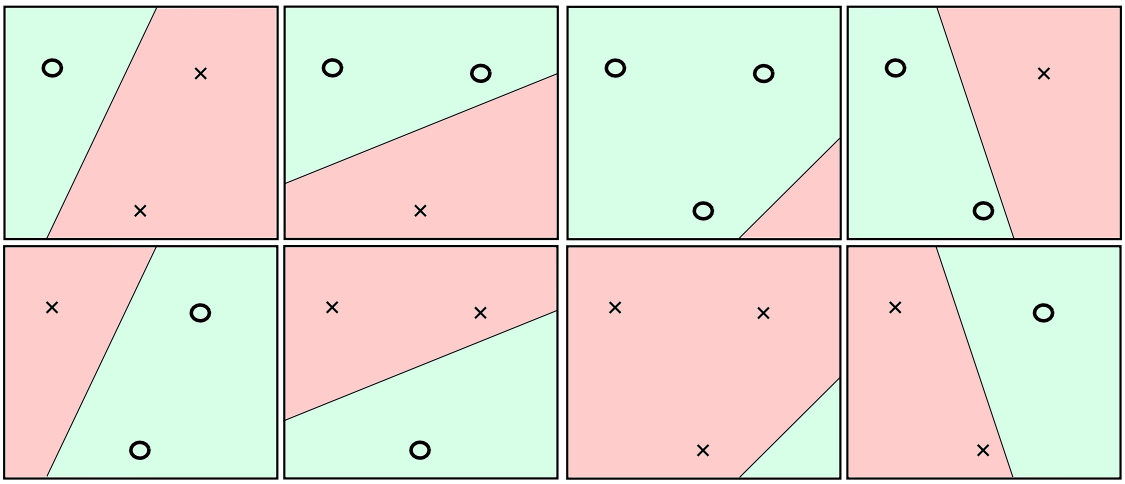

Figure 6.1: Three points shattered by a set of linear classifiers.

As a first step to familiarise ourselves with the new notions, we study the growth function and VC dimension of realisations of NNs with one neuron and the Heaviside function as activation function. This situation was discussed before in the third point of Example 6.1.

We have the following theorem:

Theorem 6.2 ([1, Theorem 3.4])**.** *Let $d \in \mathbb{N}$ and $\varrho = \mathbb{1}_{\mathbb{R}^+}$. Let $\mathcal{SN}(d)$ be the set of realisations of neural networks with two layers, d-dimensional input, one neuron in the first layer and one dimensional output and the weights in the second layer satisfy $(A_2, b_2) = (1, 0)$. Then $\mathcal{SN}(d)$ shatters a set of points $(x_i)_{i=1}^m \subset \mathbb{R}^d$ if and only if*

$$(x_1, 1), (x_2, 1), \ldots, (x_m, 1) \tag{6.1}$$

are linearly independent points. In particular, $\mathrm{VCdim}(\mathcal{SN}(d)) = d + 1$.

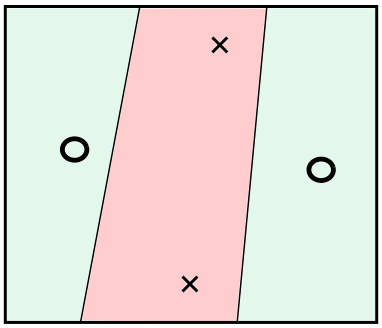

Figure 6.2: Four points which cannot be classified in every possible way by a single linear classifier. The classification sketched above requires at least sums of two linear classifiers.

Proof. Assume first, that $(x_i)_{i=1}^m$ is such that it is shattered by $\mathcal{SN}(d)$ and assume towards a contradiction that (6.1) are not linearly independent.

Then we have that for every $v \in \{0,1\}^m$ there exists a neural network Φ^v, such that, for all $j \in \{1,\dots,m\}$,

$$\mathrm{R}(\Phi^v)(x_j) = v_j.$$

Moreover, since (6.1) are not linearly independent there exist $(\alpha_j)_{j=1}^m \subset \mathbb{R}$ such that, without loss of generality,

$$\sum_{j=1}^{m-1} \alpha_j \begin{pmatrix} x_j \\ 1 \end{pmatrix} = \begin{pmatrix} x_m \\ 1 \end{pmatrix}.$$

Let $v \in \{0,1\}^m$ be such that, for $j \in \{1,\dots,m-1\}$, $v_j = 1 - \mathbb{1}_{\mathbb{R}^+}(\alpha_j)$ and $v_m = 1$. Then,

$$\begin{aligned} \mathrm{R}\left(\Phi^v\right)(x_m) &= \varrho\left(\begin{bmatrix} A_1^v & b_1^v \end{bmatrix}\begin{bmatrix} x_m & 1 \end{bmatrix}\right) \\ &= \varrho\left(\sum_{j=1}^{m-1} \alpha_j \cdot (A_1^v x_j + b_1^v)\right) = 0, \end{aligned}$$

where the last equality is because $\mathbb{1}_{\mathbb{R}^+}(A_1^v x_j + b_1^v) = v_j = 1 - \mathbb{1}_{\mathbb{R}^+}(\alpha_j)$. This produces the desired contradiction.

If, on the other hand (6.1) are linearly independent, then the matrix

$$X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix}$$

has rank m. Hence, for every $v \in \{0,1\}^m$ there exists a vector $\begin{bmatrix} A_1^v & b_1^v \end{bmatrix} \in \mathbb{R}^{1,d+1}$ such that $X\begin{bmatrix} A_1^v & b_1^v \end{bmatrix}^T = v$. Setting $\Phi^v \coloneqq ((A_1^v, b_1^v), (1,0))$ yields the claim. □

In establishing bounds on the VC dimension of a set of neural networks, the activation function plays a major role. For example, we have the following lemma.

Lemma 6.3 ([1, Lemma 7.2]). *Let* $H \coloneqq \{x \mapsto \mathbb{1}_{\mathbb{R}^+}(\sin(ax)) \colon a \in \mathbb{R}\}$. *Then* $\mathrm{VCdim}(H) = \infty$.

Proof. Let $x_i \coloneqq 2^{i-1}$, for $i \in \mathbb{N}$. Next, we will show that, for every $k \in \mathbb{N}$, the set $\{x_1,\dots,x_k\}$ is shattered by H.

The argument is based on the following bit-extraction technique: Let $b \coloneqq \sum_{j=1}^{k} b_j 2^{-j} + 2^{-k-1}$. Setting $a \coloneqq 2\pi b$, we have that

$$\begin{aligned}\mathbb{1}_{\mathbb{R}^+}(\sin(ax_i)) &= \mathbb{1}_{\mathbb{R}^+}\left(\sin\left(2\pi\sum_{j=1}^{k} b_j 2^{-j} x_i + 2\pi 2^{-k-1} x_i\right)\right) \\ &= \mathbb{1}_{\mathbb{R}^+}\left(\sin\left(2\pi\sum_{j=1}^{i-1} b_j 2^{-j} x_i + 2\pi\sum_{j=i}^{k} b_j 2^{-j} x_i + 2\pi 2^{-k-1} x_i\right)\right) =: \mathrm{I}(x_i).\end{aligned}$$

Since $\sum_{j=1}^{i-1} b_i 2^{-j} x_i \in \mathbb{N}$, we have by the 2π periodicity of $\sin$ that

$$\begin{aligned}\mathrm{I}(x_i) &= \mathbb{1}_{\mathbb{R}^+}\left(\sin\left(2\pi\sum_{j=i}^{k} b_j 2^{-j} x_i + 2\pi 2^{-k-1} x_i\right)\right) \\ &= \mathbb{1}_{\mathbb{R}^+}\left(\sin\left(b_i\pi + \pi\cdot\left(\sum_{j=i+1}^{k} b_j 2^{i+1-j} + 2^{i-k}\right)\right)\right).\end{aligned}$$

Since $\left(\sum_{j=i+1}^{k} b_j 2^{i+1-j} + 2^{i-k}\right) \in (0,1)$, we have that

$$\mathrm{I}(x_i) = \begin{cases} 0 & \text{if } b_i = 1, \\ 1 & \text{else.} \end{cases}$$

Since b was chosen arbitrary, this shows that $\mathrm{VCdim}(H) \geq k$ for all $k \in \mathbb{N}$. □

In the previous two results (Theorem 6.2, Lemma 6.3), we observed that the VC dimension of sets of realisations of NNs depends on their size and also on the associated activation function. We have the following result, that we state without proof:

Denote, $d, L, M \in \mathbb{N}$ by $\mathcal{NN}_{d,L,M}$ the set of neural networks with d dimensional input, L layers and at most M weights.

Theorem 6.4 ([1, Theorem 8.8]). *Let $d, \ell, p \in \mathbb{N}$, and $\varrho : \mathbb{R} \to \mathbb{R}$ be a piecewise polynomial with at most ℓ pieces of degree at most p. Let, for $L, M \in \mathbb{N}$,*

$$H \coloneqq \{\mathbb{1}_{\mathbb{R}^+} \circ \mathrm{R}(\Phi) \colon \Phi \in \mathcal{NN}_{d,L,M}\}.^{a}$$

Then, for all $L, M \in \mathbb{N}$,

$$\mathrm{VCdim}(H) \lesssim ML\log_2(M) + ML^2.$$

[a] We are a bit sloppy with the notation here. In [1, Theorem 8.8] the result only applies to sets of neural networks that all have the same M indices of weights potentially non-zero.

6.2 Lower bounds on approximation rates

We will see next that the bound on the VC dimension of sets of neural networks of Theorem 6.4 implies a lower bound on the approximation capabilities of neural networks. The argument below follows [41, Section 4.2].

We first show the following auxiliary result.

Lemma 6.5. *Let $d, k \in \mathbb{N}$, $K \subset \mathbb{R}^d$, $H \subset \{h : K \to \mathbb{R}\}$ be such that, for $\epsilon > 0$, $\{x_1, \ldots, x_k\} \subset K$ and every $b \in \{0,1\}^k$, there exists $h \in H$ such that*

$$h(x_i) = \epsilon b_i, \quad \text{for all } i = 1, \ldots, k. \tag{6.2}$$

Let $G \subset \{g : K \to \mathbb{R}\}$ be such that for every $h \in H$, there exists a $g \in G$ satisfying

$$\sup_{x \in K} |g(x) - h(x)| < \epsilon/2. \tag{6.3}$$

Then

$$\mathrm{VCdim}(\{\mathbb{1}_{\mathbb{R}^+} \circ (g - \epsilon/2) : g \in G\}) \geq k. \tag{6.4}$$

Proof. Choose for any $b \in \{0,1\}^k$ an associated $h_b \in H$ according to (6.2) and g_b according to (6.3).

Then $|g_b(x_i) - b_i| < \epsilon/2$ and therefore $g_b(x_i) - \epsilon/2 > 0$ if $b_i = 1$ and $g_b(x_i) - \epsilon/2 < 0$ otherwise. Hence

$$\mathbb{1}_{\mathbb{R}^+}(g_b - \epsilon/2)(x_i) = b_i,$$

which yields the claim. □

Remark 6.6. *Lemma 6.5 and Theorem 6.4 allow an interesting observation about approximation by NNs. Indeed, if a set of functions H is sufficiently large so that* (6.2) *holds, and NNs with M weights and L layers achieve an approximation error less than $\epsilon > 0$ for every function in H, then $ML\log_2(M) + ML^2 \gtrsim k$.*

We would now like to establish a lower bound on the size of neural networks that approximate regular functions well. Considering functions $f \in C^s([0,1]^d)$ with $\|f\|_{C^s} \leq 1$, we we would, in view of Remark 6.6, like to find out which value of k is achievable for any given ϵ.

We begin by constructing one bump function with a finite C^n norm.

Lemma 6.7. *For every $n, d \in \mathbb{N}$, there exists a constant $C > 0$, such that, for every $\epsilon > 0$, there exists a smooth function $f_\epsilon \in C^n(\mathbb{R}^d)$ with*

$$\mathrm{supp}\ f_\epsilon \subset [-C\epsilon^{1/n}, C\epsilon^{1/n}]^d, \tag{6.5}$$

such that $f_\epsilon(0) = \epsilon$ and $\|f\|_{C^n(\mathbb{R}^d)} \leq 1$.

Proof. The function

$$\tilde{f}(x) \coloneqq \begin{cases} e^{1-1/(1-|x|^2)} & \text{for } |x| < 1, \\ 0 & \text{else.} \end{cases}$$

is smooth and supported in $[-1,1]^d$ and $\tilde{f}(0) = 1$. We set

$$f_\epsilon(x) \coloneqq \epsilon \tilde{f}\left(\frac{\epsilon^{-1/n}}{1 + \|\tilde{f}\|_{C^n}} x\right).$$

Then $f_\epsilon(0) = \epsilon$, $\mathrm{supp}\, f_\epsilon \subset [-(1 + \|\tilde{f}\|_{C^n})\epsilon^{-1/n}, (1 + \|\tilde{f}\|_{C^n})\epsilon^{-1/n}]^d$, and $\|f_\epsilon\|_{C^n} \leq 1$ by the chain rule. □

Adding up multiple, shifted versions of the function of Lemma 6.7 yields sets of functions that satisfy (6.2). Concretely, we have the following lemma.

Lemma 6.8. *Let $n, d \in \mathbb{N}$. There exists $C > 0$ such that, for every $\epsilon > 0$, there are $\{x_1, \ldots, x_k\}$ with $k \geq C\epsilon^{-d/n}$ such that, for every $b \in \{0,1\}^k$ there is $f_b \in C^n([0,1]^d)$ with $\|f\|_{C^n} \leq 1$ and*

$$f_b(x_i) = \epsilon b_i.$$

Proof. Let, for $C > 0$ as in (6.5), $\{x_1, \ldots, x_k\} \coloneqq 2C\epsilon^{1/n}\mathbb{Z}^d \cap [0,1]^d$. Clearly, $k \geq C'\epsilon^{-d/n}$ for a constant $C' > 0$.

Let $b \in \{0,1\}^k$. Now set, for f_ϵ as in Lemma 6.7,

$$f_b \coloneqq \sum_{i=1}^{k} b_i f_\epsilon(\cdot - x_i).$$

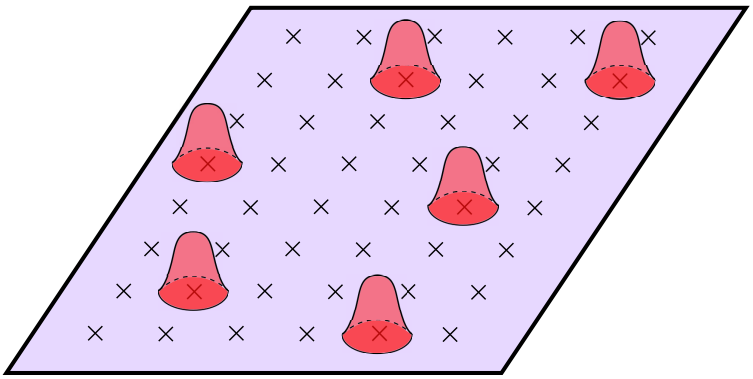

Figure 6.3: Illustration of f_b from Lemma 6.8 on $[0,1]^2$.

By the properties of f_ϵ, we have that $f_\epsilon(\cdot - x_i)$ vanishes on every x_j for $j \neq i$ and hence

$$f_b(x_i) = \epsilon b_i, \text{ for all } i = 1, \dots, k.$$

The construction of f_b is depicted in Figure 6.3. Moreover, since $\operatorname{supp} f_\epsilon(\cdot - x_i) \subset x_i + [-C\epsilon^{1/n}, C\epsilon^{1/n}]^d$, we have that $\operatorname{supp} f_\epsilon(\cdot - x_i) \cap \operatorname{supp} f_\epsilon(\cdot - x_j) = \varnothing$ if $i \neq j$. Hence $\|f_b\|_{C^n} = \sup_{i=1,\dots,k} \|f_\epsilon(\cdot - x_i)\|C^n \leq 1$. □

Combining all observations until here, yields the following result.

Theorem 6.9. *Let $n, d \in \mathbb{N}$. Let $\varrho : \mathbb{R} \to \mathbb{R}$ be piecewise polynomial. Assume that, for all $\epsilon > 0$, there exist $M(\epsilon), L(\epsilon) \in \mathbb{N}$ such that*

$$\sup_{f \in C^n([0,1]^d), \|f\|_{C^n} \leq 1} \inf_{\Phi \in \mathcal{NN}_{d, L(\epsilon), M(\epsilon)}} \|f - \mathrm{R}(\Phi)\|_\infty \leq \epsilon/2,$$

then

$$(M(\epsilon)+1)L(\epsilon)\log_2(M(\epsilon)+1) + (M(\epsilon)+1)L(\epsilon)^2 \gtrsim \epsilon^{-d/n}.$$

Proof. Let $H \coloneqq \{h \in C^n([0,1]^d : \|h\|_{C^n} \leq 1)$ and $G \coloneqq \{\mathrm{R}(\Phi) : \Phi \in \mathcal{NN}_{d, L(\epsilon), M(\epsilon)}\}$.

H satisfies (6.2) with $k \geq C\epsilon^{-d/n}$ due to Lemma 6.8 and G satisfies (6.3) by assumption. Hence,

$$\mathrm{VCdim}(\{\mathbb{1}_{\mathbb{R}^+} \circ (g - \epsilon/2) : g \in G\}) \geq k. \tag{6.6}$$

Moreover,

$$\{g - \epsilon/2 : g \in G\} \subseteq \{\mathrm{R}(\Phi) : \Phi \in \mathcal{NN}_{d, L(\epsilon), M(\epsilon)+1}\}.$$

Hence

$$\mathrm{VCdim}\left(\left\{\mathbb{1}_{\mathbb{R}^+} \circ \mathrm{R}(\Phi) : \Phi \in \mathcal{NN}_{d, L(\epsilon), M(\epsilon)+1}\right\}\right) \geq C\epsilon^{-d/n}.$$

An application of Theorem 6.4 yields the result. □

Remark 6.10. *Theorem 6.9 shows that to achieve a uniform error of $\epsilon > 0$ over sets of C^n regular functions requires a number of weights M and layers L such that*

$$ML\log_2(M) + ML^2 \geq \epsilon^{-d/n}.$$

If we require L to only grow like $\log_2(\epsilon)$ then this demonstrates that the rate of Theorem 3.19/ Remark 3.20 is optimal.

For the case, that L is arbitrary, [1, Theorem 8.7] yields an upper bound on the VC dimension of

$$\widetilde{H} \coloneqq \left\{ \mathbb{1}_{\mathbb{R}^+} \circ \mathrm{R}(\Phi) \colon \Phi \in \bigcup_{\ell=1}^{\infty} \mathcal{NN}_{d,\ell,M} \right\}. \tag{6.7}$$

of the form

$$\mathrm{VCdim}\left(\widetilde{H}\right) \lesssim M^2. \tag{6.8}$$

Using (6.8) yields the following result:

Theorem 6.11. *Let $n, d \in \mathbb{N}$. Let $\varrho : \mathbb{R} \to \mathbb{R}$ be piecewise polynomial. Assume that, for all $\epsilon > 0$, there exist $M(\epsilon) \in \mathbb{N}$ such that*

$$\sup_{f \in C^n([0,1]^d), \|f\|_{C^n} \leq 1} \inf_{\Phi \in \bigcup_{\ell=1}^{\infty} \mathcal{NN}_{d,\ell,M(\epsilon)}} \|f - \mathrm{R}(\Phi)\|_{\infty} \leq \epsilon/2,$$

then

$$M(\epsilon) \gtrsim \epsilon^{-d/(2n)}.$$

Proof. The proof is the same as for Theorem 6.9, using (6.8) instead of Theorem 6.4. □

Remark 6.12. *Comparing Theorem 6.9 and Theorem 6.11, we see that approximation by NNs with arbitrarily many layers can potentially achieve double the rate of that with restricted or only slowly growing number of layers.*

Indeed, at least for the ReLU activation function, the lower bound of Theorem 6.11 is sharp. It could be shown in [41], that ReLU realisations of NNs with unrestricted numbers of layers achieve approximation fidelity $\epsilon > 0$ using only $\mathcal{O}(\epsilon^{-d/(2n)})$ many weights, uniformly over the unit ball of $C^n([0,1]^d)$.

7 Spaces of realisations of neural networks

As a final step of our analysis of deep neural networks from a functional analytical point of view, we would like to understand set-topological aspects of sets of realisations of NNs. What we are analysing in this section are sets of neural networks of a *fixed architecture*. We first define the notion of an architecture.

Definition 7.1. *A vector $S = (N_0, N_1, \ldots, N_L) \in \mathbb{N}^{L+1}$ is called* architecture of a neural network $\Phi = ((A_1, b_1), \ldots, (A_L, b_L))$ *if $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ for all $\ell = 1, \ldots, L$. We denote by $\mathcal{NN}(S)$ the* set of neural networks with architecture S *and, for an activation function $\varrho : \mathbb{R} \to \mathbb{R}$, we denote by*

$$\mathcal{RNN}_{\varrho}(S) \coloneqq \{\mathrm{R}(\Phi) \colon \Phi \in \mathcal{NN}(S)\}$$

the set of realisations of neural networks with architecture S.

For any architecture S, $\mathcal{NN}(S)$ is a finite dimensional vector space on which we use the norm

$$\|\Phi\|_{\mathrm{total}} \coloneqq |\Phi|_{\mathrm{scaling}} + |\Phi|_{\mathrm{shift}} \coloneqq \max_{\ell=1}^{L} \|A_\ell\|_{\infty} + \max_{\ell=1}^{L} \|b_\ell\|_{\infty}.$$

Now we have that, for a given architecture $S = (d, N_1, \ldots, N_L) \in \mathbb{N}^{L+1}$, a compact set $K \subset \mathbb{R}^d$, and for a continuous activation function $\varrho : \mathbb{R} \to \mathbb{R}$:

$$\mathcal{RNN}_{\varrho}(S) \subset L^p(K),$$

for all $p \in [1, \infty]$. In this context, we can ask ourselves about the properties of $\mathcal{RNN}_{\varrho}(S)$ as a subset of the normed linear spaces $L^p(K)$.

The results below are based on the following observation about the realisation map:

Theorem 7.2 ([23, Proposition 4]). *Let $\Omega \subset \mathbb{R}^d$ be compact and let $S = (d, N_1, \dots, N_L) \in \mathbb{N}^{L+1}$ be a neural network architecture. If the activation function $\varrho : \mathbb{R} \to \mathbb{R}$ is continuous, then the map*

$$\begin{aligned} \mathrm{R} \colon \mathcal{NN}(S) &\to L^\infty(\Omega) \\ \Phi &\mapsto \mathrm{R}(\Phi) \end{aligned}$$

is continuous. Moreover, if ϱ is locally Lipschitz continuous, then R is locally Lipschitz continuous.

7.1 Network spaces are not convex

We begin by analysing the simple question if, for a given architecture S, the set $\mathcal{RNN}_\varrho(S)$ is star-shaped. We start by fixing the notion of a centre and of star-shapedness.

Definition 7.3. *Let Z be a subset of a linear space. A point $x \in Z$ is called a* centre of Z *if, for every $y \in Z$ it holds that*

$$\{tz + (1-t)y \colon t \in [0,1]\} \subset Z.$$

A set is called star-shaped *if it has at least one centre.*

The following proposition follows directly from the definition of a neural network:

Proposition 7.4. *Let $S = (d, N_1, \dots, N_L) \in \mathbb{N}^{L+1}$. Then $\mathcal{RNN}_\varrho(S)$ is scaling invariant, i.e. for every $\lambda \in \mathbb{R}$ it holds that $\lambda f \in \mathcal{RNN}_\varrho(S)$ if $f \in \mathcal{RNN}_\varrho(S)$, and hence $0 \in \mathcal{RNN}_\varrho(S)$ is a centre of $\mathcal{RNN}_\varrho(S)$.*

Knowing that $\mathcal{RNN}_\varrho(S)$ is star-shaped with centre 0, we can also ask ourselves if $\mathcal{RNN}_\varrho(S)$ has more than this one centre. It is not hard to see that also every constant function is a centre. The following theorem yields an upper bound on the number of centres.

Theorem 7.5 ([23, Proposition C.4]). *Let $S = (N_0, N_1, \dots, N_L)$ be a neural network architecture, let $\Omega \subset \mathbb{R}^{N_0}$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Then $\mathcal{RNN}_\varrho(S)$ contains at most $\sum_{\ell=1}^{L}(N_{\ell-1}+1)N_\ell$ linearly independent centres, where $N_0 = d$.*

Proof. Let $M^* \coloneqq \sum_{\ell=1}^{L}(N_{\ell-1}+1)N_\ell$. We first observe that $M^* = \dim(\mathcal{NN}(S))$.

Assume towards a contradiction, that there are functions $(g_i)_{i=1}^{M^*+1} \subset \mathcal{RNN}_\varrho(S) \subset L^2(\Omega)$ that are linearly independent.

By the Theorem of Hahn-Banach, there exist $(g_i')_{i=1}^{M^*+1} \subset (L^2(\Omega))'$ such that $g_i'(g_j) = \delta_{i,j}$, for all $i, j \in \{1, \dots, L+1\}$. We define

$$T \colon L^2(\Omega) \to \mathbb{R}^{M^*+1}, \qquad g \mapsto \begin{pmatrix} g_1'(g) \\ g_2'(g) \\ \vdots \\ g_{M+1}'(g) \end{pmatrix}.$$

Since T is continuous and linear, we have that $T \circ \mathrm{R}$ is locally Lipschitz continuous by Theorem 7.2. Moreover, since the $(g_i)_{i=1}^{M^*+1}$ are linearly independent, they span an M^*+1 dimensional linear space V and $T(V) = \mathbb{R}^{M^*+1}$.

Next we would like to establish that $\mathcal{RNN}_\varrho(S) \supset V$. Let $g \in V$ then

$$g = \sum_{\ell=1}^{M^*+1} a_\ell g_\ell,$$

for some $(a_\ell)_{\ell=1}^{M^*+1} \subset \mathbb{R}$. We show by induction that $\tilde{g}^{(m)} \coloneqq \sum_{\ell=1}^{m} a_\ell g_\ell \in \mathcal{RNN}_\varrho(S)$ for every $m \leq M^* + 1$. This is obviously true for $m = 1$. Moreover, we have that $\tilde{g}^{(m+1)} = a_{m+1} g_{m+1} + \tilde{g}^{(m)}$. Hence the induction step holds true if $a_{m+1} = 0$. If $a_{m+1} \neq 0$, then we have that

$$g^{(m+1)} = 2a_{m+1} \left(\frac{1}{2} g_{m+1} + \frac{1}{2a_{m+1}} \widetilde{g}^{(m)} \right), \tag{7.1}$$

By Proposition 7.4 $\widetilde{g}^{(m)}/(a_{m+1}) \in V$. Additionally, g_{m+1} is a centre of $\mathcal{RNN}_\varrho(S)$. Therefore, we have that $\frac{1}{2} g_{m+1} + \frac{1}{2a_{m+1}} \widetilde{g}^{(m)} \in \mathcal{RNN}_\varrho(S)$. By Proposition 7.4, we conclude that $\widetilde{g}^{(m+1)} \in \mathcal{RNN}_\varrho(S)$. Hence $V \subset \mathcal{RNN}_\varrho(S)$. Therefore, $T \circ \mathrm{R}(\mathcal{NN}(S)) \supseteq T(V) = \mathbb{R}^{M^*+1}$.

It is a well known fact of basic analysis that there does not exist a surjective and locally Lipschitz continuous map from $\mathbb{R}^n$ to $\mathbb{R}^{n+1}$ for any $n \in \mathbb{N}$. This yields the contradiction. □

For a convex set X, the line between any two points of X is a subset of X. Hence, every point of a convex set is a centre. This yields the following corollary.

Corollary 7.6. *Let $S = (N_0, N_1, \ldots, N_L)$ be a neural network architecture, let $\Omega \subset \mathbb{R}^{N_0}$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. If $\mathcal{RNN}_\varrho(S)$ contains more than $\sum_{\ell=1}^{L}(N_{\ell-1}+1)N_\ell$ linearly independent functions, then $\mathcal{RNN}_\varrho(S)$ is not convex.*

Remark 7.7. *It was shown in [23, Theorem 2.1] that the only Lipschitz continuous activation functions such that $\mathcal{RNN}_\varrho(S)$ contains* not *more than $\sum_{\ell=1}^{L}(N_{\ell-1}+1)N_\ell$ linearly independent functions are affine linear functions.*

Additionally, it can be shown that Corollary 7.6 holds for locally Lipschitz functions as well. In this case, $\mathcal{RNN}_\varrho(S)$ necessarily contains more than $\sum_{\ell=1}^{L}(N_{\ell-1}+1)N_\ell$ linearly independent functions if the activation function is not a polynomial.

In addition to the non-convexity of $\mathcal{RNN}_\varrho(S)$, we will now show that, under mild assumptions on the activation function, $\mathcal{RNN}_\varrho(S)$ is also *very* non-convex. Let us first make the notion of convexity quantitative.

Definition 7.8. *A subset X of a metric space is called* r-convex, *if $\bigcup_{x \in X} B_r(x)$ is convex.*

By Proposition 7.4, it is clear that $\mathcal{RNN}_\varrho(S) + B_r(0) = r\left(\mathcal{RNN}_\varrho(S) + B_1(0)\right)$. Hence,

$$\mathcal{RNN}_\varrho(S) + B_r(0) = r/r' \cdot \left(\mathcal{RNN}_\varrho(S) + B_{r'}(0)\right),$$

for every $r, r' > 0$. Therefore, $\mathcal{RNN}_\varrho(S)$ is r-convex for one $r > 0$ if and only if $\mathcal{RNN}_\varrho(S)$ is r-convex for every $r > 0$.

With this observation we can now prove the following result.

Proposition 7.9 ([23, Theorem 2.2.])**.** *Let $S \in \mathbb{N}^{L+1}$, $\Omega \subset \mathbb{R}^{N_0}$ be compact, and $\varrho \in C^1$ be discriminatory and such that $\mathcal{RNN}_\varrho(S)$ is not dense in $C(\Omega)$. Then there does not exist an $r > 0$ such that $\mathcal{RNN}_\varrho(S)$ is r-convex.*

Proof. By the discussion leading up to Proposition 7.9 we can assume, towards a contradiction that $\mathcal{RNN}_\varrho(S)$ is r-convex for every $r > 0$.

We have that

$$\mathrm{co}(\mathcal{RNN}_\varrho(S)) \subset \bigcap_{r>0} (\mathcal{RNN}_\varrho(S) + B_r(0)) \subset \bigcap_{r>0} (\overline{\mathcal{RNN}_\varrho(S)} + B_r(0)) \subset \overline{\mathcal{RNN}_\varrho(S)}.$$

Therefore $\mathrm{co}(\overline{\mathcal{RNN}_\varrho(S)}) = \overline{\mathrm{co}(\mathcal{RNN}_\varrho(S))} \subset \overline{\mathcal{RNN}_\varrho(S)}$ and thus we conclude that $\overline{\mathcal{RNN}_\varrho(S)}$ is convex.

We now aim at producing a contradiction by showing that $\overline{\mathcal{RNN}_\varrho(S)} = C(\Omega)$. We show this for $L = 2$, and $N_2 = 1$ only, the general case is demonstrated in [23, Theorem 2.2.] (there also the differentiability of ϱ is used).

Per assumption, for every $a \in \mathbb{R}^{N_1}$, $t \in \mathbb{R}$,

$$x \mapsto \varrho(ax - t) \in \overline{\mathcal{RNN}_\varrho(S)}.$$

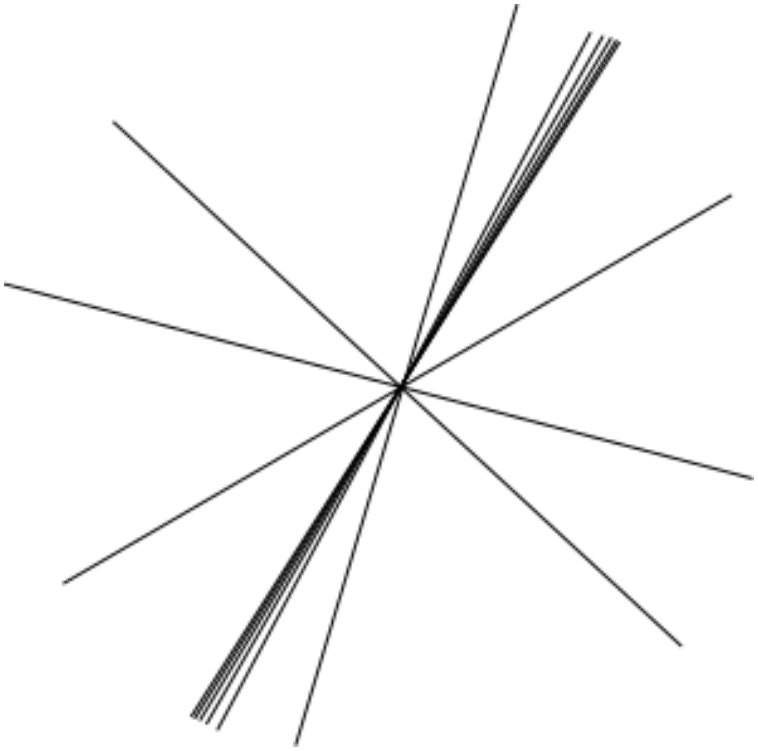

Figure 7.1: Sketch of the set of realisations of neural networks with a fixed architecture. This set is star-shaped, having 0 in the centre. It is not r-convex for any r and hence we see multiple holes between different rays. It is not closed, which means that there are limit points outside of the set.

By the same argument applied in the proof of Theorem 7.5 in (7.1), we have that for all sequences $(a_\ell)_{\ell=1}^\infty \subset \mathbb{R}^{N_1}$, $(b_\ell)_{\ell=1}^\infty \subset \mathbb{R}$, and $(t_\ell)_{\ell=1}^\infty \subset \mathbb{R}$ the function

$$g^{(m)}(x) \coloneqq \sum_{\ell=1}^{m} b_\ell \varrho(a_\ell x - t_\ell)$$

satisfies $g^{(m)} \subset \overline{\mathcal{RNN}_\varrho(S)}$ for all $m \in \mathbb{N}$.

By Theorem 2.4, we have that

$$\overline{\left\{\sum_{\ell=1}^{m} b_\ell \varrho(a_\ell \cdot - t_\ell) \colon (a_\ell)_{\ell=1}^\infty \subset \mathbb{R}^{N_1}, (b_\ell)_{\ell=1}^\infty, (t_\ell)_{\ell=1}^\infty \subset \mathbb{R}\right\}} = C(\Omega)$$

and hence $C(\Omega) \subset \overline{\mathcal{RNN}_\varrho(S)}$ which yields the desired contradiction. □

7.2 Network spaces are not closed

The second property that we would like to understand is closedness. To make this more precise, we need to decide on a norm first. We will now study closedness in the uniform norm.

Theorem 7.10. *Let $L \in \mathbb{N}$, $S = (N_0, N_1, \dots, N_{L-1}, 1) \in \mathbb{N}^{L+1}$, where $N_1 \geq 2$, $\Omega \in \mathbb{R}^d$ compact with nonempty interior, and $\varrho \in C^2 \setminus C^\infty$. Then $\mathcal{RNN}_\varrho(S)$ is not closed in L^∞.*

Proof. Since $\varrho \in C^2 \setminus C^\infty$, we have that there exists $k \in \mathbb{N}$ such that $\varrho \in C^k$ and $\varrho \notin C^{k+1}$. It is not hard to see that therefore $\mathcal{RNN}_\varrho(S) \subset C^k(\Omega)$ and the map

$$F : \mathbb{R}^d \to \mathbb{R} \colon x \mapsto F(x) = \varrho'(x_1)$$

is not in $C^k(\mathbb{R}^d)$. Therefore, since Ω has non-empty interior, there exists $t \in \mathbb{R}^d$ so that $F(\cdot - t) \notin C^k(\Omega)$ and thus $F(\cdot - t) \notin \mathcal{RNN}_\varrho(S)$.

Assume for now that $S = (N_0, 2, 1)$. The general statement follows by extending the networks below to neural networks with architecture $(N_0, 2, 1, \dots, 1, 1)$ by concatenating with the neural networks from Proposition 2.11. To artificially increase the width of the networks and produce neural networks of architecture S one can simply zero-pad the weight and shift matrices without altering the associated realisations.

We define the neural network

$$\Phi_n \coloneqq \left(\left(\begin{pmatrix} 1 & 0_{1\times(N_0-1)} \\ 1 & 0_{(N_1-1)\times 1} \end{pmatrix}, \begin{pmatrix} 1/n \\ 0 \end{pmatrix} \right), ([n, -n], 0) \right),$$

and observe that for every $x \in \Omega$

$$|\mathrm{R}(\Phi_n)(x) - \varrho'(x_1)| = |n(\varrho(x_1 + 1/n) - \varrho(x_1)) - \varrho'(x_1)| \leq \sup_{z\in[-B,B]} |\varrho''(z)|/n,$$

by the mean value theorem, where $B > 0$ is such that $\Omega \subset [-B, B]^d$. Therefore, $\mathrm{R}(\Phi_n) \to F$ in $L^\infty(\Omega)$ and hence $\mathcal{RNN}_\varrho(S)$ is not closed. □

Remark 7.11. *Theorem 7.10 holds in much more generality. In fact, a similar statement holds for various types of activation functions, see [23, Theorem 3.3]. Surprisingly, the statement does not hold for the ReLU activation function, [23, Theorem 3.8].*

Theorem 7.10, should be contrasted to the following result that shows that subsets of the set of realisations of neural networks with bounded weights are always closed.

Proposition 7.12. *Let $S \in \mathbb{N}^{L+1}$, $\Omega \subset \mathbb{R}^{N_0}$ be compact, and ϱ be continuous. For $C > 0$, we denote by*

$$\mathcal{RNN}^C \coloneqq \{\mathrm{R}(\Phi) \colon \Phi \in \mathcal{NN}(S), \|\Phi\|_{\mathrm{total}} \leq C\}$$

the set of realisations of neural networks with weights bounded by C. Then $\mathcal{RNN}^C$ is a closed subset of $C(\Omega)$.

Proof. By the Theorem of Heine-Borel, we have that

$$\{\Phi \in \mathcal{NN}(S) \colon \|\Phi\|_{\mathrm{total}} \leq C\}$$

is compact. Hence the result follows by Theorem 7.2. □

Combining Theorem 7.10 and Proposition 7.12 yields the following observation: Consider a function $g \in \overline{\mathcal{RNN}_\varrho(S)} \setminus \mathcal{RNN}_\varrho(S)$ and a sequence $\Phi_n \in \mathcal{NN}(S)$ so that

$$\mathrm{R}(\Phi_n) \to g.$$

Then $\|\Phi_n\|_{\mathrm{total}} \to \infty$ since if $\|\Phi_n\|_{\mathrm{total}}$ would remain bounded, then $g \in \overline{\mathcal{RNN}_\varrho(S)^C} = \mathcal{RNN}_\varrho(S)^C \subset \mathcal{RNN}_\varrho(S)$.

References

[1] M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.

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

[3] R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

[4] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.

[5] E. K. Blum and L. K. Li. Approximation theory and feedforward networks. *Neural networks*, 4(4):511–515, 1991.

[6] C. K. Chui and H. N. Mhaskar. Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, 4:12, 2018.

[7] A. Cohen and R. DeVore. Approximation of high-dimensional parametric pdes. *Acta Numerica*, 24:1–159, 2015.

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

[9] R. A. DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[10] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

[11] C. L. Frenzen, T. Sasao, and J. T. Butler. On the number of segments needed in a piecewise linear approximation. *Journal of Computational and Applied mathematics*, 234(2):437–446, 2010.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

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

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

[15] A. N. Kolmogorov. The representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Doklady Akademii Nauk SSSR*, 108(2):179–182, 1956.

[16] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.*, 6(6):861–867, 1993.

[17] V. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3):81–91, 1999.

[18] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.

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

[20] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[21] E. Novak and H. Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *Journal of Complexity*, 25(4):398–404, 2009.

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

[23] P. C. Petersen, M. Raslan, and F. Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *arXiv preprint arXiv:1806.08459*, 2018.

[24] P. C. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.*, 180:296–330, 2018.

[25] F. V. Philipp Petersen. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv preprint arXiv:2112.12555*, 2021.

[26] G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz")*, 1980-1981.

[27] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *Int. J. Autom. Comput.*, 14(5):503–519, 2017.

[28] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[29] W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991.

[30] W. Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 2006.

[31] I. Safran and O. Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2979–2987, 2017.

[32] J. Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

[33] I. Schoenberg. Cardinal interpolation and spline functions. *Journal of Approximation theory*, 2(2):167–206, 1969.

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

[35] Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

[36] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

[37] M. Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

[38] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[39] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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

[41] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference On Learning Theory*, pages 639–649, 2018.