FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. et Bc. Vít Fojtík

Approximation of Functions Continuous on Compact Sets by Layered Neural Networks

Department of Theoretical Computer Science and Mathematical Logic

Supervisor of the master thesis: Ing. František Hakl, CSc.
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2022

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

Above all, I would like to thank my supervisor, Ing. František Hakl, CSc., for his ideas and insights, as well as his long-lasting, patient support. His incredibly thorough and rigorous approach to research not only elevated the quality of this thesis, but also hopefully impacted my future work-style. Further thanks to Daniel Kršek for his readiness to engage in helpful discussions, Petra Štarmanová for her readiness to engage in useless discussions and Tomáš Kult for being around.

Title: Approximation of Functions Continuous on Compact Sets by Layered Neural Networks

Author: Bc. et Bc. Vít Fojtík

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Ing. František Hakl, CSc., Department of Machine Learning, Institute of Computer Science, Czech Academy of Sciences

Abstract: Despite abundant research into neural network applications, many areas of the underlying mathematics remain largely unexplored. The study of neural network expressivity is vital for understanding their capabilities and limitations. However, even for shallow networks this topic is far from solved. We provide an upper bound on the number of neurons of a shallow neural network required to approximate a function continuous on a compact set with given accuracy. Dividing the compact set into small polytopes, we approximate the indicator function of each of them by a neural network and combine these into an approximation of the target function. This method, inspired by a specific proof of the Stone-Weierstrass Theorem, is more general than previous bounds of this character, with regards to approximation of continuous functions. Also, it is purely constructive.

Keywords: Neural Networks, Approximation, Space Complexity of Networks

Contents

Notation Index

$\mathbb{N}$	The set of all natural numbers (excluding 0)
$\mathbb{R}$	The set of all non-negative real numbers
d	A fixed natural number larger than 1
$B(x,\varepsilon)$	The d-dimensional open ball centred at $\boldsymbol{x} \in \mathbb{R}^d$ with radius $\varepsilon > 0$
$\langle \boldsymbol{a}, \boldsymbol{x} \rangle$	The standard scalar product of vectors $\boldsymbol{a}$ and $\boldsymbol{x}$
$\overline{A}$	The closure of a set A
1_A	The indicator function of a set A
K	A fixed compact subset of $\mathbb{R}^d$
$C(K)$	The space of continuous functions from K to $\mathbb{R}$
$\lVert\cdot\rVert$	The Euclidean norm (on $\mathbb{R}^d$)
$\lVert\cdot\rVert_\infty$	The sup norm (on a space of functions)
ρ	The rectifier function, $\rho : \mathbb{R} \to \mathbb{R}$, $\rho(x) = x^+ = \max(0, x)$
S^{d-1}	The d-dimensional unit sphere, $S^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \lVert\boldsymbol{x}\rVert = 1\}$
$f\|_A$	The restriction of a function f to a set A
$M^\top$	The transpose of a matrix M
$\mathcal{P}(A)$	The set of all subsets of a set A

Introduction

Even though neural networks have been a dominant area of research for years now, we still lack systematic theoretical understanding and many fundamental questions remain unanswered. One of the more studied theoretical topics is expressive power of shallow (one-hidden-layer) networks, where several versions of approximational universality have been proved. However, no bounds found in the literature are general and explicit enough to allow us in practice, given a continuous function, to estimate the sufficient and necessary size of a shallow network required to approximate the function with given accuracy. The main goal of the thesis is to provide such a bound.

Given a compact set $K \subseteq \mathbb{R}^d$, a continuous function $f : K \to \mathbb{R}$ and $\varepsilon > 0$, we construct an upper bound on the least number h such that there exists a neural network g that has a single hidden layer consisting of h neurons and that satisfies $\|f - g\|_\infty < \varepsilon$. Our bound depends on the input dimension $d \geq 2$, on the diameter of the set K denoted diam K, on the sup norm of f, $\|f\|_\infty$, and of course on ε. Furthermore, complexity of the target function f is expressed by the inverse modulus of continuity,

$$\omega^{-1}(f, \varepsilon) = \sup \left\{\delta' > 0 \mid \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in K : \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \delta' \Rightarrow |f(\boldsymbol{y}_1) - f(\boldsymbol{y}_2)| \leq \varepsilon \right\}.$$

Denoting $\delta = \omega^{-1}\left(f, \frac{\varepsilon}{2}\right)$, one of the bounds we obtained is

$$h \leq \left(6\sqrt{d+1}\frac{\operatorname{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{10(d+1)^4\frac{\operatorname{diam}(K)}{\delta}}.$$

This holds for any K, f and ε.

Previous bounds of this nature are formulated for f contained in Sobolev spaces. The Sobolev space $W^{k,p}(K)$ is the set of functions that have bounded weak derivatives of all orders up to k in the L^p norm (see Subsection 1.2 for more details).

For $f : [0,1]^d \to \mathbb{R}$ lying in the unit ball of the space $W^{k,p}\left([0,1]^d\right)$, Maiorov [1] and Mhaskar [2] provide bounds of

$$C_1\left(\frac{1}{\varepsilon}\right)^{\frac{d-1}{k}} \leq h \leq C_2\left(\frac{1}{\varepsilon}\right)^{\frac{d}{k}}.$$

Note that here dependence on the norm of f is hidden in the assumption that f lies in the unit ball.

In order for any of these bounds to be applicable, the function f needs to be at least Lipschitz (or satisfy a Hölder condition) — They cannot be used for general continuous functions. Also, Sobolev spaces require a very nice underlying set. The bounds cannot be applied for example for K not full-dimensional, or even for many reasonable choices of K. However, in situations where both are applicable, these bounds surpass our result.

Most previous bounds are formulated for networks with a sigmoidal activation function and others have been stated for ReLU networks. Our bound uses the exponential activation, $\exp(t) = e^t$. This is because we use the property that the

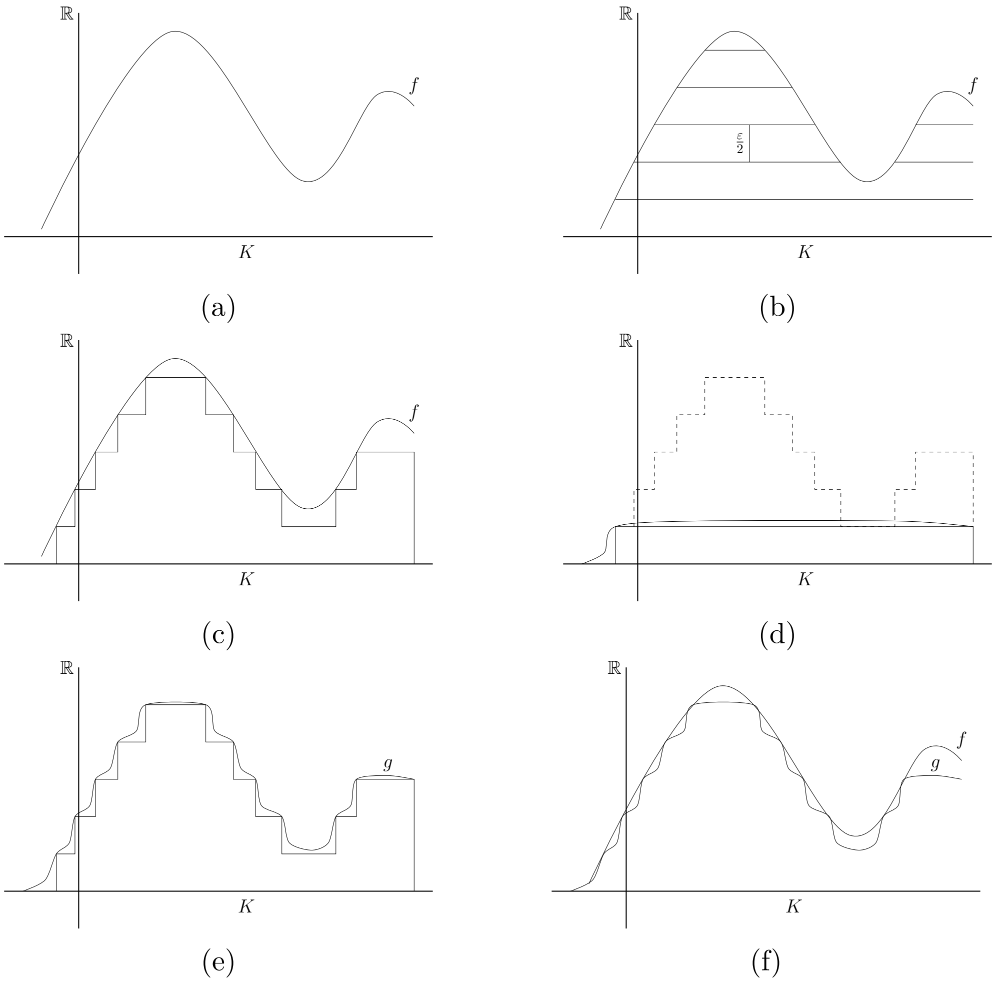

Figure 1: (a) We are given a function f continuous on a compact set K. (b) We divide the graph of f into horizontal slices. (c) Sum of indicator functions of the slices approximates f. (d) We approximate the indicator function of each slice by a neural network. (e) Then, we sum these approximants. (f) The sum approximates the original function f.

product of two functions representing such a neural network is again a function representing a neural network. Even though it is possible to transfer the bound to sigmoidal or ReLU activations, it was beyond the scope of the thesis to provide an efficient transition — In Subsection 2.2 we provide a simple method of transfer that results in a significant increase of the bound.

Our construction is inspired by a proof of the Stone-Weierstrass Theorem, in particular by Brosowski and Deutsch [3]. As illustrated in Figure 1, we divide the graph of the target function f into horizontal slices of height $\frac{\varepsilon}{2}$. We approximate the indicator function of each of these slices and sum them to get an approximation of f.

The non-trivial step here is approximating the indicator functions of the slice sets, as these can be complex. Our solution, shown in Figure 2, is to divide K into congruent polytopes, approximate the indicator function of each of them and take the product of those approximants that correspond to polytopes intersecting the set.

The indicator function of a single polytope is then approximated by taking

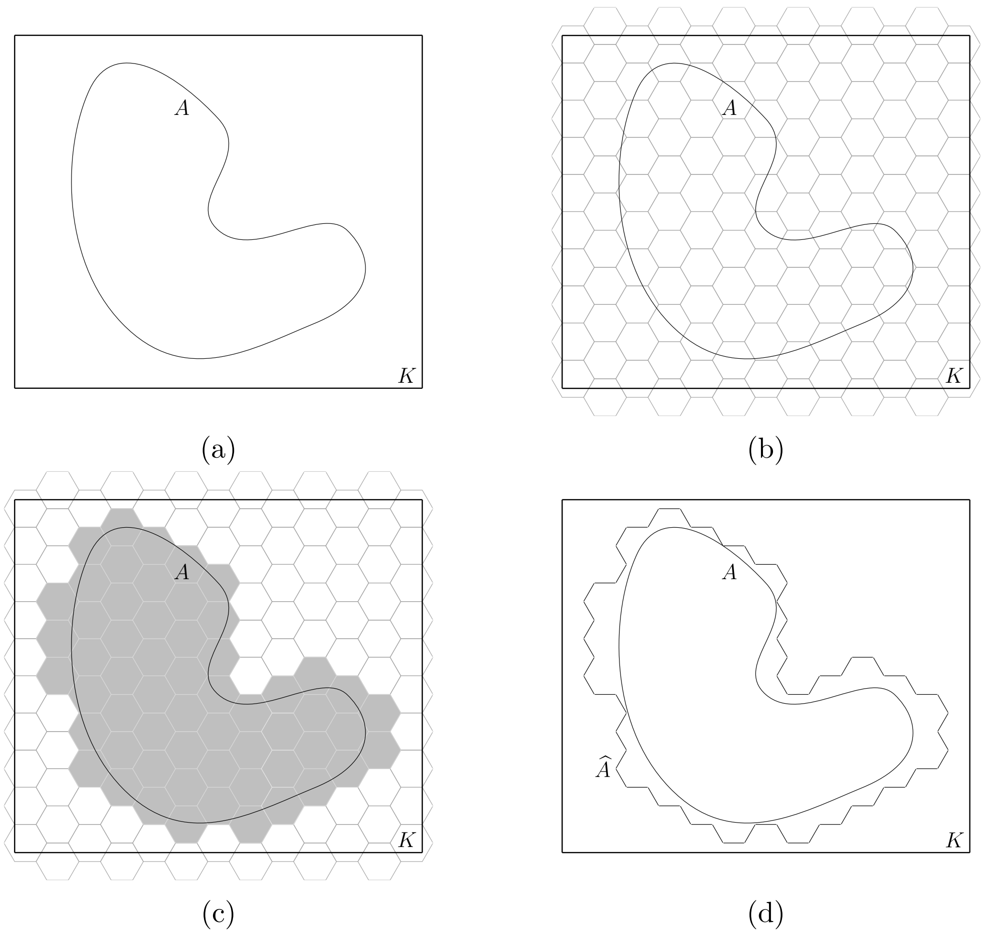

Figure 2: (a) We want to approximate the indicator function of a set $A \subseteq K$. (b) We divide K into regular polytopes and approximate the indicator of each of them. (c) We take the product of all approximants that correspond to polytopes intersecting A. (d) The result approximates the indicator of the set $\widehat{A}$ defined as the union of polytopes that intersect A, which in turn is an approximation of A.

an exponential function for each of its facets and exponentiating their average in a specific way.

As for the structure of the thesis, in Chapter 1 we survey some previous results on the expressive power of shallow neural networks. First, we look at different ways of proving shallow networks are universal approximators — Section 1.1 introduces three such methods. We illustrate each method by applying it to the ReLU activation function. Thereafter, in Section 1.2 we survey some previous bounds on network size found in the literature.

However, the main focus lies in Chapter 2, which presents the construction described above. We proceed in the opposite direction from the one seen here, starting with approximation of a polytope's indicator function in Section 2.1. Section 2.2 then contains the rest of the construction. More precisely, in Subsection 2.2.1 we survey some results from the theory of lattices, in order to make the decomposition of K into polytopes as efficient as possible — This corresponds to an open problem called the Lattice Covering Problem. Finally, Subsection 2.2.2 puts this all together to approximate a continuous function on a compact set.

To conclude Chapter 2 we explore an alternative approach in Section 2.3. If we are content with approximating the target function everywhere on K except for a set of small measure, we can use the Vitali Covering Theorem to get a sequence of small enough disjoint balls that cover most of K. Then, we approximate each of these balls by a polytope and approximate each polytope's indicator by a neural network as in Section 2.1 — This is why we set Section 2.1 apart from the rest of the original construction. Subsequently, a weighted sum of these neural networks approximates the target function. However, to fully complete this construction we would need to bound the number of balls in the Vitali Covering Theorem (or, more precisely, the minimal radius). Such a bound is not found in the literature and it is beyond the scope of the thesis to construct one.

Subsection 2.3.1 studies approximations of a ball by a polytope. There are two independent proofs stating how many facets of a polytope we need to approximate a convex body with given accuracy, but one of them, by Dudley [4], is hidden in terms of a complicated theory, while the other, by Bronshteyn and Ivanov [5], is very brief. Therefore, as we have not found a clearer reformulation of either in the literature, we created one for the special case of a ball. This also allows us to specify the bound explicitly, without any non-specific constants. We then use the results in 2.3.2 to finish the sketch of the construction.

1. Expressivity of Shallow Neural Networks

In this chapter, we survey some of the most important among the multitude of results concerning approximation properties of shallow neural networks. First, we review classical methods of proving approximability of classes of functions on compact sets by shallow neural networks. Thereafter, we look at bounds on network size in relation to approximation accuracy.

In almost all of the thesis, we focus on shallow (one-hidden-layer) networks with a ridge activation function, that is, a function of the form $\varphi(\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b)$, where $\varphi : \mathbb{R} \to \mathbb{R}$. We adopt the following formalisms.

Definition 1. *Let H be a set of functions from K to $\mathbb{R}$, sometimes called a dictionary. Define the affine span of H as*

$$\operatorname{span} H = \left\{ c_0 + \sum_{i=1}^{n} c_i h_i \mid c_0 \in \mathbb{R}, n \in \mathbb{N}, \forall i \leq n : c_i \in \mathbb{R}, h_i \in H \right\}.$$

For any function $\varphi : \mathbb{R} \to \mathbb{R}$ and $k \in \mathbb{N}$, the set of (d-dimensional) functions represented by a k-hidden-layer neural network with activation φ is defined inductively as follows:

$$G_\varphi^{(0)} = \operatorname{span} \Pi,$$

where Π is the set of all d projection functions, and

$$G_\varphi^{(k+1)} = \operatorname{span} \varphi \circ G_\varphi^{(k)}.$$

Here, $\varphi \circ G_\varphi^{(k)}$ denotes the set of all compositions $\left\{ \varphi \circ g \mid g \in G_\varphi^{(k)} \right\}$.

Remark. We will write simply G_φ instead of $G_\varphi^{(1)}$ and call it the set of representables for short. In other words,

$$G_\varphi = \left\{ c_0 + \sum_{i=1}^{m} c_i \varphi(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + b_i) \mid m \in \mathbb{N}, c_0 \in \mathbb{R}, \forall i \leq m : \boldsymbol{a}_i \in \mathbb{R}^d, c_i, b_i \in \mathbb{R} \right\}.$$

Example. Probably the most well-known examples of activation functions are sigmoidal functions and the rectifier. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is sigmoidal if it is continuous and if $\lim_{t \to -\infty} \sigma(t) = 0$ and $\lim_{t \to \infty} \sigma(t) = 1$. Then, we say the set

$$H_\sigma = \left\{ \sigma\left(\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b \right) \mid \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

is a sigmoidal dictionary. We have $G_\sigma = \operatorname{span} H_\sigma$ — G_σ can be seen as a free vector space over H_σ.

Let $\rho : \mathbb{R} \to \mathbb{R}$ be the rectifier function, $\rho(t) = \max(0, t)$. Then,

$$H_\rho = \left\{ \max\left(0, \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b\right) \mid \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

is the ReLU (rectified linear unit) dictionary. Again, $G_\rho = \operatorname{span} H_\rho$.

The central result of this chapter is the following, proved independently by Cybenko [6], Hornik et al. [7] and Funahashi [8].

Theorem 1 (Universal Approximation Theorem)**.** *Let $K \subseteq \mathbb{R}^d$ be compact and let $\sigma : \mathbb{R} \to \mathbb{R}$ be a sigmoidal function. Then, the set of functions represented by a one-hidden-layer neural network with activation σ is dense in the space of continuous functions on K with the supremum norm, $\overline{G_\sigma} = C(K)$.*

1.1 Approximation Methods

In the following, we review several methods of showing that G_σ (or a different G_φ) is dense in the space $C(K)$ of real continuous functions on K. We can restate the claim of Theorem 1 in the following way: Any function in $C(K)$ can be approximated with arbitrary precision by a function represented by a neural network. Or, more formally, for all $\varepsilon > 0$ and all $f \in C(K)$ there exists $g \in G_\varphi$ such that $\|f - g\|_\infty < \varepsilon$.

1.1.1 Stone-Weierstrass Theorem

One way to prove density of the set of representables G_φ generated by a function φ is to show that affine combinations of φ can approximate a one-dimensional function h (or a set of functions) for which we know that G_h is dense. A similar method was employed by Hornik et al. [7]. We formalised this by proving the following simple claim.

Theorem 2. *Let $K \subseteq \mathbb{R}^d$ be compact, $h : \mathbb{R} \to \mathbb{R}$ such that G_h is dense in the space of continuous functions on K, $C(K)$, and $\varphi : \mathbb{R} \to \mathbb{R}$. If for all $\delta > 0$ and all $u < v \in \mathbb{R}$ there exist $n \in \mathbb{N}$ and $\{a_j\}_{j=1}^n, \{b_j\}_{j=1}^n, \{c_j\}_{j=0}^n \subseteq \mathbb{R}$ such that*

$$\sup_{t\in[u,v]} \left| h(t) - \left(c_0 + \sum_{j=1}^n c_j \varphi\left(a_j t + b_j\right) \right) \right| < \delta,$$

then G_φ is dense in $C(K)$.

Proof. For $f \in C(K)$ and $\varepsilon > 0$, take $m \in \mathbb{N}$, $\{\boldsymbol{\alpha}_i\}_{i=1}^m \subseteq \mathbb{R}^d$, $\{\beta_i\}_{i=1}^m, \{\gamma_i\}_{i=0}^m \subseteq \mathbb{R}$ such that

$$\left\| f(\boldsymbol{x}) - \left(\gamma_0 + \sum_{i=1}^m \gamma_i h(\langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i) \right) \right\|_\infty < \varepsilon.$$

For each $i \leq m$, define $u_i = \inf_{\boldsymbol{x}\in K} \langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i$ and $v_i = \sup_{\boldsymbol{x}\in K} \langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i$. Then, $\delta_i = \frac{\varepsilon}{m|\gamma_i|}$, there exist coefficients $\{a_j^i\}_{j=1}^{n_i}$, $\{b_j^i\}_{j=1}^{n_i}$, $\{c_j^i\}_{j=0}^{n_i}$ such that

$$\sup_{t\in[u_i,v_i]} \left| h(t) - \left(c_0^i + \sum_{j=1}^{n_i} c_j^i \varphi\left(a_j^i t + b_j^i\right) \right) \right| < \delta_i.$$

Denoting by φ_i the function

$$\varphi_i(\boldsymbol{x}) = c_0^i + \sum_{j=1}^{n_i} c_j^i \varphi(a_j^i(\langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i) + b_j^i),$$

we have

$$\sup_{\boldsymbol{x}\in K} \left| \gamma_i h\left(\langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i\right) - \gamma_i \varphi_i(\boldsymbol{x}) \right| < \frac{\varepsilon}{m}.$$

Together, this implies

$$\begin{aligned} \left\| f(\boldsymbol{x}) - \left(\gamma_0 + \sum_{i=1}^m \gamma_i \varphi_i(\boldsymbol{x}) \right) \right\|_\infty \leq & \left\| f(\boldsymbol{x}) - \left(\gamma_0 + \sum_{i=1}^m \gamma_i h(\langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i) \right) \right\|_\infty + \\ & \sum_{i=1}^m \quad \left\| \gamma_i h(\langle \boldsymbol{\alpha_i}, \boldsymbol{x} \rangle + \beta_i) - \gamma_i \varphi_i(\boldsymbol{x}) \right\|_\infty < 2\varepsilon. \end{aligned}$$

Since $(\gamma_0 + \sum_{i=1}^m \gamma_i \varphi_i(\boldsymbol{x})) \in G_\varphi$, we get $f \in \overline{G_\varphi}$. $\square$

This proposition is useful in case we can prove density of G_h for a function h that is not particularly suitable as an activation function of neural nets.

The next claim, proved in this form by Chen et al. [9], shows that for sigmoidal dictionaries we can use Theorem 2 with any continuous function h. Furthermore, it gives a precise bound on approximation accuracy.

Definition 2. *Let $K \subseteq \mathbb{R}^d$. Denote the modulus of continuity of $f : K \to \mathbb{R}$ by*

$$\omega(f, \delta) = \sup \{|f(\boldsymbol{x}) - f(\boldsymbol{y})| \mid \boldsymbol{x}, \boldsymbol{y} \in K \ \& \ \|\boldsymbol{x} - \boldsymbol{y}\| < \delta\}$$

Theorem 3. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $\lim_{t\to\infty} \sigma(t) = 1$, $\lim_{t\to-\infty} \sigma(t) = 0$ and $\sigma(0) = 1$. Then for any $h \in C([0, 1])$:*

$$\|h - g_n\|_\infty \leq \omega\left(h, \frac{1}{n}\right)\left(4 + 2 \sup_{t\in\mathbb{R}} |\sigma(t)|\right),$$

where

$$g_n(t) = h(0) + \sum_{i=1}^{n} \left(h\left(\frac{i}{n}\right) - h\left(\frac{i-1}{n}\right)\right) \sigma(l_n(nt - i)),$$

where l_n is the smallest positive integer such that for $t \leq -l_n$: $|\sigma(t)| < \frac{1}{n}$ and for $t \geq l_n$: $|1 - \sigma(t)| \leq \frac{1}{n}$.

Remark. This implies a similar result for all $C([a, b])$, although with a potentially different accuracy.

The proof of the theorem uses the fact that any sigmoidal σ approximates the Heaviside step function $\sigma_0 = 1_{[0,\infty]}$ in the sense that for $t \neq 0$:

$$\lim_{k\to\infty} \sigma(kt) = \sigma_0(t).$$

To show that neural networks with any sigmoidal activation function are dense in $C(K)$ it now suffices to find a continuous $h : \mathbb{R} \to \mathbb{R}$ such that G_h is dense in $C(K)$. This can be done using the Stone–Weierstrass Theorem (see for example [10, Theorem 7.32]).

Theorem 4 (Stone–Weierstrass)**.** *Let $K \subseteq \mathbb{R}^d$ be compact and let G be a vector subspace of $C(K)$ satisfying*

(i) G is closed under multiplication,

(ii) G contains a constant, non-zero function,

(iii) for all $\boldsymbol{x} \neq \boldsymbol{y} \in K$ there exists $h \in G$ such that $h(\boldsymbol{x}) \neq h(\boldsymbol{y})$.

Then, G is dense in $C(K)$.

We can easily verify that all three conditions are satisfied for the exponential dictionary. There are also other options, such as the dictionary of all polynomial functions (even though it is not generated by a single function, the main ideas remain the same).

Corollary 5. Let $K \subseteq \mathbb{R}^d$ be compact. Then, $G_{\exp}$ is dense in $C(K)$, where $\exp(t) = e^t$ is the exponential function.

Combining Theorems 2 and 3 with Corollary 5, we get:

Corollary 6. Let $K \subseteq \mathbb{R}^d$ be compact and let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous such that $\lim_{t\to\infty} \sigma(t) = 1$, $\lim_{t\to-\infty} \sigma(t) = 0$ and $\sigma(0) = 1$. Then, the set of representables G_σ is dense in $C(K)$.

The ReLU Dictionary

We illustrate the method by proving that neural networks with the rectified linear unit activation, $\rho(t) = \max(0, t)$, are dense in $C(K)$. This follows from the fact that a one-dimensional ReLU network can represent a sigmoidal function, for example $\varphi(t) = \rho(t) - \rho(t-1)$. Since $G_\varphi \subseteq G_\rho$, density of G_φ implies density of G_ρ. However, we constructed a variation of the proof adapted directly for ρ in order to provide an example of the method.

To that end we prove the following, which corresponds to Theorem 3 — That the rectifier can approximate any continuous function on the one-dimensional unit interval. Similar proofs are known, but we created our own for clarity. Notice the approximation accuracy is better than in the sigmoidal case.

Theorem 7. *For all $h \in C([0,1])$ and all $n \in \mathbb{N}$,*

$$\left\| h(t) - \left(h(0) + \sum_{i=1}^{n} c_i \rho(t + b_i) \right) \right\|_\infty \leq 2\omega\left(h, \frac{1}{2n}\right),$$

where $b_i = -\frac{i-1}{n}$, $c_1 = n\left(h\left(\frac{1}{n}\right) - h(0)\right)$ and $c_i = n\left(h\left(\frac{i}{n}\right) - 2h\left(\frac{i-1}{n}\right) + h\left(\frac{i-2}{n}\right)\right)$ for $i \in \{2, \ldots, n\}$.

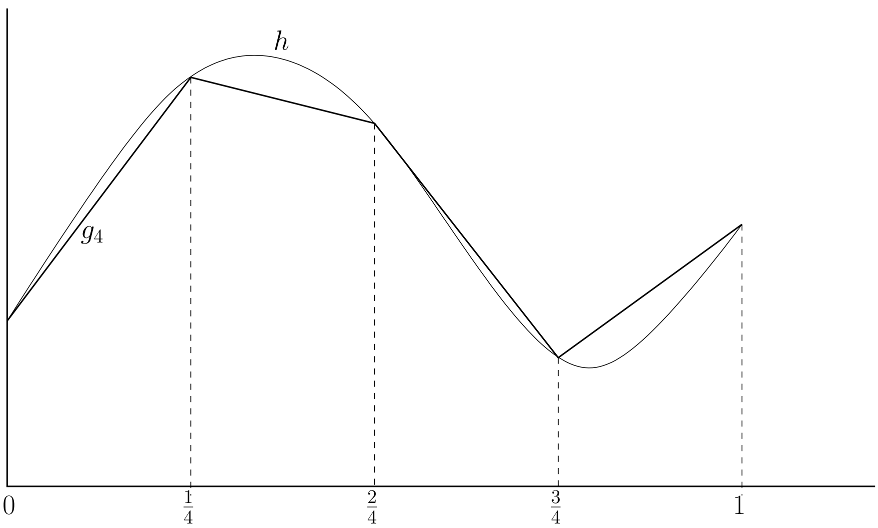

Figure 1.1: The function g_4 is a polygonal chain with 4 vertices approximating h.

Proof. Denote $g_n(t) = h(0) + \sum_{i=1}^{n} c_i \rho(t + b_i)$. We will show that g_n is a polygonal chain as in Figure 1.1.

For $j \in \{1, \ldots, n\}$ and $t \in \left[\frac{j-1}{n}, \frac{j}{n}\right)$, $\rho(t + b_i)$ equals 0 for $i > j$ and $t + b_i$ for $i \leq j$. Therefore,

$$g_n(t) = h(0) + \sum_{i=1}^{j} c_i (t + b_i).$$

We will show by induction that

$$h(0) + \sum_{i=1}^{j} c_i (t + b_i) = n\left(h\left(\frac{j}{n}\right) - h\left(\frac{j-1}{n}\right)\right)t - (j-1)h\left(\frac{j}{n}\right) + jh\left(\frac{j-1}{n}\right).$$

For $j = 1$ this equals $n\left(h\left(\frac{1}{n}\right) - h(0)\right)t + h(0)$. As for the induction step,

$$
\begin{aligned}
h(0) + \sum_{i=1}^{j} c_i\left(t + b_i\right) + c_{j+1}\left(t + b_{j+1}\right) = \\
= \; & n\left(h\left(\frac{j}{n}\right) - h\left(\frac{j-1}{n}\right)\right)t - (j-1)h\left(\frac{j}{n}\right) + jh\left(\frac{j-1}{n}\right) + \\
& n\left(h\left(\frac{j+1}{n}\right) - 2h\left(\frac{j}{n}\right) + h\left(\frac{j-1}{n}\right)\right)\left(t - \frac{j}{n}\right) \\
= \; & n\left(h\left(\frac{j+1}{n}\right) - h\left(\frac{j}{n}\right)\right)t - jh\left(\frac{j+1}{n}\right) + (j+1)h\left(\frac{j}{n}\right).
\end{aligned}
$$

Hence, g_n is a polygonal chain with vertices $\left(\frac{i}{n}, h\left(\frac{i}{n}\right)\right)$, $i \le n$, as shown in Figure 1.1.

For all i, $\omega\left(h, \frac{1}{2n}\right) \ge \frac{1}{2}\left|h\left(\frac{i}{n}\right) - h\left(\frac{i-1}{n}\right)\right|$, since

$$
\begin{aligned}
\left|h\left(\frac{i}{n}\right) - h\left(\frac{i-1}{n}\right)\right| &\le \left|h\left(\frac{i-1}{n}\right) - h\left(\frac{i-1}{n} + \frac{1}{2n}\right)\right| + \left|h\left(\frac{i}{n} - \frac{1}{2n}\right) - h\left(\frac{i}{n}\right)\right| \\
&= \lim_{t \to \frac{1}{2n}^-}\left|h\left(\frac{i-1}{n}\right) - h\left(\frac{i-1}{n} + t\right)\right| + \lim_{t \to \frac{1}{2n}^-}\left|h\left(\frac{i}{n} - t\right) - h\left(\frac{i}{n}\right)\right| \\
&\le \omega\left(h, \frac{1}{2n}\right) + \omega\left(h, \frac{1}{2n}\right).
\end{aligned}
$$

Therefore, for all $t \in \left[\frac{i-1}{n}, \frac{i-1}{n} + \frac{1}{2n}\right)$:

$$
\begin{aligned}
|h(t) - g_n(t)| \le & \left|h(t) - h\left(\frac{i-1}{n}\right)\right| + \left|h\left(\frac{i-1}{n}\right) - g_n(t)\right| \\
& \le \omega\left(h, \frac{1}{2n}\right) + \frac{1}{2}\left|h\left(\frac{i}{n}\right) - h\left(\frac{i-1}{n}\right)\right| \le 2\omega\left(h, \frac{1}{2n}\right).
\end{aligned}
$$

Similarly for $t \in \left(\frac{i-1}{n} + \frac{1}{2n}, \frac{i}{n}\right]$ and by continuity the inequality holds for all $t \in [0, 1]$. $\square$

Therefore, the rectifier can approximate any one-dimensional function on an interval, including the exponential, and combining Theorem 7 with Theorem 2 and Corollary 5, we get the following.

Corollary 8. Let $K \subseteq \mathbb{R}^d$ be compact and let $\rho : \mathbb{R} \to \mathbb{R}$ denote the rectifier function. Then, the set of representables G_ρ is dense in $C(K)$.

1.1.2 Convolution

Apart from more direct approaches, such as the one in Subsection 1.1.1, proofs based on integral approximation are common in approximation theory. Two such methods are presented in the following two subsections.

The convolution method uses the fact that a sequence of convolutions of the goal function can be shown to converge to the function and we can approximate the convolutions by the dictionary in question. The approach was first adopted in this way by Xu et al. [11], although the proof by Funahashi [8] proceeded similarly.

Theorem 9. *For $h \in L^1(\mathbb{R}^d)$ (a function Lebesgue-absolutely integrable on $\mathbb{R}^d$) such that $\int h(\boldsymbol{x})\, d\boldsymbol{x} = 1$, let $h_n(\boldsymbol{x}) = n^d h(n\boldsymbol{x})$. Furthermore, let $f : \mathbb{R}^d \to \mathbb{R}$ be bounded and uniformly continuous. Then,*

$$f * h_n \xrightarrow[n\to\infty]{\|\cdot\|_\infty} f,$$

where $f * h_n(\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{x} - \boldsymbol{y}) h_n(\boldsymbol{y})\, d\boldsymbol{y}$ is the convolution of f and h_n.

We will derive the function h from our activation function.

Definition 3. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be continuous. Define the convolution kernel $h_\varphi : \mathbb{R}^d \to \mathbb{R}$ as*

$$h_\varphi(\boldsymbol{x}) = \frac{1}{\alpha_{d-1}} \int_{S^{d-1}} \varphi(\langle \boldsymbol{x}, \boldsymbol{u} \rangle)\, dS^{d-1}(\boldsymbol{u}),$$

where α_{d-1} is the surface area of the d-dimensional unit sphere S^{d-1}, $\alpha_{d-1} = \int_{S^{d-1}} dS^{d-1}(\boldsymbol{u})$.

The following theorem forms the basis of this approach.

Theorem 10. *Let $a > 0$, $K = [-a, a]^d$ and let $\varphi : \mathbb{R} \to \mathbb{R}$ be uniformly continuous on $\mathbb{R}$. If*

(i) $h_\varphi \in L^1(\mathbb{R}^d)$ and

(ii) $\int_{\mathbb{R}^d} h_\varphi(\boldsymbol{x})\, d\boldsymbol{x} \neq 0$,

then, the set of representables G_φ is dense in $C(K)$.

Remark. The proof of Theorem 10 uses the convergence from Theorem 9, approximating the convolutions by an integral sum of h_g, which in turn is approximated by a linear combination of g.

Complications arise in directly applying Theorem 10 to sigmoidal functions. Since h_φ is a radial function (depending only on the distance from the origin), to satisfy condition (i) φ has to go to 0 towards infinity quickly enough (faster than t^{-d}). Sigmoidal functions, however, tend to 1 by definition. Furthermore, it is clear that $h_\varphi \equiv 0$ for φ odd, which is often the case for $\varphi(t) = 2\sigma(t) - 1$ with the most popular sigmoidal functions, implying $h_\sigma \equiv \frac{1}{2} \notin L^1(\mathbb{R})$.

We will instead use the function $\varphi(t) = \sigma(1 + t) + \sigma(1 - t) - 1$, where σ is a sigmoidal function. Then φ is even and since φ can be represented by a one-dimensional σ-network, clearly $G_\varphi \subseteq G_\sigma$, so density of G_φ implies density of G_σ.

The following technical lemma from [11] facilitates verification of the conditions of Theorem 10.

Lemma 11. *Let d be odd and let $\varphi : \mathbb{R} \to \mathbb{R}$ be even and uniformly continuous for which there exists $c > 0$ such that for all $p > d - 2$: $|\varphi(t)| \leq \frac{c}{|t|^p}$. Then:*

1. *$h_\varphi \in L^1$ if and only if for all $j \in \left\{0, \ldots, \frac{d-3}{2}\right\}$:*

$$\int_0^\infty \varphi(t) t^{2j}\, dt = 0.$$

2. It holds that

$$\int_{\mathbb{R}^d} h_\varphi(\boldsymbol{x})\ d\boldsymbol{x} = -2\alpha_{d-2}\tau_d \int_0^\infty \varphi(t)t^{d-1}\ dt,$$

where $\tau_d = \int_0^1 t(1-t)^{\frac{d-3}{2}}\ dt$.

A similar claim holds for d even. The proof of the lemma is somewhat complicated, but it allows us to formulate the following result.

Theorem 12 (Xu, Light, Cheney [11]). *Let $a > 0$, $K = [-a,a]^d$ and let $\varphi : \mathbb{R} \to \mathbb{R}$ be even and uniformly continuous. Suppose there exists $c > 0$ such that for all $p > d-2$: $|\varphi(t)| \leq \frac{c}{|t|^p}$ and that*

$$\int_0^\infty \varphi(t)t^{d-1}\ dt \neq 0.$$

Then, the set of representables G_φ is dense in $C(K)$.

Remark. While the conditions are not satisfied for sigmoidal functions in general, they hold for some of the most commonly used, for example taking the logistic function

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

and $\varphi(t) = \sigma(1+t) + \sigma(1-t) - 1$.

The ReLU Dictionary

The convolution method is easy to apply to the rectified linear unit, $\rho(t) = \max(0,t)$, using Theorem 12. Again, we cannot use ρ itself, since it does not vanish toward infinity, so we will use instead the triangle function

$$\varphi(t) = \rho(t+1) - 2\rho(t) + \rho(t-1),$$

which is a polygonal chain with vertices $(-1,0)$, $(0,1)$ and $(1,0)$. It is clearly representable by a one-dimensional ρ-network, so proving density of G_φ in $C(K)$ suffices to prove density of G_ρ.

The inequality $|\varphi(t)| \leq \frac{1}{|t|^p}$ holds for any $p > d-2$, since $\varphi(t) \leq 1$ on $[-1,1]$ and $\varphi(t) = 0$ elsewhere. Furthermore, for $t \in [0,1]$, $\varphi(t) = 1-t$, so

$$\int_0^\infty \varphi(t)t^{d-1}\ dt = \int_0^1 \varphi(t)t^{d-1}\ dt = \int_0^1 (1-t)t^{d-1}\ dt = \frac{1}{d} - \frac{1}{d+1} \neq 0.$$

Therefore we can apply Theorem 12:

Corollary 13. Let $a > 0$, $K = [-a,a]^d$ and $\rho(t) = \max(0,t)$. The set of representables G_ρ is dense in $C(K)$.

1.1.3 Dual Spaces

Another integral-based approach to neural network approximation uses dual space theory. This method was employed by Cybenko [6].

First, let us review some simple notions from functional analysis.

Definition 4. *Let X be a real vector space. The dual space X^* of X is the space of all bounded linear mappings from X to $\mathbb{R}$ (bounded linear functionals on X).*

Let $V \subseteq X$, the annihilator of V is the set $V^0 \subseteq X^*$ defined as

$$V^0 = \{h \in X^* \mid \forall v \in V : h(v) = 0\}.$$

The following simple theorem forms the basis of this approach to approximation. For a proof see for example [12, Theorem 1.18].

Theorem 14. *Let X be a real vector space. Then a subspace $V \subseteq X$ is dense in X if and only if the annihilator of V is trivial, that is, $V^0 = \{\mathbf{0}\}$, where $\mathbf{0}$ is the constant zero functional.*

Therefore, to prove density of the set of representables it is sufficient to study its annihilator. This is mainly useful in spaces whose dual has a manageable representation.

The following theorem, reformulating Theorem 1 of Cybenko [6], specifies this for the space of continuous functions on the d-dimensional unit cube, using the fact that the dual of $C([0,1]^d)$ is isomorphic to the space of Radon measures on $[0,1]^d$ with bounded variation.

Theorem 15 (Cybenko)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a continuous function such that for all Radon measures μ on $[0,1]^d$ with bounded variation it holds that if for all $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$:*

$$\int_{[0,1]^d} \varphi(\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b) \; d\mu(\boldsymbol{x}) = 0,$$

then $\mu \equiv 0$. Then, the set of representables G_φ is dense in $C([0,1]^d)$.

The rather technical proof that the annihilator of a sigmoidal dictionary is trivial can be found as Lemma 1 in [6] — See Theorem 18 for a similar proof.

While this method is less constructive than the previous two, it can be easily generalized to other cases. For example, we can get a similar result for radial basis activation functions, functions depending only on the distance from a certain point.

Theorem 16. *Let $K \subseteq \mathbb{R}^d$ be compact and $\varphi : \mathbb{R} \to [0,1]$ a continuous function such that $\lim_{t\to\pm\infty} \varphi(t) = 0$ and $\varphi(0) = 1$. Then:*

(i) The following implication holds for all Radon measures μ on K with bounded variation: If for all $\boldsymbol{b} \in K$ and all $a \in \mathbb{R}$:

$$\int_K \varphi\left(a \left\|\boldsymbol{x} - \boldsymbol{b}\right\|\right) \; d\mu(\boldsymbol{x}) = 0,$$

then $\mu \equiv 0$.

(ii) The set

$$\operatorname{span} \; \{\varphi\left(a \left\|\boldsymbol{x} - \boldsymbol{b}\right\|\right) \mid \boldsymbol{b} \in K, a \in \mathbb{R}\}$$

is dense in $C(K)$.

This approach is also easily transferable to other function spaces. The case of L^p spaces — spaces of functions with Lebesgue integrable p-th power, whose dual is known to be L^q for q such that $\frac{1}{p}+\frac{1}{q}=1$ — is a corollary of the case of $C(K)$: Density of G_φ in $C(K)$ implies density of G_φ in all $L^p(K)$, because $C(K)$ is dense in $L^p(K)$ when K is compact. However, the following theorem formulates the approach directly for $L^p(K)$.

Theorem 17. *Let $K \subseteq \mathbb{R}^d$ be compact, $p, q \in [1, \infty)$ such that $\frac{1}{p}+\frac{1}{q}=1$ and $\varphi : \mathbb{R} \to \mathbb{R}$. The set G_φ is dense in $L^p(K)$ if for all $h \in L^q(K)$ the following implication holds: If for all $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$:*

$$\int_K \varphi\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle + b\right) h(\boldsymbol{x})\ d\boldsymbol{x} = 0,$$

then $h \equiv 0$ almost everywhere.

However, the proof that sigmoidal functions satisfy the required implication is again non-trivial.

The ReLU Dictionary

Once more we illustrate the method by applying it to ReLU networks G_ρ, $\rho(t) = \max(0, t)$. We will show that ρ satisfies the condition of Theorem 15. We adapted the proof from the one for sigmoidal functions.

Theorem 18. *Let $K \subseteq \mathbb{R}^d$ be compact and let μ be a Radon measure on K with bounded variation. If for all $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$*

$$\int_K \rho\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle + b\right)\ d\mu(\boldsymbol{x}) = 0,$$

then $\mu \equiv 0$.

Proof. Define μ as zero outside K. We will show that the Fourier transform of μ,

$$F_\mu(\boldsymbol{a}) = \int_{R^d} e^{i\langle \boldsymbol{a}, \boldsymbol{x}\rangle}\ d\mu(\boldsymbol{x}),$$

where i is the imaginary unit, equals zero for all $\boldsymbol{a} \in \mathbb{R}^d$.

We will approximate the integral by a Riemann sum. Let $b = \inf_{\boldsymbol{x}\in K} \langle \boldsymbol{a}, \boldsymbol{x}\rangle$, $c = \sup_{\boldsymbol{x}\in K} \langle \boldsymbol{a}, \boldsymbol{x}\rangle$ and for $n \in \mathbb{N}$ and $j \in \{0, \ldots, n\}$ define $t_j^n = b + j\frac{c-b}{n}$. Then, by the Left Hand Riemann Sum Theorem,

$$\begin{aligned} F_\mu(\boldsymbol{a}) &= \lim_{n\to\infty} \sum_{j=1}^{n} \int_{t_{j-1}^n \le \langle \boldsymbol{a}, \boldsymbol{x}\rangle < t_j^n} e^{it_{j-1}^n} d\mu(\boldsymbol{x}) \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} e^{it_{j-1}^n} \mu\left(\left\{t_{j-1}^n \le \langle \boldsymbol{a}, \boldsymbol{x}\rangle < t_j^n\right\}\right). \end{aligned}$$

We will show that for all $j \le n$, $\mu\left(\left\{t_{j-1}^n \le \langle \boldsymbol{a}, \boldsymbol{x}\rangle < t_j^n\right\}\right) = 0$. Fixing one such j, define for all $m \in \mathbb{N}$

$$\begin{aligned} s_m = \int_K\ & m\rho\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle - \left(t_{j-1}^n - \frac{1}{m}\right)\right) - m\rho\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle - t_{j-1}^n\right) \\ & - m\rho\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle - \left(t_j^n - \frac{1}{m}\right)\right) + m\rho\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle - t_j^n\right)\ d\mu(\boldsymbol{x}). \end{aligned}$$

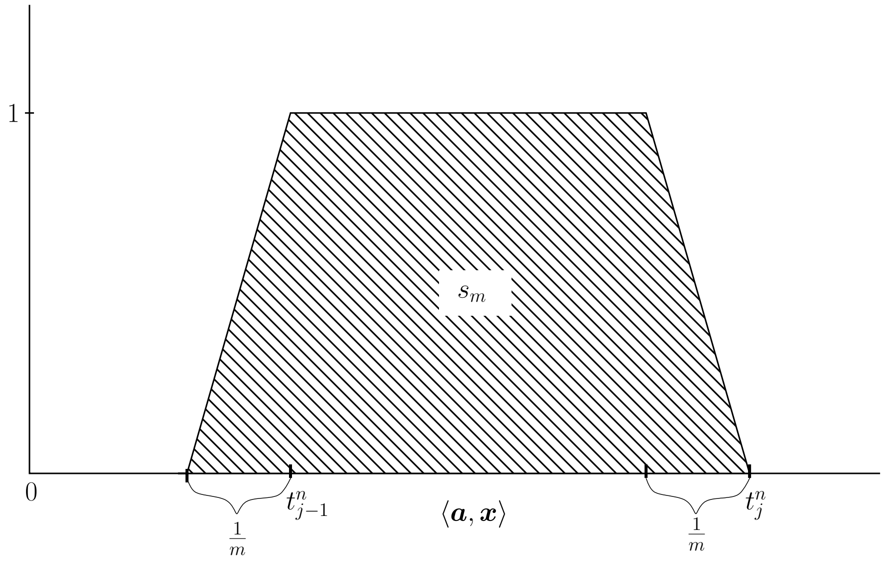

Figure 1.2: The sequence $\{s_m\}_{m=1}^{\infty}$ approximates the measure of the set $\left\{\boldsymbol{x} \in \mathbb{R}^d \mid \langle \boldsymbol{a}, \boldsymbol{x} \rangle \in \left[t^n_{j-1}, t^n_j\right)\right\}$.

The definition of s_m is illustrated in Figure 1.2. By the Lebesgue Dominated Convergence Theorem,

$$s_m \xrightarrow[m\to\infty]{} \int_K 1_{\left[t^n_{j-1}, t^n_j\right)}(\langle \boldsymbol{a}, \boldsymbol{x} \rangle)\ d\mu(\boldsymbol{x}) = \mu\left(\left\{t^n_{j-1} \leq \langle \boldsymbol{a}, \boldsymbol{x} \rangle < t^n_j\right\}\right),$$

where $1_A(u)$ is the indicator function. At the same time, $s_m = 0$ for all m by the claim's assumption that $\int_K \rho\left(\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b\right)\ d\mu(\boldsymbol{x}) = 0$. Altogether, $F_\mu(\boldsymbol{a}) = 0$ for all $\boldsymbol{a} \in \mathbb{R}^d$ and, by uniqueness of the Fourier transform, μ is the constant zero measure. □

Together with Theorem 15 we get the desired result.

Corollary 19. The set of representables G_ρ is dense in $C\left([0,1]^d\right)$, where $\rho(t) = \max(0, t)$.

1.1.4 Non-Polynomial Activation Functions

We have seen several methods of proving the Universal Approximation Theorem, Theorem 1, focusing on sigmoidal and ReLU activation functions. However, Leshno et al. [13] proved a stronger version of the theorem, stating that G_φ is dense in $C(K)$ if and only if φ is not Lebesgue-equivalent to a polynomial function. While the method at the heart of this proof is the one of Subsection 1.1.1, showing φ can approximate any one-dimensional continuous function and using the Stone-Weierstrass Theorem, in this general version the one-dimensional step is more complicated.

Theorem 20 (Leshno et al. [13])**.** *Let $K \subseteq \mathbb{R}^d$ be compact and let $\varphi : \mathbb{R} \to \mathbb{R}$ be locally bounded such that the set of points of discontinuity of φ has Lebesgue measure zero. Then, the following are equivalent:*

(i) G_φ is dense in $C(K)$,

(ii) φ is not Lebesgue-equivalent to a polynomial.

Proof. One implication is simple: If φ is a polynomial of degree k, then any $h \in G_\varphi$ is a polynomial (in d variables) of degree at most k. The set of polynomials of bounded degree is not dense in $C(K)$.

In the other direction, denote by G^1_φ the set of functions representable by a neural network with one-dimensional input. Leshno et al. [13] divide the proof into steps similar to the following:

1. If G^1_φ is dense in $C([0,1])$, then G_φ is dense in $C(K)$.

   - This follows from theorems of Subsection 1.1.1, namely Theorem 2 and Corollary 5.

2. If φ is non-polynomial and smooth on $\mathbb{R}$, $\varphi \in C^\infty(\mathbb{R})$, then G^1_φ is dense in $C([0,1])$.

   - Since φ is not a polynomial, there exists $t_0 \in \mathbb{R}$ such that all derivatives of φ are non-zero at t_0. Using this it can be shown that φ can approximate any polynomial, so G^1_φ is dense in $C([0,1])$ by the Weierstrass Approximation Theorem.

3. Denoting by $C_c^\infty(\mathbb{R})$ the set of all smooth real functions with compact support, for all $h \in C_c^\infty(\mathbb{R})$ the function $\varphi * h$ is an element of the closure of G^1_φ.

   - Here, $\varphi * h$ is the convolution of φ and h as defined in Subsection 1.1.2, $\varphi * h(t) = \int \varphi(t-s)h(s)ds$. The proof of this step is non-trivial.

4. If there exists $h \in C_c^\infty(\mathbb{R})$ such that $\varphi * h$ is non-polynomial, then G^1_φ is dense in $C([0,1])$.

   - By 3., $\varphi * h \in \overline{G^1_\varphi}$, which implies $\overline{G^1_{\varphi * h}} \subseteq \overline{G^1_\varphi}$. Since $\varphi * h \in C^\infty$ non-polynomial, by 2. $G^1_{\varphi * h}$ is also dense in $C([0,1])$, so G^1_φ is dense in $C([0,1])$.

5. If for all $h \in C_c^\infty(\mathbb{R})$: $\varphi * h$ is a polynomial, then there exists $k \in \mathbb{N}$ such that all $\varphi * h$ are polynomials of degree at most k.

   - Again, this step is a bit technical: Denote by $C_c^\infty([a,b])$ the set of all functions in $C_c^\infty(\mathbb{R})$ with support in $[a,b]$. Because $C_c^\infty([a,b]) = \varphi * C_c^\infty([a,b])$, we can write $C_c^\infty([a,b])$ as the union over k of sets of all $\varphi * h$ of degree at most k. By the Baire Category Theorem, this sequence of sets has to stabilize, meaning all $\varphi * h$ are of uniformly bounded degree.

6. If for all $h \in C_c^\infty(\mathbb{R})$ the function $\varphi * h$ is a polynomial, then φ is a polynomial.

   - This follows from 5. using results from distribution theory.

Altogether, if φ is non-polynomial, then by 6. some $\varphi * h$ is non-polynomial, so by 4. G^1_φ is dense in $C([0,1])$ and by 1. G_φ is dense in $C(K)$.

□

Remark. We changed a few details from the original proof, as it used some potentially misleading notation: G_φ is there defined to be dense in $C(\mathbb{R}^d)$ if it is dense in $C(K)$ for all $K \subseteq \mathbb{R}^d$ compact. However, for example for σ sigmoidal, G_σ is dense in $C(\mathbb{R}^d)$ in this sense, but it is never dense in $C(\mathbb{R}^d)$ in the sense that for all $f : \mathbb{R}^d \to \mathbb{R}$ continuous and $\varepsilon > 0$ there exists $g \in G_\sigma$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x}) - g(\boldsymbol{x})| < \varepsilon$.

1.2 Bounds on Network Size

In this section we survey some previous bounds on expressive power of neural networks. Basic expressivity results are usually stated in the form "Networks of complexity m can approximate some class of functions with accuracy ε", or (equivalently) "In order to approximate functions with accuracy ε by a neural network, the minimal required complexity is m". For our purposes, complexity of a neural network will be synonymous with the number of hidden units. However, in more recent literature other measures of complexity are often considered, such as the number of non-zero weights (as in [14]) or an approximation thereof by the sum of weight sizes (see for example [15] for further motivation).

Definition 5. *For $\varphi : \mathbb{R} \to \mathbb{R}$ and $m \in \mathbb{N}$ denote by $G_{\varphi,m}$ the set of functions represented by neural networks with activation φ having at most m neurons in the hidden layer (including a potential constant term),*

$$G_{\varphi,m} = \left\{ \sum_{i=1}^{m} c_i \varphi(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle + b_i) \mid \forall i \leq m : \boldsymbol{a}_i \in \mathbb{R}^d, c_i, b_i \in \mathbb{R} \right\}.$$

Note that this definition implies that the complexity of a function $g \in G_\varphi$ is the *least* number of neurons necessary to express g, even though g may have several representations of different sizes.

Definition 6. *Let X be a vector space and $A \subseteq X$. Define the error of approximation of A by B in the space X as*

$$E_X(A, B) = \sup_{a\in A} \inf_{b\in B} \|a - b\|_X .$$

We will write $E_X(a, B)$ instead of $E_X(\{a\}, B)$.

Remark. Consequently, $E_X(A, B) \leq \varepsilon$ is equivalent to: For all $a \in A$ there exists $b \in B$ such that $\|a - b\|_X \leq \varepsilon$. That is, any member of A can be approximated by some member of B within an error of ε. Conversely, $E_X(A, B) \geq \varepsilon$ can be restated as: There exists $a \in A$ such that for all $b \in B$: $\|a - b\|_X \geq \varepsilon$. That is, some element of A cannot by approximated by B better than with an error of ε.

1.2.1 Sobolev Spaces

Most bounds on network complexity found in the literature are expressed in terms of Sobolev spaces and because these are often presented in a somewhat advanced manner, we give a quick summary of some important facts. The bounds use Sobolev spaces to quantify complexity of the approximated function — Functions belonging to higher order Sobolev spaces are more regular (smoother in case of continuous derivatives) and therefore easier to approximate.

Recall that for $p \in [1, \infty)$ we define $L^p(K)$ as the space of all functions $f : K \to \mathbb{R}$ such that the integral $\int_K |f|^p \, d\lambda^d$ is finite, together with the norm $\|f\|_p = \left(\int_K |f|^p \, d\lambda^d\right)^{\frac{1}{p}}$. For $p = \infty$, $L^\infty(K)$ is the space of functions bounded almost everywhere on K with $\|\cdot\|_\infty$ defined as the essential supremum. On a compact set it holds that $L^p(K) \supseteq L^q(K)$ for $p \leq q$ and that $C(K)$ is a dense subset of all $L^p(K)$.

Sobolev spaces are typically defined on the whole $\mathbb{R}^d$ or on well-behaved open subsets, which usually means sets with a nice boundary called Lipschitz domains. A Lipschitz domain is a non-empty connected open set U such that the boundary of U can be locally represented as the graph of a Lipschitz function. Previous bounds are usually formulated on balls or cubes, which satisfy these conditions, but there exist many natural sets that do not.

Definition 7. *Let $U \subseteq \mathbb{R}^d$ be a Lipschitz domain. For $k \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty]$, the Sobolev space $W^{k,p}(U)$ is the space of all functions $f : K \to \mathbb{R}$ such that all weak derivatives of f of order up to k have a finite L^p norm.*

Remark. A function h is an i-th weak derivative of f, $i \leq d$, if for all $\psi \in C^\infty(U)$ with compact support: $\int_U f \frac{\partial}{\partial x_i} \psi \; d\lambda^d = - \int_U h\psi \; d\lambda^d$.

For $k = 0$, $W^{0,p}(U) = L^p(U)$. For $k_1 < k_2$ we have $W^{k_1,p} \supsetneq W^{k_2,p}$.

Alternatively, $W^{k,p}(U)$ can be defined as completion of the space of functions that have all derivatives up to order k bounded in the L^p norm. However, it is important to note that weak derivatives lack some properties of classical derivatives. The space is equipped with a norm based on L^p norms of derivatives of order up to k:

$$\|f\|_{W^{k,p}} = \left(\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|_p^p \right)^{\frac{1}{p}},$$

or for $p = \infty$:

$$\|f\|_{W^{k,\infty}} = \max_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|_\infty .$$

As we are mainly interested in continuous functions and the sup norm, we will define a more intuitive type of spaces and relate it to Sobolev spaces.

Definition 8. *Let $U \subseteq \mathbb{R}^d$ be open. For $k \in \mathbb{N} \cup \{0\}$ the Hölder space $C^{k,1}(U)$ is the space of all $f : U \to \mathbb{R}$ such that derivatives of f of order up to k are continuous and derivatives of order k are Lipschitz continuous.*

Denoting by $C^k(U)$ the space of functions with continuous derivatives of order up to k (where $C^0(U) = C(U)$), we have

$$C^0(U) \supsetneq C^{0,1}(U) \supsetneq C^1(U) \supsetneq C^{1,1}(U) \supsetneq \dots .$$

The following are consequences of one of Sobolev Embedding Theorems.

Lemma 21. *Let $U \subseteq \mathbb{R}^d$ be a Lipschitz domain. Then:*

1. *For all $k \geq 1$: $W^{k,\infty}(U) = C^{k-1,1}(U)$.*

2. *If $k_1 \geq k_2 \geq 0$ and $p \in [1, \infty)$ such that $k_2 < k_1 - \frac{d}{p}$, then $W^{k_1,p}(U) \subseteq W^{k_2,\infty}(U)$. In particular, if $f \in W^{k,p}(U)$ for $p(k-1) > d$, then f is Lipschitz continuous.*

3. *If $f \in W^{k,p}(U)$ for $pk > d$, then f is continuous.*

Sobolev spaces are not usually defined on non-open sets. However, for K that is a closed ball or a hypercube, under assumptions on k and p (in particular that $kp > d$) it holds that any function in the Sobolev space defined on the interior of K has a uniquely determined extension, along with its derivatives, onto the boundary of K. Then, $W^{k,p}(K)$ can be seen as the same space as $W^{k,p}(\operatorname{int} K)$.

We formulate the bounds for $K = [0, 1]^d$, but the same bounds with potentially different constants hold for $K = \overline{B(\mathbf{0}, 1)}$, or other similar sets. The bounds are usually presented normalized in the sense that f is assumed to satisfy $\|f\|_{W^{k,p}} \leq 1$.

Definition 9. *Denote by $B^{k,p}$ the closed unit ball in $W^{k,p}([0,1]^d)$, that is,*

$$B^{k,p} = \left\{ f \in W^{k,p}\left([0,1]^d\right) \;\middle|\; \|f\|_{W^{k,p}} \leq 1 \right\}.$$

Remark. Note that not only are the bounds formulated only on very nice sets K, but they also require $k \geq 1$. With respect to $C\left([0,1]^d\right)$, we need the approximated function to be at least Lipschitz in order for the bounds to be applicable.

1.2.2 Bounds on Error of Approximation

Now we look at some lower and upper bounds on approximation error, which correspond to lower and upper bounds on the number of necessary neurons. For a more detailed survey see [16] and [14].

We will begin with lower bounds. In this thesis we focus only on ridge activations, so that the resulting function is a combination of some $\varphi(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + b)$. If we allow φ to vary, we get a surprisingly useful lower bound. Denote by R_m the set of all m-term combinations of continuous ridge functions,

$$R_m = \left\{ \sum_{i=1}^{m} c_i \varphi_i(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) \mid \forall i \leq m : \boldsymbol{a}_i \in \mathbb{R}^d, c_i \in \mathbb{R}, \varphi_i \in C(\mathbb{R}) \right\}.$$

If $\varphi : \mathbb{R} \to \mathbb{R}$ is continuous, then clearly $G_{\varphi,m} \subseteq R_m$, so for any X containing R_m and for $f \in X$:

$$E_X\left(f, G_{\varphi,m}\right) = \inf_{g \in G_{\varphi,m}} \|f - g\|_X \geq \inf_{g \in R_m} \|f - g\|_X = E_X\left(f, R_m\right).$$

Therefore, if we can bound $E_X(f, R_m)$ from below, we get a bound also on $E_X(f, G_{\varphi,m})$. Maiorov [1] provides such bounds for approximation by ridge functions for $X = L^2$. Note that the upper bound does not imply anything about $G_{\varphi,m}$, it only means we can get no better lower bounds using this approach.

Theorem 22 (Maiorov [1])**.** *Let $k \geq 1$, then there exist $C_1, C_2 > 0$ such that for all $m \geq 1$*

$$C_1 m^{-\frac{k}{d-1}} \leq E_{L^2}\left(B^{k,2}, R_m\right) \leq C_2 m^{-\frac{k}{d-1}}.$$

Remark. In other words, the lower bound implied for any $\varphi \in C(\mathbb{R})$ can be stated as: There exists $f \in B^{k,2}$ such that for all $\varepsilon > 0$ and $g \in G_{\varphi,h}$: If $\|f-g\|_2 < \varepsilon$, then

$$h \geq \widetilde{C}_1 \left(\frac{1}{\varepsilon}\right)^{\frac{d-1}{k}}.$$

It was even proved by Maiorov et al. [17] that the set of all f that force this bound is large (in measure), so it is not just some pathological case.

It is simple to verify that the same bound holds for any $p \in [2, \infty]$.

This lower bound on approximation error is in general not tight — For some σ sigmoidal it holds that the number of necessary neurons is at least $C\left(\frac{1}{\varepsilon}\right)^{\frac{d}{k}}$. However, Maiorov and Pinkus [18] define a sigmoidal function that attains this bound.

Theorem 23 (Maiorov and Pinkus [18])**.** *There exists $\sigma \in C(R)$ sigmoidal and strictly increasing such that for all $k \geq 1$ and $p \in [1, \infty]$ there exists $C > 0$ such that for all $m \geq 1$*

$$E_{L^p}\left(B^{k,p}, G_{\sigma,m}\right) \leq C m^{-\frac{k}{d-1}}.$$

Remark. However, even though the function is C^∞, it is constructed using separability of the space $C([-1,1])$ and it is definitely not usable as an activation function in practice. The conclusions that the authors draw from this result are that (i) sigmoidality, monotonicity and smoothness are not sufficiently strong properties to rule out pathological functions, and that (ii) these properties do not impede approximation, as there is a function having these properties that attains the best possible degree of approximation.

A stronger lower bound on error of approximation can be derived if the parameters of the approximating network, the coefficients, weights and thresholds, depend continuously on the approximated function f. While this is usually not the case in practice, results based on DeVore et al. [19] imply that for such networks the lower bound increases to

$$C m^{-\frac{k}{d}}.$$

Next, we turn our attention to upper bounds on approximation error. Probably the most important bound in this direction was proved by Mhaskar [2].

Theorem 24 (Mhaskar [2])**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function satisfying that there exists an open interval $U \subseteq \mathbb{R}$ such that $\varphi|_U \in C^\infty(U)$ and φ is non-polynomial on U. Then, for all $k \geq 1$ and $p \in [1, \infty]$ there exists $C > 0$ such that for all $m \geq 1$*

$$E_{L^p}\left(B^{k,p}, G_{\varphi,m}\right) \leq C m^{-\frac{k}{d}}.$$

Remark. In other words, for φ smooth the following holds: For all $f \in B^{k,p}$ and $\varepsilon > 0$ there exists $g \in G_{\varphi,h}$ for

$$h \leq \widetilde{C} \left(\frac{1}{\varepsilon}\right)^{\frac{d}{k}}$$

such that $\|f-g\|_p < \varepsilon$.

The proof uses the known fact that for all $f \in W^{k,p}$ and all $n \in \mathbb{N}$, there exists a (d-variate) polynomial p_n of degree at most n such that $\|f - p_n\| \leq Cn^{-k} \|f\|_{W^{k,p}}$. The polynomial is then approximated using derivatives of φ.

Because of the assumptions on φ, the previous bound cannot be applied to the ReLU activation or the Heaviside function $1_{[0,\infty)}$ that are either non-smooth or polynomial on all intervals. Petrushev [20] proved, among other results, that these functions satisfy the same bound for certain values of k.

Theorem 25 (Petrushev [20])**.** *For $n \in \mathbb{N} \cup \{0\}$, define*

$$\sigma_n(t) = \begin{cases} t^n, & t \geq 0; \\ 0, & t < 0. \end{cases}$$

Then, for all $1 \leq k \leq n + 1 + \frac{d-1}{2}$ there exists $C > 0$ such that for all $m \geq 1$:

$$E_{L^2}\left(B^{k,2}, G_{\sigma_n,m}\right) \leq Cm^{-\frac{k}{d}}.$$

Remark. As the Heaviside function and the rectifier equal σ_0 and σ_1, the bounds hold for them for k up to $\frac{d+1}{2}$ and $\frac{d+3}{2}$, respectively.

As for more recent results, probably the most notable in a direction similar to the previous is the work of Yarotsky [14], who proves several lower and upper bounds for deeper neural networks. One of these results specifies a bound on the size of one neural network architecture that approximates all functions in $B^{k,\infty}$ with some setting of weights.

Theorem 26 (Yarotsky [14])**.** *Let $k \geq 1$ and $\varepsilon \in (0,1)$. Then, there exists $c > 0$ and a neural network architecture N with ReLU activation of depth at most*

$$c\left(\log\left(\frac{1}{\varepsilon}\right) + 1\right)$$

and with at most

$$c\left(\frac{1}{\varepsilon}\right)^{\frac{d}{n}}\left(\log\left(\frac{1}{\varepsilon}\right) + 1\right)$$

neurons and connections (non-zero weights), such that for all $f \in B^{k,\infty}$ there exists a weight setting of N represented by a function $g \in G_\rho$ such that $\|f - g\|_\infty < \varepsilon$.

On the other hand, if the architecture is allowed to depend on the approximated function f, he proves that a network of constant depth and the number of neurons bounded by

$$\frac{c}{\varepsilon \log\left(\frac{1}{\varepsilon}\right)}$$

that approximates f. However, this result is formulated only for one-dimensional functions f, that is, for any Lipschitz function on $[0,1]$.

2. A Constructive Upper Bound on Network Size

We construct an upper bound on the number of neurons in the hidden layer necessary to approximate a continuous function on a compact set. We achieve this using the approach from Subsection 1.1.1, working with networks with exponential activation and then approximating the exponentials by the given sigmoidal function in one dimension.

The construction is inspired by a specific proof of the Stone-Weierstrass Theorem by Brosowski and Deutsch [3]. We use a lattice to divide the compact set into polytopes on each of which the target function's range lies in a small interval. We approximate the constant function on each polytope by a neural network with exponential activation and combine then to get an approximation of the target function.

At the end of this chapter we consider an alternative approach — Approximating the target function everywhere except a set of small measure. We give an overview of the potential construction and specify what theoretical results are required to complete it.

2.1 Approximation of the Indicator Function of a Polytope

In this thesis, by a polytope we understand the convex hull of finitely many points in $\mathbb{R}^d$. For a detailed overview of basic definitions see [21, Section 4.2].

Definition 10. *For $d, m \in \mathbb{N}$, $d \geq 2$, denote by $\mathbb{P}^d_m$ the set of all d-dimensional polytopes having m facets.*

Remark. Because the definition requires the polytopes to be full-dimensional, $\mathbb{P}^d_m$ is non-empty only for $m \geq d + 1$.

The following theorem, stating that we can approximate the indicator function of any polytope, is original, but it is loosely inspired by Lemma 1 from [3].

Definition 11. *Let $P \in \mathbb{P}^d_m$, $\boldsymbol{x} \in P$ and $\alpha \in \mathbb{R}$. Denote by $\alpha *_{\boldsymbol{x}} P$ the dilation of P by α centred at $\boldsymbol{x}$,*

$$\alpha *_{\boldsymbol{x}} P = \alpha(P - \boldsymbol{x}) + \boldsymbol{x}.$$

Recall that $G_{\exp,h}$ denotes the set of functions representable by a shallow neural network with the exponential activation function, $\exp(t) = e^t$, that has h units in the hidden layer.

Theorem 27. *Let $K \subseteq \mathbb{R}^d$ be compact, $P \in \mathbb{P}^d_m$ such that $P \cap K \neq \emptyset$, $\boldsymbol{x}_0 \in int\ P$, $\alpha > 1$ and $\varepsilon > 0$. There exists $g \in G_{\exp}$ such that*

- $g|_K : K \to [0, 1]$,
- *for $\boldsymbol{y} \in P \cap K$: $g(\boldsymbol{y}) > 1 - \varepsilon$,*

- *for* $\boldsymbol{y} \in K \setminus (\alpha *_{\boldsymbol{x}_0} P)$*:* $g(\boldsymbol{y}) < \varepsilon$*,*
- $g \in G_{\exp,h}$ *for*

$$h = \left(\frac{3e}{m}\right)^m \left(\frac{2}{\varepsilon}\right)^{6m\log_2 m \frac{diam(P)diam(K)}{(\alpha-1)q^2}},$$

where q *is the minimal distance from* $\boldsymbol{x}_0$ *to a facet of* P*.*

Proof. Denote by $H(\boldsymbol{a}, b)$ the hyperplane perpendicular to $\boldsymbol{a} \in S^{d-1}$ shifted by $b\boldsymbol{a}$, that is,

$$H(\boldsymbol{a}, b) = \left\{\boldsymbol{y} \in \mathbb{R}^d \mid \langle \boldsymbol{a}, \boldsymbol{y}\rangle - b = 0\right\}.$$

Let $\{\boldsymbol{a}_i\}_{i=1}^m$ be outer unit normal vectors of P's facets and let $\{b_i^1\}_{i=1}^m$ and $\{b_i^2\}_{i=1}^m$ be real numbers such that the supporting hyperplanes of facets of P and $\alpha *_{\boldsymbol{x}_0} P$ are of the form $H(\boldsymbol{a}_i, b_i^1)$ and $H(\boldsymbol{a}_i, b_i^2)$, respectively. That is, $P = \{\boldsymbol{y} \mid \forall i \leq m : \langle \boldsymbol{a}_i, \boldsymbol{y}\rangle - b_i^1 \leq 0\}$ and $\alpha *_{\boldsymbol{x}_0} P = \{\boldsymbol{y} \mid \forall i \leq m : \langle \boldsymbol{a}_i, \boldsymbol{y}\rangle - b_i^2 \leq 0\}$. Take $b_i = \frac{b_i^1 + b_i^2}{2}$. The whole situation is illustrated in Figure 2.1.

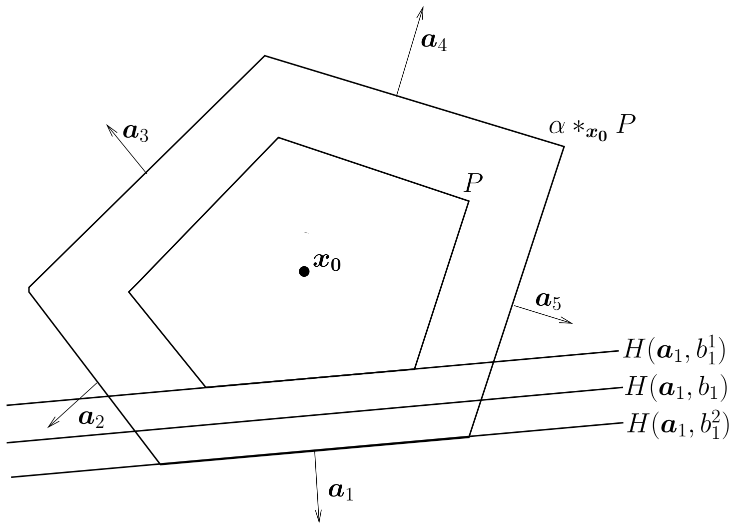

Figure 2.1: Facets of polytopes P and $\alpha *_{\boldsymbol{x}_0} P$ lie on hyperplanes of the form $H(\boldsymbol{a}_i, b_i^1)$ and $H(\boldsymbol{a}_i, b_i^2)$, respectively. Hyperplanes $H(\boldsymbol{a}_i, b_i)$ then correspond to facets of the polytope $\frac{\alpha}{2} *_{\boldsymbol{x}_0} P$. Outer normal unit vectors of facets of all polytopes are $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. b_i^1, b_i and b_i^2 are oriented distances from the corresponding hyperplanes to the point $\mathbf{0}$.

For each facet of P we will define an exponential function that is larger than $m\gamma$ for some γ on the outer side of $H(\boldsymbol{a}_i, b_i^2)$ (w.r.t. $\boldsymbol{a}_i$) and smaller than $\frac{\gamma}{2}$ on the inner side of $H(\boldsymbol{a}_i, b_i^1)$, also normalizing them to range inside $[0, 1]$ on K. This way, taking $p(\boldsymbol{x})$ as the average of these exponentials, p is larger than γ outside $\alpha *_{\boldsymbol{x}_0} P$ and smaller than $\frac{\gamma}{2}$ inside P. This will allow us to exponentiate p in order to get the required bounds.

If $\langle \boldsymbol{a}_i, \boldsymbol{y}\rangle - b_i^1 = 0$ for some $\boldsymbol{y}$, then $b_i^2 = \langle \boldsymbol{a}_i, \alpha(\boldsymbol{y} - \boldsymbol{x}_0) + \boldsymbol{x}_0\rangle$, so

$$b_i^2 - b_i^1 = \langle \boldsymbol{a}_i, \alpha(\boldsymbol{y} - \boldsymbol{x}_0) + \boldsymbol{x}_0\rangle - \langle \boldsymbol{a}_i, \boldsymbol{y}\rangle = (\alpha - 1)(b_i^1 - \langle \boldsymbol{a}_i, \boldsymbol{x}_0\rangle).$$

$b_i^1 - \langle \boldsymbol{a}_i, \boldsymbol{x}_0\rangle = |b_i^1 - \langle \boldsymbol{a}_i, \boldsymbol{x}_0\rangle|$ is the distance between hyperplanes $H(\boldsymbol{a}_i, b_i^1)$ and $H(\boldsymbol{a}_i, \langle \boldsymbol{a}_i, \boldsymbol{x}_0\rangle)$, which equals the distance from $\boldsymbol{x}_0$ to the i-th facet of P. Denote this distance by q_i and let $q = \min_{i \le m} q_i$ be the minimal distance from $\boldsymbol{x}_0$ to a facet of P. We get

$$b_i^2 - b_i^1 = (\alpha - 1) q_i \ge (\alpha - 1) q. \tag{2.1}$$

For each $i \le m$ take $s_i = \frac{\log(2m) q_i}{(\alpha-1)q^2}$[1]. Then, by (2.1) we have

$$e^{s_i\left(b_i^2 - b_i\right)} = (2m)^{\frac{(\alpha-1)q_i^2}{2(\alpha-1)q^2}} \ge (2m)^{\frac{1}{2}} = \sqrt{2m} \tag{2.2}$$

and

$$e^{s_i\left(b_i^1 - b_i\right)} = (2m)^{\frac{-(\alpha-1)q_i^2}{2(\alpha-1)q^2}} \le (2m)^{-\frac{1}{2}} = \frac{1}{\sqrt{2m}}. \tag{2.3}$$

Taking $\boldsymbol{y}_1 \in P$ and $\boldsymbol{y}_2 \notin \alpha *_{\boldsymbol{x}_0} P$ — which means $\langle \boldsymbol{a}_i, \boldsymbol{y}_1 \rangle - b_i^1 \le 0$ for all i, and $\langle \boldsymbol{a}_j, \boldsymbol{y}_2 \rangle - b_j^2 > 0$ for some $j \le m$ — we get from (2.2) and (2.3):

$$\begin{aligned} e^{s_j(\langle \boldsymbol{a}_j, \boldsymbol{y}_2 \rangle - b_j)} &> e^{s_j\left(b_j^2 - b_j\right)} \ge \sqrt{2m} = 2\frac{m}{\sqrt{2m}} \\ &\ge 2\sum_{i=1}^{m} e^{s_i\left(b_i^1 - b_i\right)} > 2\sum_{i=1}^{m} e^{s_i(\langle \boldsymbol{a}_i, \boldsymbol{y}_1 \rangle - b_i)}. \end{aligned} \tag{2.4}$$

Let

$$z = (2m)^{\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}$$

and define

$$p(\boldsymbol{x}) = \frac{1}{mz}\sum_{i=1}^{m} e^{s_i(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i)} = \frac{1}{m}\sum_{i=1}^{m} (2m)^{\frac{q_i\left(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i\right) - \operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}.$$

Then, $p|_K$ ranges in $[0,1]$: p is clearly non-negative. For any $\boldsymbol{y} \in K$ and $i \le m$, taking $\boldsymbol{y}' \in P \cap K \ne \emptyset$ we have $\langle \boldsymbol{a}_i, \boldsymbol{y}' \rangle \le b_i^1 < b_i$, so $\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - b_i < \langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - \langle \boldsymbol{a}_i, \boldsymbol{y}' \rangle \le |\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - \langle \boldsymbol{a}_i, \boldsymbol{y}' \rangle|$. This is the distance between hyperplanes $H(\boldsymbol{a}_i, \langle \boldsymbol{a}_i, \boldsymbol{y} \rangle)$ and $H(\boldsymbol{a}_i, \langle \boldsymbol{a}_i, \boldsymbol{y}' \rangle)$, which is at most $\|\boldsymbol{y} - \boldsymbol{y}'\|$. However, since both $\boldsymbol{y}$ and $\boldsymbol{y}'$ are elements of K, their distance is at most $\operatorname{diam}(K)$. Altogether, we get $\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - b_i < \operatorname{diam}(K)$. Combined with the fact that $q_i < \operatorname{diam}(P)$, this implies

$$(2m)^{\frac{q_i\left(\langle \boldsymbol{a}_i, \boldsymbol{y} \rangle - b_i\right)}{(\alpha-1)q^2}} < (2m)^{\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}} = z,$$

so $p \le 1$.

Furthermore, by (2.4) for any $\boldsymbol{y}_1 \in P$ and $\boldsymbol{y}_2 \notin \alpha *_{\boldsymbol{x}_0} P$ we have

$$p(\boldsymbol{y}_2) > \frac{\sqrt{2}}{\sqrt{m}z} = 2\frac{1}{z\sqrt{2m}} > 2p(\boldsymbol{y}_1).$$

Next, we manipulate p to get the required bounds. Denote

$$\gamma = \frac{\sqrt{2}}{\sqrt{m}z} = \sqrt{\frac{2}{m}}\,(2m)^{-\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}},$$

[1]This definition is motivated by a later result, Lemma 34, where it will be useful to have $s_i = cq_i$, where c is common for all i. We could instead take $s_i = \frac{\log(2m)}{(\alpha-1)q}$ for all i, slightly reducing the bound of this theorem — See the remark after this proof.

so that $p < \frac{\gamma}{2}$ on P and $p > \gamma$ outside $\alpha *_{\boldsymbol{x}_0} P$. If $K \setminus (\alpha *_{\boldsymbol{x}_0} P)$ is empty, the theorem can be satisfied by a constant function, so assume $K \setminus (\alpha *_{\boldsymbol{x}_0} P) \neq \emptyset$. This implies for some i that $\operatorname{diam}(K) \geq b_i^2 - b_i^1 \geq (\alpha - 1)q$. Because $\operatorname{diam}(P) > 2q$ and $m \geq 2$, we have $\gamma < 4^{-2} = \frac{1}{16}$.

Let $k = \left\lfloor \frac{\sqrt{2}}{\gamma} - 1 \right\rfloor$, meaning $k\gamma \in \left[\sqrt{2} - \gamma, \sqrt{2}\right) \subseteq \left(\sqrt{2} - \frac{1}{16}, \sqrt{2}\right)$, and define

$$n = \left\lceil \frac{-5\log(\varepsilon)}{2\log(2)} \right\rceil.$$

Then,

$$n \geq \frac{5\log(\varepsilon)}{4\log\left(\frac{1}{\sqrt{2}}\right)} > \frac{\log(\varepsilon)}{\log\left(\frac{\sqrt{2}}{2}\right)} \geq \frac{\log(\varepsilon)}{\log\left(\frac{k\gamma}{2}\right)}, \tag{2.5}$$

which implies $\left(\frac{k\gamma}{2}\right)^n < \varepsilon$, and also

$$n \geq \frac{-\log(\varepsilon)}{\frac{4}{5}\log\left(\sqrt{2}\right)} > \frac{-\log(\varepsilon)}{\log\left(\sqrt{2} - \frac{1}{16}\right)} \geq \frac{-\log(\varepsilon)}{\log(k\gamma)}, \tag{2.6}$$

meaning $\frac{1}{(k\gamma)^n} < \varepsilon$.

Finally, take

$$g = (1 - p^n)^{k^n}.$$

As in the proof of [3, Lemma 1], for $\boldsymbol{y} \in P \cap K$ we get by Bernoulli's inequality and by (2.5):

$$g(\boldsymbol{y}) \geq 1 - (kp(\boldsymbol{y}))^n > 1 - \left(\frac{k\gamma}{2}\right)^n > 1 - \varepsilon$$

and for $\boldsymbol{y} \in K \setminus (\alpha *_{\boldsymbol{x}_0} P)$, using again Bernoulli's inequality and (2.6):

$$\begin{aligned} g(\boldsymbol{y}) &= \frac{(kp(\boldsymbol{y}))^n}{(kp(\boldsymbol{y}))^n}(1 - p(\boldsymbol{y})^n)^{k^n} \leq \frac{1}{(kp(\boldsymbol{y}))^n}(1 - p(\boldsymbol{y})^n)^{k^n}(1 + k^n p(\boldsymbol{y})^n) \\ &\leq \frac{1}{(kp(\boldsymbol{y}))^n}(1 - p(\boldsymbol{y})^n)^{k^n}(1 + p(\boldsymbol{y})^n)^{k^n} = \frac{1}{(kp(\boldsymbol{y}))^n}(1 - p(\boldsymbol{y})^{2n})^{k^n} \\ &\leq \frac{1}{(kp(\boldsymbol{y}))^n} < \frac{1}{(k\gamma)^n} < \varepsilon. \end{aligned}$$

To summarize, we have defined the function g as

$$g(\boldsymbol{x}) = \left(1 - \left(\frac{1}{m}\sum_{i=1}^{m}(2m)^{\frac{q_i(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - b_i) - \operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}\right)^n\right)^{k^n},$$

where $n = \left\lceil \frac{-5\log(\varepsilon)}{2\log(2)} \right\rceil$ and $k = \left\lfloor \sqrt{m}(2m)^{\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}} - 1 \right\rfloor$. As n and k are both positive integers, it follows that $g \in G_{\exp}$. We have shown that $g|_K$ ranges in $[0, 1]$ and that for $\boldsymbol{y} \in P \cap K$: $g(\boldsymbol{y}) > 1 - \varepsilon$ and for $\boldsymbol{y} \in K \setminus (\alpha *_{\boldsymbol{x}_0} P)$: $g(\boldsymbol{y}) < \varepsilon$.

As for the number of hidden units, we know that $p \in G_{\exp,m}$. As a consequence of the multinomial theorem, the n-th power of an m-term expression contains at most $\binom{n+m-1}{m-1}$ terms (see e.g. [22, Chapter 5]). The number of terms in $g = (1 - p^n)^{k^n}$ is at most the number of terms in $(1 - p)^{nk^n}$. Therefore, $g \in G_{\exp,h}$ for

$$h = \binom{nk^n + m}{m} \leq \left(\frac{(nk^n + m)e}{m}\right)^m = e^m \left(\frac{nk^n}{m} + 1\right)^m.$$

Furthermore,
$$n \leq -\frac{5\log\varepsilon}{2\log 2} + 1 \leq \frac{5}{2}\log_2\left(\frac{2}{\varepsilon}\right)$$
and
$$\begin{aligned} k &\leq \frac{\sqrt{2}}{\gamma} = \sqrt{m}(2m)^{\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}} \\ &= 2^{(1+\log_2 m)\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}+\frac{1}{2}\log_2 m} \\ &\leq 2^{2\log_2 m\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}. \end{aligned}$$
Together,
$$k^n \leq \left(\frac{2}{\varepsilon}\right)^{5\log_2 m\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}$$
and therefore
$$\begin{aligned} h &\leq e^m\left(\frac{5\log_2\left(\frac{2}{\varepsilon}\right)}{2m}\left(\frac{2}{\varepsilon}\right)^{5\log_2 m\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}} + 1\right)^m \\ &\leq \left(\frac{3e}{m}\right)^m\left(\frac{2}{\varepsilon}\right)^{5m\log_2 m\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}+m} \\ &\leq \left(\frac{3e}{m}\right)^m\left(\frac{2}{\varepsilon}\right)^{6m\log_2 m\frac{\operatorname{diam}(P)\operatorname{diam}(K)}{(\alpha-1)q^2}}. \end{aligned}$$
□

Remark. Instead taking $s_i = \frac{\log(2m)}{(\alpha-1)q}$ for all i in the proof produces a slightly better bound of
$$h = \left(\frac{3e}{m}\right)^m\left(\frac{2}{\varepsilon}\right)^{6m\log_2 m\frac{\operatorname{diam}(K)}{(\alpha-1)q}}.$$
However, our version is preferable for the purposes of Section 2.2.

2.2 Constructive Approximation of a Continuous Function

We will use the preceding approximation of a polytope's indicator function to construct an approximation of a continuous function. We will do this by dividing the underlying compact set into small enough polytopes and approximating the value of the function on each of them. In Subsection 2.2.1 we look at ways of efficiently dividing sets into polytopes and in Subsection 2.2.2 we put this together with Theorem 27 to finalize the construction.

2.2.1 Lattices

In the rest of the construction we want to divide the space into polytopes such that each fits in a sphere of a given radius. The problem of finding the most economical way to cover the Euclidean space with balls of equal radius is known as the Covering Problem and in general it is open. For our purposes, regular arrangements of ball centres, lattices, are of particular interest. We will see that while an optimal lattice is unknown in all but a few dimensions, we are able to define a somewhat reasonable one. In this subsection we survey some results on lattice coverings. We will proceed to decompose the space into polytopes using the lattice points as their centres.

We are interested in point arrangements such that balls around each of them cover the whole $\mathbb{R}^d$. On the other hand, we want to keep the covering as sparse as possible, so we do not want points to be too close to each other. These two dual notions are formalised in the following definition.

Definition 12. *Let $\delta > 0$. A set of points $X \subseteq \mathbb{R}^d$ forms*

- *a δ-packing if for all $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in X$ the balls $B\left(\boldsymbol{x}_1, \frac{\delta}{2}\right)$ and $B\left(\boldsymbol{x}_2, \frac{\delta}{2}\right)$ are disjoint,*
- *a δ-covering (of $\mathbb{R}^d$) if $\mathbb{R}^d \subseteq \bigcup_{\boldsymbol{x} \in X} \overline{B(\boldsymbol{x}, \delta)}$,*
- *a δ-net on $\mathbb{R}^d$ if it is both a δ-covering of $\mathbb{R}^d$ and a δ-packing.*

To illustrate consequences of the definition we give some simple bounds on the number of points in a ball of radius R. Even though we constructed our own proof, similar bounds are known.

Theorem 28. *Let $R > 0$, $\delta > 0$, and let $X \subseteq \mathbb{R}^d$. Then,*

(i) if X is a δ-packing,

$$|X \cap B(\mathbf{0}, R)| \leq \left(\frac{2R}{\delta} + 1\right)^d;$$

(ii) if X is a δ-covering of $\mathbb{R}^d$, define $\widetilde{X} = \{\boldsymbol{x} \in X \mid B(\boldsymbol{x}, \delta) \cap B(\mathbf{0}, R) \neq \emptyset\}$. Then $B(\mathbf{0}, R) \subseteq \bigcup_{\boldsymbol{x} \in \widetilde{X}} \overline{B(\boldsymbol{x}, \delta)}$ and

$$\left|\widetilde{X}\right| \geq \left(\frac{R}{\delta}\right)^d.$$

Proof. (i) For all $\boldsymbol{x} \in X \cap B(\mathbf{0}, R)$ the balls $B\left(\boldsymbol{x}, \frac{\delta}{2}\right)$ are pairwise disjoint subsets of $B\left(\mathbf{0}, R + \frac{\delta}{2}\right)$. Denoting by λ^d the d-dimensional Lebesgue measure, this means that

$$\lambda^d\left(B\left(\mathbf{0}, R + \frac{\delta}{2}\right)\right) \geq |X \cap B(\mathbf{0}, R)|\, \lambda^d\left(B\left(\mathbf{0}, \frac{\delta}{2}\right)\right)$$

and so

$$|X \cap B(\mathbf{0}, R)| \leq \frac{\lambda^d\left(B\left(\mathbf{0}, R + \frac{\delta}{2}\right)\right)}{\lambda^d\left(B\left(\mathbf{0}, \frac{\delta}{2}\right)\right)} = \left(\frac{R + \frac{\delta}{2}}{\frac{\delta}{2}}\right)^d.$$

(ii) Since $\underline{B(\mathbf{0}, R) \subseteq \mathbb{R}^d} \subseteq \bigcup_{\boldsymbol{x} \in X} \overline{B(\boldsymbol{x}, \delta)}$, from the definition of $\widetilde{X}$ it follows that $B(\mathbf{0}, R) \subseteq \bigcup_{\boldsymbol{x} \in \widetilde{X}} \overline{B(\boldsymbol{x}, \delta)}$. Similarly to (i):

$$\lambda^d(B(\mathbf{0}, R)) \leq \left|\widetilde{X}\right| \lambda^d(\overline{B(\mathbf{0}, \delta)}),$$

meaning

$$\left|\widetilde{X}\right| \geq \frac{\lambda^d(B(\mathbf{0}, R))}{\lambda^d(\overline{B(\mathbf{0}, \delta)})} = \left(\frac{R}{\delta}\right)^d.$$

□

Lattices and Voronoi Tessellations

The main goal of this section is to divide a compact set into small regular subsets. Given a set of vertices, we can achieve this by taking sets of points closest to one vertex.

Definition 13. *Let $X \subseteq \mathbb{R}^d$. The Voronoi tessellation of X is the map $V : X \to \mathcal{P}(\mathbb{R}^d)$ defined for $\boldsymbol{x} \in X$ by*

$$V(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^d \mid \forall \boldsymbol{x}' \in X \setminus \{\boldsymbol{x}\} : \|\boldsymbol{y} - \boldsymbol{x}\| \leq \|\boldsymbol{y} - \boldsymbol{x}'\|\}.$$

The set $V(\boldsymbol{x})$ is called the Voronoi cell of $\boldsymbol{x}$.

As the Voronoi tessellation of a general $X \subseteq \mathbb{R}^d$ can be wild, we will focus only on regular sets of vertex points.

Definition 14. *A set $X \subseteq \mathbb{R}^d$ is called a (geometric) lattice if*

1. *X is a subgroup of the additive group $(\mathbb{R}^d, +, -, \mathbf{0})$, that is, if it is closed under addition and subtraction, and contains $\mathbf{0}$, and*
2. *there exists $\alpha, \beta > 0$ such that X is an α-packing and a β-covering.*

We define the packing radius and the covering radius of X respectively as

$$\alpha_0 = \sup\{\alpha > 0 \mid X \text{ is a } 2\alpha\text{-packing}\} = \frac{1}{2} \inf_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in X} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| = \frac{1}{2} \inf_{\mathbf{0} \neq \boldsymbol{x} \in X} \|\boldsymbol{x}\|$$

and

$$\beta_0 = \inf\{\beta > 0 \mid X \text{ is a } \beta\text{-covering}\} = \sup_{\boldsymbol{y} \in \mathbb{R}^d} \inf_{\boldsymbol{x} \in X} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

The following lemma summarizes some well known facts about lattices, see e.g. [23, Sections 1.1.2 and 2.1.2].

Lemma 29. *Let $X \subseteq \mathbb{R}^d$ be a lattice. Then*

(i) the additive group $(X, +, -, \mathbf{0})$ is isomorphic to $(\mathbb{Z}^d, +, -, \mathbf{0})$. In particular, X is countable and has a basis of size d;

(ii) the Voronoi cells $V(\boldsymbol{x})$ for $\boldsymbol{x} \in X$ are polytopes and they are congruent in the sense that for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$: $V(\boldsymbol{x}_1) = V(\boldsymbol{x}_2) - \boldsymbol{x}_2 + \boldsymbol{x}_1$;

(iii) the packing radius α_0 of X is equal to the inradius of $V(\boldsymbol{x})$ and the covering radius β_0 of X is the circumradius of $V(\boldsymbol{x})$ for any $\boldsymbol{x} \in X$.

Example. The simplest example of a lattice is the cubic lattice $\mathbb{Z}^d$. This is clearly an additive subgroup of $\mathbb{R}^d$ and it is also a 1-packing and a $\frac{\sqrt{d}}{2}$-covering, as the length of the main diagonal of a unit d-cube is $\sqrt{d}$. Voronoi cells of $\mathbb{Z}^d$ are hypercubes of edge length 1 — For example, $V(\mathbf{0})$ is the hypercube with vertices of the form $\left(\pm\frac{1}{2}, \ldots, \pm\frac{1}{2}\right)^\top$.

Definition 15. *Let X be a lattice and let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ be its basis. The matrix $M = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)^\top$ is called a generator matrix of X.*

Example. For many lattices it is easier to specify a basis as vectors in $\mathbb{R}^{d+1}$ that all lie on the same hyperplane. The corresponding generating matrix is then of the shape $d \times (d+1)$.

For example, the simplectic lattice A_d is typically defined via the basis $\boldsymbol{u}_i = \boldsymbol{e}_{i+1} - \boldsymbol{e}_1$, where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d+1}$ is the standard basis of $\mathbb{R}^{d+1}$. The vectors $\boldsymbol{e}_i$ are vertices of a regular d-simplex and the basis vectors $\boldsymbol{u}_i$ corresponds to some edges of the simplex. This is illustrated in Figure 2.2 for $d = 2$. The Voronoi tessellation of A_2 is the well known hexagonal tiling of the plane. A_d can also be defined as those vectors in $\mathbb{Z}^{d+1}$ whose coordinates sum to 0.

Lattice Covering Density

For a given $\delta > 0$, we would like to find a lattice δ-covering that is as sparse as possible. This is known as the Lattice Covering Problem and it is an open problem for all but a few dimensions.

Definition 16. *Let M be a generator matrix for a lattice X. The determinant of the lattice X is defined as* $\det X = \det(MM^\top)$.

Remark. The determinant of X corresponds to the square of the volume of the parallelotope with vertices $\sum_{i \in I} \boldsymbol{u}_i$, $I \subseteq \{1, \ldots, d\}$, called the fundamental parallelotope of X. The determinant does not depend on the choice of the basis. Figure 2.2c illustrates how copies of the fundamental parallelotope fill the whole space.

Definition 17. *Let X be a lattice and $\beta_0 > 0$ its covering radius. The covering density of X is defined as*

$$\Theta(X) = \frac{\lambda^d(B(\mathbf{0}, \beta_0))}{\sqrt{\det X}}.$$

The Lattice Covering Problem is then formulated as finding the least dense lattice in $\mathbb{R}^d$.

Example. In two dimensions, the covering radius of $\mathbb{Z}^2$ is $\frac{\sqrt{2}}{2}$ and the fundamental parallelotope is a unit square (of volume 1). The density of $\mathbb{Z}^2$ is therefore $\frac{\pi}{2}$. It can be shown that A_2 has a density of $\frac{2\pi}{3\sqrt{3}}$ and Kershner [24] showed that this is the smallest possible covering density in $\mathbb{R}^2$.

The Permutohedral Lattice

We will now define a lattice that is optimal in low dimensions. We do so by specifying a basis, again on a hyperplane in $\mathbb{R}^{d+1}$

Definition 18. *Denote $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^{d+1}$. The d-dimensional permutohedral lattice $A_d^* \subseteq \mathbb{R}^{d+1}$ is generated by the basis $\left\{\boldsymbol{e}_i - \frac{1}{d+1}\mathbf{1}\right\}_{i=1}^{d}$, where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d+1}$ is the standard basis of $\mathbb{R}^{d+1}$.*

Let us fix for the rest of the thesis a linear isometry Φ from the hyperplane $\left\{(y_1, \ldots, y_{d+1})^\top \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} y_i = 0\right\}$ to $\mathbb{R}^d$. We will denote by $\mathcal{A}_d^* \subseteq \mathbb{R}^d$ the image of A_d^* under Φ.

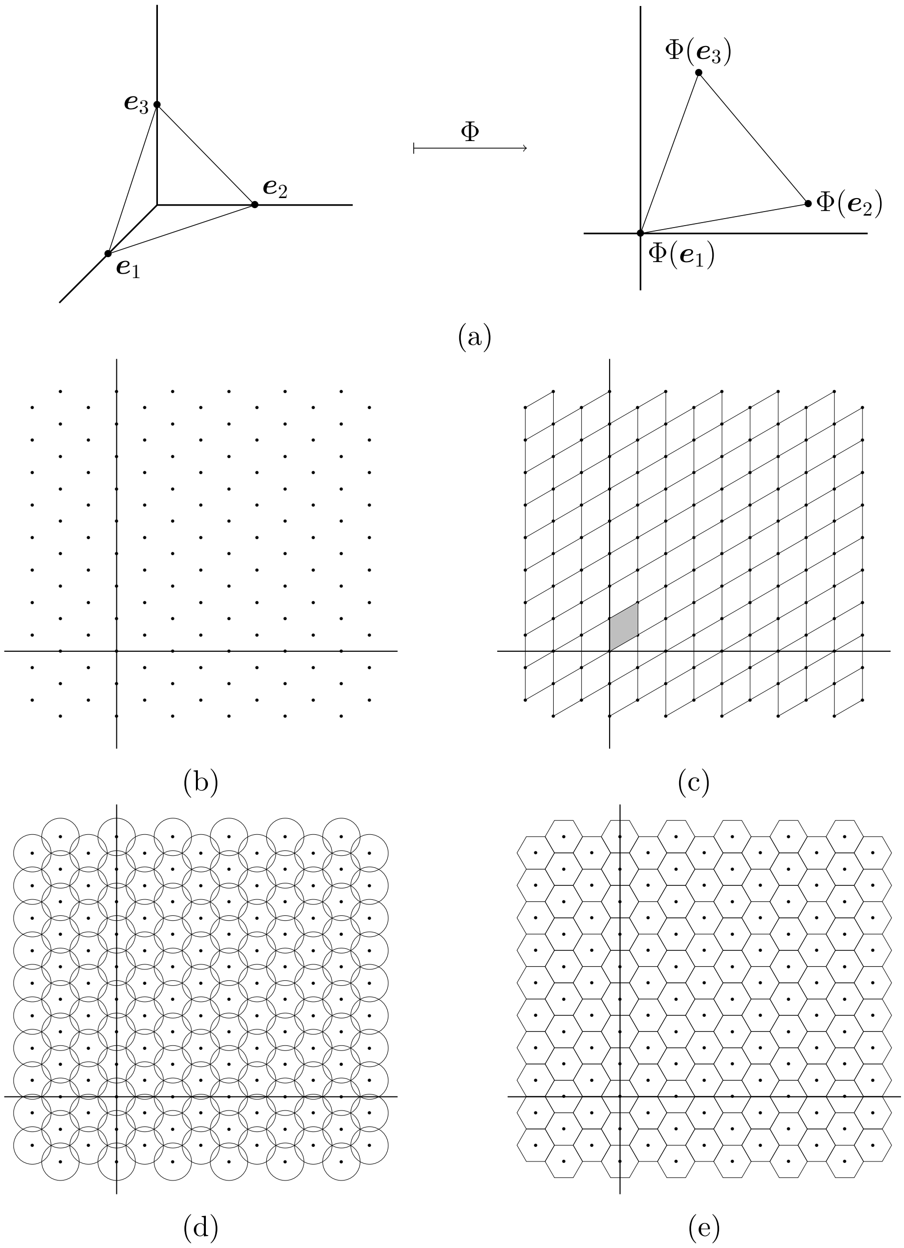

Figure 2.2: (a) Points $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are vertices of a regular 2-simplex, an equilateral triangle. The vectors $\boldsymbol{e}_2 - \boldsymbol{e}_1$ and $\boldsymbol{e}_3 - \boldsymbol{e}_1$ form a basis of A_2. In $\mathbb{R}^2$ this corresponds to $\Phi(\boldsymbol{e}_2) - \Phi(\boldsymbol{e}_1)$ and $\Phi(\boldsymbol{e}_3) - \Phi(\boldsymbol{e}_1)$, where Φ is a linear isometry from the hyperplane $\left\{(y_1, y_2, y_3)^\top \in \mathbb{R}^3 \mid \sum_{i=1}^3 y_i = 1\right\}$ to $\mathbb{R}^2$. (b) Points of A_2 in $\mathbb{R}^2$. (c) Decomposition of $\mathbb{R}^2$ into copies of the fundamental parallelotope of A_2, shown in grey. The parallelotope depends on the choice of basis, but its volume does not. (d) The covering radius of A_2. (e) The Voronoi tessellation of A_2 is the hexagonal tiling of the plane.

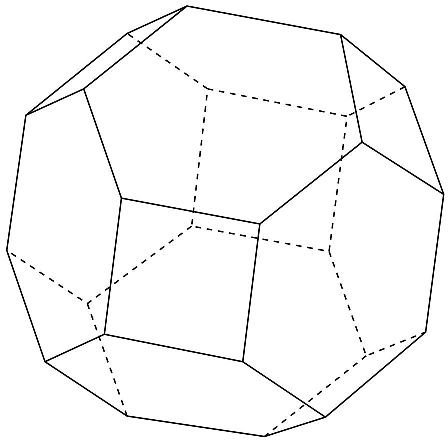

Figure 2.3: The three-dimensional permutohedron of order 4 is known as the truncated octahedron and it is the Voronoi cell of A_3^*. It has 8 hexagonal faces and 6 square faces, totalling $2^4 - 2 = 14$.

Remark. The notation A_d^* comes from the fact that the permutohedral lattice is dual to the simplectic lattice A_d.

A_d^* can be defined as the orthogonal projection of $\mathbb{Z}^{d+1}$ onto the hyperplane $\left\{(y_1, \ldots, y_{d+1})^\top \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} y_i = 0\right\}$, the basis above corresponding to the projection of the standard basis of $\mathbb{R}^{d+1}$. See [25] for a detailed overview of the permutohedral lattice.

While the definition does not give much insight into the structure of A_d^*, the lattice is well studied and the following lemma summarizes some known facts (see [23, Section 4.6.6]).

Lemma 30. *The permutohedral lattice A_d^* has covering density equal to*

$$\Theta(A_d^*) = \lambda^d\left(B(\mathbf{0}, 1)\right)\sqrt{d+1}\left(\frac{d(d+2)}{12(d+1)}\right)^{\frac{d}{2}}.$$

The packing and covering radii are

$$\alpha_0 = \frac{1}{2}\sqrt{\frac{d}{d+1}}, \quad \beta_0 = \sqrt{\frac{d(d+2)}{12(d+1)}}$$

and $\det A_d^* = \frac{1}{d+1}$. *The Voronoi cells of A_d^* are permutohedra of order $d+1$, having $2^{d+1} - 2$ facets.*

Remark. A permutohedron of order $d+1$ is a d-dimensional polytope having $(d+1)!$ vertices. Explicit definitions of permutohedra are again typically given in $\mathbb{R}^{d+1}$, this time on the hyperplane $\left\{(y_1, \ldots, y_{d+1})^\top \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} y_i = \frac{n(n+1)}{2}\right\}$, as the convex hull of all coordinate permutations of the vector $(1, \ldots, d+1)^\top$. Permutohedra of order 3 are hexagons, Figure 2.3 shows a permutohedron of order 4.

Example. For $d = 2$, $A_2^* \cong A_2$ is the lattice whose Voronoi tessellation is the hexagonal tiling, shown in Figure 2.2. A_3^* is called the body centred cubic lattice and it appears in crystallography. Its Voronoi tessellation, known as the bitruncated cubic honeycomb, consists of truncated octahedra, shown in Figure 2.3. A basis in $\mathbb{R}^3$ is given by $\left(-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)^\top$, $\left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right)^\top$ and $\left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)^\top$.

Remark. A_d^* is known to be optimal for the Lattice Covering Problem for $d \leq 5$, but is not optimal in general — For $d = 24$ there is a lattice specific for this dimension called the Leech lattice E_{24} such that $\Theta(E_{24}) < \Theta(A_{24}^*)$. In high dimensions, constructions based on low-dimensional lattices produce lattices that significantly surpass A_d^* (see [26]). However, A_d^* is relatively simple to define and to work with, so we will use it as a good approximation.

The dual problem to the Covering Problem, the Sphere Packing Problem, is probably more well-known and studied. It asks to maximize the packing density of a (lattice) point arrangement, defined analogously to the covering density. In the cases of $d = 2$ and $d = 3$, the Sphere Packing Problem is known as the Honeycomb Conjecture and the Kepler Conjecture, respectively, and in both cases the simplectic lattice A_d has been proved optimal (even including non-lattice arrangements) by Hales in [27] and [28]. Among lattice packings, A_d is known to be optimal for $d \leq 7$.

In dimensions 8 and 24, lattices specific for the dimensions, E_8 and E_{24}, have been proved to be optimal by Viazovska [29] and Cohn et al. [30]. Accordingly, the self-dual Leech lattice E_{24} surpasses A_{24}^* in the Covering Problem. On the other hand, $E_8^* \cong E_8$ is not optimal for the Covering Problem, as $\Theta(A_8^*) < \Theta(E_8)$.

Bounds on the Number of Lattice Points in a Ball

Finally, we proved the following theorem, which converts the results into a form useful for our construction.

Theorem 31. *Let $X \subseteq \mathbb{R}^d$ be a lattice with packing and covering radii $\alpha_0 > 0$ and $\beta_0 > 0$, let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ be a basis of X and let $\delta > 0$. Then, δX is a lattice with packing radius $\delta\alpha_0$ and covering radius $\delta\beta_0$ and for all $R > 0$:*

$$|\delta X \cap B(\mathbf{0}, R)| \leq \frac{\lambda^d\left(B\left(\mathbf{0}, 1\right)\right)}{\sqrt{\det(X)}} \left(\frac{R}{\delta} + \sum_{i=1}^{d} \|\boldsymbol{u}_i\|\right)^d.$$

Proof. The first half of the claim is trivial. As for the bound, copies of the fundamental parallelotope given by $\delta\boldsymbol{u}_1, \ldots, \delta\boldsymbol{u}_d$, shifted by members of δX, fill the whole space $\mathbb{R}^d$. Boundaries of these polytopes can be redistributed so that we get a decomposition of $\mathbb{R}^d$ into disjoint sets such that each contains exactly one lattice point. One such set is called a fundamental region (see [23, Section 1.1.2]). Denote by $L(\boldsymbol{x})$ the fundamental region containing $\boldsymbol{x} \in \delta X$. We have

$$\lambda^d(L(\boldsymbol{x})) = \sqrt{\det(\delta X)} = \sqrt{\det\left((\delta M)(\delta M)^\top\right)} = \delta^d \sqrt{\det(X)},$$

where M is the generator matrix of X with rows $\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_d^\top$.

Denote $\widetilde{X} = \{\boldsymbol{x} \in \delta X \mid L(\boldsymbol{x}) \cap B(\mathbf{0}, R) \neq \emptyset\}$. If $\boldsymbol{x} \in \delta X \cap B(\mathbf{0}, R)$, then clearly $\boldsymbol{x} \in \widetilde{X}$, so $\widetilde{X} \supseteq X \cap B(\mathbf{0}, R)$.

Also, for any $\boldsymbol{x} \in \widetilde{X}$: $L(\boldsymbol{x}) \subseteq \overline{B(\mathbf{0}, R + \operatorname{diam} L(\boldsymbol{x}))}$. Since $L(\boldsymbol{x})$ corresponds to a parallelotope except for boundaries, diam $L(\boldsymbol{x})$ is the length of the longest main diagonal, which is of the form $\left\|\sum_{i=1}^d \pm\delta\boldsymbol{u}_i\right\| \leq \delta \sum_{i=1}^d \|\boldsymbol{u}_i\|$.

Then,

$$\bigcup_{\boldsymbol{x}\in\widetilde{X}} L(\boldsymbol{x}) \subseteq \overline{B\left(\mathbf{0}, R+\delta\sum_{i=1}^{d}\|\boldsymbol{u}_i\|\right)},$$

which means

$$\lambda^d\left(B\left(\mathbf{0}, R+\delta\sum_{i=1}^{d}\|\boldsymbol{u}_i\|\right)\right) \geq \sum_{\boldsymbol{x}\in\widetilde{X}} \lambda^d(L(\boldsymbol{x})) = \left|\widetilde{X}\right| \lambda^d(L(\mathbf{0})).$$

Altogether,

$$|\delta X \cap B(\mathbf{0}, R)| \leq \left|\widetilde{X}\right| \leq \frac{\lambda^d\left(B\left(\mathbf{0}, R+\delta\sum_{i=1}^{d}\|\boldsymbol{u}_i\|\right)\right)}{\delta^d\sqrt{\det(X)}}.$$

□

Remark. By the nature of the proof, the same bound holds for balls not centred on $\mathbf{0}$.

The theorem also formally justifies why we are interested in the least dense covering: Let X and Y be lattices with covering radii β_X and β_Y, respectively, such that $\Theta(X) < \Theta(Y)$. By definition, this means

$$\begin{aligned}\lambda^d(B(\mathbf{0},1))\frac{\beta_X^d}{\sqrt{\det X}} &< \lambda^d(B(\mathbf{0},1))\frac{\beta_Y^d}{\sqrt{\det Y}} \quad \Leftrightarrow\\ \frac{\lambda^d\left(B\left(\mathbf{0},R\right)\right)}{\left(\frac{\delta}{\beta_X}\right)^d\sqrt{\det(X)}} &< \frac{\lambda^d\left(B\left(\mathbf{0},R\right)\right)}{\left(\frac{\delta}{\beta_Y}\right)^d\sqrt{\det(Y)}},\end{aligned}$$

which is almost the bound from Theorem 31, except for the diameter of the fundamental region, for δ equal to $\frac{\delta}{\beta_X}$ and $\frac{\delta}{\beta_Y}$.

Corollary 32. Let $\delta > 0$ and define $X = \delta\sqrt{\frac{12(d+1)}{d(d+2)}}\mathcal{A}_d^*$. Then, the lattice X has covering radius δ, packing radius $\delta\sqrt{\frac{3}{d+2}}$ and for all $R > 0$:

$$|X \cap B(\mathbf{0}, R)| \leq \lambda^d(B(\mathbf{0},1))\sqrt{d+1}\left(\frac{\sqrt{d+1}R}{\sqrt{12}\delta} + \sqrt{d+1}\right)^d.$$

Proof. By Lemma 30, the covering radius of $\mathcal{A}_d^*$ is $\sqrt{\frac{d(d+2)}{12(d+1)}}$ and $\det\mathcal{A}_d^* = \frac{1}{d+1}$. By the proof of Theorem 31, we can replace the sum of basis norms in the bound by the diameter of the fundamental parallelotope, which can be shown to be $\sqrt{\frac{d(d+2)}{d+1}} < \sqrt{d+1}$. Therefore,

$$\begin{aligned}|X \cap B(\mathbf{0}, R)| &\leq \lambda^d(B(\mathbf{0},1))\sqrt{d+1}\left(\sqrt{\frac{d(d+2)}{12(d+1)}}\frac{R}{\delta} + \sqrt{d+1}\right)^d\\ &\leq \lambda^d(B(\mathbf{0},1))\sqrt{d+1}\left(\sqrt{\frac{d+1}{12}}\frac{R}{\delta} + \sqrt{d+1}\right)^d.\end{aligned}$$

□

Corollary 33. Let $\delta > 0$ and define $X = \delta\frac{2}{\sqrt{d}}\mathbb{Z}^d$. Then, the lattice X has covering radius δ, packing radius $\delta\frac{1}{\sqrt{d}}$ and for all $R > 0$:

$$|X \cap B(\mathbf{0}, R)| \leq \lambda^d(B(\mathbf{0}, 1)) \left(\frac{\sqrt{d}R}{2\delta} + \sqrt{d}\right)^d.$$

Proof. The covering radius of $\mathbb{Z}^d$ is $\frac{\sqrt{d}}{2}$ and $\det \mathbb{Z}^d = 1$. The fundamental parallelotope is the unit cube. This way we get

$$|X \cap B(\mathbf{0}, R)| \leq \lambda^d(B(\mathbf{0}, 1)) \left(\frac{\sqrt{d}R}{2\delta} + \sqrt{d}\right)^d.$$

□

Remark. In comparison, bounds given by Theorem 28 are

$$|X_1 \cap B(\mathbf{0}, R)| \leq \left(\frac{\sqrt{d+2}R}{\sqrt{3}\delta} + 1\right)^d$$

for $X_1 = \delta\sqrt{\frac{12(d+1)}{d(d+2)}}\mathcal{A}_d^*$ and

$$|X_2 \cap B(\mathbf{0}, R)| \leq \left(\frac{\sqrt{d}R}{\delta} + 1\right)^d$$

for $X_2 = \delta\frac{2}{\sqrt{d}}\mathbb{Z}^d$. In low dimensions these bounds may be more efficient than the ones given by the corollaries.

Bound on the Number of Exponentiated Lattice Terms

Before finishing the construction, in the next lemma we proved a bound on the number of exponential functions needed in our particular situation, which will allow us to significantly decrease the final bound.

Definition 19. *Let X be a lattice.. Two elements $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in X$ are neighbouring, if $V(\boldsymbol{x}_1)$ and $V(\boldsymbol{x}_2)$ share a facet.*

Remark. Let $\boldsymbol{a} \in S^{d-1}$ be an outer unit normal vector of a facet of $V(\boldsymbol{x})$ and let q be the distance from $\boldsymbol{x}$ to that facet. Then, $\boldsymbol{x} + 2q\boldsymbol{a}$ is a neighbouring element of $\boldsymbol{x}$ in X.

Lemma 34. *Let $X \subseteq \mathbb{R}^d$ be a lattice and $\boldsymbol{x} \in X$. Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in S^{d-1}$ be outer unit normal vectors of facets of $V(\boldsymbol{x}) \in \mathbb{P}_m^d$, let $q_1, \ldots, q_m > 0$ be distances from $\boldsymbol{x}$ to the corresponding facets and define for some $c_0, \ldots, c_m \in \mathbb{R}$:*

$$g(\boldsymbol{y}) = c_0 + \sum_{i=1}^{m} c_i e^{q_i\langle \boldsymbol{a}_i, \boldsymbol{y}\rangle}.$$

Then, for all $n \in \mathbb{N}$: $g^n \in G_{\exp,h}$, where

$$h = (2n+1)^d.$$

Proof. For all $i \leq m$, $q_i \boldsymbol{a}_i$ are members of the lattice $\frac{1}{2}X$ that neighbour the point $\boldsymbol{0}$. Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d$ be a basis of $\frac{1}{2}X$. It follows that all lattice points neighbouring $\boldsymbol{0}$ have coordinates in $\{-1, 0, 1\}^d$ with respect to the basis.

Let $\varphi : \mathbb{Z}^d \to \frac{1}{2}X$ be the coordinate function, $\varphi(z_1, \dots, z_d) = \sum_{i=1}^d z_i \boldsymbol{u}_i$. Let $Z_k^d = \left\{\boldsymbol{z} \in \mathbb{Z}^d \mid \|\boldsymbol{z}\|_\infty \leq k\right\} = \{-k, \dots, k\}^d$. By the previous, there exist $\boldsymbol{z}_1, \dots, \boldsymbol{z}_m \in Z_1^d$ such that

$$q_i \boldsymbol{a}_i = \varphi(\boldsymbol{z}_i).$$

Additionally, set $\boldsymbol{z}_0 = \boldsymbol{0}$. Then,

$$g(\boldsymbol{y}) = \sum_{i=0}^{m} c_i e^{\langle \varphi(\boldsymbol{z}_i), \boldsymbol{y} \rangle}.$$

Any additive term of g^n is of the form

$$\prod_{j=1}^{n} c_{i_j} e^{\left\langle \varphi\left(\boldsymbol{z}_{i_j}\right), \boldsymbol{y} \right\rangle} = \left(\prod_{j=1}^{n} c_{i_j}\right) e^{\left\langle \sum_{j=1}^{n} \varphi\left(\boldsymbol{z}_{i_j}\right), \boldsymbol{y} \right\rangle} = \left(\prod_{j=1}^{n} c_{i_j}\right) e^{\left\langle \varphi\left(\sum_{j=1}^{n} \boldsymbol{z}_{i_j}\right), \boldsymbol{y} \right\rangle},$$

where $i_j \in \{0, \dots, m\}$ for each j. Since $\left\|\boldsymbol{z}_{i_j}\right\|_\infty \leq 1$ for all j, we have

$$\left\|\sum_{j=1}^{n} \boldsymbol{z}_{i_j}\right\|_\infty \leq \sum_{j=1}^{n} \left\|\boldsymbol{z}_{i_j}\right\|_\infty \leq n,$$

so $\sum_{j=1}^{n} \boldsymbol{z}_{i_j} \in Z_n^d$. That is, there exist $\boldsymbol{z}_1, \dots, \boldsymbol{z}_h \in Z_n^d$, $\boldsymbol{z}_i \neq \boldsymbol{z}_j$, and $\tilde{c}_1, \dots \tilde{c}_h \in \mathbb{R}$ such that

$$g^n(\boldsymbol{y}) = \sum_{i=1}^{h} \tilde{c}_i e^{\langle \varphi(\boldsymbol{z}_i), \boldsymbol{y} \rangle},$$

which implies an injective function the terms of g^n to Z_n^d. Therefore, the number of terms in g^n is at most $\left|Z_n^d\right| = (2n+1)^d$.

□

Remark. It can be easily verified that in the case of A_d^* this bound is attained.

2.2.2 Construction

In the proof of the following theorem, we took inspiration in the final proof in [3]. Recall that the Voronoi cell of a set X at $\boldsymbol{x} \in X$ is the set of all points of $\mathbb{R}^d$ closest to $\boldsymbol{x}$ among members of X (Definition 13). Also recall that $G_{\exp,k}$ denotes the set of functions representable by a shallow neural network with the exponential activation function, $\exp(t) = e^t$, that has k units in the hidden layer (Definition 5).

Definition 20. *Let $f : K \to \mathbb{R}$ and $\varepsilon > 0$. Denote by $\omega^{-1}(f, \varepsilon)$ the inverse modulus of continuity of f at ε, that is,*

$$\omega^{-1}(f, \varepsilon) = \sup\left\{\delta' > 0 \mid \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in K : \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \delta' \Rightarrow |f(\boldsymbol{y}_1) - f(\boldsymbol{y}_2)| \leq \varepsilon\right\}.$$

Theorem 35. *Let $K \subseteq \mathbb{R}^d$ be compact, let $f : K \to \mathbb{R}$ be continuous and let $\widetilde{\varepsilon} \in (0,1)$.*

Define $C = \max_{\boldsymbol{y}\in K} f(\boldsymbol{y}) - \min_{\boldsymbol{y}\in K} f(\boldsymbol{y})$, let $n = \left\lfloor \frac{2C}{\widetilde{\varepsilon}} \right\rfloor - 1$ and define $\delta = \omega^{-1}\left(f, \frac{\widetilde{\varepsilon}}{2}\right)$ as the inverse of the modulus of continuity of f at $\frac{\widetilde{\varepsilon}}{2}$.

Let $X \subseteq \mathbb{R}^d$ be a lattice with covering radius $\frac{\delta}{3}$ such that the Voronoi cells of X have m facets, $V(\boldsymbol{x}) \in \mathbb{P}^d_m$, and let q be the packing radius of X. Let

$$k = \left|\{\boldsymbol{x} \in X \mid V(\boldsymbol{x}) \cap K \neq \emptyset\}\right|,$$

Then, there exists $g \in G_{\exp,h}$ such that $\|f - g\|_\infty < \widetilde{\varepsilon}$, where

$$h = (6k)^d \left(\frac{4kn}{\widetilde{\varepsilon}}\right)^{5d \log_2 m^{\frac{2\delta \, diam(K)}{3q^2}} + d}.$$

Proof. Take $\widetilde{f} = f - \min_K f$, so that the codomain of $\widetilde{f}$ lies in $[0, C]$. Also, let $\varepsilon = \frac{\widetilde{\varepsilon}}{2}$, let $\widetilde{X} = \{\boldsymbol{x} \in X \mid V(\boldsymbol{x}) \cap K \neq \emptyset\}$ and for all $\boldsymbol{x} \in \widetilde{X}$ define

$$v_{\boldsymbol{x}} = \frac{1}{2}\left(\max_{\boldsymbol{y}\in V(\boldsymbol{x})} \widetilde{f}(\boldsymbol{y}) + \min_{\boldsymbol{y}\in V(\boldsymbol{x})} \widetilde{f}(\boldsymbol{y})\right).$$

Then, $V(\boldsymbol{x}) \subseteq \overline{B\left(\boldsymbol{x}, \frac{\delta}{3}\right)}$, so for $\boldsymbol{y}_1, \boldsymbol{y}_2 \in V(\boldsymbol{x})$: $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \delta$, which means $\left|\widetilde{f}(\boldsymbol{y}_1) - \widetilde{f}(\boldsymbol{y}_2)\right| \leq \varepsilon$. Therefore, for all $\boldsymbol{y} \in V(\boldsymbol{x})$: $\left|v_{\boldsymbol{x}} - \widetilde{f}(\boldsymbol{y})\right| \leq \frac{\varepsilon}{2}$.

For all $\boldsymbol{x} \in X$ define $g_{\boldsymbol{x}}$ as in Theorem 27 such that $g_{\boldsymbol{x}}|_K : K \to [0,1]$, for $\boldsymbol{y} \in V(\boldsymbol{x}) \cap K$: $g_{\boldsymbol{x}}(\boldsymbol{y}) > 1 - \frac{\varepsilon}{kn}$ and for $\boldsymbol{y} \in K \setminus (2 *_{\boldsymbol{x}} V(\boldsymbol{x}))$: $g_{\boldsymbol{x}}(\boldsymbol{y}) < \frac{\varepsilon}{kn}$.

For all $i \in \{1, \ldots, n\}$ define $X_i = \{\boldsymbol{x} \in \widetilde{X} \mid v_{\boldsymbol{x}} \geq i\varepsilon\}$, let

$$p_i = 1 - \prod_{\boldsymbol{x}\in X_i} (1 - g_{\boldsymbol{x}})$$

and take

$$\widetilde{g} = \varepsilon \sum_{i=1}^{n} p_i.$$

Then, p_i approximates the indicator function of $\bigcup_{\boldsymbol{x}\in X_i} V(\boldsymbol{x})$ in the following sense: $p_i|_K : K \to [0,1]$. Let $\boldsymbol{x}_0 \in X_i$ and $\boldsymbol{y} \in V(\boldsymbol{x}) \cap K$. Then, $1 - g_{\boldsymbol{x}_0}(\boldsymbol{y}) < \frac{\varepsilon}{kn}$ and for $\boldsymbol{x} \neq \boldsymbol{x}_0 \in X_i$: $g_{\boldsymbol{x}}(\boldsymbol{y}) \leq 1$, so

$$p_i(\boldsymbol{y}) > 1 - \frac{\varepsilon}{kn} \geq 1 - \frac{\varepsilon}{n}.$$

For $\boldsymbol{y} \in K \setminus \bigcup_{\boldsymbol{x}\in X_i} (2 *_{\boldsymbol{x}} V(\boldsymbol{x}))$, $1 - g_{\boldsymbol{x}}(\boldsymbol{y}) > 1 - \frac{\varepsilon}{kn}$ for all $\boldsymbol{x} \in X_i$, so by Bernoulli's inequality

$$p_i(\boldsymbol{y}) < 1 - \left(1 - \frac{\varepsilon}{kn}\right)^{|X_i|} \leq 1 - 1 + \frac{\varepsilon}{n}\frac{|X_i|}{k} \leq \frac{\varepsilon}{n}.$$

If $\boldsymbol{y}_1 \in (2 *_{\boldsymbol{x}_2} V(\boldsymbol{x}_2)) \setminus V(\boldsymbol{x}_2)$, then for all $\boldsymbol{y}_2 \in V(\boldsymbol{x}_2)$:

$$\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \leq \|\boldsymbol{y}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{x}_2 - \boldsymbol{y}_2\| \leq \frac{2}{3}\delta + \frac{1}{3}\delta = \delta.$$

Therefore, $|f(\boldsymbol{y}_1) - v_{\boldsymbol{x}_2}| \leq \frac{\varepsilon}{2}$. However, if $\boldsymbol{y}_1 \in V(\boldsymbol{x}_1)$, then also $|f(\boldsymbol{y}_1) - v_{\boldsymbol{x}_1}| \leq \frac{\varepsilon}{2}$, which means that $|v_{\boldsymbol{x}_1} - v_{\boldsymbol{x}_2}| \leq \varepsilon$. As a consequence, for any $i < j \leq n$, for all

$\boldsymbol{x}_1 \notin X_i$ and $\boldsymbol{x}_2 \in X_j$ we have $v_{\boldsymbol{x}_1} < i\varepsilon$ and $v_{\boldsymbol{x}_2} \geq (i+1)\varepsilon$, so by the previous $V(\boldsymbol{x}_1) \cap (2 *_{\boldsymbol{x}_2} V(\boldsymbol{x}_2)) = \emptyset$.

Altogether, for $\boldsymbol{x} \in \widetilde{X}$ and $i \leq n$ such that $\boldsymbol{x} \in X_i \setminus X_{i+1}$ — that is, $v_{\boldsymbol{x}} \in [i\varepsilon, (i+1)\varepsilon)$ — there are i indices j such that p_j is close to 1 on $V(\boldsymbol{x})$, $n-i-1$ js such that p_j is close to 0 on $V(\boldsymbol{x})$ and one index, $i+1$, such that $p_{i+1} \in [0,1]$ on $V(\boldsymbol{x})$.

Take $\boldsymbol{y} \in V(\boldsymbol{x}) \cap K$. On one hand, for $j \leq i+1$: $p_j(\boldsymbol{y}) \leq 1$ and for $j > i+1$: $p_j(\boldsymbol{y}) < \frac{\varepsilon}{n}$, so

$$\begin{aligned}\widetilde{g}(\boldsymbol{y}) = \varepsilon \sum_{j=1}^{i+1} p_j(\boldsymbol{y}) + \varepsilon \sum_{j=i+2}^{n} p_j(\boldsymbol{y}) &< (i+1)\varepsilon + \varepsilon(n-i-1)\frac{\varepsilon}{n} \\ &\leq (i+1)\varepsilon + \varepsilon^2\end{aligned}$$

On the other hand, for $j \leq i$: $p_j(\boldsymbol{y}) > 1 - \frac{\varepsilon}{n}$ and for $j > i$: $p_j(\boldsymbol{y}) \geq 0$, so

$$\begin{aligned}\widetilde{g}(\boldsymbol{y}) = \varepsilon \sum_{j=1}^{i} p_j(\boldsymbol{y}) + \varepsilon \sum_{j=i+1}^{n} p_j(\boldsymbol{y}) &> \varepsilon i(1 - \frac{\varepsilon}{n}) \\ &\geq i\varepsilon - \varepsilon^2.\end{aligned}$$

That is, $\widetilde{g}(\boldsymbol{y}) \in (i\varepsilon - \varepsilon^2, (i+1)\varepsilon + \varepsilon^2)$. Because $v_{\boldsymbol{x}} \in [i\varepsilon, (i+1)\varepsilon)$, we get $|v_{\boldsymbol{x}} - \widetilde{g}(\boldsymbol{y})| < \varepsilon + \varepsilon^2$. We also know that $\left|v_{\boldsymbol{x}} - \widetilde{f}(\boldsymbol{y})\right| \leq \frac{\varepsilon}{2}$, hence (because $\varepsilon < \frac{1}{2}$)

$$\left|\widetilde{f}(\boldsymbol{y}) - \widetilde{g}(\boldsymbol{y})\right| < \left(1 + \varepsilon + \frac{1}{2}\right)\varepsilon < 2\varepsilon = \widetilde{\varepsilon}.$$

Finally, take $g = \widetilde{g} + \min_{\boldsymbol{y} \in K} f(\boldsymbol{y})$. Then, $\|f(\boldsymbol{y}) - g(\boldsymbol{y})\|_\infty < \widetilde{\varepsilon}$.

As for the number of terms, denote $\kappa = \left\lfloor \sqrt{m}(2m)^{\frac{\operatorname{diam}(V(\mathbf{0}))\operatorname{diam}(K)}{q^2}} - 1 \right\rfloor$ (note that the packing radius q of X equals the least distance from $\boldsymbol{x}$ to a facet of $V(\boldsymbol{x})$ for all $\boldsymbol{x} \in X$) and $\nu = \left\lceil \frac{5}{2} \log_2\left(\frac{kn}{\varepsilon}\right)\right\rceil$. Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ denote outer normal unit vectors of facets of $V(\mathbf{0})$ and let $q_1, \ldots, q_m$ be distances from $\mathbf{0}$ to corresponding facets of $V(\mathbf{0})$. By the proof of Theorem 27, $g_{\boldsymbol{x}}$ is for all $\boldsymbol{x} \in \widetilde{X}$ of the form

$$\left(1 - \left(\sum_{i=1}^{m} c_i^{\boldsymbol{x}} e^{s\langle q_i \boldsymbol{a}_i, \boldsymbol{y}\rangle}\right)^{\nu}\right)^{\kappa^\nu}.$$

for some constants $c_i^{\boldsymbol{x}}$ and a common s. Then,

$$p_j(\boldsymbol{y}) = 1 - \prod_{\boldsymbol{x} \in X_j} \left(1 - \left(1 - \left(\sum_{i=1}^{m} c_i^{\boldsymbol{x}} e^{s\langle q_i \boldsymbol{a}_i, \boldsymbol{y}\rangle}\right)^{\nu}\right)^{\kappa^\nu}\right).$$

In turn, $\frac{1}{\varepsilon}\widetilde{g}$ is a sub-expression of

$$1 - \prod_{\boldsymbol{x} \in \widetilde{X}} \left(1 - \left(1 - \left(\sum_{i=1}^{m} c_i^{\boldsymbol{x}} e^{s\langle q_i \boldsymbol{a}_i, \boldsymbol{y}\rangle}\right)^{\nu}\right)^{\kappa^\nu}\right),$$

which contains at most as many exponential terms as

$$\left(1 - \sum_{i=1}^{m} e^{s\langle q_i \boldsymbol{a}_i, \boldsymbol{y}\rangle}\right)^{\nu\kappa^\nu |\widetilde{X}|} = \left(1 - \sum_{i=1}^{m} e^{s\langle q_i \boldsymbol{a}_i, \boldsymbol{y}\rangle}\right)^{\nu\kappa^\nu k}. \tag{2.7}$$

By Lemma 34, the number of terms in (2.7) is at most

$$h = (2\nu\kappa^{\nu}k + 1)^d$$

As in Theorem 27,

$$\nu \le \frac{5}{2}\log_2\left(\frac{2kn}{\varepsilon}\right)$$

and

$$\kappa^{\nu} \le \left(\frac{2kn}{\varepsilon}\right)^{5\log_2 m \frac{\mathrm{diam}(V(\mathbf{0}))\mathrm{diam}(K)}{q^2}}.$$

Since $V(\mathbf{0}) \subseteq B\left(\mathbf{0}, \frac{\delta}{3}\right)$, $\mathrm{diam}(V(\mathbf{0})) \le \frac{2\delta}{3}$. Together,

$$\begin{aligned} h &\le \left(5\log_2\left(\frac{2kn}{\varepsilon}\right)\left(\frac{2kn}{\varepsilon}\right)^{5\log_2 m \frac{2\delta\mathrm{diam}(K)}{3q^2}} k + 1\right)^d \\ &\le (6k)^d\left(\frac{2kn}{\varepsilon}\right)^{5d\log_2 m\frac{2\delta\mathrm{diam}(K)}{3q^2}+d}. \end{aligned}$$

□

Remark. If $f : K \to \mathbb{R}$ is Lipschitz continuous with a constant $L > 0$, that is, if for all $\boldsymbol{x}, \boldsymbol{y} \in K$

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le L\,\|\boldsymbol{x} - \boldsymbol{y}\|,$$

then for any $\varepsilon > 0$ the inverse modulus of continuity is bounded by $\frac{\varepsilon}{L}$.

Putting together Theorem 35 with the permutohedral lattice and Lemma 30, we get the following bound.

Corollary 36. Let $d \ge 2$, let $K \subseteq \mathbb{R}^d$ be compact, let $f : K \to \mathbb{R}$ be continuous and let $\varepsilon \in (0, 1)$.

Denote by $\delta = \omega^{-1}(f, \frac{\varepsilon}{2})$ the inverse modulus of continuity of f at $\frac{\varepsilon}{2}$. Then, there exists $g \in G_{\exp,h}$ such that $\|f - g\|_{\infty} < \varepsilon$, where

$$h = \left(6\sqrt{d+1}\frac{\mathrm{diam}(K)\,\|f\|_{\infty}}{\delta\varepsilon^2}\right)^{10(d+1)^4\frac{\mathrm{diam}(K)}{\delta}}.$$

Proof. Let $X = \frac{\delta}{3}\sqrt{\frac{12(d+1)}{d(d+2)}}\mathcal{A}_d^*$. By Corollary 32, X has covering radius $\frac{\delta}{3}$ and packing radius $q = \frac{\delta}{\sqrt{3(d+2)}}$. By Lemma 30, the number of facets of $V(\boldsymbol{x})$ is $m = 2^{d+1} - 2$. Let $\widetilde{X} = \{\boldsymbol{x} \in X \mid V(\boldsymbol{x}) \cap K \ne \emptyset\}$. There exists $\boldsymbol{y}_0 \in \mathbb{R}^d$ such that $K \subseteq B(\boldsymbol{y}_0, R)$ for $R = \frac{1}{2}\mathrm{diam}\ K$. Then, $\widetilde{X} \subseteq B\left(\boldsymbol{y}_0, R + \frac{\delta}{3}\right)$ and, denoting $k = \left|\widetilde{X}\right|$, again by Corollary 32

$$\begin{aligned} k \le \left|X \cap B\left(\boldsymbol{y}_0, R + \frac{\delta}{3}\right)\right| &\le \lambda^d(B(\mathbf{0}, 1))\sqrt{d+1}\left(\frac{\sqrt{3(d+1)}(R + \frac{\delta}{3})}{2\delta} + \sqrt{d+1}\right)^d \\ &= \lambda^d(B(\mathbf{0}, 1))(d+1)^{\frac{d+1}{2}}\left(\frac{\sqrt{3}R}{2\delta} + \frac{\sqrt{3}+6}{6}\right)^d \\ &\le 6(d+1)^{\frac{d+1}{2}}\left(\frac{3R}{2\delta}\right)^d \end{aligned}$$

Then, taking $C = \max_{\boldsymbol{y}\in K} f(\boldsymbol{y}) - \min_{\boldsymbol{y}\in K} f(\boldsymbol{y})$ and $n = \left\lfloor \frac{2C}{\varepsilon} \right\rfloor - 1 \leq \frac{4\|f\|_\infty}{\varepsilon}$, Theorem 35 gives us

$$\begin{aligned}
h &\leq (6k)^d \left(\frac{2kn}{\varepsilon}\right)^{5d\log_2\left(2^{d+1}-2\right)\frac{6(d+2)\delta\mathrm{diam}(K)}{3\delta^2}+d} \\
&\leq (6k)^d \left(\frac{2kn}{\varepsilon}\right)^{10d(d+1)(d+2)\frac{\mathrm{diam}(K)}{\delta}+d} \\
&\leq \left(\frac{2kn}{\varepsilon}\right)^{10(d+1)^3\frac{\mathrm{diam}(K)}{\delta}} \\
&\leq \left(6\sqrt{d+1}\frac{\mathrm{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{10(d+1)^4\frac{\mathrm{diam}(K)}{\delta}}.
\end{aligned}$$

□

We get a similar bound using $\mathbb{Z}^d$ instead of $\mathcal{A}_d^*$.

Corollary 37. Let $d \geq 2$, let $K \subseteq \mathbb{R}^d$ be compact, let $f : K \to \mathbb{R}$ be continuous and let $\varepsilon \in (0, 1)$.

Denote by $\delta = \omega^{-1}(f, \frac{\varepsilon}{2})$ the inverse modulus of continuity of f at $\frac{\varepsilon}{2}$. Then, there exists $g \in G_{\exp,h}$ such that $\|f - g\|_\infty < \varepsilon$, where

$$h = \left(8\sqrt{d}\frac{\mathrm{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{60d^3\log_2(d)\frac{\mathrm{diam}(K)}{\delta}}.$$

Proof. Let $X = \delta\frac{2}{3\sqrt{d}}\mathbb{Z}^d$. By Corollary 32, X has covering radius $\frac{\delta}{3}$ and packing radius $q = \frac{\delta}{3\sqrt{d}}$. The number of facets of $V(\boldsymbol{x})$ is $m = 2d$. Again, let $\widetilde{X} = \{\boldsymbol{x} \in X \mid V(\boldsymbol{x}) \cap K \neq \emptyset\}$. There exists $\boldsymbol{y}_0 \in \mathbb{R}^d$ such that $K \subseteq B(\boldsymbol{y}_0, R)$ for $R = \frac{1}{2}\mathrm{diam}\ K$. Then, $\widetilde{X} \subseteq B\left(\boldsymbol{y}_0, R + \frac{\delta}{3}\right)$ and, denoting $k = \left|\widetilde{X}\right|$, by Corollary 33

$$\begin{aligned}
k \leq \left|X \cap B\left(\boldsymbol{y}_0, R + \frac{\delta}{3}\right)\right| &\leq \lambda^d(B(\mathbf{0}, 1))\left(\frac{3\sqrt{d}(R+\frac{\delta}{3})}{2\delta} + \sqrt{d}\right)^d \\
&= \lambda^d(B(\mathbf{0}, 1))d^{\frac{d}{2}}\left(\frac{3R}{2\delta} + \frac{3}{2}\right)^d \\
&\leq 6d^{\frac{d}{2}}\left(\frac{2R}{\delta}\right)^d
\end{aligned}$$

Then, taking $C = \max_{\boldsymbol{y}\in K} f(\boldsymbol{y}) - \min_{\boldsymbol{y}\in K} f(\boldsymbol{y})$ and $n = \left\lfloor \frac{2C}{\varepsilon} \right\rfloor - 1 \leq \frac{4\|f\|_\infty}{\varepsilon}$, Theorem 35 gives us

$$
\begin{aligned}
h \leq\ & (6k)^d \left(\frac{2kn}{\varepsilon}\right)^{5d\log_2(2d)\frac{18d\delta\operatorname{diam}(K)}{3\delta^2}+d} \\
\leq\ & (6k)^d \left(\frac{2kn}{\varepsilon}\right)^{30d^2(\log_2(d)+1)\frac{\operatorname{diam}(K)}{\delta}+d} \\
\leq\ & \left(\frac{2kn}{\varepsilon}\right)^{60d^2\log_2(d)\frac{\operatorname{diam}(K)}{\delta}} \\
\leq\ & \left(8\sqrt{d}\frac{\operatorname{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{60d^3\log_2(d)\frac{\operatorname{diam}(K)}{\delta}} .
\end{aligned}
$$

□

Transition to a Sigmoidal Activation

As the exponential function is not commonly used as an activation function in neural networks, it would be desirable to transfer the bound to sigmoidal (or ReLU) functions. However, the straightforward method along the lines of Subsection 1.1.1 leads to the following unsatisfactory bound. This is due to dependence on outer weight size of the exponential neural network in addition to the number of neurons. See the Conclusion for possible improvements.

Theorem 38. *Let $d \geq 2$, let $K \subseteq \mathbb{R}^d$ be compact, let $f : K \to \mathbb{R}$ be continuous and let $\varepsilon \in (0,1)$. Also let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $\lim_{t\to\infty}\sigma(t) = 1$, $\lim_{t\to-\infty}\sigma(t) = 0$ and $\sigma(0) = 1$.*

Denote by $\delta = \omega^{-1}(f, \frac{\varepsilon}{2})$ the inverse modulus of continuity of f at $\frac{\varepsilon}{2}$. Then, there exists $g \in G_{\sigma,l}$ such that $\|f - g\|_\infty < \varepsilon$, where

$$
l = (5+\tau)\frac{\|f\|_\infty}{\varepsilon} h^{\frac{1}{d}h^{\frac{1}{d}}+2},
$$

where h is the bound from Corollary 36.

Proof. By the proofs of Theorems 35 and 27, the function from Corollary 36 is of the form

$$
\widetilde{g}(\boldsymbol{y}) = \varepsilon\sum_{i=j}^{n}\left(1 - \prod_{\boldsymbol{x}\in X_j}\left(1-\left(1-\left(\sum_{i=1}^{m}\frac{1}{m}e^{s(\langle q_i\boldsymbol{a}_i,\boldsymbol{y}\rangle-\beta_i^{\boldsymbol{x}})}\right)^{\nu}\right)^{\kappa^{\nu}}\right)\right),
$$

where $s(\langle q_i\boldsymbol{a}_i, \boldsymbol{y}\rangle - \beta_{i,j}^{\boldsymbol{x}}) \leq 0$ for all $\boldsymbol{x}$, i and $\boldsymbol{y} \in K$. Multiplying two exponentials that range in $(0,1]$ produces another one such, so we can write

$$
\widetilde{g}(\boldsymbol{y}) = \sum_{i=1}^{h}\widetilde{c}_i e^{\langle\widetilde{\boldsymbol{a}}_i,\boldsymbol{y}\rangle-\widetilde{b}_i},
$$

where for all i: $e^{\langle\widetilde{\boldsymbol{a}}_i,\cdot\rangle-\widetilde{b}_i} : K \to (0,1]$. Denote $A_i = \{\langle\widetilde{\boldsymbol{a}}_i,\boldsymbol{y}\rangle \mid \boldsymbol{y} \in K\} = \langle\widetilde{\boldsymbol{a}}_i, K\rangle \subseteq \mathbb{R}$ and $g_i : A_i \to (0,1]$, $g_i(t) = e^{t-\widetilde{b}_i}$. That is,

$$
\widetilde{g}(\boldsymbol{y}) = \sum_{i=1}^{h}\widetilde{c}_i g_i(\langle\widetilde{\boldsymbol{a}}_i,\boldsymbol{y}\rangle).
$$

Then, for all i and $\delta > 0$, $\omega(g_i, \delta) \leq \delta$. Let $n_i = \left\lceil \frac{|\widetilde{c}_i| h(4+\tau)}{\varepsilon} \right\rceil$, where $\tau = \sup_{t\in\mathbb{R}} |\sigma(t)|$. By Theorem 3, there exists $\widehat{g} \in G_{\sigma,n}$ such that

$$\|g_i - \widehat{g}_i\|_\infty < \omega\left(g_i, \frac{1}{n_i}\right)(4+\tau) \leq \frac{1}{n_i}(4+\tau) \leq \frac{\varepsilon}{|\widetilde{c}_i| h(4+\tau)}(4+\tau) = \frac{\varepsilon}{|\widetilde{c}_i| h}.$$

Let $\widehat{g}(\boldsymbol{y}) = \sum_{i=1}^{h} \widetilde{c}_i \widehat{g}_i(\langle \widetilde{\boldsymbol{a}}_i, \boldsymbol{y}\rangle)$. Then,

$$\begin{aligned}
\|\widetilde{g} - \widehat{g}\|_\infty^K &\leq \sum_{i=1}^{h} |\widetilde{c}_i| \, \|g_i(\langle \widetilde{\boldsymbol{a}}_i, \cdot\rangle) - \widehat{g}_i(\langle \widetilde{\boldsymbol{a}}_i, \cdot\rangle)\|_\infty^K \\
&= \sum_{i=1}^{h} |\widetilde{c}_i| \, \|g_i - \widehat{g}_i\|_\infty^{A_i} \\
&< \sum_{i=1}^{h} |\widetilde{c}_i| \frac{\varepsilon}{|\widetilde{c}_i| h} \\
&= \varepsilon.
\end{aligned}$$

As a consequence of the multinomial theorem, for all $i \leq h$:

$$|\widetilde{c}_i| \leq \varepsilon n \left(\frac{k\nu\kappa^\nu}{m}\right)^{k\nu\kappa^\nu}.$$

Also, by the proof of Theorem 35, $h = (2k\nu\kappa^\nu + 1)^d$. Together, $\widehat{g} \in G_{\sigma,l}$, where

$$\begin{aligned}
l &\leq \sum_{i=1}^{h} n_i \leq \sum_{i=1}^{h} \left(\frac{|\widetilde{c}_i| h(4+\tau)}{\varepsilon} + 1\right) \\
&= h + \frac{h(4+\tau)}{\varepsilon} \sum_{i=1}^{h} |\widetilde{c}_i| \\
&\leq h + \frac{h^2(4+\tau)}{\varepsilon} \varepsilon n \left(\frac{k\nu\kappa^\nu}{m}\right)^{k\nu\kappa^\nu} \\
&\leq (5+\tau)\frac{\|f\|_\infty}{\varepsilon} h^{\frac{1}{d}h^{\frac{1}{d}}+2}.
\end{aligned}$$

□

2.3 Towards a Bound on Approximation Almost Everywhere

An alternative approach to the construction presented in Subsection 2.2.2 would be to create a function represented by a neural network that approximates the target function everywhere except a set of small measure. This could lead not only to a cheaper approximation with respect to the number of hidden units, but also even general measurable functions could potentially be approximated in this manner via Lusin's Theorem (see e.g. [10, Chapter 2]).

In this section we lay the groundwork for further research in this direction. However, some steps are beyond the scope of this thesis. The main idea is to approximate the target function by a sum of functions constant on balls. The balls are then approximated by polytopes and the indicator functions of the polytopes are approximated by a neural network as per 27.

Given a continuous function $f : K \to \mathbb{R}$ and $\varepsilon > 0$, define a set of balls

$$\mathcal{V}_f = \left\{ B(\boldsymbol{x}, r) \mid \boldsymbol{x} \in K \ \&\ r \in \left(0, \omega^{-1}(f, \varepsilon)\right) \right\},$$

where ω^{-1} is the inverse modulus of continuity defined in Definition 20. The set $\mathcal{V}_f$ is a Vitali covering of K, meaning that for every $\boldsymbol{y} \in K$ and $\gamma > 0$ there exists $B \in \mathcal{V}_f$ such that $\boldsymbol{y} \in B$ and $\lambda^d(B) < \gamma$. As such, we can apply the Vitali Covering Theorem found in this form in [31, Lemma 3.9 and Corollary 3.10].

Theorem 39 (Vitali Covering Theorem)**.** *Let $K \subseteq \mathbb{R}^d$ be a set of finite measure and let $\mathcal{V}$ be a Vitali covering of K by balls. Then for every $\delta > 0$ there exists a finite sequence $B_1, \ldots, B_n \in \mathcal{V}$ of disjoint balls such that*

$$\lambda^d \left(K \setminus \bigcup_{i=1}^{n} B_i \right) < \delta.$$

For a given $\delta > 0$, let $B(\boldsymbol{x}_1, r_1), \ldots, B(\boldsymbol{x}_n, r_n) \in \mathcal{V}_f$ be the sequence of balls from the theorem. Then, the function $\widehat{f}$ defined as

$$\widehat{f} = \sum_{i=1}^{n} f(\boldsymbol{x}_i) 1_{B(\boldsymbol{x}_i, r_i)},$$

where $1_{B(\boldsymbol{x}_i, r_i)}$ is the indicator function, approximates f on $\bigcup_{i=1}^{n} B_i$: If $\boldsymbol{y} \in B(\boldsymbol{x}_i, r_i)$ for some i, then $\|\boldsymbol{y} - \boldsymbol{x}_i\| < \omega^{-1}(f, \varepsilon)$ and therefore $\varepsilon > |f(\boldsymbol{x}_i) - f(\boldsymbol{y})| = \left|\widehat{f}(\boldsymbol{y}) - f(\boldsymbol{y})\right|$.

2.3.1 Approximation of the Unit Ball by a Polytope

It was proved independently by Dudley [4] and Bronshteyn and Ivanov [5] that any convex body can be approximated by a convex polytope having m vertices with accuracy $\frac{c}{m^{\frac{2}{d-1}}}$ (in the Hausdorff metric). We will use the formulation by Bronshteyn and Ivanov. However, because the proof in the original paper is very brief, we adapted it and expounded it in the rest of the section for the special case of approximation of the unit ball. This results in a simplified proof and also allows us to specify the multiplicative constant. The first lemma of this section corresponds to Lemma 1 from the paper.

Lemma 40. *For all $\gamma \in (0, 1)$ there exists $m \in \mathbb{N}$, $m \leq \sqrt{2\pi d} \left(\frac{8}{\gamma\sqrt{16-\gamma^2}} \right)^{d-1}$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \in S^{d-1}$ such that $S^{d-1} \subseteq \bigcup_{i=1}^{m} B(\boldsymbol{y}_i, \gamma)$. That is, there is a γ-covering of the d-dimensional unit sphere having m points.*

Proof. Let m be the largest number such that there exist points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in S^{d-1}$ so that the balls $B(\boldsymbol{x}_i, \frac{\gamma}{2})$ are pairwise disjoint (that is, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ form a γ-packing). Such a maximal number exists because we can clearly find one such point, but not arbitrarily many.

Then, the points form a γ-covering — Suppose by contradiction there is a point $\boldsymbol{y} \in S^{d-1}$ such that for all $i \leq m$: $\|\boldsymbol{y} - \boldsymbol{x}_i\| > \gamma$. This means the balls

$B\left(\boldsymbol{y}, \frac{\gamma}{2}\right)$ and $B\left(\boldsymbol{x}_i, \frac{\gamma}{2}\right)$ are disjoint and $\{\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ is a γ-packing of size $m+1$.

Next we construct a bound on the number m. Let $\mathcal{S}$ denote the surface area (the $(d-1)$-dimensional Hausdorff measure in $\mathbb{R}^d$). Since the sets $S^{d-1} \cap B\left(\boldsymbol{x}_i, \frac{\gamma}{2}\right)$ are pairwise disjoint and their union is a subset of the sphere, the sum of their surface areas is at most the surface of the sphere. All $\mathcal{S}\left(S^{d-1} \cap B\left(\boldsymbol{x}_i, \frac{\gamma}{2}\right)\right)$ being equal, we get $m\mathcal{S}\left(S^{d-1} \cap B\left(\boldsymbol{x}_i, \frac{\gamma}{2}\right)\right) \leq \mathcal{S}\left(S^{d-1}\right)$, or

$$m \leq \frac{\mathcal{S}\left(S^{d-1}\right)}{\mathcal{S}\left(S^{d-1} \cap B\left(\boldsymbol{x}_1, \frac{\gamma}{2}\right)\right)}.$$

Let U be the orthogonal projection of $S^{d-1} \cap B\left(\boldsymbol{x}_1, \frac{\gamma}{2}\right)$ onto the tangent hyperplane to S^{d-1} at $\boldsymbol{x}_1$. Then, $\mathcal{S}(U) \leq \mathcal{S}\left(S^{d-1} \cap B\left(\boldsymbol{x}_1, \frac{\gamma}{2}\right)\right)$ and therefore

$$m \leq \frac{\mathcal{S}\left(S^{d-1}\right)}{\mathcal{S}(U)}. \tag{2.8}$$

U is a $(d-1)$-dimensional ball of some radius $r > 0$. We will show that $r = \frac{\gamma}{2}\sqrt{1 - \frac{\gamma^2}{16}}$.

Choose a point $\boldsymbol{z} \in S^{d-1} \cap \partial B\left(\boldsymbol{x}_1, \frac{\gamma}{2}\right)$ on the intersection of the unit sphere and the boundary of $B\left(\boldsymbol{x}_1, \frac{\gamma}{2}\right)$ and denote by α the angle between $\boldsymbol{x}_1$ and $\boldsymbol{z}$ — The situation is illustrated in Figure 2.4. Since $\mathbf{0}$, $\boldsymbol{x}_1$ and $\boldsymbol{z}$ form an isosceles triangle with sides $\frac{\gamma}{2}$, 1 and 1, we can easily get that $\sin\left(\frac{\alpha}{2}\right) = \frac{\gamma}{4}$.

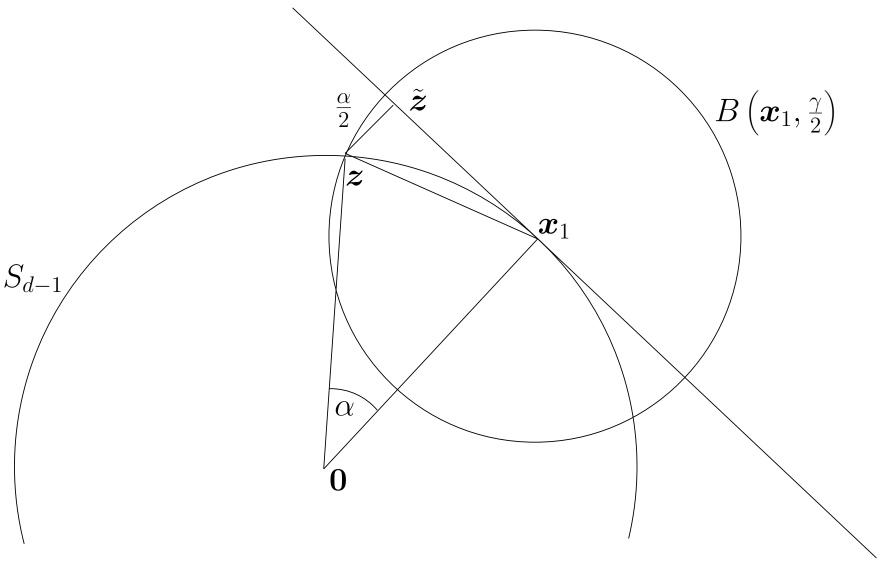

Figure 2.4: The radius of the projection, $\|\tilde{\boldsymbol{z}} - \boldsymbol{x}_1\|$, can be calculated by means of elementary geometry.

$\frac{\alpha}{2}$ is also the angle at $\boldsymbol{x}_1$ in the right triangle formed by $\boldsymbol{x}_1$, $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$, where $\tilde{\boldsymbol{z}}$ is the projection of $\boldsymbol{z}$ to the tangent hyperplane (the complement to $\frac{\pi}{2}$ of this angle is equal to the pair of angles in the isosceles triangle $\mathbf{0}$, $\boldsymbol{x}_1$, $\boldsymbol{z}$). So,

$\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\| = \sin\left(\frac{\alpha}{2}\right)\|\boldsymbol{z} - \boldsymbol{x}_1\| = \frac{\gamma}{8}$. Finally, we apply the Pythagorean theorem to the same triangle to get

$$\|\tilde{\boldsymbol{z}} - \boldsymbol{x}_1\| = \sqrt{\|\boldsymbol{z} - \boldsymbol{x}_1\|^2 - \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2} = \sqrt{\left(\frac{\gamma}{2}\right)^2 - \left(\frac{\gamma^2}{8}\right)^2} = \frac{\gamma}{2}\sqrt{1 - \frac{\gamma^2}{16}}. \quad (2.9)$$

Next,

$$\mathcal{S}\left(S^{d-1}\right) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

and

$$\mathcal{S}(U) = \frac{\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)} r^{d-1},$$

which together with (2.8) and (2.9) means that

$$m \leq 2\sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\left(\frac{\gamma}{2}\sqrt{1 - \frac{\gamma^2}{16}}\right)^{1-d} \leq \sqrt{2\pi d}\left(\frac{8}{\gamma\sqrt{16 - \gamma^2}}\right)^{d-1}.$$

□

Definition 21. *A point $\boldsymbol{y} \in \mathbb{R}^d \setminus \overline{B(\mathbf{0}, 1)}$ illuminates a point $\boldsymbol{z} \in S^{d-1}$ if the line segment $[\boldsymbol{y}, \boldsymbol{z})$ does not intersect S^{d-1}.*

The following lemma corresponds to Lemma 3 of [5].

Lemma 41. *Let $\delta > 0$, $\boldsymbol{x} \in S^{d-1}$ and let $\boldsymbol{y} = \boldsymbol{x} + \delta\boldsymbol{x}$. The set of all points illuminated by $\boldsymbol{y}$ is $S^{d-1} \cap \overline{B(\boldsymbol{x}, r)}$, where $r = \sqrt{\frac{2\delta}{1+\delta}}$.*

Proof. A point $\boldsymbol{z} \in S^{d-1}$ is illuminated by $\boldsymbol{y}$ if and only if the tangent hyperplane to S^{d-1} at $\boldsymbol{z}$ intersects the line segment $[\boldsymbol{x}, \boldsymbol{y}]$. Let $\boldsymbol{z}$ be a point such that the tangent hyperplane at $\boldsymbol{z}$ contains $\boldsymbol{y}$. We will show that $\|\boldsymbol{x} - \boldsymbol{z}\| = \sqrt{\frac{2\delta}{1+\delta}}$.

Denote by α the angle at $\mathbf{0}$ in the right triangle $\mathbf{0}$, $\boldsymbol{z}$, $\boldsymbol{y}$, as shown in Figure 2.5. Then,

$$\cos(\alpha) = \frac{\|\boldsymbol{z}\|}{\|\boldsymbol{y}\|} = \frac{1}{1+\delta}.$$

Considering the triangle $\mathbf{0}$, $\boldsymbol{x}$, $\frac{\boldsymbol{x}+\boldsymbol{z}}{2}$ we get

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\| = \sin\left(\frac{\alpha}{2}\right) = \sqrt{\frac{1 - \cos(\alpha)}{2}} = \sqrt{\frac{\delta}{2(1+\delta)}}.$$

Therefore, $\|\boldsymbol{x} - \boldsymbol{z}\| = \sqrt{\frac{2\delta}{1+\delta}}$.

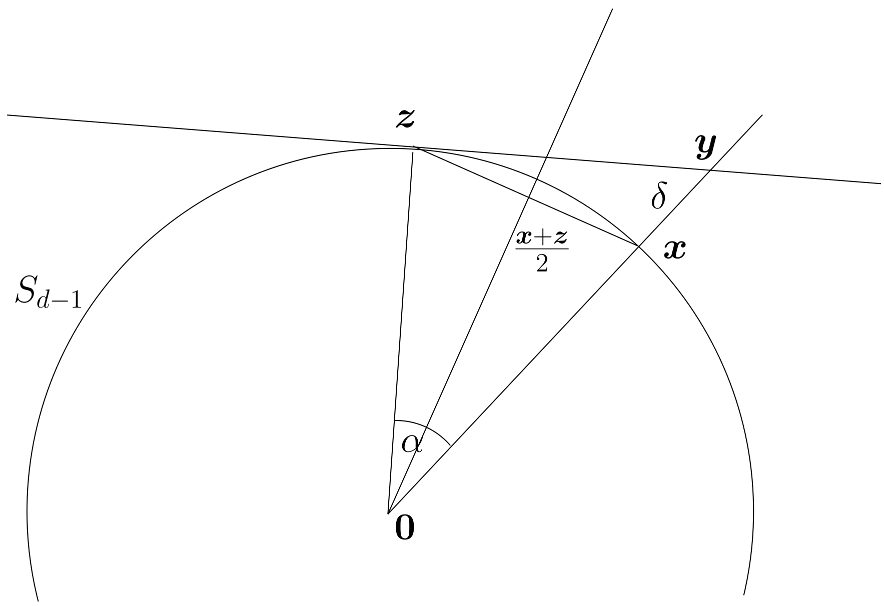

Figure 2.5: Calculation of the distance from $\boldsymbol{x}$ to $\boldsymbol{z}$.

□

The left-to-right implication of the next lemma was inspired by Lemma 4 from [5].

Lemma 42. *Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \in \mathbb{R}^d \setminus \overline{B(\mathbf{0}, 1)}$. Every point of S^{d-1} is illuminated by some $\boldsymbol{y}_i$ if and only if S^{d-1} is contained in the convex hull of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$.*

Proof. Denote by N the convex hull of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ and first suppose there is a point $\boldsymbol{z} \in S^{d-1} \setminus N$. Since N is convex, there exists a hyperplane L separating $\boldsymbol{z}$ from N. Among the points of S^{d-1} that are on the same side of N as $\boldsymbol{z}$, there exists one, $\boldsymbol{x}$, whose tangent hyperplane, $\tilde{L}$, is parallel to L. Because all the points y_i are on the same side of $\tilde{L}$ as S^{d-1}, the hyperplane does not intersect any of the line segments $[\boldsymbol{y}_i, \frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|}]$ and so the point $\boldsymbol{x}$ is not illuminated by any of them.

Conversely, suppose $\boldsymbol{x} \in S^{d-1}$ is not illuminated by any of the vertices. Then, the tangent hyperplane at $\boldsymbol{x}$ does not intersect any of the line segments $[\boldsymbol{y}_i, \frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|}]$ and so the hyperplane does not intersect the convex hull N. Therefore, $\boldsymbol{x} \notin N$. □

The following theorem is the main result of this subsection and the proof is based on the main theorem of [5].

Theorem 43. *Let $\delta \in (0, 1)$. There exists $m \in \mathbb{N}$, $m \leq \sqrt{2\pi d}\left(\frac{4(1+\delta)}{\sqrt{7\delta^2+8\delta}}\right)^{d-1}$, and a polytope P having m vertices such that $B(\mathbf{0}, 1) \subseteq P \subseteq \overline{B(\mathbf{0}, 1+\delta)}$.*

Proof. Let $\gamma = \sqrt{\frac{2\delta}{1+\delta}}$. By Lemma 40 there exists a γ-covering of S^{d-1} of $m \leq \sqrt{2\pi d}\left(\frac{8}{\gamma\sqrt{16-\gamma^2}}\right)^{d-1}$ points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. For all $i \leq m$ let $\boldsymbol{y}_i = \boldsymbol{x}_i + \delta \boldsymbol{x}_i$. By Lemma 41, all points in $S^{d-1} \cap \bigcup_{i=1}^m B(\boldsymbol{x}_i, \gamma)$ are illuminated by some $\boldsymbol{y}_i$. However, since the points form a γ-covering, $S^{d-1} \cap \bigcup_{i=1}^m B(\boldsymbol{x}_i, \gamma) = S^{d-1}$ and therefore by

Lemma 42 S^{d-1} is contained in the convex hull of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. That in turn is clearly contained in $\overline{B(\mathbf{0}, 1+\delta)}$.

By definition of γ,

$$\sqrt{2\pi d}\left(\frac{8}{\gamma\sqrt{16-\gamma^2}}\right)^{d-1} = \sqrt{2\pi d}\left(\frac{8}{\sqrt{\frac{2\delta}{1+\delta}}\sqrt{16-\frac{2\delta}{1+\delta}}}\right)^{d-1}$$
$$= \sqrt{2\pi d}\left(\frac{4(1+\delta)}{\sqrt{7\delta^2+8\delta}}\right)^{d-1}.$$

□

Remark. For $\delta \in (0,1)$, $\frac{4(1+\delta)}{\sqrt{7\delta^2+8\delta}} < \frac{2.1}{\sqrt{\delta}}$ and therefore

$$m \leq \sqrt{2\pi d}\left(\frac{2.1}{\sqrt{\delta}}\right)^{d-1}.$$

Decreasing the range of allowed values for δ, we can lower the constant to any value higher than $\sqrt{2}$.

While many results describe polytopes by the number of their vertices, for our purposes the number of facets is more useful. In the proof of the following claim we use the fact that those two characteristics are dual.

Corollary 44. For all $r_2 > r_1 > 0$, $r_2 < 2r_1$, there exists $m \in \mathbb{N}$ such that

$$m \leq \sqrt{2\pi d}\left(\frac{2.1}{\sqrt{\frac{r_2-r_1}{r_1}}}\right)^{d-1},$$

and that there is a polytope $P \in \mathbb{P}^d_m$ satisfying $B(\mathbf{0}, r_1) \subseteq P \subseteq \overline{B(\mathbf{0}, r_2)}$.

Proof. Take $\delta = \frac{r_2-r_1}{r_1}$ (then, $\delta \in (0,1)$) and apply Theorem 43 to get a polytope P having m vertices satisfying $B(\mathbf{0}, 1) \subseteq P \subseteq \overline{B(\mathbf{0}, 1+\delta)}$. Then the polar of P, $P^\circ = \{\boldsymbol{x} \in \mathbb{R}^d \mid \forall \boldsymbol{y} \in P : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq 1\}$, is a convex polytope having m facets, $P^\circ \in \mathbb{P}^d_m$, see e.g. [21, Lemma 2.4.5]. That Lemma also implies that if $B(\mathbf{0}, 1) \subseteq P \subseteq \overline{B(\mathbf{0}, 1+\delta)}$, then $B(\mathbf{0}, \frac{1}{1+\delta}) \subseteq P^\circ \subseteq \overline{B(\mathbf{0}, 1)}$. Multiplying by $r_1(1+\delta)$ we get $B(\mathbf{0}, r_1) \subseteq r_1(1+\delta)P^\circ \subseteq \overline{B(\mathbf{0}, r_2)}$. □

Remark. By simple translation, Theorem 44 can be applied to any pair of concentric balls.

For completeness, in the following lemma we construct an upper bound on approximation accuracy to complement Theorem 43. Even though the proof of the lemma is original, we created it based on the concluding remark of Bronshteyn and Ivanov in [5].

Lemma 45. *Let $P \in \mathbb{P}^d_m$ and $r > 0$ such that $B(\mathbf{0}, r) \subseteq P \subseteq \overline{B(\mathbf{0}, 1)}$. Then,*

$$r < 1 - \frac{1}{2m^{\frac{2}{d-1}}}.$$

Proof. First, let P be a polytope having m vertices such that $B(\mathbf{0},1) \subseteq P \subseteq \overline{B(\mathbf{0},1+\delta)}$ for some $\delta > 0$. Take a vertex $\boldsymbol{y}$ of P and denote $\boldsymbol{x} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$. By Lemma 41, the set of points illuminated by $\boldsymbol{y}$ is a subset of $S^{d-1} \cap \overline{B(\boldsymbol{x},r)}$, where $r = \sqrt{\frac{2\delta}{1+\delta}}$. The surface area of $S^{d-1} \cap \overline{B(\boldsymbol{x},r)}$ is less than the surface area of the sphere $\partial B(\boldsymbol{x},r)$. By Lemma 42, any point on S^{d-1} is illuminated by some vertex. Altogether,

$$\mathcal{S}(S^{d-1}) < m\mathcal{S}(\partial B(\boldsymbol{x},r)),$$

which means

$$m > \frac{\mathcal{S}(S^{d-1})}{\mathcal{S}(\partial B(\boldsymbol{x},r))} = \frac{1}{r^{d-1}} = \left(\frac{1}{2} + \frac{1}{2\delta}\right)^{\frac{d-1}{2}},$$

or in other words,

$$\delta > \frac{1}{2m^{\frac{2}{d-1}} - 1}. \tag{2.10}$$

Now let $P \in P_m^d$ and $r > 0$ such that $B(\mathbf{0},r) \subseteq P \subseteq \overline{B(\mathbf{0},1)}$. Then, taking the polar $P^\circ = \{\boldsymbol{x} \in \mathbb{R}^d \mid \forall \boldsymbol{y} \in P : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 1\}$ as in the proof of Corollary 44, we have $B(\mathbf{0},1) \subseteq P \subseteq \overline{B(\mathbf{0},\frac{1}{r})}$. By (2.10), this implies that

$$\frac{1}{r} - 1 > \frac{1}{2m^{\frac{2}{d-1}} - 1},$$

which yields

$$r < 1 - \frac{1}{2m^{\frac{2}{d-1}}}.$$

□

Similarly to Corollary 44, we can reformulate this in the following way:

Corollary 46. Let $r_2 > r_1 > 0$ and let $P \in \mathbb{P}_m^d$ be a polytope satisfying $B(\mathbf{0},r_1) \subseteq P \subseteq \overline{B(\mathbf{0},r_2)}$. Then,

$$m > \left(\frac{1}{\sqrt{2\frac{r_2-r_1}{r_2}}}\right)^{d-1}.$$

Proof. Take $P' = \frac{1}{r_2}P$. Then, $B\left(\mathbf{0}, \frac{r_1}{r_2}\right) \subseteq P \subseteq \overline{B(\mathbf{0},1)}$ and by Lemma 45

$$\frac{r_1}{r_2} < 1 - \frac{1}{2m^{\frac{2}{d-1}}}.$$

In other words,

$$m > \left(\frac{1}{\sqrt{2\frac{r_2-r_1}{r_2}}}\right)^{d-1}.$$

□

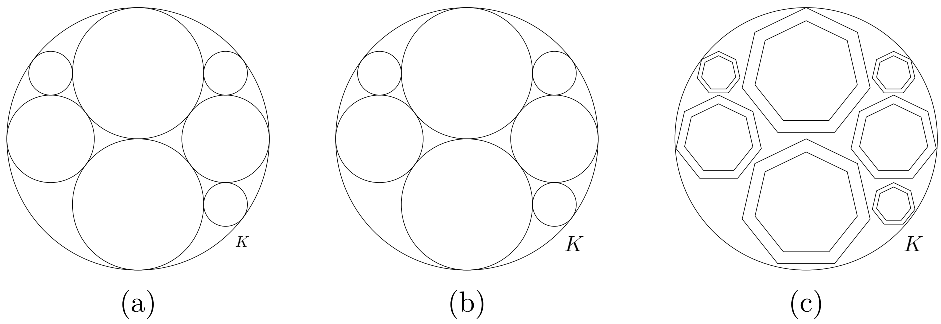

Figure 2.6: (a) Based on the proof the Vitali Covering Theorem, we can construct n disjoint balls (here $n = 7$) that cover the set K except for a set of small size. (b) We approximate each ball by a pair of concentric polytopes and for each we use Theorem 27 to construct a function small outside the outer polytope and close to 1 inside the inner polytope. (c) A weighted sum of these functions approximates the target function on the union of the inner polytopes. This set, denoted A_k, is shown in grey and we can increase its size either by increasing the number of balls n, or by improving the approximation of the balls by polytopes, expressed by k.

2.3.2 Sketch of a Bound on Almost-Everywhere Approximation

At this point, we have almost everything we need to formulate a bound on the number of neurons required to approximate a function everywhere except on a small set. However, we would first need a bound on the number of balls required in the Vitali Covering Theorem (Theorem 39).

Definition 22. *Let $K \subseteq \mathbb{R}^d$ be compact and let $R > 0$ and $\delta > 0$. Denote by $n_K(R, \delta)$ the least number of disjoint balls of radius at most R required to cover K up to a set of measure less than δ,*

$$n_K(R,\delta) = \min \Big\{ n \in \mathbb{N} \;\Big|\; \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d \ \exists r_1, \ldots, r_n \in (0, R] : \\ \lambda^d \left(K \setminus \bigcup_{i=1}^{n} B(\boldsymbol{x}_i, r_i) \right) < \delta \ \&\ \forall i \neq j \leq n : B(\boldsymbol{x}_i, r_i) \cap B(\boldsymbol{x}_j, r_j) = \emptyset \Big\}.$$

While the Vitali Covering Theorem implies that $n_K(R, \delta)$ is finite for all K, R and δ, we have not found any bounds on the number in the literature and it is beyond the scope of this thesis to create one. To be more precise, we need a lower bound on the radius of the smallest ball, which can then be used to get a bound on the number of balls. We proved the following theorem, into which such a bound can be inserted. The construction is illustrated in Figure 2.6.

Theorem 47. *Let $K \subseteq \mathbb{R}^d$, let $f : K \to \mathbb{R}$ be continuous, let $\varepsilon > 0$ and let $\delta > 0$.*

Let $B(\boldsymbol{x}_1, r_1), \ldots, B(\boldsymbol{x}_n, r_n)$ be disjoint balls such that $r_i \leq \frac{1}{2}\omega^{-1}(f, \varepsilon)$ for all i and that $\lambda^d \left(K \setminus \bigcup_{i=1}^{n} B(\boldsymbol{x}_i, r_i)\right) < \delta$. Denote $r = \min_{i \leq n} r_i$.

Then, for every $k \in \mathbb{N}$ there exists $A_k \subseteq K$ and $g_k \in G_{\exp, h_k}$, where

$$h_k = n \left(\frac{4n \left\| \widetilde{f} \right\|_\infty}{\varepsilon} \right)^{14\pi d (5dk)^d \frac{diam(K)}{r}}$$

such that

$$\lambda^d (K \setminus A_k) < \delta + \frac{2}{k}\lambda^d \left(\bigcup_{i=1}^n B(\boldsymbol{x}_i, r_i)\right)$$

and for all $\boldsymbol{y} \in A_k$*:* $|f(\boldsymbol{y}) - g_k(\boldsymbol{y})| < \varepsilon$.

Proof. Assume $k \geq 2$ — otherwise the claim is trivial — and take $\widetilde{f} = f + \min_{\boldsymbol{y}\in K} f(\boldsymbol{y})$, so that $\min_{\boldsymbol{y}\in K} \widetilde{f}(\boldsymbol{y}) = 0$ and $\max_{\boldsymbol{y}\in K} \widetilde{f}(\boldsymbol{y}) = \left\|\widetilde{f}\right\|_\infty \leq 2\|f\|_\infty$.

Define $R_1 = \sqrt[d]{1 - \frac{2}{k}}$ and $R_2 = 1$. Then, the balls $B(\mathbf{0}, R_1)$ and $B(\mathbf{0}, R_2)$ satisfy

$$\lambda^d (B(\mathbf{0}, R_2) \setminus B(\mathbf{0}, R_1)) \leq \frac{2}{k}\lambda^d(B(\mathbf{0}, R_2)). \tag{2.11}$$

Let $R_3 = \sqrt[d]{1 - \frac{1}{k}}$. By Corollary 44, there exists a polytope $P \in \mathbb{P}^d_m$ such that $B(\mathbf{0}, R_1) \subseteq P \subseteq \overline{B(\mathbf{0}, R_3)}$, where

$$m \leq \sqrt{2\pi d}\left(\frac{2.1}{\sqrt{\frac{R_3 - R_1}{R_1}}}\right)^{d-1} = \sqrt{2\pi d}\left(\frac{4.5}{\sqrt[d]{\frac{k-1}{k-2}} - 1}\right)^{\frac{d-1}{2}}.$$

Consequently, $B(\mathbf{0}, R_1) \subseteq P \subseteq \frac{1}{R_3}P \subseteq \overline{B(\mathbf{0}, R_2)}$.

We assume that for all $i \leq n$: $K \cap B(\boldsymbol{x}_i, r_i) \neq \emptyset$ — otherwise, omitting unnecessary balls will only decrease the final bound. For each $i \leq n$ define

$$P_i = r_i P + \boldsymbol{x}_i$$

and let $A_k = K \cap \bigcup_{i=1}^n P_i$. We have $B(\boldsymbol{x}_i, R_1 r_i) \subseteq P_i \subseteq \frac{1}{R_3} *_{\boldsymbol{x}_i} P \subseteq \overline{B(\mathbf{0}, r_i)}$ and by (2.11):

$$\begin{aligned}
\lambda^d(K \setminus A_k) &\leq \lambda^d\left(K \setminus \bigcup_{i=1}^n B(\boldsymbol{x}_i, r_i)\right) + \lambda^d\left(\bigcup_{i=1}^n B(\boldsymbol{x}_i, r_i) \setminus A_k\right) \\
&< \delta + \sum_{i=1}^n \lambda^d (B(\boldsymbol{x}_i, r_i) \setminus P_i) \\
&\leq \delta + \sum_{i=1}^n \lambda^d (B(\boldsymbol{x}_i, r_i) \setminus B(\boldsymbol{x}_i, R_1 r_i)) \\
&\leq \delta + \sum_{i=1}^n \frac{2}{k}\lambda^d (B(\boldsymbol{x}_i, r_i)) \\
&= \delta + \frac{2}{k}\lambda^d\left(\bigcup_{i=1}^n B(\boldsymbol{x}_i, r_i)\right).
\end{aligned}$$

Define for each $i \leq n$

$$v_i = \frac{1}{2}\left(\max_{\boldsymbol{y}\in B(\boldsymbol{x}_i, r_i)} \widetilde{f}(\boldsymbol{y}) + \min_{\boldsymbol{y}\in B(\boldsymbol{x}_i, r_i)} \widetilde{f}(\boldsymbol{y})\right).$$

Then, as $r_i \leq \frac{1}{2}\omega^{-1}\left(\widetilde{f}, \varepsilon\right)$, for all $\boldsymbol{y}_1, \boldsymbol{y}_2 \in K \cap B(\boldsymbol{x}_i, r_i)$: $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \omega^{-1}\left(\widetilde{f}, \varepsilon\right)$, so $\left|\widetilde{f}(\boldsymbol{y}_1) - \widetilde{f}(\boldsymbol{y}_2)\right| \leq \varepsilon$. Therefore, $\left|\widetilde{f}(\boldsymbol{y}_1) - v_i\right| \leq \frac{\varepsilon}{2}$.

By Theorem 27 for each $i \leq n$ there exists $g_i \in G_{\exp}$ such that $g_i|_K : K \to [0,1]$, for $\boldsymbol{y} \in P_i \cap K$: $g_i(\boldsymbol{y}) > 1 - \frac{\varepsilon}{2n\|\widetilde{f}\|_\infty}$ and for $\boldsymbol{y} \in K \setminus \left(\frac{1}{R_3} *_{\boldsymbol{x}_i} P_i\right)$: $g_i(\boldsymbol{y}) < \frac{\varepsilon}{2n\|\widetilde{f}\|_\infty}$. Define

$$\widetilde{g}(\boldsymbol{y}) = \sum_{i=1}^{n} v_i g_i(\boldsymbol{y})$$

Then, for $\boldsymbol{y} \in K \cap P_i$:

$$\widetilde{g}(\boldsymbol{y}) > v_i \left(1 - \frac{\varepsilon}{2n \left\|\widetilde{f}\right\|_\infty}\right) \geq v_i - \frac{\varepsilon}{2}$$

and also, since $\boldsymbol{y} \in K \setminus \left(\frac{1}{R_3} *_{\boldsymbol{x}_j} P_j\right)$ for all $j \neq i$,

$$\widetilde{g}(\boldsymbol{y}) < v_i + \sum_{j \neq i} v_j \frac{\varepsilon}{2n \left\|\widetilde{f}\right\|_\infty} \leq v_i + n\frac{\varepsilon}{2n} \leq v_i + \frac{\varepsilon}{2},$$

which together means $\left|\widetilde{f}(\boldsymbol{y}) - \widetilde{g}(\boldsymbol{y})\right| \leq \left|\widetilde{f}(\boldsymbol{y}) - v_i\right| + |v_i - \widetilde{g}(\boldsymbol{y})| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. So defining $g = \widetilde{g} - \max_{\boldsymbol{y} \in K} f(\boldsymbol{y})$, for all $\boldsymbol{y} \in A_k$ we have $|f(\boldsymbol{y}) - g(\boldsymbol{y})| < \varepsilon$.

Lastly, we bound the number of neurons. The remark after Theorem 27 gives us a bound for one g_i of

$$\left(\frac{3e}{m}\right)^m \left(\frac{4n\left\|\widetilde{f}\right\|_\infty}{\varepsilon}\right)^{6m\log_2 m \frac{\operatorname{diam}(K)}{\left(\frac{1}{R_3}-1\right)q_i}} \leq \left(\frac{4n\left\|\widetilde{f}\right\|_\infty}{\varepsilon}\right)^{7m^2 \frac{\operatorname{diam}(K)}{\left(\frac{1}{R_3}-1\right)q_i}},$$

where q_i is the smallest distance from $\boldsymbol{x}_i$ to a facet of P_i. Define $r = \min_{i \leq n} r_i$. Then, for all i: $q_i \geq r$. Using the Laurent series at ∞, it can be shown for all $t \geq 0$ and $d \in \mathbb{N}$ that

$$\frac{\sqrt[d]{t}}{\sqrt[d]{t+1} - \sqrt[d]{t}} = \frac{1}{\sqrt[d]{1 + \frac{1}{t}} - 1} \in \left[dt, dt + \frac{d-1}{2}\right).$$

Therefore, we have

$$\begin{aligned} m^2 &\leq 2\pi d \left(\frac{4.5\sqrt[d]{k-2}}{\sqrt[d]{k-1} - \sqrt[d]{k-2}}\right)^{d-1} \\ &\leq 2\pi d \left(4.5\left(d(k-2) + \frac{d-1}{2}\right)\right)^{d-1} \\ &\leq 2\pi d\,(5dk)^{d-1}, \end{aligned}$$

and also

$$\left(\frac{1}{R_3} - 1\right) = \frac{\sqrt[d]{k} - \sqrt[d]{k-1}}{\sqrt[d]{k-1}} \geq \frac{1}{d(k-1) + \frac{d-1}{2}} \geq \frac{1}{dk}.$$

Together, $g \in G_{\exp,h}$ for

$$h \leq n \left(\frac{4n\left\|\widetilde{f}\right\|_\infty}{\varepsilon}\right)^{14\pi d (5dk)^d \frac{\operatorname{diam}(K)}{r}}.$$

$\square$

Conclusion

Neural networks abound both in science and in everyday life and there has been an unceasing, rapid increase in their use for many years now. In practical applications of neural networks, specific bounds on the necessary number of neurons could prove invaluable for assessing resource requirements. Currently, such deliberations are mostly based on heuristics and trial-and-error methods, but a better understanding of the theory could lead to a significant increase in efficiency. Existing bounds do not easily lend themselves to such uses. The aim of this thesis was to bring theoretical results a step closer to practical applicability.

We provided a bound that applies to any function f continuous on a compact set K. Given a desired accuracy ε, we have shown the minimal number of neurons h required to approximate f satisfies

$$h \leq \left(6\sqrt{d+1}\frac{\operatorname{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{10(d+1)^4\frac{\operatorname{diam}(K)}{\delta}}, \tag{2.12}$$

or alternatively

$$h \leq \left(8\sqrt{d}\frac{\operatorname{diam}(K)\,\|f\|_\infty}{\delta\varepsilon^2}\right)^{60d^3\log_2(d)\frac{\operatorname{diam}(K)}{\delta}},$$

where δ is the inverse modulus of continuity at $\frac{\varepsilon}{2}$.

Previous bounds of this nature are only applicable to specific sets K such as balls and cubes and they require the approximated function to have continuous derivatives of order up to k and Lipschitz continuous derivatives of order k. Under such conditions, the bound by Mhaskar [2] gives

$$h \leq C\left(\frac{1}{\varepsilon}\right)^{\frac{d}{k+1}}. \tag{2.13}$$

Here, C is independent of ε, but it is allowed to depend on the dimension d, as well as the size of K.

Focusing on continuous functions, our bound is strictly more general. The bound (2.13) can however also be applied to some non-continuous functions.

Given a function $f : \overline{B(\mathbf{0},1)} \to \mathbb{R}$ such that $\|f\|_{W^{1,\infty}} \leq 1$, we have $\|f\|_\infty \leq 1$ and f is Lipschitz continuous with a Lipschitz constant that is at most 1. Then, $\delta \geq \varepsilon$ and (2.12) reduces to

$$h \leq \left(6\sqrt{d+1}\frac{2}{\varepsilon}\right)^{20(d+1)^4\frac{1}{\varepsilon}} = \widetilde{C}\left(\frac{1}{\varepsilon}\right)^{60(d+1)^4\frac{1}{\varepsilon}}.$$

In comparison, (2.13) gives

$$h \leq C\left(\frac{1}{\varepsilon}\right)^{d}$$

However, the results of this thesis are only the first step in this direction and several ways of improving the bound could be considered. One is simply a more careful numerical analysis of the number of neurons. While this would probably improve the multiplicative constant and the number in the exponent, some ingredients of the bound cannot be reduced without changing the method: In the exponent, we can get no less than $d^2\sqrt{d}\log d$, as one d is given by exponentiation and one comes from the number of lattice points inside a sphere, the $\sqrt{d}$ results from the fact the exponent contains 1 over the packing radius of the lattice — For the lattices we have seen, this is at least $\frac{\sqrt{d}}{\delta}$. The $\log d$ corresponds to the logarithm of the number of facets of one Voronoi cell, which is at least $d+1$.

As for our choice of lattice, we focused mainly on the number of lattice points in a sphere. However, the number of facets is also an important factor, as is the packing radius. From (2.12) and (2.3.2) we can see that the choice of lattice did not have a great impact on the bound. We might be able to get a lattice that performs better asymptotically in d, but for low dimensions A_d^* is the best lattice known. Therefore, it would be probably better to focus on the simple cubic lattice $\mathbb{Z}^d$, for which we can make some further optimizations that might help surpass (2.12).

Another topic for further research is transfer of the bound to sigmoidal activations. The direct approach from Theorem 38 could probably be improved to some extent, but we would like to consider a different approach. Notice that nowhere in the construction itself we use the fact that we are working with exponential functions — The proofs would work only with slight modifications for any transition function, even though the result would not correspond to a neural network, but to a linear combination of products of the function. If we started with linear functions, the result would then be a polynomial and we could use previous results on approximation of polynomials by sigmoidal networks.

Alternatively, the important property of the exponential functions we start with in the proof is that each is large on some half-space and small on almost the complement. A product of two such exponentials is then a function large on some other half-space whose normal vector is sum of the original two. If these properties in the final function are sufficient to guarantee approximation, we might be able to replace each exponential in the result by one sigmoidal function that is large on the corresponding half-space.

Also, a clear bottleneck of the method is the approximation of one polytope, from which comes the $\frac{\operatorname{diam}(K)}{\delta}$ in the exponent. It might be worthwhile to consider improvements of this proof, since the current bound for one polytope might be far larger than necessary.

Many other possible alternatives could be investigated, especially when it comes to the method of approximating the slice sets of the function. For example for concave functions, these level sets are convex, so each could be approximated by a single polytope.

Eventually, since our bound is constructive and the proofs are more or less elementary, the method could in theory be used to construct the actual approximating network. Should the bound be improved, this might also be worth exploring further.

Bibliography

[1] V. Maiorov. On Best Approximation by Ridge Functions. *Journal of Approximation Theory*, 99(1):68–94, 1999.

[2] H. Mhaskar. Neural Networks for Optimal Approximation of Smooth and Analytic Functions. *Neural computation*, 8(1):164–177, 1996.

[3] B. Brosowski and F. Deutsch. An Elementary Proof of the Stone-Weierstrass Theorem. *Proceedings of the American Mathematical Society*, pages 89–92, 1981.

[4] R. M. Dudley. Metric Entropy of Some Classes of Sets with Differentiable Boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974.

[5] E. M. Bronshteyn and L. D. Ivanov. The Approximation of Convex Sets by Polyhedra. *Siberian Mathematical Journal*, 16(5):852–853, 1975.

[6] G. Cybenko. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[7] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural networks*, 2(5):359–366, 1989.

[8] K. Funahashi. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural networks*, 2(3):183–192, 1989.

[9] T. Chen, H. Chen, and R. Liu. A Constructive Proof and an Extension of Cybenko's Approximation Theorem. In *Computing science and statistics*, pages 163–168. Springer, 1992.

[10] W. Rudin et al. *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York, 1976.

[11] Y. Xu, W. A. Light, and E. W. Cheney. Constructive Methods of Approximation by Ridge Functions and Radial Functions. *Numerical Algorithms*, 4(2):205–223, 1993.

[12] W. Rudin. *Functional Analysis*. McGraw-Hill Science, Engineering & Mathematics, 1991.

[13] M. Leshno, V. Lin, A. Pinkus, and S. Schocken. Multilayer Feedforward Networks with a Nonpolynomial Activation Function can Approximate Any Function. *Neural networks*, 6(6):861–867, 1993.

[14] D. Yarotsky. Error Bounds for Approximations with Deep ReLU Networks. *Neural Networks*, 94:103–114, 2017.

[15] V. Kůrková. Limitations of Shallow Networks Representing Finite Mappings. *Neural Computing and Applications*, 31(6):1783–1792, 2019.

[16] A. Pinkus. Approximation Theory of the MLP Model in Neural Networks. *Acta numerica*, 8:143–195, 1999.

[17] V. Maiorov, R. Meir, and J. Ratsaby. On the Approximation of Functional Classes Equipped with a Uniform Measure Using Ridge Functions. *Journal of approximation theory*, 99(1):95–111, 1999.

[18] V. Maiorov and A. Pinkus. Lower Bounds for Approximation by MLP Neural Networks. *Neurocomputing*, 25(1-3):81–91, 1999.

[19] R. DeVore, R. Howard, and C. Micchelli. Optimal Nonlinear Approximation. *Manuscripta mathematica*, 63(4):469–478, 1989.

[20] P. Petrushev. Approximation by Ridge Functions and Neural Networks. *SIAM Journal on Mathematical Analysis*, 30(1):155–189, 1998.

[21] R. Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Number 151. Cambridge university press, 2014.

[22] W. Feller. Xii. In *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1950.

[23] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*, volume 290. Springer Science & Business Media, 2013.

[24] R. Kershner. The Number of Circles Covering a Set. *American Journal of mathematics*, 61(3):665–671, 1939.

[25] J. Baek and A. Adams. Some Useful Properties of the Permutohedral Lattice for Gaussian Filtering. *other words*, 10(1):0, 2009.

[26] H. Davenport. The Covering of Space by Spheres. *Rendiconti del Circolo Matematico di Palermo*, 1(1):92–107, 1952.

[27] T. C. Hales. The Honeycomb Conjecture. *Discrete & Computational Geometry*, 25(1):1–22, 2001.

[28] T. C. Hales. A Proof of the Kepler Conjecture. *Annals of mathematics*, 162(3):1065–1185, 2005.

[29] M. S. Viazovska. The Sphere Packing Problem in Dimension 8. *Annals of Mathematics*, pages 991–1015, 2017.

[30] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. Viazovska. The Sphere Packing Problem in Dimension 24. *Annals of Mathematics*, pages 1017–1033, 2017.

[31] E. M. Stein and R. Shakarchi. *Real Analysis.* Princeton University Press, 2009.

List of Figures