



KATHOLISCHE UNIVERSITÄT
EICHSTÄTT-INGOLSTADT



MIDS

Lipschitz and Approximation Bounds for Neural Networks

by

Paul Geuchen

A cumulative dissertation submitted to the
Faculty of Mathematics and Geography / Department of Mathematics
in partial fulfillment of the requirements
for the degree of
Doctor of Mathematics (Dr. rer. nat.)

Catholic University of Eichstätt-Ingolstadt

Eichstätt, 2026

Defended on 20 October 2025

Reviewer: Prof. Dr. Felix Voigtlaender

Co-Reviewer: Prof. Dr. Helmut Bölcskei

Contents

1. Introduction	7
1.1. Preliminaries	9
1.1.1. Notation	9
1.1.2. Fully-connected feedforward neural networks	11
1.1.3. A mathematical framework of supervised machine learning	13
2. The Lipschitz constant of random neural networks	15
2.1. The Lipschitz constant of a neural network as a measure of adversarial robustness	15
2.2. Setting and preliminaries	18
2.3. Upper and lower bounds for the Lipschitz constant of random neural networks . .	21
2.3.1. Upper bounds	21
2.3.2. The lower bound	24
2.4. Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks	26
2.4.1. Pointwise estimates	26
2.4.2. The upper bound in the case of zero-bias networks	27
2.4.3. The lower bound in the case of zero-bias networks	28
2.4.4. Bounds in the case of networks with biases	29
3. Approximation properties of complex-valued neural networks	33
3.1. Complex-valued neural networks	33
3.2. Approximation theory of neural networks	34
3.3. Optimal approximation using complex-valued neural networks	36
3.4. On best approximation by multivariate ridge functions	41
3.5. Universal approximation with complex-valued deep narrow neural networks	44
Bibliography	47
A. Additional work regarding Chapter 2	55
A.1. Proof of Theorem 2.3.2	55
A.2. Generalization of [36] to other initializations	58
A.2.1. Upper bound	58
A.2.2. Lower bound	63
B. Full texts of the papers	72
B.1. Upper and lower bounds for the Lipschitz constant of random neural networks . .	72
B.2. Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks	133
B.3. Optimal approximation using complex-valued neural networks	239
B.4. On best approximation by multivariate ridge functions	297
B.5. Universal approximation using complex-valued deep narrow neural networks	369

Acknowledgments

First and foremost, I would like to express my greatest gratitude to my supervisor Felix Voigtlaender. His mathematical brilliance, constant support, and positive attitude were the main reasons why my PhD was such an enjoyable time for me. I am also very grateful to Sjoerd Dirksen, who supervised me during my two-months research stay at Utrecht University in 2024. Despite his busy schedule, he always made time for discussions, which has led to a fruitful collaboration and which made the stay in Utrecht one of the highlights of my PhD. Moreover, I would like to thank his PhD student, Patrick Finke, for the great time we spent together in Utrecht and Aachen. I really enjoyed working with him and appreciated how well we got along. I would also like to acknowledge Olov Schavemaker and Palina Salanevich—my two other collaborators from Utrecht University.

I would then like to thank all the members of the Mathematical Institute for Machine Learning and Data Science who have accompanied me during the last years. Especially, I would like to mention Thomas Jahn and Hannes Matt—my working group colleagues—as well as Dominik Stöger and Andrei Caragea. I am particularly indebted to Andrei for proofreading this thesis. I would also like to thank the secretaries at MIDS for always being open to a nice chat, bringing some welcome variety into the workday. In this regard, I would also like to thank Johannes Gückel, Simon Mader, and Merve İbiş Bozyel for all the great lunch and coffee breaks we shared.

Apart from the people that have directly accompanied me as colleagues over the past years, there are many more people in my life that are very important to me and that were with me during my academic journey. There isn't enough space here to name them all, but I am thinking of all the great friends I made during my studies in Aachen. Especially I would like to mention David Bartusel and Timm Gilles who also proofread this thesis. I am also thinking of Milena and her unwavering support over the past three years. She has spent countless hours in trains just to be with me and her trust in my skills and abilities has played a major role in the successful completion of this thesis.

But most importantly, I am thinking of my family and my parents in particular. The sort of support that I receive from my family is really the most unconditional support you could possibly wish for and I know they will always be by my side, no matter what happens.

Financial support

For the entire duration of his PhD, the author was supported by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1.

Preface

This thesis is a cumulative thesis based on five publications. While all of these publications deal with the mathematical foundations of machine learning, they focus on two different sub-areas. The main part of this thesis contains a general introduction (see Chapter 1), and then discusses the five publications, divided into the two sub-areas (see Chapters 2 and 3). Both Chapters 2 and 3 have a similar structure, starting with a brief introduction to the sub-area and then proceeding by discussing the individual publications. The appendix contains a postponed proof, some additional work done by the author regarding Section 2.3 (see Chapter A), and the full texts of the five papers (see Chapter B). The five publications on which this thesis is based are cited using the same numeric style as the other references. However, to make them stand out, their citation numbers are shown in bold.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of Chapters 1 to 3, the author used ChatGPT v.4.0 (OpenAI, San Francisco, USA) and DeepL Translator (DeepL SE, Cologne, Germany) for improvement of readability, grammar check, and spelling check. After usage, the author reviewed and edited the content as needed and takes full responsibility for the content of this thesis.

1. Introduction

Over the past decade, machine learning—and deep learning in particular—has significantly influenced both scientific research and everyday technology. Its applications range from imaging tasks such as image classification [48] or object detection [38], as well as game-playing agents [59] and speech recognition [42] to, most notably in recent years, natural language processing [86]. The development of easily accessible and widely applicable large language models such as ChatGPT [9] and Google Gemini [32] has ensured that deep learning is no longer relegated to academic research, but is a part of everyday life for the general public.

The term *machine learning* refers to the “study of algorithms that can learn from experience” [96, p. 2]. More precisely, it describes the general approach of tackling a task not by implementing explicit instructions (such as in traditional programming), but by implementing a model that can learn from given data and generalize to unseen data. This process of learning from data, the so called *training*, is, from the mathematical perspective, an optimization process, where the parameters of the underlying model are adapted in order to minimize a given objective function. Popular machine learning models, aside from those used in deep learning, are linear regression, logistic regression, support vector machines, decision trees, and random forests.

The term *deep learning* refers to the “subset of machine learning concerned with models based on many-layered neural networks” [96, p. 27]. The fundamental prototype of these models is the *fully-connected feedforward (artificial) neural network* (FNN), historically known as the *multilayer perceptron*, a model which consists of sequential compositions of affine transformations, each followed by a nonlinear activation function. Notably, foundational research on this model by pioneers such as McCulloch and Pitts [57] and Rosenblatt [72], motivated by modeling the neural activity in the human brain, can be traced back to the mid-20th century. Although modern deep learning architectures such as convolutional neural networks [50] and transformers [86] have evolved quite a bit from the classic FNN model, they still are fundamentally composed of variants and extensions of the classical FNN building blocks.

Another typical element of deep learning is the fact that the training of the models, i.e., the minimization of the given objective function, is usually done via gradient-based methods, i.e., variants of gradient descent. This practice is motivated not only by the empirical success of such methods in finding solutions that generalize well to unseen data, but also by the fact that the compositional structure of deep learning models allows for an efficient computation of the gradients using the chain rule. This computational procedure is commonly referred to as *backpropagation* [73].

The overwhelming success and significance of deep learning in practical applications has consequently led to an increased interest in investigating the *theoretical* properties of the models and methods used in this field. On the one hand, a solid theoretical understanding of deep learning models is essential in order to guarantee the reliability of these models, which is particularly important in security-relevant applications; on the other hand, this interest is based on pure scientific curiosity. Remarkably, the success of deep learning is not entirely explainable using classical mathematical theory. For example, the minimized objective function in deep learning is typically non-convex, so it is not clear why gradient-based methods can successfully minimize it at all, which has led to a variety of publications addressing that question [1, 26, 45]. Moreover, empirical work has shown that in particular heavily overparameterized models have the tendency to generalize well to unseen data, contradicting classical theory which might suggest that such models *overfit*,

i.e., adapt too strongly to the training data and consequently fail to produce reliable predictions on unseen data. Apparently, gradient-based methods have a tendency to converge towards solutions with good generalization properties, a phenomenon now widely studied under the term *implicit bias* [15, 43, 76].

These two previously described aspects are only examples for the questions and challenges that deep learning raises to the mathematical community. The present thesis is concerned with two other topics, namely *Lipschitz continuity* (see Chapter 2) and *approximation theory* (see Chapter 3). Here, we give a short overview of these two areas and refer to Sections 2.1 and 3.2 for more detailed introductions.

The *Lipschitz constant* of a machine learning model quantifies how sensitively the model’s output responds to changes in its input; specifically, it bounds how far apart two outputs can be, relative to the distance between their corresponding inputs. The interest in the Lipschitz constant of deep learning models is based on several aspects, one of which is *adversarial robustness*. Empirical work has shown that trained neural networks that on the one hand have good generalization properties might on the other hand actually be very susceptible to adversarial attacks [39, 79]. This means that one can generate small changes to an input, in the case of vision tasks even changes so subtle as to be unrecognizable to the human eye, such that the perturbed input is classified incorrectly by the neural network. Since this phenomenon raises questions regarding the reliability of neural networks, it has become the focus of extensive research. This research can be broadly divided into two main directions: one focuses on improving the robustness of neural networks against adversarial attacks, often by adapting the training process [25, 54]; the other is concerned with developing increasingly effective adversarial attacks designed to mislead neural networks [3, 61]. In addition to the variety of practical works addressing adversarial robustness, there is also a significant amount of theoretical publications examining this issue [6, 31, 84]. The Lipschitz constant may be viewed as a measure of *worst-case* adversarial robustness; this means in particular that a small Lipschitz constant is indicative of the non-existence of adversarial examples in the entire input space.

In the present thesis, we focus on neural networks with ReLU activation function (see Definition 2.2.1) and with *random* parameters that are drawn according to a generalization of the popular He initialization [44], and prove upper and lower bounds for their Lipschitz constants that hold with high probability. Since it is typical to initialize the parameters of a network at random before training it, the study of neural networks with random parameters is often an important starting point towards the theoretical explanation of empirically observed phenomena in deep learning. In particular, we expect that some of the techniques used in the present thesis for networks with random parameters can be lifted to the study of trained networks.

Approximation theory is, as the term already suggests, a purely theoretical concept. Roughly speaking, it deals with the question of how well a given function class $\mathcal{F}_1$ can be approximated by elements of another function class $\mathcal{F}_2$. In the context of deep learning, the class $\mathcal{F}_2$ typically consists of neural networks of a given architecture and the class $\mathcal{F}_1$ is a classical function class such as the class of continuous functions or Sobolev functions. Approximation theory mostly deals with the question whether an approximating function exists and not necessarily with the question how such a function can be found. This means that, in the context of deep learning, approximation theory does *not* study the training of neural networks but only answers the question if there theoretically exists a neural network that is able to approximate a given function sufficiently well. Notably, in approximation theory one is not only interested in positive results, but also in negative ones, i.e., proofs for the fact that a given function from $\mathcal{F}_1$ can *not* be approximated below some error threshold by elements of $\mathcal{F}_2$. Good approximation-theoretic properties of deep learning models may be viewed as a *necessary* (but by no means sufficient) condition for deep learning to work well in practical applications.

The study of the approximation-theoretic properties of neural networks dates back to the early 1990s [5, 19, 52]. However, the success of deep learning in recent years has reignited the interest in

this topic [68, 75, 95]. In the present thesis, we put a specific emphasis on the approximation properties of *complex-valued* neural networks, which are defined similarly to classical neural networks, with the exception that the weights and biases of such networks are allowed to be complex-valued and the activation functions are maps from $\mathbb{C}$ to $\mathbb{C}$; see Section 3.1 for an introduction. We prove various novel results regarding the approximation properties of such networks. More specifically, we investigate the approximation of Sobolev functions by shallow complex-valued neural networks and establish optimal approximation rates both in the case of continuous weight selection (see Section 3.3) and in the case of unrestricted weight selection (see Section 3.4). Moreover, we analyze the universality of complex-valued deep narrow neural networks, i.e., the possibility to approximate arbitrary continuous functions on compact sets by sets consisting of complex-valued neural networks with an arbitrary number of hidden layers but restricted width (see Section 3.5).

In both Chapters 2 and 3, we focus on the study of classical fully-connected feedforward neural networks. At first glance, this may seem restrictive, especially regarding the fact that the models that are popular nowadays are based on significant refinements of classical FNN architectures. However, modern models typically retain fundamental components of classical FNNs as building blocks. Therefore, developing a deeper theoretical understanding of these foundational models is a crucial step towards the analysis of more complex architectures.

1.1. Preliminaries

1.1.1. Notation

We start by fixing some notation that is used in Chapters 1 to 3. We note that the notation introduced here might partly differ from those used in the papers included in Chapter B (see the notation sections of the individual papers for in situ conventions).

General

- For a mathematical statement A , we let

$$\mathbb{1}_A := \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{if } A \text{ is false.} \end{cases}$$

- The letters $\mathbb{R}$ and $\mathbb{C}$ denote the set of real and complex numbers, respectively. By $\mathbb{N}$ we denote the set of integers strictly greater than zero and we let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.
- For two quantities A and B , we write $A \lesssim B$ if there exists an absolute constant $C > 0$ with $A \leq C \cdot B$. We write $A \asymp B$ if both $A \lesssim B$ and $B \lesssim A$. We write $A \lesssim_\alpha B$ if the implied constant may depend on some parameter α .
- For $x \in \mathbb{R}$, we let $x_+ = \max\{0, x\}$.
- We let $|K|$ denote the cardinality of a set K .
- For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$ their inner product. For $p \in [1, \infty)$, we denote by $\|x\|_p = \left(\sum_{j=1}^d |x_j|^p \right)^{1/p}$ the ℓ^p -norm of x . Moreover, we let $\|x\|_\infty = \max_{j \in \{1, \dots, d\}} |x_j|$ denote the ℓ^∞ -norm of x .
- We denote by $B^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ the closed d -dimensional unit ball and by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$.
- For a matrix $A \in \mathbb{R}^{m \times n}$ and $j \in \{1, \dots, m\}$, we denote by $A_{j,:} \in \mathbb{R}^{1 \times n}$ the j th row of A .

- If $A \in \mathbb{R}^{m \times n}$ and $p, q \in [1, \infty]$, we denote the $(\ell^p \rightarrow \ell^q)$ -operator norm of A by

$$\|A\|_{p \rightarrow q} = \sup_{x \in \mathbb{R}^n, \|x\|_p=1} \|Ax\|_q.$$

- We let $\text{diag}(\alpha_1, \dots, \alpha_d) \in \mathbb{R}^{d \times d}$ denote the diagonal matrix with entries $\alpha_1, \dots, \alpha_d$ on the diagonal.
- For a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ and $K \subseteq \mathcal{X}$, we denote by $f|_K$ the restriction of f onto K .

Probability theory

- For a real-valued random variable X , we denote by $\mathbb{E}[X]$ its expectation. Its sub-gaussian norm is denoted by $\|X\|_{\psi_2}$, i.e.,

$$\|X\|_{\psi_2} = \inf \{t > 0 : \mathbb{E} [\exp (X^2/t^2)] \leq 2\};$$

see [87, Section 2.5] for more details. If X is a d -dimensional real-valued random vector, we set

$$\|X\|_{\psi_2} = \sup_{x \in \mathbb{S}^{d-1}} \|\langle X, x \rangle\|_{\psi_2}. \quad (1.1)$$

- We write $X \sim \mathcal{N}(0, \sigma^2)$ if X follows a normal distribution with zero mean and variance σ^2 . If X is a d -dimensional real-valued random vector, we write $X \sim \mathcal{N}(0, \sigma^2 \cdot I_d)$ if $X_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for $j \in \{1, \dots, d\}$.
- We write $X \sim \text{Unif}\{0, 1\}$ if X follows a Bernoulli distribution with parameter $\frac{1}{2}$.
- For two random variables X and Y , we write $X \stackrel{d}{=} Y$ if they are equal in distribution.
- For $\emptyset \neq K \subseteq \mathbb{R}^d$, we let

$$w(K) := \mathbb{E}_{g \sim \mathcal{N}(0, I_d)} \left[\sup_{x \in K} \langle g, x \rangle \right]$$

denote the Gaussian width of K .

Functional analysis

- For a set $K \subseteq \mathbb{R}^d$, we call $\Lambda \subseteq K$ an ε -net of K if for every $x \in K$ there exists $x^* \in \Lambda$ satisfying $\|x - x^*\|_2 \leq \varepsilon$. We then define the ε -covering number of K as

$$\mathcal{N}(K, \varepsilon) := \min\{|\Lambda| : \Lambda \text{ is } \varepsilon\text{-net of } K\} \in \mathbb{N}_0 \cup \{\infty\}.$$

- For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $x = (x_1, \dots, x_d) \in \mathbb{F}^d$, and a multiindex $\mathbf{k} \in \mathbb{N}_0^d$, we set

$$|\mathbf{k}| := \sum_{j=1}^d \mathbf{k}_j \quad \text{and} \quad x^{\mathbf{k}} := \prod_{j=1}^d x_j^{\mathbf{k}_j}.$$

- For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, a measurable subset $K \subseteq \mathbb{F}^d$, and a measurable function $f : K \rightarrow \mathbb{F}^d$, we denote for $p \in [1, \infty)$ the L^p -norm of f by

$$\|f\|_{L^p(K; \mathbb{F})} = \left(\int_K |f(x)|^p \, dx \right)^{1/p} \quad \text{and} \quad \|f\|_{L^\infty(K; \mathbb{F})} = \text{ess sup}_{x \in K} |f(x)|$$

with ess sup denoting the essential supremum. For $p \in [1, \infty]$, we let

$$L^p(K; \mathbb{F}) = \left\{ f : K \rightarrow \mathbb{F} \text{ measurable} : \|f\|_{L^p(K; \mathbb{F})} < \infty \right\}.$$

- For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $p \in [1, \infty]$, we write

$$L^p_{\text{loc}}(\mathbb{F}^d; \mathbb{F}) = \left\{ f : \mathbb{F}^d \rightarrow \mathbb{F} : f|_K \in L^p(K; \mathbb{F}) \text{ for every compact } K \subseteq \mathbb{F}^d \right\}.$$

- For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $K \subseteq \mathbb{F}^d$, we denote by $C(K; \mathbb{F})$ the set of continuous functions from K to $\mathbb{F}$.
- For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, an open set $U \subseteq \mathbb{F}^d$, and $r \in \mathbb{N}_0 \cup \{\infty\}$, we denote by $C^r(U; \mathbb{F})$ the set of r times continuously differentiable functions from U to $\mathbb{F}$. In the case $\mathbb{F} = \mathbb{C}$, differentiability is to be understood in the sense of real variables, i.e., by identifying $\mathbb{C}$ with $\mathbb{R}^2$. For $f \in C^r(U; \mathbb{F})$, we let

$$\|f\|_{C^r(U; \mathbb{F})} := \sup_{\mathbf{k} \in \mathbb{N}_0^d, |\mathbf{k}| \leq r} \|\partial^{\mathbf{k}} f\|_{L^\infty(U; \mathbb{F})}$$

with $\partial^{\mathbf{k}} f$ denoting the partial derivative of f according to the multiindex $\mathbf{k}$. Moreover, if $\overline{U}$ denotes the closure of U (with respect to the Euclidean topology), we set

$$C^r(\overline{U}; \mathbb{F}) := \left\{ f : \overline{U} \rightarrow \mathbb{F} : f|_U \in C^r(U; \mathbb{F}), \partial^{\mathbf{k}}(f|_U) \text{ uniformly continuous for all } |\mathbf{k}| \leq r \right\}.$$

For $f \in C^r(\overline{U}; \mathbb{F})$, we let $\|f\|_{C^r(\overline{U}; \mathbb{F})} = \|f|_U\|_{C^r(U; \mathbb{F})}$.

- For $d, r \in \mathbb{N}$ and $p \in [1, \infty]$, we call a function $f \in L^p(B^d; \mathbb{R})$ an L^p -Sobolev function of regularity r if for every $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq r$ the derivative $\partial^{\mathbf{k}} f$ exists (in the weak sense) on the interior of B^d and is itself contained in $L^p(B^d; \mathbb{R})$. We then set

$$\|f\|_{W_d^{r,p}} := \begin{cases} \left(\sum_{\mathbf{k} \in \mathbb{N}_0^d, |\mathbf{k}| \leq r} \|\partial^{\mathbf{k}} f\|_{L^p(B^d; \mathbb{R})}^p \right)^{1/p} & \text{if } p < \infty, \\ \max_{\mathbf{k} \in \mathbb{N}_0^d, |\mathbf{k}| \leq r} \|\partial^{\mathbf{k}} f\|_{L^\infty(B^d; \mathbb{R})} & \text{if } p = \infty. \end{cases}$$

We denote by $\mathcal{B}(W_d^{r,p})$ the set of L^p -Sobolev functions $f \in L^p(B^d; \mathbb{R})$ of regularity r with $\|f\|_{W_d^{r,p}} \leq 1$.

1.1.2. Fully-connected feedforward neural networks

As fully-connected feedforward neural networks (FNNs) are the central object of the present thesis, we define them formally in this section.

Definition 1.1.1 (Fully-connected feedforward neural network (FNN)). Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $\varrho : \mathbb{F} \rightarrow \mathbb{F}$, $L \in \mathbb{N}$, and $N_0, \dots, N_{L+1} \in \mathbb{N}$. For each $\ell \in \{0, \dots, L\}$ let $V^{(\ell)} : \mathbb{F}^{N_\ell} \rightarrow \mathbb{F}^{N_{\ell+1}}$ be an $\mathbb{F}$ -affine map. Set

$$\Phi := V^{(L)} \circ \varrho \circ V^{(L-1)} \circ \dots \circ \varrho \circ V^{(1)} \circ \varrho \circ V^{(0)} : \mathbb{F}^{N_0} \rightarrow \mathbb{F}^{N_{L+1}}, \quad (1.2)$$

where the application of ϱ is to be understood componentwise, i.e.,

$$\varrho(x_1, \dots, x_{N_\ell}) = (\varrho(x_1), \dots, \varrho(x_{N_\ell})) \quad \text{for every } \ell \in \{1, \dots, L\}.$$

Then Φ is called a fully-connected feedforward $\mathbb{F}$ -valued neural network with

- activation function ϱ ,
- input dimension N_0 ,

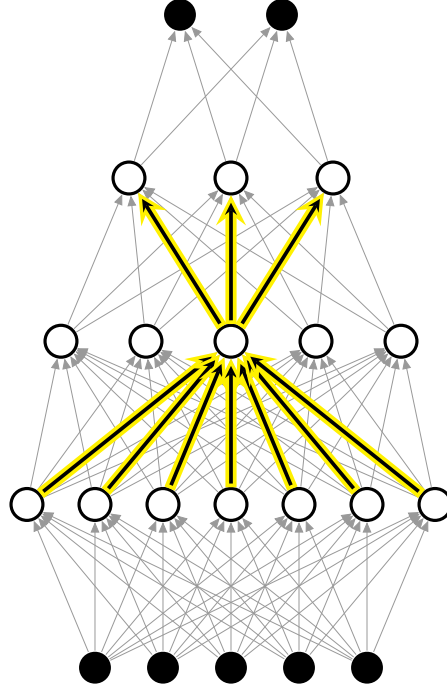


Figure 1.1.: Graphical representation of an FNN with 5 input neurons, 2 output neurons, and 3 hidden layers with 7, 5, and 3 neurons. The input and output neurons are drawn as dots to indicate that the activation function is not applied at these neurons. For the third neuron in the second hidden layer, all its incoming and outgoing connections are highlighted in yellow.

- output dimension N_{L+1} ,
- L hidden layers,
- width $W := \max_{\ell \in \{1, \dots, L\}} N_\ell$.

Φ is called *shallow* if $L = 1$ and *deep* if $L > 1$. Since the maps $V^{(\ell)}$ are $\mathbb{F}$ -affine, there exist $W^{(\ell)} \in \mathbb{F}^{N_{\ell+1} \times N_\ell}$ and $b^{(\ell)} \in \mathbb{F}^{N_{\ell+1}}$ with $V^{(\ell)}x = W^{(\ell)}x + b^{(\ell)}$ for every $x \in \mathbb{F}^{N_\ell}$ and $\ell \in \{0, \dots, L\}$. The matrices $W^{(\ell)}$ are called the *weight matrices* of Φ and the entries of these matrices are called the *weights* of Φ . The vectors $b^{(\ell)}$ are called the *bias vectors* of Φ and their entries the *biases*. We use the abbreviations RVNN and CVNN to refer to real-valued and complex-valued FNNs, respectively.

Typically, an FNN is visualized using multiple *layers* of stacked so-called *neurons*; see Figure 1.1. Let us formalize this idea. Let Φ be an FNN as in (1.2). Take $\ell \in \{1, \dots, L\}$ and $j \in \{1, \dots, N_\ell\}$. Then, the j th neuron in the ℓ th hidden layer of Φ refers to the map

$$\mathbb{F}^{N_{\ell-1}} \rightarrow \mathbb{F}, \quad x \mapsto \varrho \left(W_{j,:}^{(\ell-1)} x + b_j^{(\ell-1)} \right),$$

where we recall that $W_{j,:}^{(\ell-1)}$ denotes the j th row of $W^{(\ell-1)}$. By stacking all such neurons for $j \in \{1, \dots, N_\ell\}$, we obtain the mapping

$$\mathbb{F}^{N_{\ell-1}} \rightarrow \mathbb{F}^{N_\ell}, \quad x \mapsto \left(\varrho \circ V^{(\ell-1)} \right) (x),$$

which we refer to as the ℓ th *hidden layer* of Φ . Moreover, for each $j \in \{1, \dots, N_{L+1}\}$, the map

$$\mathbb{F}^{N_L} \rightarrow \mathbb{F}, \quad x \mapsto W_{j,:}^{(L)} x + b_j^{(L)}$$

is called the j th *output neuron* of Φ and the map

$$\mathbb{F}^{N_L} \rightarrow \mathbb{F}^{N_{L+1}}, \quad x \mapsto V^{(L)}(x)$$

is referred to as the *output layer* of Φ . Lastly, for each $j \in \{1, \dots, N_0\}$, the map $\mathbb{F}^{N_0} \rightarrow \mathbb{F}, x \mapsto x_j$ is called an *input neuron* of Φ and the identity map $\mathbb{F}^{N_0} \rightarrow \mathbb{F}^{N_0}$ is the *input layer* of Φ . We say that the ℓ th hidden layer consists of M neurons if $N_\ell = M$.

1.1.3. A mathematical framework of supervised machine learning

In this section, along the lines of [74], we give a brief introduction to the mathematical formulation of a supervised machine learning task. This formulation is used both in Chapters 2 and 3 to motivate the study of the Lipschitz constant and of the approximation properties of neural networks, respectively.

Let $\mathcal{X}$ be a set of inputs, e.g., a set of images, and $\mathcal{Y}$ be a set of outputs. Further, let $\mathcal{D}$ be a probability distribution defined on the Cartesian product $\mathcal{X} \times \mathcal{Y}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ be a given *loss function* on $\mathcal{Y}$. In the case of a binary classification task, one typically has $\mathcal{Y} = \{0, 1\}$ together with the so-called 0-1-loss, given via $\ell(y_1, y_2) = \mathbb{1}_{y_1 \neq y_2}$. In the case of a regression task, a typical choice would be $\mathcal{Y} = \mathbb{R}$ together with the absolute value loss $\ell(y_1, y_2) = |y_1 - y_2|$ or the squared loss $\ell(y_1, y_2) = (y_1 - y_2)^2$.

The goal then is to find a function $f = f_{\mathcal{D}} : \mathcal{X} \rightarrow \mathcal{Y}$ which minimizes the *expected risk* given by

$$R_{\mathcal{D}}(f) := \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(f(x), y)]. \quad (1.3)$$

Roughly speaking, one wants to find a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ which best (in terms of the loss function ℓ) describes the relation between inputs and outputs that is given by the distribution $\mathcal{D}$.

In a realistic scenario, however, the underlying distribution $\mathcal{D}$ is *unknown*. The learner only has access to it via finitely many, independently drawn *training samples*

$$S = ((x_1, y_1), \dots, (x_N, y_N)) \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}.$$

Thus, the idea is to find a function which works well on the training samples with the hope that this function also has a small expected risk.

To this end, the learner restricts itself to a *hypothesis class* $\mathcal{H} \subseteq \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$ and then seeks to minimize the *empirical risk*

$$L_S(h) := \frac{1}{N} \cdot \sum_{j=1}^N \ell(h(x_j), y_j)$$

over all hypotheses $h \in \mathcal{H}$. Typically, one picks a hypothesis class that is parameterizable by finitely many parameters, so that the search for an empirical risk minimizer can be accomplished by numerical algorithms. We can use the strong law of large numbers [27, Theorem 8.3.5] to get that for every fixed hypothesis h with $R_{\mathcal{D}}(h) < \infty$,

$$L_S(h) \rightarrow R_{\mathcal{D}}(h) \quad \text{almost surely} \quad (1.4)$$

as the number N of training samples goes to infinity. This backups the intuition that a hypothesis with small empirical risk also has a small expected risk.

This approach, however, has the disadvantage that it only holds in a pointwise sense, i.e., for every fixed hypothesis h , and that the convergence speed in (1.4) is unknown. The notion of *probably-approximately-correct (PAC) learnability* addresses these two aspects.

Definition 1.1.2 (cf. [74, Definition 3.4]). Let $\mathcal{H} \subseteq \{g : \mathcal{X} \rightarrow \mathcal{Y}\}$ be a hypothesis class. $\mathcal{H}$ is called *PAC learnable* if there exists a function $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$ and a learning algorithm

$$\mathcal{A} : \bigcup_{N=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^N \rightarrow \mathcal{H}$$

with the following property. For every $\varepsilon, \delta \in (0, 1)$, every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and every $N \geq m_{\mathcal{H}}(\varepsilon, \delta)$, if $S = ((x_1, y_1), \dots, (x_N, y_N)) \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$, then with probability at least $1 - \delta$,

$$R_{\mathcal{D}}(\mathcal{A}(S)) \leq \inf_{h' \in \mathcal{H}} R_{\mathcal{D}}(h') + \varepsilon.$$

Definition 1.1.2 states that a hypothesis class is PAC learnable if there exists a learning algorithm such that *every* given distribution, i.e., every given “task”, can be learned arbitrarily well by applying this algorithm to finitely many i.i.d. drawn training samples, as long as sufficiently many samples are drawn. Here, “arbitrarily well” means that the expected risk of the output of the learning algorithm can get arbitrarily close to the best achievable expected risk over all hypotheses from the given class, with probability arbitrarily close to 1. Hence, PAC learnability of a hypothesis class does not necessarily imply that the output of the corresponding learning algorithm has small expected risk with high probability; the magnitude of this expected risk crucially depends on the quantity

$$\inf_{h' \in \mathcal{H}} R_{\mathcal{D}}(h'),$$

called the *approximation error* of $\mathcal{H}$; see Section 3.2.

Note that in the case of empirical risk minimization, the corresponding learning algorithm is given by

$$\mathcal{A}(S) \in \operatorname{argmin}_{h' \in \mathcal{H}} L_S(h') \quad \text{for } S = ((x_1, y_1), \dots, (x_N, y_N)),$$

where we assume for simplicity that an empirical risk minimizer exists. In the case of the 0-1-loss, the *Fundamental Theorem of Statistical Learning* provides a complete characterization of hypothesis classes that are PAC learnable via empirical risk minimization using the notion of *VC dimension* [74, Theorem 6.7].

Deep learning is a special case of machine learning, where the hypothesis class $\mathcal{H}$ typically consists of a class of neural networks with a given architecture, e.g., a prescribed amount of hidden layers of a certain width. The training of a neural network corresponds to the search for an empirical risk minimizer, usually done using numerical algorithms such as variants of gradient descent. However, we would like to add that *stochastic* gradient descent may be viewed as an algorithm that directly minimizes the expected risk $R_{\mathcal{D}}(\cdot)$, rather than the empirical risk [74, Section 14.5].

2. The Lipschitz constant of random neural networks

This chapter is based on the following two papers.

- “Upper and lower bounds for the Lipschitz constant of random neural networks” (co-authored with Dominik Stöger, Thomas Telaar, and Felix Voigtlaender), published in *Information and Inference: A Journal of the IMA* 14.2 in 2025, henceforth referred to as [36];
- “Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks” (co-authored with Sjoerd Dirksen, Patrick Finke, Dominik Stöger, and Felix Voigtlaender), appeared as a preprint with the identifier *arXiv:2506.19695* in 2025, henceforth referred to as [23].

2.1. The Lipschitz constant of a neural network as a measure of adversarial robustness

In this section, we give a brief motivation on the connection between the Lipschitz constant of a machine learning model and its adversarial robustness.

As was already introduced in Chapter 1, adversarial robustness describes the susceptibility of machine learning models to small, often handcrafted perturbations of an input. In the context of deep learning, this phenomenon was first studied ten years ago [39, 79]. In these seminal papers, it is shown that popular convolutional neural networks such as AlexNet [48] and GoogleLeNet [78] are highly vulnerable to such perturbations. After these first works on adversarial robustness in deep learning, there was a wave of research on the subject in the subsequent years, primarily focusing on the development of increasingly sophisticated adversarial attacks on the one hand [3, 14, 18, 61, 65], and defenses against such attacks on the other hand [25, 54, 66, 98]. The fact that newly proposed defenses were in many cases soon circumvented by more sophisticated attacks sparked the hypothesis that an inherent trade-off may exist between adversarial robustness and model performance on test data [77, 82]. Nevertheless, other works such as [93] argue that this trade-off is not fundamental and may be avoided.

The fact that an adversary can easily fool deep learning models to misclassify inputs raises serious concerns regarding the reliability and safety of deep learning algorithms in practical applications, especially in security-critical ones. In this regard, it is particularly important to mention that adversarial examples for a model can even be generated if the model itself and its training data are unknown [8, 64]. Even if it is not possible to feed data directly into the model but one can only feed data into the model via a sensor, e.g., a camera, it is possible to generate adversarial examples [49]. The topic’s relevance motivates its theoretical study, with a specific emphasis on the question of whether adversarial examples are *inherent* to deep learning models or whether they can be avoided by, e.g., using suitably adapted training algorithms.

We recall from Section 1.1.3 that, in a theoretical setup, the performance of a machine learning model $f : \mathcal{X} \rightarrow \mathcal{Y}$ with respect to a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is measured using the expected risk

$$R_{\mathcal{D}}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(f(x), y)],$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ is a given loss function. In the following, we take $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $\mathcal{Y} = \mathbb{R}$. In practical applications, the marginal distribution $\mathcal{D}^{\mathcal{X}}$ of $\mathcal{D}$ on $\mathcal{X}$ typically has a quite “thin” support. For instance, the set of “natural” images with resolution 500×500 showing cats and dogs is only a tiny subset of $\mathbb{R}^{500 \times 500}$. Adversarial attacks produce inputs that are *outside* of the support of $\mathcal{D}^{\mathcal{X}}$ and are thus *invisible* to the expected risk. Hence, the notion of expected risk is no longer sufficient in adversarial settings.

To tackle this problem, it has been proposed to replace the expected risk by the *expected adversarial risk*

$$R_{\mathcal{D},\text{adv}}(f, \varepsilon, p) := \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|x' - x\|_p \leq \varepsilon} \ell(f(x'), y) \right];$$

see [16, Section 3.2]. Here, the parameter $p \in [1, \infty]$ specifies the norm with respect to which the adversarial perturbation is measured and $\varepsilon > 0$ specifies the allowed size of the perturbation. The expected adversarial risk hence does not only capture the performance of a model on an input x itself, but on a neighborhood of this input. Thus, it takes into account not only inputs that lie inside the support of the data distribution, but also inputs that are “close” to it.

For a given model $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we define its ℓ^p -Lipschitz constant for $p \in [1, \infty]$ by

$$\text{Lip}_p(f) := \sup_{x, y \in \mathbb{R}^d, x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|_p}.$$

We further assume that the underlying loss function $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ satisfies a Lipschitz condition too, i.e., that there exists $\lambda > 0$ satisfying

$$|\ell(y_1, y_0) - \ell(y_2, y_0)| \leq \lambda \cdot |y_1 - y_2| \quad \text{for every } y_0 \in \text{supp}(\mathcal{D}^{\mathcal{Y}}) \text{ and } y_1, y_2 \in \mathbb{R}. \quad (2.1)$$

Here, $\mathcal{D}^{\mathcal{Y}}$ denotes the marginal distribution of $\mathcal{D}$ on $\mathcal{Y}$ and $\text{supp}(\mathcal{D}^{\mathcal{Y}})$ its support. For instance, in the case of a binary classification problem, it is typical to have $\text{supp}(\mathcal{D}^{\mathcal{Y}}) = \{0, 1\}$ and to consider the loss function

$$\ell_{\log}(y_1, y_0) = -y_0 \cdot \ln(\sigma(y_1)) - (1 - y_0) \cdot \ln(1 - \sigma(y_1)) \quad \text{for } y_0, y_1 \in \mathbb{R},$$

which is the composition of the sigmoid function $\sigma(t) = (1 + \exp(-t))^{-1}$ and the logistic loss. If $\text{supp}(\mathcal{D}^{\mathcal{Y}}) = \{0, 1\}$, this loss function satisfies (2.1) with $\lambda = 1$.¹

We then obtain

$$\begin{aligned} R_{\mathcal{D},\text{adv}}(f, \varepsilon, p) &= (R_{\mathcal{D},\text{adv}}(f, \varepsilon, p) - R_{\mathcal{D}}(f)) + R_{\mathcal{D}}(f) \\ &= \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|x' - x\|_p \leq \varepsilon} \ell(f(x'), y) - \ell(f(x), y) \right] + R_{\mathcal{D}}(f) \\ &\leq \lambda \cdot \mathbb{E}_{x \sim \mathcal{D}^{\mathcal{X}}} \left[\max_{\|x - x'\|_p \leq \varepsilon} |f(x') - f(x)| \right] + R_{\mathcal{D}}(f) \\ &\leq \lambda \cdot \text{Lip}_p(f) \cdot \varepsilon + R_{\mathcal{D}}(f). \end{aligned}$$

A similar computation is done in [16, Section 3.2] for the case of a multi-label classification problem. The Lipschitz constant of a machine learning model can thus be viewed as a quantity that controls the difference between the expected risk and the expected adversarial risk. Therefore, if a model has a *small* Lipschitz constant, it is likely to be adversarially robust in the sense that the expected adversarial risk is only slightly worse than the expected risk.

¹Since $\sigma'(t) = \sigma(t)(1 - \sigma(t))$, we get for $y_0, y_1 \in \mathbb{R}$ that $\frac{\partial}{\partial y_1} \ell(y_1, y_0) = -y_0 \cdot (1 - \sigma(y_1)) - (1 - y_0) \cdot (-\sigma(y_1))$. This yields $\left| \frac{\partial}{\partial y_1} \ell(y_1, y_0) \right| \leq 1$ for every $y_0 \in \{0, 1\}$ and $y_1 \in \mathbb{R}$.

A small Lipschitz constant is not only beneficial regarding adversarial robustness, it also has a regularizing effect, meaning that a model with small Lipschitz constant is less susceptible to overfitting [41]. The impact of the Lipschitz constant on robustness and generalization has sparked interest in studying neural networks with Lipschitz constraints, i.e., networks whose Lipschitz constant is artificially enforced to be small, even during training [16, 67]. Moreover, since the computation of the Lipschitz constant of a neural network is computationally difficult, several algorithms have been proposed to estimate it efficiently [29, 88].

While upper bounds on a machine learning model's Lipschitz constant have clear implications for its adversarial robustness, the significance of lower bounds is comparatively less evident. To illustrate this, let $0 < \Lambda < \text{Lip}_p(f)$. By definition of the Lipschitz constant, there exist $x_0, r \in \mathbb{R}^d$ with

$$|f(x_0 + r) - f(x_0)| \geq \Lambda \cdot \|r\|_p. \quad (2.2)$$

This does not rule out the possibility that $\|r\|_p$ is large, whereas a small norm $\|r\|_p$ is desired for an adversarial example. However, in the case where f is differentiable almost everywhere, one has

$$\text{Lip}_p(f) = \sup_{\substack{x \in \mathbb{R}^d \\ f \text{ differentiable at } x}} \|\nabla f(x)\|_{p'},$$

where $1/p + 1/p' = 1$; see for instance [23, Proposition 2.4]. Here, $\nabla f(x)$ denotes the gradient of f at x . In particular, this is the case if f is locally Lipschitz; see [28, Section 3.1.2].

Hence, if $0 < \Lambda < \text{Lip}_p(f)$, there exists $x_0 \in \mathbb{R}^d$ with

$$\|\nabla f(x_0)\|_{p'} \geq \Lambda.$$

Therefore, by duality, there exists $\nu \in \mathbb{R}^d$ with $\|\nu\|_p = 1$ and

$$\lim_{t \rightarrow 0} \frac{|f(x_0 + t\nu) - f(x_0)|}{|t|} = |\langle \nabla f(x_0), \nu \rangle| = \|\nabla f(x_0)\|_{p'} \geq \Lambda.$$

Thus, there exists $\delta > 0$ such that for $t \in (-\delta, \delta)$,

$$|f(x_0 + t\nu) - f(x_0)| \geq \Lambda \cdot |t| = \Lambda \cdot \|t\nu\|_p. \quad (2.3)$$

This is much stronger than (2.2), as it states that for a certain $x_0 \in \mathbb{R}^d$, the perturbation r in (2.2) can be picked to have arbitrarily small ℓ^p -norm.

However, we still observe the following problems with respect to adversarial examples.

- The size of δ is unknown. In particular, if δ is too small, then $\Lambda \cdot \|t\nu\|_p$ is small too and thus, the lower bound on $|f(x_0 + t\nu) - f(x_0)|$ is insignificant compared to $|f(x_0)|$ (crucially depending on the size of Λ and $|f(x_0)|$).
- $\|x_0\|_p$ can be very large, whereas inputs to machine learning models are typically scaled so that inputs with large norm normally do not occur.
- In practice, it is observed that for *many* inputs there exists a small perturbation that drastically alters the prediction of the model. A lower bound on the Lipschitz constant, however, only implies the *existence* of a point x_0 such that (2.3) is satisfied and does not imply that the property holds for *every* possible input. Some of the lower bounds established in [23, 36], however, are *pointwise* bounds on the gradients and thus show that, in the setting of [23, 36], (2.3) holds at every fixed input $x_0 \in \mathbb{R}^d$ with high probability; see Sections 2.3 and 2.4.

Hence, while (2.3) seems indicative of the existence of adversarial examples at first, the situation is in fact more subtle. However, establishing lower bounds for the Lipschitz constant of machine

learning models is, apart from scientific curiosity, still crucial towards understanding their adversarial robustness, as a large Lipschitz constant may be viewed as a *necessary* condition for the existence of adversarial examples.

There exist several papers that investigate the Lipschitz constant of deep learning models from a theoretical point of view. In [11], it is shown in a very general setting that a model with a small Lipschitz constant that interpolates i.i.d. drawn training samples is necessarily largely overparameterized. However, this does *not* imply that an overparameterized model necessarily has a small Lipschitz constant. Upper bounds for the Lipschitz constant of networks with random parameters were established in [12, 62]. Taken together, our results established in [23, 36] significantly improve and generalize the results of those two papers. We refer to the full texts of [23, 36] for detailed comparisons. Moreover, there exists a variety of papers studying the existence of adversarial examples in different settings, both in random [6, 10, 20, 21, 60] and in trained neural networks [7, 31, 84, 91].

2.2. Setting and preliminaries

Before studying the papers [23, 36] in depth, we start with a description of the setting and with a few preliminary observations regarding the gradient of ReLU networks, an important tool for our study of the Lipschitz constant of random neural networks.

We start by defining the ReLU activation function, which is indisputably one of the most popular activation functions in practical applications.

Definition 2.2.1 (Rectified Linear Unit (ReLU)). The *Rectified Linear Unit* (ReLU) is the function

$$\text{ReLU} : \mathbb{R} \rightarrow \mathbb{R}, \quad \text{ReLU}(x) = \max\{0, x\}.$$

An FNN that uses the ReLU as the activation function is called a ReLU neural network.

In the following assumption, we describe the random neural networks that are considered in the present chapter.

Assumption 2.2.2. In both papers [23, 36], we study ReLU neural networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with d input neurons, a single output neuron, with L hidden layers, and N neurons in each hidden layer; see Figure 2.1. We let $W^{(0)} \in \mathbb{R}^{N \times d}$, $W^{(1)}, \dots, W^{(L-1)} \in \mathbb{R}^{N \times N}$, and $W^{(L)} \in \mathbb{R}^{1 \times N}$ denote the weight matrices associated with Φ and $b^{(0)}, \dots, b^{(L-1)} \in \mathbb{R}^N$, $b^{(L)} \in \mathbb{R}$ denote the bias vectors of Φ . We assume that the matrices $W^{(\ell)}$ as well as the vectors $b^{(\ell)}$ are randomly chosen in the following way:

- The entries of $W^{(0)}, b^{(0)}, \dots, W^{(L)}, b^{(L)}$ are jointly independent.

- We have

$$W_{i,j}^{(\ell)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N), \quad 0 \leq \ell < L \quad \text{and} \quad W_{1,j}^{(L)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$$

for all suitable indices i, j .

- At this point, we do not impose any restrictions (apart from independence) on the distribution of the biases. However, note that mild assumptions on the distribution of the biases are required for some of the results presented in this section, which will be mentioned in the individual statements.

We call a network zero-bias if all its biases are initialized to zero.

This random initialization is heavily inspired by the popular He initialization [44] with the following two differences:

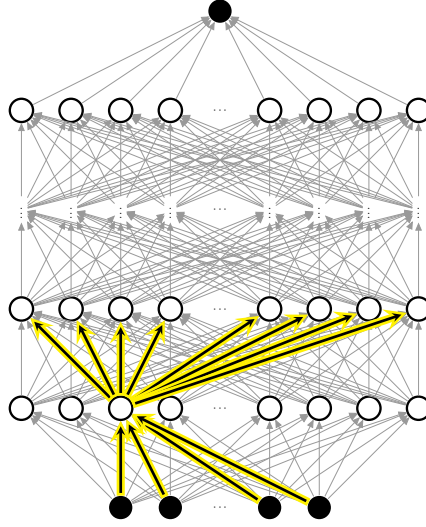


Figure 2.1.: The FNNs considered in Chapter 2. Note that all the hidden layers are assumed to consist of the same number of neurons.

- In [44], the entries of $W^{(0)}$ are drawn according to $\mathcal{N}(0, 2/d)$ and the entries of $W^{(L)}$ are drawn according to $\mathcal{N}(0, 1/N)$.
- The biases in [44] are set to zero, whereas we allow more general distributions for the entries of the bias vectors.

The random initialization that we consider in [23, 36] is a common choice in theoretical works about random ReLU networks; see [1, 12]. Moreover, at least in the case of zero biases, we note that the function Φ for the two initialization schemes only differs by a factor of $1/\sqrt{d}$. Hence, the results in [23, 36] readily imply bounds for the classical He initialization as well.

The random initialization used in [23, 36] is natural, since, at least in the case of zero biases, each layer is *isometric in expectation*. Indeed, for $k \in \mathbb{N}$ let $W \in \mathbb{R}^{N \times k}$ be a random matrix with $W_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N)$. For $x \in \mathbb{R}^k \setminus \{0\}$ we then observe

$$\begin{aligned}
 \mathbb{E} \left[\|\text{ReLU}(Wx)\|_2^2 \right] &= \sum_{j=1}^N \mathbb{E} \left[\mathbb{1}_{(Wx)_j > 0} \cdot (Wx)_j^2 \right] \\
 &= \frac{1}{2} \cdot \sum_{j=1}^N \left(\mathbb{E} \left[\mathbb{1}_{(Wx)_j > 0} \cdot (Wx)_j^2 \right] + \mathbb{E} \left[\mathbb{1}_{(Wx)_j < 0} \cdot (Wx)_j^2 \right] \right) \\
 &\stackrel{W \sim -W}{=} \frac{1}{2} \cdot \sum_{j=1}^N \left(\mathbb{E} \left[\mathbb{1}_{(Wx)_j > 0} \cdot (Wx)_j^2 \right] + \mathbb{E} \left[\mathbb{1}_{(Wx)_j < 0} \cdot (-Wx)_j^2 \right] \right) \\
 &= \frac{1}{2} \cdot \sum_{j=1}^N \mathbb{E} \left[\mathbb{1}_{(Wx)_j \neq 0} \cdot (Wx)_j^2 \right] = \frac{1}{2} \cdot \sum_{j=1}^N \mathbb{E} \left[(Wx)_j^2 \right] \\
 &\stackrel{(Wx)_j \sim \mathcal{N}(0, 2\|x\|_2^2/N)}{=} \frac{1}{2} \cdot \sum_{j=1}^N \frac{2\|x\|_2^2}{N} = \|x\|_2^2.
 \end{aligned}$$

If $x = 0$, we trivially have $\mathbb{E} \left[\|\text{ReLU}(Wx)\|_2^2 \right] = 0 = \|x\|_2^2$.

2. The Lipschitz constant of random neural networks

We introduce some additional notation: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU network with L hidden layers of width N . For an input $x \in \mathbb{R}^d$, we set $\Phi^{(-1)}(x) = x$ and for $\ell \in \{0, \dots, L-1\}$, we define recursively

$$\begin{aligned} D^{(\ell)}(x) &:= \text{diag} \left(\mathbb{1}_{(W^{(\ell)}\Phi^{(\ell-1)}(x)+b^{(\ell)})_1 > 0}, \dots, \mathbb{1}_{(W^{(\ell)}\Phi^{(\ell-1)}(x)+b^{(\ell)})_N > 0} \right) \in \mathbb{R}^{N \times N}, \\ \Phi^{(\ell)}(x) &:= D^{(\ell)}(x) \cdot \left(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)} \right) = \text{ReLU}(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}). \end{aligned} \quad (2.4)$$

The analysis of the Lipschitz constant of random ReLU networks in [23, 36] is based on the gradient of such networks: For a ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ as above and an input $x \in \mathbb{R}^d$, we let

$$\bar{\nabla}\Phi(x) := \left(W^{(L)}D^{(L-1)}(x) \dots D^{(0)}(x)W^{(0)} \right)^T \in \mathbb{R}^d \quad (2.5)$$

be the formal² gradient of Φ at x . The relation between the formal gradient and the Lipschitz constant of random ReLU networks is established in the following theorem.

Theorem 2.2.3 ([23, Theorem 2.8]). *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network as in Assumption 2.2.2 and $p, p' \in [1, \infty]$ with $1/p + 1/p' = 1$. Let*

$$M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla\Phi(x) = \bar{\nabla}\Phi(x)\}.$$

Moreover, fix $x_0 \in \mathbb{R}^d \setminus \{0\}$. Then, almost surely with respect to the randomness in Φ ,

$$\|\bar{\nabla}\Phi(x_0)\|_{p'} \leq \text{Lip}_p(\Phi) = \sup_{x \in M_\Phi} \|\bar{\nabla}\Phi(x)\|_{p'} \leq \sup_{x \in \mathbb{R}^d} \|\bar{\nabla}\Phi(x)\|_{p'}$$

and, in case Φ is zero-bias,

$$\|\bar{\nabla}\Phi(x_0)\|_{p'} \leq \text{Lip}_p(\Phi) = \sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} \|\bar{\nabla}\Phi(x)\|_{p'} \leq \sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla}\Phi(x)\|_{p'}.$$

Theorem 2.2.3 is the central tool used in [23, 36] to obtain upper and lower bounds for the Lipschitz constant of random ReLU neural networks.

We provide an explanation for why it suffices to take inputs from $\mathbb{S}^{d-1}$ in the zero-bias case in Theorem 2.2.3. To this end, let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU network with L hidden layers. On $\mathbb{R}^d$, we can define the equivalence relation

$$x \sim y \iff D^{(\ell)}(x) = D^{(\ell)}(y) \text{ for all } \ell \in \{0, \dots, L-1\}.$$

Each equivalence class is called an *activation region* of Φ . Note that the formal gradient remains constant inside a given activation region. If Φ is zero-bias, all activation regions are invariant with respect to scaling by positive scalars. This follows easily by the definition of the matrices $D^{(\ell)}(x)$. Therefore, the activation regions are cones originating from the origin and it suffices to study inputs from $\mathbb{S}^{d-1}$ in Theorem 2.2.3. If the biases are non-zero, the activation regions are (potentially infinitely large) convex polygons tessellating the input space. Therefore, it is no longer sufficient to only study inputs from the unit sphere to estimate the Lipschitz constant of Φ . We refer to Figure 2.2 for an illustration of the situation.

²Since the ReLU is not differentiable at 0, ReLU networks are not necessarily everywhere differentiable either. Hence, the gradient defined here is not an actual gradient (since it is defined for every input x) but rather a formal gradient which is sufficient for our analysis of the Lipschitz constant.

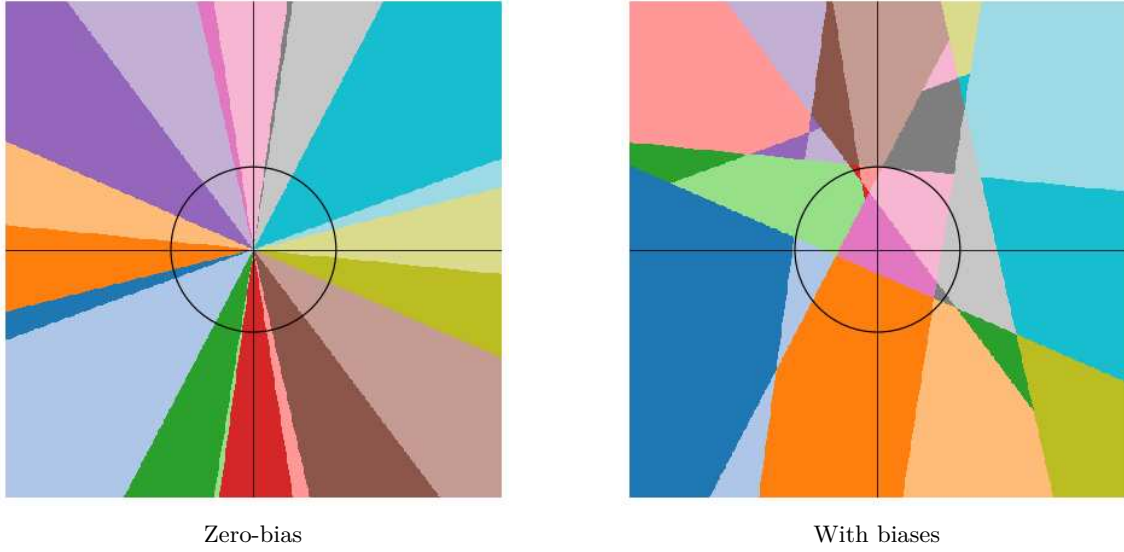


Figure 2.2.: Activation regions of ReLU networks. In the zero-bias case (left), the activation regions are cones and therefore invariant with respect to scaling by positive scalars. Thus, all activation regions can be observed on the unit sphere. With biases (right), this invariance is lost, and the regions are convex polygons tessellating the input space. Therefore, not all activation patterns are necessarily found on the unit sphere.

2.3. Upper and lower bounds for the Lipschitz constant of random neural networks

The paper *Upper and lower bounds for the Lipschitz constant of random neural networks* [36], co-authored with Dominik Stöger, Thomas Telaar, and Felix Voigtlaender, provides upper and lower bounds for the ℓ^2 -Lipschitz constant of ReLU networks with random weights and biases that hold in expectation and with high probability. The paper specifically distinguishes between the case of shallow and deep neural networks. Notably, the techniques from [36] can also be applied to obtain bounds for the ℓ^2 -Lipschitz constant of random ReLU neural networks where the weights follow *arbitrary* sub-gaussian distributions; see Section A.2. This section is split into the discussion of the upper bounds (see Section 2.3.1) and the discussion of the lower bounds (see Section 2.3.2) shown in [36].

2.3.1. Upper bounds

In this section, we discuss the upper bounds (both in the shallow and in the deep case) that are established in [36] for ReLU networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with random parameters. In our derivations, we focus on upper bounding $\text{Lip}_2(\Phi)$ in expectation and note that high-probability statements can be obtained in a similar fashion.

Note that, according to Theorem 2.2.3, in order to get an upper bound on $\text{Lip}_2(\Phi)$ it suffices to bound $\sup_{x \in \mathbb{R}^d} \|\nabla \Phi(x)\|_2$. Using the definition of the formal gradient, we get

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \|\nabla \Phi(x)\|_2 &= \sup_{x \in \mathbb{R}^d, z \in B^d} \langle \nabla \Phi(x), z \rangle \\ &= \sup_{x \in \mathbb{R}^d, z \in B^d} \left\langle (W^{(L)})^T, D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)} z \right\rangle. \end{aligned}$$

If we condition on $W^{(0)}, \dots, W^{(L-1)}$ as well as $b^{(0)}, \dots, b^{(L-1)}$ and only consider the randomness with respect to $W^{(L)}$, which is possible due to independence, we get

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] = w(\mathcal{L}),$$

where we set

$$\mathcal{L} := \left\{ D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}z : x \in \mathbb{R}^d, z \in B^d \right\} \subseteq \mathbb{R}^N$$

and recall that $w(\mathcal{L})$ denotes the Gaussian width of $\mathcal{L}$. A classical way of bounding the Gaussian width of a set from above is by using Dudley's inequality, which yields

$$w(\mathcal{L}) \lesssim \int_0^{\text{diam}(\mathcal{L})} \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} \, d\varepsilon; \quad (2.6)$$

see [87, Theorem 8.1.10]. This way, the problem of bounding the Lipschitz constant of random ReLU networks is transferred to bounding the *covering numbers* of the set $\mathcal{L}$. We remark that $\mathcal{L}$ is a random set, since it depends on the weights $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$.

We start by studying the case of *deep* networks, since it is a little simpler than the case of shallow networks (for which we prove a somewhat stronger bound). Moreover, the approach in the case of deep networks works for shallow networks as well. However, using refined techniques in the case of shallow networks, one can in that case improve the bound by a logarithmic factor; see Theorem 2.3.3.

We write

$$\mathcal{D} := \left\{ \left(D^{(0)}(x), D^{(1)}(x), \dots, D^{(L-1)}(x) \right) : x \in \mathbb{R}^d \right\}$$

for the set consisting of all tuples of D -matrices that can be encountered when passing inputs through the network. Note that there exists a natural one-to-one correspondence between $\mathcal{D}$ and the set of activation regions of Φ . The cardinality of $\mathcal{D}$ can naively be bounded by 2^{LN} , but using VC dimension based arguments, one can, in the case of $N > d + 2$, show that it can in fact be bounded by $(eN/(d+1))^{L(d+1)}$; see [36, Lemma 5.7]. Notably, the cardinality of $\mathcal{D}$ hence does *not* increase exponentially in N . By definition, we get

$$\mathcal{L} = \bigcup_{(C^{(0)}, \dots, C^{(L-1)}) \in \mathcal{D}} \left(C^{(L-1)}W^{(L-1)} \dots C^{(0)}W^{(0)}B^d \right)$$

which implies for every $\varepsilon > 0$ that

$$\mathcal{N}(\mathcal{L}, \varepsilon) \leq \sum_{(C^{(0)}, \dots, C^{(L-1)}) \in \mathcal{D}} \mathcal{N}(C^{(L-1)}W^{(L-1)} \dots C^{(0)}W^{(0)}B^d, \varepsilon).$$

We set $\Lambda := \|W^{(0)}\|_{2 \rightarrow 2} \dots \|W^{(L-1)}\|_{2 \rightarrow 2}$ and it is then easy to see that for $\varepsilon \in (0, \Lambda]$,

$$\mathcal{N}(\mathcal{L}, \varepsilon) \leq |\mathcal{D}| \cdot \mathcal{N}(\Lambda \cdot B^d, \varepsilon) \leq \left(\frac{eN}{d+1} \right)^{L(d+1)} \cdot \left(\frac{3\Lambda}{\varepsilon} \right)^d;$$

see [36, Lemma 5.8]. Further, since $\text{diam}(\mathcal{L}) \leq 2\Lambda$, one can show using (2.6) that

$$w(\mathcal{L}) \lesssim \Lambda \cdot \sqrt{d \cdot L \cdot \ln(N/d)}.$$

Lastly, from standard bounds on the operator norm of Gaussian matrices [87, Theorem 7.3.1],

$$\mathbb{E}_{W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}} \Lambda \leq (2\sqrt{2})^L$$

and thus overall

$$\begin{aligned} \mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] &= \mathbb{E}_{W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}} \left[\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] \right] \\ &\lesssim \mathbb{E}_{W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}} \left[\Lambda \cdot \sqrt{d \cdot L \cdot \ln(N/d)} \right] \\ &\lesssim (2\sqrt{2})^L \cdot \sqrt{d \cdot L \cdot \ln(N/d)}. \end{aligned}$$

This leads to the following result.

Theorem 2.3.1 ([36, Theorem 3.3]). *There exist absolute constants $C, c > 0$ such that, if $N > d+2$ and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network with L hidden layers of width N with random weights and biases following Assumption 2.2.2, then*

$$(1) \quad \text{Lip}_2(\Phi) \leq C \cdot (3\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}$$

with probability at least $(1 - 2\exp(-d))((1 - 2\exp(-c \cdot N))_+)^L$ and

$$(2) \quad \mathbb{E} [\text{Lip}_2(\Phi)] \leq C \cdot (2\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}.$$

We compare the result above to a naive bound that can be obtained by directly using the submultiplicativity of the operator norm. Precisely, we get for an arbitrary input $x \in \mathbb{R}^d$ that

$$\begin{aligned} \|\bar{\nabla} \Phi(x)\|_2 &\leq \|W^{(0)}\|_{2 \rightarrow 2} \cdot \underbrace{\|D^{(0)}(x)\|_{2 \rightarrow 2}}_{\leq 1} \cdots \|W^{(L-1)}\|_{2 \rightarrow 2} \cdot \underbrace{\|D^{(L-1)}(x)\|_{2 \rightarrow 2}}_{\leq 1} \cdot \|(W^{(L)})^T\|_2 \\ &\leq \|W^{(0)}\|_{2 \rightarrow 2} \cdots \|W^{(L-1)}\|_{2 \rightarrow 2} \cdot \|(W^{(L)})^T\|_2 \end{aligned}$$

and hence

$$\begin{aligned} \mathbb{E} [\text{Lip}_2(\Phi)] &\leq \mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] \leq \mathbb{E} [\|W^{(0)}\|_{2 \rightarrow 2}] \cdots \mathbb{E} [\|W^{(L-1)}\|_{2 \rightarrow 2}] \cdot \mathbb{E} [\|(W^{(L)})^T\|_2] \\ &\lesssim (2\sqrt{2})^L \cdot \sqrt{N}, \end{aligned} \tag{2.7}$$

where the last step uses [87, Thm. 7.3.1, Ex. 3.1.4], noting that $W^{(L)}$ has N entries. Theorem 2.3.1 improves this naive bound in the regime $N \gtrsim Ld \ln(N/d)$ by showing that the upper bound only depends logarithmically on the width N and not polynomially (as in (2.7)).

Moreover, the upper bound stated in Theorem 2.3.1 depends *exponentially* on the number of hidden layers L . However, in the given generality, where arbitrary distributions for the entries of the bias vectors are allowed, this exponential dependence is *unavoidable*, as the following theorem states. Note that Theorem 2.3.2 is a novel result which does not appear in either of the papers [23, 36].

Theorem 2.3.2. *There exist absolute constants $C, c > 0$ such that the following holds. For $d, N, L \in \mathbb{N}$ with $N \geq C \cdot \ln(eNL)$ and $d \geq C$ there exists a deterministic choice of bias vectors $b^{(0)}, \dots, b^{(L-1)} \in \mathbb{R}^N$ such that for a random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with L hidden layers of width N satisfying Assumption 2.2.2 with these constant biases and an arbitrary $x_0 \in \mathbb{S}^{d-1}$,*

$$\text{Lip}_2(\Phi) \geq \|\bar{\nabla} \Phi(x_0)\|_2 \geq (1.4)^L \cdot \sqrt{d}$$

with probability at least $1 - \exp(-c \cdot \min\{N, d\})$.

We prove Theorem 2.3.2 in Section A.1. We note that it has been known since 2020 that in the regime $N \gtrsim d^4 \ln^4(N)L$, the Lipschitz constant in the case of *zero-bias* networks following Assumption 2.2.2 can be upper bounded by $\sqrt{d \ln(N)}$ — *without* any dependence on the number of hidden layers L [12, Theorem B.5]. The exact dependence on L in the case of more general biases is further investigated in [23]; see also Section 2.4.

In the case of shallow networks, even the logarithmic dependence on the width N can be removed.

Theorem 2.3.3 ([36, Theorem 3.1]). *There exist absolute constants $C, c > 0$ such that the following holds. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network of arbitrary width $N \in \mathbb{N}$ with random weights and biases following Assumption 2.2.2, then*

$$(1) \quad \text{Lip}_2(\Phi) \leq C \cdot \sqrt{d}$$

with probability at least $(1 - 2 \exp(-\min\{d, N\})) \cdot (1 - 2 \exp(-c \cdot \max\{d, N\}))_+$ and

$$(2) \quad \mathbb{E}[\text{Lip}_2(\Phi)] \leq C \cdot \sqrt{d}.$$

Remarkably, Theorem 2.3.3 does not require any assumptions on N and d or on the distribution of the biases. Its proof works by bounding the covering numbers of $\mathcal{L}$ directly (without using the bound on the cardinality of $\mathcal{D}$) using a VC dimension based argument [87, Theorem 8.3.18].

2.3.2. The lower bound

In addition to providing upper bounds on the ℓ^2 -Lipschitz constant of random ReLU networks, [36] contains lower bounds on the Lipschitz constant as well. Again, the analysis is split into the case of shallow and the case of deep networks.

In the case of shallow networks, it is possible to show that for every *deterministic* shallow ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, the Lipschitz constant of Φ can be *lower* bounded by the Lipschitz constant of the corresponding *linear* network $\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., the network that arises from Φ by omitting the ReLU activation [36, Proposition 4.8]. More precisely,

$$\text{Lip}_2(\Phi) \geq \frac{1}{2} \cdot \text{Lip}_2(\tilde{\Phi}) \tag{2.8}$$

for every shallow ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$. Note that this is slightly counterintuitive: Since the ReLU itself is Lipschitz continuous with Lipschitz constant 1, one would rather expect the converse inequality $\text{Lip}_2(\Phi) \lesssim \text{Lip}_2(\tilde{\Phi})$. However, the latter inequality does not hold deterministically since there exist shallow ReLU networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ such that Φ is not constant, but $\tilde{\Phi}$ is; see the proof of [36, Proposition 4.7].

Equation (2.8) reduces the task of lower bounding the Lipschitz constant of a random shallow ReLU network to lower bounding the norm of the product of a Gaussian matrix and a Gaussian vector. This leads to the following result.

Theorem 2.3.4 ([36, Theorem 3.2]). *There exist absolute constants $c, c_1 > 0$ such that the following holds. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network of width N and with random weights and biases following Assumption 2.2.2, then*

$$(1) \quad \text{Lip}_2(\Phi) \geq \frac{1}{4\sqrt{2}} \cdot \sqrt{d}$$

with probability at least $(1 - 2 \exp(-c_1 \cdot N))_+ (1 - 2 \exp(-c_1 \cdot d))_+$ and

$$(2) \quad \mathbb{E}[\text{Lip}_2(\Phi)] \geq c \cdot \sqrt{d} \quad \text{if } d, N > \frac{\ln(2)}{c_1}.$$

Just like the upper bound for shallow networks in Theorem 2.3.3, the lower bound does *not* require any assumptions on N and d (except for $\min\{N, d\} \gtrsim 1$). Consequently, for random shallow ReLU networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ following Assumption 2.2.2 we overall get

$$\text{Lip}_2(\Phi) \asymp \sqrt{d}$$

with high probability and in expectation, where the implied constants are absolute.

The setting of deep networks is more difficult than studying shallow networks, since an inequality analogous to (2.8) does not hold. Indeed, if such an inequality were true for deep networks with arbitrary biases, then the ℓ^2 -Lipschitz constant of these networks would increase exponentially in the number of hidden layers L , which is known to be false in the case of zero biases [12]. Thus, lower bounding $\text{Lip}_2(\Phi)$ necessarily requires a different approach for deep networks than for shallow networks.

Following Theorem 2.2.3, a lower bound on $\text{Lip}_2(\Phi)$ can be obtained by lower bounding the Euclidean norm of the gradient of Φ at a single, fixed, non-zero input $x_0 \in \mathbb{R}^d \setminus \{0\}$. This leads to lower bounding the expression

$$\|\nabla \Phi(x_0)\|_2 = \left\| W^{(L)} D^{(L-1)}(x_0) W^{(L-1)} \dots D^{(0)}(x_0) W^{(0)} \right\|_2.$$

Under the assumption that all the entries of the bias vectors follow *symmetric* distributions, the following can be shown: For every $\ell \in \{0, \dots, L-1\}$, the matrix $\sqrt{N} D^{(\ell)}(x_0) W^{(\ell)}$ has independent, isotropic, and sub-gaussian rows (with bounded subgaussian norm), conditioned on $W^{(0)}, \dots, W^{(\ell-1)}, b^{(0)}, \dots, b^{(\ell-1)}$; see [36, Lemma 6.4]. An iterative application of the *matrix deviation inequality* [53, Theorem 3] under the additional assumption $N \gtrsim dL^2$ then gives us

$$\text{for all } y \in \mathbb{R}^d : \quad \left\| D^{(L-1)}(x_0) W^{(L-1)} \dots D^{(0)}(x_0) W^{(0)} y \right\|_2 \gtrsim \|y\|_2$$

with high probability; see [36, Proposition 6.5]. Note that this is equivalent to the smallest singular value of $V := D^{(L-1)}(x_0) W^{(L-1)} \dots D^{(0)}(x_0) W^{(0)}$ being larger than an absolute constant. Lastly, the final layer $W^{(L)}$ can be incorporated by using the singular value decomposition of V and the rotation invariance of a Gaussian random vector. This leads to the following theorem.

Theorem 2.3.5 ([36, Theorem 3.4]). *There exist absolute constants $C, c, c_1 > 0$ with the following property: If $N \geq C \cdot dL^2$ and if $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network with L hidden layers of width N and with random weights and symmetric biases following Assumption 2.2.2, then*

$$(1) \quad \text{Lip}_2(\Phi) \geq \frac{1}{4} \sqrt{d}$$

with probability at least $(1 - \frac{1}{2^N} - \exp(-N/(C \cdot L^2)))^L \cdot (1 - 2 \exp(-c_1 \cdot d))_+$ and

$$(2) \quad \mathbb{E} [\text{Lip}_2(\Phi)] \geq c \cdot \sqrt{d} \quad \text{if } N \geq C \cdot L^2 \ln(4L) \text{ and } d > \frac{\ln(2)}{c_1}.$$

We emphasize that Theorem 2.3.5 requires the biases to be drawn from symmetric distributions, which is not assumed in Theorems 2.3.1, 2.3.3 and 2.3.4. Note, however, that this assumption is quite natural, since it includes common initialization choices such as the zero initialization or a Gaussian initialization with zero mean. Moreover, Theorem 2.3.5 requires the inequality $N \gtrsim dL^2$, which is not needed in Theorems 2.3.1, 2.3.3 and 2.3.4 either.

Combining Theorems 2.3.1 and 2.3.5, we get that under the assumption $N \gtrsim dL^2$,

$$\sqrt{d} \lesssim \text{Lip}_2(\Phi) \lesssim (3\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}$$

with high probability and in expectation. Hence, the upper and lower bounds established in [36] in the case of deep networks differ by a factor that is exponential in the number of hidden layers L but only logarithmic in the width N . This gap between upper and lower bounds is further investigated and in fact decreased in [23]. Moreover, [23] provides lower bounds for the Lipschitz constant in the case of a possibly non-symmetric distribution of the biases; see Section 2.4.

2.4. Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks

The paper *Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks* [23], co-authored with Sjoerd Dirksen, Patrick Finke, Dominik Stöger, and Felix Voigtlaender, improves the results from [36] by

- (1) removing the exponential dependence on the number of hidden layers L in the upper bound for the case of zero-bias networks and for the case of networks with biases following symmetric and absolutely continuous distributions with bounded probability density function,
- (2) extending the analysis to the study of the ℓ^p -Lipschitz constants $\text{Lip}_p(\Phi)$ for every $p \in [1, \infty]$.

In the following, we discuss the paper and its main results in some depth. However, since [23] already contains proof sketches, we provide only a very high-level overview of the proofs.

2.4.1. Pointwise estimates

The main idea to estimate the Lipschitz constant in [23] is again to translate it to the norms of the formal gradient using Theorem 2.2.3. Therefore, the analysis in [23] starts with a *pointwise* estimation of the norm of the gradient of random ReLU networks. We recall that in [36] it is shown that for fixed $x_0 \in \mathbb{R}^d$, in the case of symmetrically distributed biases, and under the condition $N \gtrsim dL^2$,

$$\|\bar{\nabla}\Phi(x_0)\|_2 \gtrsim \sqrt{d}$$

with high probability. In [23], this result is extended to arbitrary ℓ^p -norms; see Theorem 2.4.1 below.

The proof in [36] for $p = 2$ makes heavy use of the matrix deviation inequality [53, Theorem 3]. This approach, however, is quite tailored to the ℓ^2 -norm as this norm appears in the matrix deviation inequality. The proof of Theorem 2.4.1 thus requires a different technique. It uses that, for a fixed input $x_0 \in \mathbb{R}^d$, it is in distribution possible to replace the matrices $D^{(\ell)}(x_0)$ that appear in the definition of the formal gradient (see (2.5)) by *independent* diagonal matrices $\bar{D}^{(\ell)} \in \mathbb{R}^{N \times N}$ where the entries on the diagonal are i.i.d. $\text{Unif}\{0, 1\}$ -distributed; see [23, Proposition 2.11]. One can then estimate the norm of the formal gradient by conditioning on the matrices $\bar{D}^{(\ell)}$ and by computing the norm of the matrix product arising that way iteratively. This leads to the following result.

Theorem 2.4.1 ([23, Theorem 3.4]). *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases, with L hidden layers of width N as in Assumption 2.2.2, with $N \geq C \cdot L^2 \cdot \ln(eL)$, and $d \geq C$ and let $x_0 \in \mathbb{R}^d \setminus \{0\}$ be fixed.*

- (1) *If $q \in [1, 2]$, we have*

$$\mathbb{P}\left(c \cdot d^{1/p} \leq \|\bar{\nabla}\Phi(x_0)\|_q \leq C \cdot d^{1/q}\right) \geq 1 - \exp(-c \cdot \min\{d, N/L^2\}).$$

(2) If $q \in (2, c \cdot \ln(d))$, we have

$$\mathbb{P}\left(c \cdot \sqrt{q} \cdot d^{1/q} \leq \|\bar{\nabla}\Phi(x_0)\|_q \leq C \cdot \sqrt{q} \cdot d^{1/q}\right) \geq 1 - \exp(-c \cdot \min\{q \cdot d^{2/q}, N/L^2\}).$$

(3) If $q \geq \ln(d)$, we have

$$\mathbb{P}\left(c \cdot \sqrt{\ln(d)} \leq \|\bar{\nabla}\Phi(x_0)\|_q \leq C \cdot \sqrt{\ln(d)}\right) \geq 1 - \exp(-c \cdot \min\{\ln(d), N/L^2\}).$$

Notably, the behavior of $\|\bar{\nabla}\Phi(x_0)\|_q$ is equivalent to that of $\|g\|_q$ for a random vector $g \in \mathbb{R}^d$ with $g \sim \mathcal{N}(0, I_d)$.

By combining Theorems 2.2.3 and 2.4.1, we immediately obtain that for fixed $p \in [1, \infty]$,

$$\text{Lip}_p(\Phi) \gtrsim \begin{cases} d^{1-1/p} & \text{if } p \in [2, \infty], \\ \sqrt{\frac{p}{p-1}} \cdot d^{1-1/p} & \text{if } 2 < \frac{p}{p-1} < c \cdot \ln(d), \\ \sqrt{\ln(d)} & \text{if } \frac{p}{p-1} \geq \ln(d), \end{cases} \quad (2.9)$$

with high probability, where $c > 0$ is an absolute constant. We recall from [36] that in the case of shallow networks and $p = 2$, the pointwise estimate already provides a *sharp* estimate on the Lipschitz constant, since $\text{Lip}_2(\Phi) \asymp \sqrt{d}$ in that case. One might be tempted to think that the same behavior is true for every $p \in [1, \infty]$, i.e., that

$$\text{Lip}_p(\Phi) \asymp \begin{cases} d^{1-1/p} & \text{if } p \in [2, \infty], \\ \sqrt{\frac{p}{p-1}} \cdot d^{1-1/p}, & \text{if } 2 < \frac{p}{p-1} < c \cdot \ln(d), \\ \sqrt{\ln(d)} & \text{if } \frac{p}{p-1} \geq \ln(d). \end{cases}$$

Surprisingly, however, further investigations show that there is a significant difference between the regimes $p \in [1, 2)$ and $p \in [2, \infty]$.

2.4.2. The upper bound in the case of zero-bias networks

In order to establish an upper bound for $\text{Lip}_p(\Phi)$ in the case of zero-bias networks, Theorem 2.2.3 shows that it suffices to bound

$$\sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla}\Phi(x)\|_{p'},$$

where $1/p + 1/p' = 1$, i.e., it suffices to consider inputs from the unit sphere instead of taking inputs from the entire input space. The boundedness of the sphere enables a *covering-based* approach. This means that, as the first step, we bound the norm of the gradient on an ε -net of $\mathbb{S}^{d-1}$. As the second step, we then control the deviation of that norm when moving from an arbitrary point on the sphere to a close net point *uniformly* over all inputs from the sphere. A similar approach was already used in [12] to show the upper bound of $\sqrt{d \ln(N)}$ in the case $p = 2$.

In [23], we, at first, focus on the case $p = 2$ too. We can use Theorem 2.4.1(1) together with a union bound to establish a bound on the ℓ^2 -norm of the formal gradient on an ε -net. The deviation of the gradients when moving from an arbitrary point on $\mathbb{S}^{d-1}$ to a close net point can be controlled using results on *random hyperplane tessellations* that first appeared in [63]; see [23, Proposition 2.13]. In the end, we use the standard inequalities

$$\|\cdot\|_2 \leq \|\cdot\|_p \quad \text{for } p \in [1, 2] \quad \text{and} \quad d^{1/p-1/2} \|\cdot\|_2 \leq \|\cdot\|_p \quad \text{for } p \in [2, \infty] \quad (2.10)$$

to obtain the following result.

Theorem 2.4.2 ([23, Corollary 4.2]). *There exist absolute constants $C, c > 0$ satisfying the following. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random zero-bias ReLU network with L hidden layers of width N as in Assumption 2.2.2 and if we assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot (\ln^2(N/d) \cdot L^3 + \ln^3(N/d)L^2),$$

then, with probability at least $1 - \exp(-c \cdot d \ln(N/d))$,

$$\text{Lip}_p(\Phi) \leq C \cdot \begin{cases} \sqrt{d \cdot \ln(N/d)} & \text{if } p \in [1, 2], \\ d^{1-1/p} \cdot \sqrt{\ln(N/d)} & \text{if } p \in [2, \infty], \end{cases} \quad \text{for all } p \in [1, \infty].$$

This result reproduces the bound of [12, Theorem B.5] in the regime $N \gtrsim dL^3$ (up to logarithmic factors), whereas [12, Theorem B.5] holds in the setting $N \gtrsim d^4L$ (up to logarithmic factors). In the natural regime $L \ll d$, our bound thus holds for a larger set of parameters.

Since a zero-bias network in particular has symmetrically distributed biases, we can compare these upper bounds to the lower bounds from (2.9). We notice that, in the case $p \in [2, \infty]$, the upper bound obtained in Theorem 2.4.2 is, up to a logarithmic factor, at par with the lower bound obtained in (2.9). Thus, in the regime $p \in [2, \infty]$, taking the supremum over the entire input space only results in an additional logarithmic factor compared to only considering the norm at a single fixed point. In the case $p \in [1, 2)$, on the other hand, there is a significant gap between the two bounds. Most notably, in the case $p = 1$, we observe the lower bound of $\sqrt{\ln(d)}$ and the upper bound of $\sqrt{d \ln(N/d)}$. At first, this seems like a shortcoming of the proof technique of Theorem 2.4.2, as the proof relies on bounding the ℓ^2 -Lipschitz constant and then using the standard inequalities in (2.10). Surprisingly, however, it turns out that the upper bound of $\sqrt{d \ln(N/d)}$ in the regime $p \in [1, 2)$ is indeed sharp up to a factor of $\sqrt{L \ln(N/d)}$, which will be discussed in the next section.

2.4.3. The lower bound in the case of zero-bias networks

In this section, we consider a lower bound on the ℓ^1 -Lipschitz constant of zero-bias networks, which directly implies the same lower bound for every $p \in [1, 2)$ since $\|\cdot\|_1 \geq \|\cdot\|_q$ for every $q \in [1, \infty]$. As we recall from Theorem 2.2.3, we have

$$\text{Lip}_1(\Phi) = \sup_{x \in \mathbb{S}^{d-1} \cap M_\Phi} \|\bar{\nabla} \Phi(x)\|_\infty$$

with $M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla \Phi(x) = \bar{\nabla} \Phi(x)\}$. Similarly to the approach in Section 2.3.1, we can rewrite

$$\begin{aligned} \sup_{x \in \mathbb{S}^{d-1} \cap M_\Phi} \|\bar{\nabla} \Phi(x)\|_\infty &\geq \sup_{x \in \mathbb{S}^{d-1} \cap M_\Phi} \langle \bar{\nabla} \Phi(x), e_1 \rangle \\ &= \sup_{x \in \mathbb{S}^{d-1} \cap M_\Phi} \left\langle (W^{(L)})^T, D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}e_1 \right\rangle, \end{aligned}$$

with e_1 denoting the first standard basis vector in $\mathbb{R}^d$. Hence, by conditioning on the weights $W^{(0)}, \dots, W^{(L-1)}$ and only considering the randomness over $W^{(L)}$, we obtain

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{S}^{d-1} \cap M_\Phi} \|\bar{\nabla} \Phi(x)\|_\infty \right] \geq w(\mathcal{L}'),$$

where

$$\mathcal{L}' := \left\{ D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}e_1 : x \in \mathbb{S}^{d-1} \cap M_\Phi \right\}.$$

Reference	range	Condition on N, d, L	Type	Bound
[23, Corollary 3.6]	$p \in [2, \infty]$	$N \gtrsim L^2, d \gtrsim 1$	$\gtrsim$	$d^{1-1/p}$
[23, Corollary 4.2]	$p \in [2, \infty]$	$N \gtrsim dL^3$	$\lesssim$	$d^{1-1/p} \sqrt{\ln(N/d)}$
[23, Theorem 5.1]	$p \in [1, 2)$	$N \geq C^L \cdot dL^{4L+10}, d \gtrsim L$	$\gtrsim$	$\sqrt{d/L}$
[23, Corollary 4.2]	$p \in [1, 2)$	$N \gtrsim dL^3$	$\lesssim$	$\sqrt{d \ln(N/d)}$

Table 2.1.: Overview of the bounds from [23] in the case of zero-bias networks. $C > 0$ is an absolute constant. For convenience, we left out logarithmic factors in the third column.

By lower bounding the covering numbers or—equivalently—the packing numbers of $\mathcal{L}'$, we can derive a lower bound for $\text{Lip}_1(\Phi)$ in expectation using *Sudakov’s minoration inequality* [87, Corollary 7.4.3]. This can be translated into a high probability bound using a general result on Gaussian concentration [87, Theorem 5.2.2]. The computation of the packing numbers of $\mathcal{L}'$ is again done by using the random hyperplane tessellation results from [63] together with decoupling techniques. The details are contained in [23, Appendix F].

Theorem 2.4.3 ([23, Theorem 5.1]). *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N as in Assumption 2.2.2. Assume $N \geq C \cdot d$,*

$$N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10}, \quad \text{and} \quad d \geq C \cdot L.$$

Then, with probability at least $1 - \exp(-c \cdot d/L)$,

$$\text{Lip}_p(\Phi) \geq c \cdot \sqrt{d/L} \quad \text{for all } p \in [1, 2).$$

In fact, the bound also holds for every $p \in [2, \infty]$, but in that case, the stronger lower bound of $d^{1-1/p}$ was obtained by studying the norm of the gradient pointwise; see (2.9). Notably, this lower bound matches the upper bound obtained in Theorem 2.4.2 up to the factor $\sqrt{L \ln(N/d)}$. However, Theorem 2.4.3 requires the hidden layer width N to grow super-exponentially in the number of hidden layers L . It is an interesting question for future research to improve this condition and to prove an analogous lower bound if N is only allowed to grow polynomially in L .

We refer to Table 2.1 for an overview of our bounds in the case of zero-bias networks.

2.4.4. Bounds in the case of networks with biases

In this section, we discuss the results derived in [23] for the case of general distributions for the bias vectors. This situation is much more difficult than the zero-bias case, as it is no longer sufficient to only consider inputs from the unit sphere. Therefore, a covering-based approach is no longer possible to derive an upper bound.

Hence, the idea is to relate the gradient of a network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with possibly non-zero biases to the gradient of the associated zero-bias network $\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., the network that arises from Φ by setting all its biases to zero. Intuitively, for inputs $x \in \mathbb{R}^d$ with very large Euclidean norm (compared to the size of the biases), the influence of the bias term is small and we expect $\nabla \Phi(x)$ to be close to $\nabla \tilde{\Phi}(x)$. A formal result which confirms this intuition is established in [23, Theorem 6.2].

In order to prove an upper bound for the ℓ^2 -Lipschitz constant of Φ , we use the decomposition

$$\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \leq \sup_{x \in R \cdot B^d} \|\bar{\nabla} \Phi(x)\|_2 + \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 + \sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \tilde{\Phi}(x)\|_2,$$

for an appropriate choice of $R > 0$. The second summand can be bounded using [23, Theorem 6.2] and the third using the already established bound for zero-bias networks; see Theorem 2.4.2. For the first summand, one can apply an approach using an ε -net, similar to the proof of Theorem 2.4.2. However, since the bias terms are now non-zero, the hyperplane tessellation results from [63] are no longer sufficient. Instead, we generalize results from [24] under the condition that the biases follow absolutely continuous distributions with bounded probability density function; see [23, Proposition 2.15]. The main result then reads as follows.

Theorem 2.4.4 ([23, Corollary 6.4]). *There exist absolute constants $C, c > 0$ with the following property: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases and with L hidden layers of width N as in Assumption 2.2.2. We assume that all the biases follow (possibly different) absolutely continuous distributions with probability density function bounded by $C_\tau > 0$ almost everywhere. Moreover, let*

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \lambda \right) \geq 1 - \exp(-d \ln(N/d))$$

for some $\lambda \geq C_\tau^{-1}$. We assume that $N \geq C \cdot d$,

$$N \geq C \cdot dL^2 \cdot (\max\{L, \ln(\lambda C_\tau N/d)\})^2 \cdot \ln(N/d), \quad \text{and} \quad N \geq C \cdot dL^3 \ln^2(N/d) \ln^2(N/(dL)).$$

Then, with probability at least $1 - \exp(-c \cdot d \ln(N/d))$,

$$\text{Lip}_p(\Phi) \leq C \cdot \begin{cases} \sqrt{d} \cdot \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right) & \text{if } p \in [1, 2], \\ d^{1-1/p} \cdot \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right) & \text{if } p \in [2, \infty], \end{cases} \quad \text{for all } p \in [1, \infty].$$

We notice the following two aspects.

- Compared to the upper bound in the zero-bias case (see Theorem 2.4.2), we observe an additional summand $\sqrt{L}$. This results from the fact that R (which is needed to relate the gradient of Φ to that of its zero-bias counterpart) grows exponentially in L . It is an interesting question for future research to improve the dependence on L .
- Moreover, we obtain an additional factor λC_τ inside of the logarithm, which is related to the exact distribution of the biases. Remarkably, λC_τ is scale-invariant, i.e., this quantity remains the same if one scales the biases with a positive factor $\alpha > 0$. Hence, Theorem 2.4.4 is independent of the scaling of the biases.

Previously stated lower bounds require the biases to be either zero or to be drawn from symmetric distributions; see Theorems 2.4.1 and 2.4.3. Using the fact that we can approximate the gradient of a network by the gradient of the associated homogeneous network for large inputs [23, Theorem 6.2], we can prove lower bounds on the Lipschitz constant of networks with *arbitrary* distributions for the biases. Precisely, we obtain the following result.

Theorem 2.4.5 ([23, Theorems 6.6 & 6.7]). *There exist absolute constants $C, c > 0$ with the following property. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N as in Assumption 2.2.2. We assume that $N \geq C \cdot dL$.*

Reference	range	Condition on biases	Type	Bound
[23, Theorem 6.7]	$p \in [2, \infty]$	arbitrary	$\gtrsim$	$d^{1-1/p}$
[23, Corollary 6.4]	$p \in [2, \infty]$	bounded pdf + symmetric	$\lesssim$	$d^{1-1/p} \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right)$
[23, Theorem 6.6]	$p \in [1, 2)$	arbitrary	$\gtrsim$	$\sqrt{d/L}$
[23, Corollary 6.4]	$p \in [1, 2)$	bounded pdf + symmetric	$\lesssim$	$\sqrt{d} \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right)$

Table 2.2.: Overview of the bounds from [23] in the case of networks with biases. For reasons of space, we do not state the required assumptions on N, d, L here, but refer for these to the individual references. The quantities λ and C_τ are related to the distribution of the biases.

(1) If $N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10}$, $N \geq C \cdot d \cdot L^5 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL))$, and $d \geq C \cdot L$, then, with probability at least $1 - \exp(-c \cdot d/L)$,

$$\text{Lip}_p(\Phi) \geq c \cdot \sqrt{d/L} \quad \text{for every } p \in [1, 2).$$

(2) If $N \geq C \cdot d \cdot L^4 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL))$ and $d \geq C$, then, with probability at least $1 - \exp(-c \cdot d)$,

$$\text{Lip}_p(\Phi) \geq c \cdot d^{1-1/p} \quad \text{for all } p \in [2, \infty].$$

We refer to Table 2.2 for an overview of the results from [23] in the case of general bias distributions. Notably, in the case of a *shallow* random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfying Assumption 2.2.2, we obtain bounds that match up to an absolute numerical constant. Precisely, by combining Theorems 2.3.3 and 2.4.5 together with the standard inequalities (2.10), we obtain under the assumptions $N \gtrsim d \ln^4(N/d)$ and $N \gtrsim d \gtrsim 1$ that with high probability,

$$\text{Lip}_p(\Phi) \asymp \begin{cases} d^{1-1/p} & \text{if } p \in [2, \infty], \\ \sqrt{d} & \text{if } p \in [1, 2], \end{cases} \quad \text{for all } p \in [1, \infty];$$

see [23, Corollary 6.8].

3. Approximation properties of complex-valued neural networks

This chapter is based on the following three papers.

- (1) “Optimal approximation using complex-valued neural networks” (co-authored with Felix Voigtlaender), published in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* in 2023, henceforth referred to as [37];
- (2) “Universal approximation using complex-valued deep narrow neural networks” (co-authored with Thomas Jahn and Hannes Matt), published in *Constructive Approximation* in 2025, henceforth referred to as [34];
- (3) “On best approximation using multivariate ridge functions with applications to generalized translation networks” (co-authored with Palina Salanevich, Olov Schavemaker, and Felix Voigtlaender), published in *Mathematical Foundations of Machine Learning* in 2025, henceforth referred to as [35].

Further, we refer to [33] for an overview paper of the approximation properties of complex-valued neural networks which in particular discusses the results from [34, 35, 37].

3.1. Complex-valued neural networks

The overwhelming success of deep learning in the last decade is based on the use of *real-valued* network architectures, i.e., architectures where all the used parameters are real-valued and where the activation functions are functions from $\mathbb{R}$ to $\mathbb{R}$. However, a glance at the literature reveals a large number of scientific papers, including very recent ones, in which *complex-valued* models are used, i.e., models in which the parameters used are complex numbers and in which the activation functions are also mappings from $\mathbb{C}$ to $\mathbb{C}$ [2, 4, 51, 70, 80, 81, 92, 97, 99]. The application areas where such complex-valued architectures are used are often characterized by the fact that the inputs that are fed into a machine learning model are *inherently* complex-valued, i.e., contain both a phase and a magnitude information.

One of these areas is *magnetic resonance imaging* (MRI) [17, 71, 85, 89]. Since the functionality of MRI is based on the generation of a magnetic field, the signals occurring in this context are naturally complex-valued, motivating the use of complex-valued deep learning models. In MRI, complex-valued models are mainly used in order to reconstruct data. While both [17, 71] propose complex-valued convolutional neural networks to tackle this challenge, [85] introduces a complex-valued generative adversarial network. In [89], on the other hand, complex-valued neural networks are used to estimate tissue parameters from MRI signals through a supervised deep learning approach.

We recall from Definition 1.1.1 that a CVNN is of the form

$$V^{(L)} \circ \varrho \circ V^{(L-1)} \circ \dots \circ \varrho \circ V^{(1)} \circ \varrho \circ V^{(0)} : \mathbb{C}^{N_0} \rightarrow \mathbb{C}^{N_{L+1}},$$

where $V^{(\ell)} : \mathbb{C}^{N_\ell} \rightarrow \mathbb{C}^{N_{\ell+1}}$ is a $\mathbb{C}$ -affine map for every $\ell \in \{0, \dots, L\}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ is the activation function. Upon identifying $\mathbb{C} \cong \mathbb{R}^2$, we may view $V^{(\ell)}$ as an $\mathbb{R}$ -affine map from $\mathbb{R}^{2N_\ell}$ to $\mathbb{R}^{2N_{\ell+1}}$. However, note crucially that the real-valued weight matrix $W^{(\ell)} \in \mathbb{R}^{(2N_{\ell+1}) \times (2N_\ell)}$ that

is associated to (the real-valued version of) $V^{(\ell)}$ is restricted to a specific structure since $V^{(\ell)}$ is $\mathbb{C}$ -affine. More precisely, $W^{(\ell)}$ consists of 2×2 -blocks, each of which is of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

with $a, b \in \mathbb{R}$.

On the other hand, while the activation function in an RVNN is a function $\mathbb{R} \rightarrow \mathbb{R}$, the activation function in a CVNN maps from $\mathbb{C}$ to $\mathbb{C}$ and may thus be viewed as a function which acts on *two* real neurons simultaneously. Hence, while CVNNs are more restrictive regarding the affine maps (as illustrated above), more versatility is allowed through the activation function. For example, many works make use of *phase-preserving* complex activation functions, i.e., functions $\phi : \mathbb{C} \rightarrow \mathbb{C}$ that satisfy

$$\phi(e^{i\theta}z) = e^{i\theta}\phi(z) \quad \text{for all } z \in \mathbb{C}, \theta \in [0, 2\pi).$$

This behavior cannot be achieved if a non-trivial real-valued activation function is applied to real and imaginary parts of a complex input separately. Two popular examples of such activation functions are the modReLU [2] and the complex cardioid [89]. For a parameter $b < 0$, the modReLU with parameter b is defined as

$$\text{modReLU}_b : \mathbb{C} \rightarrow \mathbb{C}, \quad \text{modReLU}_b(z) = \begin{cases} (|z| + b) \frac{z}{|z|} & \text{if } |z| + b \geq 0, \\ 0 & \text{if } |z| + b < 0. \end{cases} \quad (3.1)$$

The complex cardioid is defined as

$$\text{card} : \mathbb{C} \rightarrow \mathbb{C}, \quad \text{card}(z) = \frac{1}{2}(1 + \cos(\angle z))z, \quad (3.2)$$

where $\angle z$ denotes the polar angle of a complex number $z \in \mathbb{C} \setminus \{0\}$ and we set $\angle 0 := 0$.

3.2. Approximation theory of neural networks

In this section, we first contextualize the research on approximation properties of neural networks in the setting of a deeper theoretical understanding of machine learning and in particular deep learning. Then, we give a quick overview of the most fundamental approximation-theoretic results regarding neural networks.

With the notation from Section 1.1.3, we let

$$h_S := \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) \quad \text{and} \quad h_{\mathcal{D}}^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}} R_{\mathcal{D}}(h),$$

i.e., h_S minimizes the empirical risk among all hypotheses from $\mathcal{H}$ and $h_{\mathcal{D}}^*$ minimizes the expected risk among all hypotheses from $\mathcal{H}$. Here, we again assume for simplicity that both minima exist. The expected risk of h_S can then be decomposed into

$$R_{\mathcal{D}}(h_S) = \underbrace{R_{\mathcal{D}}(h_S) - R_{\mathcal{D}}(h_{\mathcal{D}}^*)}_{\text{estimation error}} + \underbrace{R_{\mathcal{D}}(h_{\mathcal{D}}^*)}_{\text{approximation error}}; \quad (3.3)$$

see also [74, Eq. 5.7]. The term $R_{\mathcal{D}}(h_S) - R_{\mathcal{D}}(h_{\mathcal{D}}^*)$ is called the *estimation error*, as it describes how well the expected risk of the empirical risk minimizer serves as an *estimate* for the optimally achievable expected risk over all hypotheses from $\mathcal{H}$. The second term $R_{\mathcal{D}}(h_{\mathcal{D}}^*) = \min_{h \in \mathcal{H}} R_{\mathcal{D}}(h)$, on the other hand, is called the *approximation error*, as it describes how well the relationship between inputs from $\mathcal{X}$ and outputs from $\mathcal{Y}$ that is given by the distribution $\mathcal{D}$ can *in theory* be

approximated by elements of $\mathcal{H}$. As the title of the present chapter suggests, we focus on the *approximation error* in this chapter.

However, the research in approximation theory of neural networks focuses less on the approximation error as considered in (3.3) (which among other things depends on the distribution $\mathcal{D}$) but is rather based on a purely functional analytic perspective. It generally focuses on the question of how well a given function $f : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ can be approximated by neural networks of a certain architecture, with a certain activation function, etc. This approximation is usually measured with respect to L^p -norms or in the sense of (locally) uniform convergence. In the following, we illustrate how this functional analytic approach may lead to a better understanding of the approximation error considered in (3.3).

As in Section 2.1, we consider the special case $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. We let

$$f^* := \operatorname{argmin}_{f: \mathcal{X} \rightarrow \mathcal{Y}} R_{\mathcal{D}}(f)$$

be a function from $\mathcal{X}$ to $\mathcal{Y}$ which minimizes the expected risk $R_{\mathcal{D}}(\cdot)$, again assuming for simplicity that such a function does exist. Note that f^* does not necessarily belong to the hypothesis class $\mathcal{H}$. Similar to Section 2.1, let us assume that for some $\lambda > 0$,

$$|\ell(y_1, y_0) - \ell(y_2, y_0)| \leq \lambda \cdot |y_1 - y_2| \quad \text{for all } y_0 \in \operatorname{supp}(\mathcal{D}^{\mathcal{Y}}) \text{ and } y_1, y_2 \in \mathbb{R}.$$

Since the approximation error is by definition lower bounded by $R_{\mathcal{D}}(f^*)$, we can also, while studying the approximation error, study the difference $R_{\mathcal{D}}(h^*) - R_{\mathcal{D}}(f^*)$ and get, for arbitrary $h \in \mathcal{H}$,

$$\begin{aligned} R_{\mathcal{D}}(h^*) - R_{\mathcal{D}}(f^*) &\leq R_{\mathcal{D}}(h) - R_{\mathcal{D}}(f^*) \leq \mathbb{E}_{(x,y) \sim \mathcal{D}} [|\ell(h(x), y) - \ell(f^*(x), y)|] \\ &\leq \lambda \cdot \mathbb{E}_{x \sim \mathcal{D}^{\mathcal{X}}} [|h(x) - f^*(x)|] \\ &\leq \lambda \cdot \sup_{x \in \operatorname{supp}(\mathcal{D}^{\mathcal{X}})} |h(x) - f^*(x)|, \end{aligned}$$

where $\mathcal{D}^{\mathcal{X}}$ denotes the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$. Thus, if for a given $\varepsilon > 0$ it is known that there exists $h \in \mathcal{H}$ with

$$\sup_{x \in \operatorname{supp}(\mathcal{D}^{\mathcal{X}})} |h(x) - f^*(x)| \leq \frac{\varepsilon}{\lambda}, \quad (3.4)$$

we get

$$R_{\mathcal{D}}(h^*) \leq R_{\mathcal{D}}(f^*) + \varepsilon$$

and thus, the approximation error as considered in (3.3) lies within ε -distance of the optimally achievable expected risk $R_{\mathcal{D}}(f^*)$. If $\operatorname{supp}(\mathcal{D}^{\mathcal{X}})$ is compact and f^* is continuous, error terms as in (3.4) are studied in the context of *universality*; see below for more details.

We note that the minimizer of the empirical risk can often not be found directly, but must be approximated using a numerical algorithm. In the case of neural networks, this algorithm is usually a gradient-based method such as stochastic gradient descent. Hence, in addition to the approximation and the estimation error, researchers also study a third error term, namely the *optimization error*, which captures the difference between the output of the numerical algorithm and the actual empirical risk minimizer.

The investigation of the approximation-theoretic properties of (real-valued) neural networks dates back to the late 1980s and early 1990s [19, 52]. The research back then focused, in particular, on the aspect of *universality*, i.e., the capability of neural networks to approximate any given continuous function up to arbitrary precision on compact sets with respect to the L^∞ -norm. Specifically, in [52] it is shown that the set of neural networks with a fixed number of hidden layers (but with unrestricted width) and with fixed activation function $\sigma \in C(\mathbb{R}; \mathbb{R})$ is universal if and only if σ is not a polynomial. This statement is nowadays known as the *universal approximation theorem*.

(*UAT*) and is often viewed as a fundamental theoretical backbone of the success of deep learning, since it states that, regarding the aspect of universality, almost every activation function is a good choice as it gives rise to universal neural networks. After the question of universality was mostly settled, the approximation-theoretic properties of neural networks were investigated from a *quantitative* perspective. Already in 1993, Barron introduced a class of functions, nowadays called the Barron class, with the property that the approximation of such functions using shallow neural networks does *not* suffer from the so-called *curse of dimensionality* [5]. This means that the number of hidden neurons needed to realize a certain approximation accuracy only grows polynomially (at most) with the input dimension. In 1996, Mhaskar examined how large a shallow network with smooth and non-polynomial activation function has to be in order to guarantee a certain approximation accuracy when approximating Sobolev functions [58]; see Section 3.3 for more details. Due to its popularity in practical applications, more recent research mostly deals with networks that employ the ReLU activation function [68, 94, 95]. Remarkably, the classical results by Barron and Mhaskar are *not* applicable to the ReLU. Since it is common to use networks with a large number of hidden layers in practical applications, it is of specific interest to examine in which way such ReLU networks have advantages compared to networks with fewer hidden layers.

The previous paragraph only gives a very brief overview of the approximation theory of neural networks. We refer to the introductions of [34, 35, 37] for more details. It is noteworthy that the theory is almost exclusively concerned with the study of *real-valued* neural networks and the complex-valued case remained largely unexplored for quite some time. Significantly, the UAT was only proved for complex-valued neural networks in 2023 [90]. It was shown that the set of shallow CVNNs with fixed activation function $\phi \in C(\mathbb{C}; \mathbb{C})$ (but with unrestricted width) is universal if and only if ϕ is not *polyharmonic*. Here, ϕ is called polyharmonic, if $\phi \in C^\infty(\mathbb{C}; \mathbb{C})$ and if there exists $m \in \mathbb{N}_0$ satisfying $\Delta^m \phi \equiv 0$, where $\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplace operator on $\mathbb{C} \cong \mathbb{R}^2$. Furthermore, the set of CVNNs with a fixed number of hidden layers $L \geq 2$ and with fixed activation function $\phi \in C(\mathbb{C}; \mathbb{C})$ is universal if and only if

- (1) ϕ is not holomorphic, and
- (2) $\bar{\phi}$ is not holomorphic, and
- (3) ϕ is not a polynomial in z and $\bar{z}$.

Remarkably, this is a richer class of functions than the set of non-polyharmonic functions which gives rise to universality in the case of shallow CVNNs. The UAT hence shows that there is a qualitative difference between the approximation-theoretic properties of real-valued and complex-valued neural networks since in the real-valued case, there exists one characterization of “good” activation functions that is valid for every number of hidden layers $L \in \mathbb{N}$. Motivated by this observation, the three papers [34, 35, 37] that this section is based on investigate in what way other classical approximation-theoretic statements of RVNNs can be transferred to the complex-valued setting and where differences exist.

3.3. Optimal approximation using complex-valued neural networks

The paper *Optimal approximation using complex-valued neural networks* [37], co-authored with Felix Voigtlaender, is concerned with the approximation of C^r -functions defined on the d -dimensional complex cube

$$\Omega_d := [-1, 1]^d + i \cdot [-1, 1]^d \subseteq \mathbb{C}^d$$

by shallow CVNNs with n hidden neurons and fixed activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$. Precisely, we seek to bound the *worst-case approximation error*, i.e.,

$$\sup_{f \in C^r(\Omega_d; \mathbb{C}), \|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \inf_{\Phi \in \mathcal{NN}_{d,n}^\phi} \|f - \Phi\|_{L^\infty(\Omega_d; \mathbb{C})}, \quad (3.5)$$

where $\mathcal{NN}_{d,n}^\phi$ denotes the set consisting of all shallow CVNNs with fixed activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$, d input neurons, a single output neuron and with n hidden neurons. A specific focus in [37] lies on the analysis of the setting of *continuous weight selection*. To describe this setting it is convenient to rewrite (3.5) as

$$\sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \inf_{\Phi \in \mathcal{NN}_{d,n}^\phi} \|f - \Phi\|_{L^\infty(\Omega_d; \mathbb{C})} = \inf_{\Theta : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathcal{NN}_{d,n}^\phi} \sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \|f - \Theta(f)\|_{L^\infty(\Omega_d; \mathbb{C})},$$

which follows from an elementary observation.¹ Moreover, each element of $\mathcal{NN}_{d,n}^\phi$ has $n(d+1)$ weights and n biases (we leave out the bias term in the last layer). Therefore, we can further rewrite

$$\begin{aligned} & \inf_{\Theta : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathcal{NN}_{d,n}^\phi} \sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \|f - \Theta(f)\|_{L^\infty(\Omega_d; \mathbb{C})} \\ &= \inf_{\text{enc} : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathbb{C}^{n(d+2)}} \sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \|f - \text{dec}(\text{enc}(f))\|_{L^\infty(\Omega_d; \mathbb{C})}, \end{aligned} \quad (3.6)$$

where $\text{dec} : \mathbb{C}^{n(d+2)} \rightarrow \mathcal{NN}_{d,n}^\phi$ denotes the *decoder map*, i.e., the map that assigns to the parameters the corresponding neural network function. In the setting of continuous weight selection, we do *not* study the quantity described in (3.6) but instead, we restrict the infimum to maps $\text{enc} : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathbb{C}^{n(d+2)}$ that are continuous with respect to some norm on $C^r(\Omega_d; \mathbb{C})$. This restriction is quite natural, since classical learning algorithms such as (stochastic) gradient descent are based on training samples $f(x_j)$ and apply continuous operations to these samples to update the network weights. Note that the resulting quantity is possibly larger than the worst-case approximation error from (3.5).

In the case of RVNNs, this problem was studied by Mhaskar in [58]. Precisely, it was shown that in the real-valued setting the worst-case approximation error can, under the restriction of continuous weight selection, be upper bounded by $n^{-r/d}$ up to a multiplicative constant that only depends on the input dimension d and the regularity r , provided that the used activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is both smooth and non-polynomial on some non-empty open interval. Moreover, [58] establishes that the approximation rate of $n^{-r/d}$ can even be achieved under the restriction that the weights connecting the input layer and the hidden layer are chosen *independently* of the particular function f which is to be approximated. Only the weights connecting the hidden layer neurons and the output neurons have to be adapted to f .

The main result from [37] translates these findings to the complex-valued setting. It reads as follows.

¹For an arbitrary map $\Theta : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathcal{NN}_{d,n}^\phi$ it is clear that

$$\sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \inf_{\Phi \in \mathcal{NN}_{d,n}^\phi} \|f - \Phi\|_{L^\infty(\Omega_d; \mathbb{C})} \leq \sup_{\|f\|_{C^r(\Omega_d; \mathbb{C})} \leq 1} \|f - \Theta(f)\|_{L^\infty(\Omega_d; \mathbb{C})}$$

so that we obtain “ $\leq$ ” by taking the infimum over all maps $\Theta : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathcal{NN}_{d,n}^\phi$. Conversely, for arbitrary $\varepsilon > 0$ we can pick a map $\Theta_\varepsilon : C^r(\Omega_d; \mathbb{C}) \rightarrow \mathcal{NN}_{d,n}^\phi$ with

$$\|f - \Theta_\varepsilon(f)\|_{L^\infty(\Omega_d; \mathbb{C})} \leq \inf_{\Phi \in \mathcal{NN}_{d,n}^\phi} \|f - \Phi\|_{L^\infty(\Omega_d; \mathbb{C})} + \varepsilon \quad \text{for all } f \in C^r(\Omega_d; \mathbb{C}).$$

Since $\varepsilon > 0$ was arbitrary, this implies “ $\geq$ ”.

Theorem 3.3.1 ([37, Theorem 3.2]). *Let $d, r \in \mathbb{N}$. Then there exists a constant $C = C(d, r) > 0$ with the following property: For any activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ that is smooth and non-polyharmonic on an open set $\emptyset \neq U \subseteq \mathbb{C}$ and for any $n \in \mathbb{N}$ there exist $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$, and $\beta_1, \dots, \beta_n \in \mathbb{C}$ such that for every $f \in C^r(\Omega_d; \mathbb{C})$ there exist $\gamma_1 = \gamma_1(f), \dots, \gamma_n = \gamma_n(f) \in \mathbb{C}$ such that*

$$\left\| f(z) - \sum_{j=1}^n \gamma_j \phi(\alpha_j^T z + \beta_j) \right\|_{L^\infty(\Omega_d; \mathbb{C})} \leq C \cdot n^{-r/(2d)} \cdot \|f\|_{C^r(\Omega_d; \mathbb{C})}$$

and such that the map

$$C^r(\Omega_d; \mathbb{C}) \rightarrow \mathbb{C}^n, \quad f \mapsto (\gamma_1(f), \dots, \gamma_n(f))$$

is linear and continuous with respect to the L^∞ -norm.

We highlight the following aspects compared to the result from [58].

- (1) The rate of $n^{-r/(2d)}$ shown in the complex setting is *equivalent* to the rate of $n^{-r/d}$ obtained in the real setting in [58] by using the identification $\mathbb{C}^d \cong \mathbb{R}^{2d}$.
- (2) While Theorem 3.3.1 only considers the approximation of C^r -functions, the same result is expected to hold for elements of the Sobolev space $W^{r, \infty}$ as well.
- (3) The assumption that the activation function is smooth and *non-polynomial* which is used in the real case, is transferred to the activation function being smooth and *non-polyharmonic* in the complex case. This is a very natural duality since the interplay between “non-polynomial” and “non-polyharmonic” was already observed in the context of the UAT for shallow networks; see Section 3.2.
- (4) Moreover, this condition is satisfied by the popular modReLU and complex cardioid activation functions as defined in (3.1) and (3.2); see [37, Appendix A.2]. Note that the analogous result in the real case [58] does *not* apply to the popular ReLU activation function.
- (5) Note that, as in the real-valued case, it is in the complex setting also possible to pick the weights connecting input and hidden layer neurons *independently* of the function f that is to be approximated and to achieve the rate of $n^{-r/(2d)}$ under the restriction of continuous weight selection.

On a broad level, the proof of Theorem 3.3.1 relies on the fact that it is possible to approximate $C^r(\Omega_d; \mathbb{C})$ -functions at the rate of s^{-r} by elements of $\mathcal{P}_s^d$, which is a direct consequence of the real-valued analogue [58, Equation (3.14)]. Here, we set

$$\mathcal{P}_s^d := \left\{ \mathbb{C}^d \rightarrow \mathbb{C}, \quad z \mapsto \sum_{\substack{\mathbf{m}, \ell \in \mathbb{N}_0^d \\ |\mathbf{m}|, |\ell| \leq s}} a_{\mathbf{m}, \ell} z^{\mathbf{m}} \bar{z}^{\ell} : a_{\mathbf{m}, \ell} \in \mathbb{C} \right\}.$$

Hence, it suffices to study the approximation of such polynomials or, on an even more basic level, the approximation of monomials

$$\Omega_d \rightarrow \mathbb{C}, \quad z \mapsto z^{\mathbf{m}} \bar{z}^{\ell} = \prod_{j=1}^d z_j^{\mathbf{m}_j} \bar{z}_j^{\ell_j},$$

for $\mathbf{m}, \ell \in \mathbb{N}_0^d$ with $|\mathbf{m}|, |\ell| \leq s$.

To this end, we make heavy use of the *Wirtinger calculus*: For an open set $\emptyset \neq U \subseteq \mathbb{C}$ and a function $f \in C^1(U; \mathbb{C})$, identifying $\mathbb{C} \cong \mathbb{R}^2$, the Wirtinger derivatives of f are defined as

$$\partial_{\text{wirt}} f := \frac{\partial f}{\partial x} - i \cdot \frac{\partial f}{\partial y} \quad \text{and} \quad \bar{\partial}_{\text{wirt}} f := \frac{\partial f}{\partial x} + i \cdot \frac{\partial f}{\partial y} \quad \text{on } U. \quad (3.7)$$

For multiindices $\mathbf{m}, \ell \in \mathbb{N}_0^d$ we write $\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell}$ for the iterated Wirtinger derivatives according to the multiindices $\mathbf{m}$ and ℓ . Then, for an activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ of sufficient regularity, a fixed point $z \in \mathbb{C}^d$, and $b \in \mathbb{C}$, a computation shows

$$\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} [w \mapsto \phi(w^T z + b)] = z^{\mathbf{m}} \bar{z}^{\ell} \cdot \left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b).$$

Plugging in $w = 0$ yields

$$\left(\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} [w \mapsto \phi(w^T z + b)] \right) \Big|_{w=0} = z^{\mathbf{m}} \bar{z}^{\ell} \cdot \left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b). \quad (3.8)$$

The idea is now to divide the equation on both sides by $\left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b)$. To justify that we are able to do so, we need the following proposition.

Proposition 3.3.2 ([37, Proposition 2.1]). *Let $\emptyset \neq U \subseteq \mathbb{C}$ be an open set and let $\phi \in C^\infty(U; \mathbb{C})$ be non-polyharmonic. Then for every $s \in \mathbb{N}_0$ there exists a point $z_s \in U$ satisfying*

$$\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} \phi(z_s) \neq 0 \quad \text{for all } \mathbf{m}, \ell \in \mathbb{N}_0 \text{ with } |\mathbf{m}|, |\ell| \leq s.$$

Hence, under the assumption that ϕ is smooth and non-polyharmonic on $U \subseteq \mathbb{C}$ and using (3.8), we can pick $b \in U$ satisfying

$$z^{\mathbf{m}} \bar{z}^{\ell} = \left(\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} [w \mapsto \phi(w^T z + b)] \right) \Big|_{w=0} \cdot \left(\left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b) \right)^{-1} \quad (3.9)$$

for every $z \in \mathbb{C}^d$ and for every $\mathbf{m}, \ell \in \mathbb{N}_0^d$ with $|\mathbf{m}|, |\ell| \leq s$. The right hand side in (3.9) is the linear combination of partial derivatives of the function $w \mapsto \phi(w^T z + b)$. Using a generalized version of *divided differences*, it can be shown that these partial derivatives can be approximated by shallow CVNNs with activation function ϕ . This leads to the following theorem about the approximation of polynomials using shallow CVNNs.

Theorem 3.3.3 ([37, Theorem 3.1]). *Let $d, s \in \mathbb{N}$, $\varepsilon > 0$, and $\phi : \mathbb{C} \rightarrow \mathbb{C}$ be smooth and non-polyharmonic on an open set $\emptyset \neq U \subseteq \mathbb{C}$. Let $\mathcal{P}' \subseteq \mathcal{P}_s^d$ be bounded² and set $n := (4s + 1)^{2d}$. Then there exist $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$, and $\beta_1, \dots, \beta_n \in \mathbb{C}$ with the property that for every $P \in \mathcal{P}'$ there exist $\gamma_1, \dots, \gamma_n \in \mathbb{C}$, such that*

$$\left\| P(z) - \sum_{j=1}^n \gamma_j \phi(\alpha_j^T z + \beta_j) \right\|_{L^\infty(\Omega_d; \mathbb{C})} \leq \varepsilon.$$

As already explained above, using the fact that C^r -functions can be approximated at the rate of s^{-r} by elements of $\mathcal{P}_s^d$, we then get Theorem 3.3.1. We refer to [37, Section 3] for proof sketches for Theorems 3.3.1 and 3.3.3 and to [37, Appendices B-D] for the full proofs of these statements.

We further note that Theorem 3.3.3 may not only be used to prove an upper bound on the approximation rate for the approximation of C^r -functions (as in Theorem 3.3.1) but can in fact also

²Boundedness is to be understood in the sense of boundedness with respect to any norm on the finite-dimensional space $\mathcal{P}_s^d$.

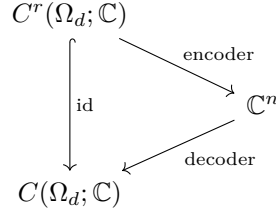


Figure 3.1.: The optimality of the approximation rate of $n^{-r/(2d)}$: Consider the setting where C^r -functions are approximated by a class of continuous functions that can be parameterized by n parameters. The encoder maps each C^r -function to the parameters of the approximating function. The decoder maps the parameters to the corresponding continuous function. Under the assumption that the encoder is continuous, at most an approximation rate of $n^{-r/(2d)}$ can be achieved and this rate is achieved by shallow CVNNs with smooth and non-polynomial activation function. Note that a shallow CVNN with d input and n hidden neurons has $n(d+2) \asymp_d n$ parameters.

be used for other function classes that are well approximated by polynomials. In [37, Appendix D], this is demonstrated by proving an approximation rate of $\nu^{-n^{1/(2d)}/17}$ for the approximation of functions $f : \Omega_d \rightarrow \mathbb{C}$ that are holomorphically extendable onto some polyellipse in $\mathbb{C}^{2d}$, where the parameter $\nu > 0$ specifies the size of that polyellipse.

Naturally, the question arises as to what extent the approximation rate of $n^{-r/(2d)}$ for the approximation of C^r -functions is optimal or whether it is possible to improve it further. Recall that in Theorem 3.3.1 it was shown that this approximation rate can be achieved under the restriction of continuous weight selection. Using results from [22], it can be proved that the rate of $n^{-r/(2d)}$ is indeed optimal under this continuity assumption; see [37, Theorem 4.1]. Note that the same is also true in the real-valued setting. Remarkably, this optimality does not only hold when approximating C^r -functions using shallow CVNNs, but it is valid whenever one approximates C^r -functions using a set of functions that can be parameterized using n parameters. Hence, under the assumption of continuous weight selection, shallow CVNNs with fixed, smooth, and non-polyharmonic activation function achieve the optimal approximation rate among all classes of functions that can be parameterized using finitely many parameters; see Figure 3.1 for a graphical illustration of the situation.

Naturally, the question arises what happens if the continuity assumption on the weight selection is dropped. Two results addressing this question are established in [37]: In [37, Theorem 4.2], it is shown that there exists a smooth and non-polyharmonic activation function with which even a rate of $n^{-r/(2d-1)}$ is achievable if the weight selection is not required to be continuous. On the other hand, [37, Theorem 4.3] establishes the existence of a smooth and non-polyharmonic activation function for which the rate of $n^{-r/(2d)}$ can (up to logarithmic factors) *not* be improved, even if a discontinuous weight selection is allowed. However, at first it remained open to establish fully general lower bounds (without the assumption of continuous weight selection) for the approximation of C^r -functions by shallow CVNNs. This question was further investigated (and in fact answered) in [35]; see Section 3.4.

The last topic that is addressed in [37] is in what way the upper bound in Theorem 3.3.1 suffers from the curse of dimensionality and whether this can be circumvented. Indeed, it is true that the curse of dimensionality appears in Theorem 3.3.1 both through the bound $n^{-r/(2d)}$ (whereby it disappears here if, e.g., $r = 2d$) and, more significantly, through the constant $C(d, r)$ appearing in the formulation of Theorem 3.3.1. However, at least under the restriction of continuous weight selection, it is possible to show that the curse of dimensionality can *not* be avoided in this setting. Since we expect this result to be of independent interest, we state it explicitly here. It is formulated

for real-valued functions since this case is expected to be of greater interest than the complex-valued case. It can easily be transferred to the complex-valued setting.

Theorem 3.3.4 ([37, Theorem 5.1]). *Let $d \in \mathbb{N}$ and let $C^{\infty,*,d}$ denote the set of all smooth functions $f \in C^\infty([-1, 1]^d; \mathbb{R})$ for which $\|f\|_{C^\infty([-1, 1]^d; \mathbb{R})} < \infty$. Let $\bar{a} : C^{\infty,*,d} \rightarrow \mathbb{R}^{2^d-1}$ be continuous with respect to some norm on $C^{\infty,*,d}$ and, moreover, let $M : \mathbb{R}^{2^d-1} \rightarrow C([-1, 1]^d; \mathbb{R})$ be an arbitrary map. Then,*

$$\sup_{\substack{f \in C^{\infty,*,d} \\ \|f\|_{C^\infty([-1, 1]^d; \mathbb{R})} \leq 1}} \|f - M(\bar{a}(f))\|_{L^\infty([-1, 1]^d; \mathbb{R})} \geq 1.$$

Theorem 3.3.4 shows that the problem of approximating C^r -functions (in fact, even functions from $C^{\infty,*,d}$) with a function class that is parameterizable using finitely many parameters and using continuous approximation methods *necessarily* suffers from the curse of dimensionality. Indeed, as Theorem 3.3.4 states, one needs at least 2^d parameters in order to be able to achieve a non-trivial worst-case approximation error.

3.4. On best approximation by multivariate ridge functions with applications to generalized translation networks

In the real-valued setting, fully general lower bounds for the worst-case approximation error for the approximation of C^r -functions by shallow real-valued neural networks were established by Maiorov et al. studying sums of *ridge functions*; see [40, 55, 56]. Such sums are functions of the form

$$\mathbb{R}^d \rightarrow \mathbb{R}, \quad x \mapsto \sum_{j=1}^n \varrho_j(a_j^T x), \quad (3.10)$$

where $\varrho_j : \mathbb{R} \rightarrow \mathbb{R}$ and $a_j \in \mathbb{R}^d$ for $j \in \{1, \dots, n\}$. Trivially, shallow RVNNs are a special case of such functions. Precisely, Maiorov et al. showed that the worst-case approximation error for the approximation of L^∞ -Sobolev functions of regularity r by functions as in (3.10) can be lower bounded by $n^{-r/(d-1)}$, even if the approximation is measured with respect to the L^1 -norm. However, this result is not directly applicable to obtain a similar lower bound for the approximation using shallow CVNNs, mostly since the activation function in a CVNN is a function $\mathbb{C} \rightarrow \mathbb{C}$, i.e., $\mathbb{R}^2 \rightarrow \mathbb{R}^2$, whereas the functions ϱ_j appearing in (3.10) are univariate, i.e., functions from $\mathbb{R}$ to $\mathbb{R}$.

This observation motivates the study of sums of *multivariate* (or, more precisely, ℓ -variate) ridge functions, i.e., functions of the form

$$\mathbb{R}^d \rightarrow \mathbb{R}, \quad x \mapsto \sum_{j=1}^n \varrho_j(A_j x), \quad (3.11)$$

where $\varrho_j : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and $A_j \in \mathbb{R}^{\ell \times d}$ for $j \in \{1, \dots, n\}$. The article *On best approximation by multivariate ridge functions with applications to generalized translation networks* [35], co-authored with Palina Salanevich, Olov Schavemaker, and Felix Voigtlaender, proves tight asymptotic bounds for the approximation of Sobolev functions using functions of the form (3.11) and applies these bounds to shallow CVNNs and shallow *generalized translation networks* (GTNs). The latter are functions of the form

$$\mathbb{R}^d \rightarrow \mathbb{R}, \quad x \mapsto \sum_{j=1}^n \tau(A_j x + b_j)$$

for a fixed activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$, weight matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$, and bias vectors $b_1, \dots, b_n \in \mathbb{R}^\ell$. The parameter ℓ is called the *activation dimension* of the GTN.

The main result of [35]—the lower bound for the approximation of Sobolev functions by sums of multivariate ridge functions—reads as follows.

Theorem 3.4.1 ([35, Theorem 1.1]). *Let $d, \ell, r \in \mathbb{N}$ with $d > \ell$ and $p, q \in [1, \infty]$. Then there exists a positive constant $c = c(d, \ell, p, q, r) > 0$ with the following property: For every $n \in \mathbb{N}$ there exists a Sobolev function $f \in \mathcal{B}(W_d^{r,p})$ such that for every choice of functions $\varrho_1, \dots, \varrho_n : \mathbb{R}^\ell \rightarrow \mathbb{R}$, and matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ with the property that $\varrho_j(A_j \cdot \bullet) \in L^1(B^d; \mathbb{R})$ for every $j \in \{1, \dots, n\}$, we have*

$$\left\| f(x) - \sum_{j=1}^n \varrho_j(A_j x) \right\|_{L^q(B^d; \mathbb{R})} \geq c \cdot n^{-r/(d-\ell)}.$$

We mention the following three aspects.

- (1) Notably, no assumptions on the functions ϱ_j themselves are needed, but only on the composite functions $\varrho_j(A_j \cdot \bullet)$. Moreover, the assumption $\varrho_j(A_j \cdot \bullet) \in L^1(B^d; \mathbb{R})$ is quite weak and is in particular satisfied whenever $\varrho_j \in L^1_{\text{loc}}(\mathbb{R}^\ell; \mathbb{R})$ and $\text{rank}(A_j) = \ell$; see [35, Proposition 2.1].
- (2) The most relevant special case of Theorem 3.4.1 is the setting $p = \infty$ and $q = 1$ since this case readily implies the claim for arbitrary $p, q \in [1, \infty]$. Indeed, due to Hölder's inequality we have

$$\|\cdot\|_{L^1(B^d; \mathbb{R})} \lesssim_{d,q} \|\cdot\|_{L^q(B^d; \mathbb{R})} \lesssim_{d,q} \|\cdot\|_{L^\infty(B^d; \mathbb{R})}$$

for every $q \in [1, \infty]$, which further implies

$$\mathcal{B}(W_d^{r,\infty}) \subseteq C \cdot \mathcal{B}(W_d^{r,p})$$

for every $p \in [1, \infty]$ and some constant $C = C(d, p, r) > 0$. Further, note that the set of functions of the form (3.11) is *invariant* with respect to scaling with positive scalars.

- (3) The proof of Theorem 3.4.1 in [35] shows that the result even holds if only C^r -functions are considered. Hence, in view of (2), we may rewrite Theorem 3.4.1 as

$$\sup_{\substack{f \in C^r(B^d; \mathbb{R}), \\ \|f\|_{C^r(B^d; \mathbb{R})} \leq 1}} \inf_{\substack{A_j \in \mathbb{R}^{\ell \times d}, \varrho_j : \mathbb{R}^\ell \rightarrow \mathbb{R}, \\ \varrho_j(A_j \cdot \bullet) \in L^1(B^d; \mathbb{R})}} \left\| f(x) - \sum_{j=1}^n \varrho_j(A_j x) \right\|_{L^1(B^d; \mathbb{R})} \gtrsim_{d,\ell,r} n^{-r/(d-\ell)}.$$

The proof of Theorem 3.4.1 follows in its main steps the proof of the lower bound for univariate ridge functions in [55], properly adjusted to the multivariate setting. As one of the key differences, we highlight that [55, Lemma 2] is replaced by a generalized version of [56, Theorem 3]; see [35, Lemma 3.5]. Moreover, while the authors of [35] were working on that paper, they discovered a gap in the proof in [55], which could be resolved by replacing an orthogonal projection by a quasi-projection. We refer to [35, Section 2.3] for the definition of that quasi-projection and to [35, Appendix A] for more details on the gap in [55].

Furthermore, it is shown in [35] that the optimal rate of $n^{-r/(d-\ell)}$ from Theorem 3.4.1 can indeed be achieved, at least in the case $q \leq p$; see [35, Theorem 1.2]. In fact, it is even shown that the optimal rate can be achieved for a fixed choice of matrices A_j and polynomial functions ϱ_j . This result reads as follows.

Theorem 3.4.2 ([35, Theorem 1.2]). *Let $d, \ell, r \in \mathbb{N}$ with $d > \ell$ and $1 \leq q \leq p \leq \infty$. Then there exists a positive constant $C = C(d, \ell, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$*

there exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ such that for any $f \in \mathcal{B}(W_d^{r,p})$ there exist polynomials $P_1, \dots, P_n : \mathbb{R}^\ell \rightarrow \mathbb{R}$ satisfying

$$\left\| f(x) - \sum_{j=1}^n P_j(A_j x) \right\|_{L^q(B^d; \mathbb{R})} \leq C \cdot n^{-r/(d-\ell)}.$$

In the remainder of this section, we discuss the applications of Theorems 3.4.1 and 3.4.2 to GTNs and CVNNs. The lower bound from Theorem 3.4.1 can be used quite directly to obtain lower bounds for the approximation of Sobolev functions using shallow GTNs and CVNNs [35, Theorem 1.3(1), Theorem 1.4(1)]. Indeed, for GTNs this follows from the fact that each shallow GTN with activation dimension ℓ and n hidden neurons is the sum of n ℓ -variate ridge functions. For CVNNs, the situation is slightly more difficult since the output of a CVNN is a complex number, whereas a multivariate ridge function maps to $\mathbb{R}$. This issue, however, can be resolved by only considering the real (or imaginary) part of a shallow CVNN: For arbitrary $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$, $\beta_1, \dots, \beta_n, \gamma_1, \dots, \gamma_n \in \mathbb{C}$, and an activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$, we compute

$$\operatorname{Re} \left(\sum_{j=1}^n \gamma_j \phi(\alpha_j^T z + \beta_j) \right) = \sum_{j=1}^n \operatorname{Re}(\gamma_j) (\operatorname{Re} \phi)(\alpha_j^T z + \beta_j) - \operatorname{Im}(\gamma_j) (\operatorname{Im} \phi)(\alpha_j^T z + \beta_j)$$

for $z \in \mathbb{C}^d$. Note that the latter is by definition the sum of $2n$ 2-variate ridge functions with inputs from $\mathbb{C}^d \cong \mathbb{R}^{2d}$. This implies the lower bound of $n^{-r/(2d-2)}$ for the approximation of Sobolev functions on B^{2d} (embedded into $\mathbb{C}^d$) by shallow CVNNs with d (complex) input neurons.

Further, the upper bound from Theorem 3.4.2 can be used to show that for every activation dimension $\ell \in \mathbb{N}$ there exists a smooth activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ such that shallow GTNs using that activation function are indeed able to achieve the optimal approximation rate of $n^{-r/(d-\ell)}$ [35, Theorem 1.3(2)]. This activation function is quite bespoke and is constructed in a way such that a countable, dense subset of $C(B^\ell; \mathbb{R})$ is realized by applying the function to shifted arguments.

As the last piece of the puzzle, it remains to show that the rate of $n^{-r/(2d-2)}$ in the case of shallow CVNNs can indeed be achieved for a proper choice of the activation function ϕ . We recall that in [37, Theorem 4.2] it was shown that there exists a smooth activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ for which the rate of $n^{-r/(2d-1)}$ can be achieved. In [35], it is shown that the same activation function even achieves the optimal rate of $n^{-r/(2d-2)}$; see [35, Theorem 1.4(2)]. The proof of that fact, however, is *not* a simple consequence of Theorem 3.4.2 (as in the case of GTNs) but requires the translation of various results from [69, Chapter 5] to the complex-valued setting; see [35, Section 5.2].

For convenience, we explicitly state the main result about the approximation of Sobolev functions by CVNNs obtained in [35].

Theorem 3.4.3 ([35, Theorem 1.4]). *The following two statements hold:*

- (1) *For any $d, r \in \mathbb{N}$ with $d \geq 2$ and any $p, q \in [1, \infty]$, there exists a constant $c = c(d, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exists a function $f : \mathbb{C}^d \rightarrow \mathbb{C}$ satisfying $\operatorname{Re}(f), \operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p})$ such that for any choice of $\phi \in L_{\text{loc}}^1(\mathbb{C}; \mathbb{C})$, complex vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$, biases $\beta_1, \dots, \beta_n \in \mathbb{C}$, and coefficients $\gamma_1, \dots, \gamma_n \in \mathbb{C}$ we have*

$$\left\| f(z) - \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \right\|_{L^q(B^{2d}; \mathbb{C})} \geq c \cdot n^{-r/(2d-2)}.$$

- (2) *There exists a smooth activation function $\phi \in C^\infty(\mathbb{C}; \mathbb{C})$ with the property that for any choice of $d, r \in \mathbb{N}$ with $d \geq 2$ and any $1 \leq q \leq p \leq \infty$ there exists a constant $C = C(d, p, q, r) > 0$ such that the following holds. For any $n \in \mathbb{N}$ there exist complex vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$*

3. Approximation properties of complex-valued neural networks

Reference	Condition on activation fct.	Continuity of weight selection	Norm	Type	Approximation Bound
[37, Theorem 3.2]	smooth & non-polyharmonic	possible	L^∞	$\lesssim_{d,r}$	$n^{-r/(2d)}$
[37, Theorem 4.1]	continuous	assumed	L^∞	$\gtrsim_{d,r}$	$n^{-r/(2d)}$
[35, Theorem 1.4(1)]	locally integrable	not assumed	L^1	$\gtrsim_{d,r}$	$n^{-r/(2d-2)}$
[35, Theorem 1.4(2)]	very special activation fct.	not assumed	L^∞	$\lesssim_{d,r}$	$n^{-r/(2d-2)}$
[37, Theorem 4.3]	$\frac{1}{1 + e^{-\operatorname{Re}(z)}}$	not assumed	L^∞	$\gtrsim_{d,r}$	$(n \cdot \ln(n))^{-r/(2d)}$

Table 3.1.: Overview of the proved approximation bounds for shallow CVNNs. r is the regularity of the approximated functions (which are assumed to be C^r), d the (complex) input dimension, and n the number of neurons in the hidden layer of the network.

such that for any function $f : \mathbb{C}^d \rightarrow \mathbb{C}$ with $\operatorname{Re}(f), \operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p})$ there exist biases $\beta_1, \dots, \beta_n \in \mathbb{C}$, and coefficients $\gamma_1, \dots, \gamma_n \in \mathbb{C}$ satisfying

$$\left\| f(z) - \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \right\|_{L^q(B^{2d}; \mathbb{C})} \leq C \cdot n^{-r/(2d-2)}.$$

Theorem 3.4.3(2) is to the best of our knowledge the first result which indicates a *superior* expressivity of CVNNs compared to RVNNs. Indeed, shallow RVNNs with $2d$ (real) input neurons can achieve an approximation rate of at most $n^{-r/(2d-1)}$ [55], while CVNNs can improve this rate to $n^{-r/(2d-2)}$. Recall from Section 3.3 that in the setting of continuous weight selection, shallow CVNNs with smooth and non-polyharmonic activation function achieve the same optimal approximation rate as shallow RVNNs with smooth and non-polynomial activation function. Hence, while there is no benefit in choosing CVNNs instead of RVNNs in the setting of continuous weight selection (as long as the approximation rate is considered), there is one if the continuity assumption is dropped. We refer to Table 3.1 for an overview of the results from [35, 37] regarding CVNNs.

3.5. Universal approximation with complex-valued deep narrow neural networks

The classical UAT, both in the real-valued case [52] and in the complex-valued case [90], studies the universality of neural networks with fixed activation function and fixed number of hidden layers. The width of the networks is not restricted. As a complementary setting and given the success of deep neural networks in practice, it is thus natural and interesting to investigate the setting of deep narrow networks, i.e., the setting where the number of hidden layers is unrestricted but the width is restricted.

In the case of RVNNs, this setting was examined in [46]. Precisely, it was shown that RVNNs with an unrestricted number of hidden layers, with width at least $n + m + 2$ (where n is the input and m the output dimension), and with fixed *non-affine* activation function $\sigma \in C(\mathbb{R}; \mathbb{R})$ that is differentiable at one point $x_0 \in \mathbb{R}$ with $\sigma'(x_0) \neq 0$, are universal. Thus, aside from the condition that there must be a point where σ is differentiable with non-vanishing derivative, there are significantly more permissible activation functions in the case of deep narrow networks than

in the case of a restricted number of hidden layers and unrestricted width, where the activation functions must be non-polynomial. As an example, the function $\sigma(x) = x^2$ does not yield universal RVNNs in the classical setting of the UAT, but it does give rise to universal RVNNs in the case of deep narrow networks.

The paper *Universal approximation with complex-valued deep narrow neural networks* [34], co-authored with Thomas Jahn and Hannes Matt, studies the universality of deep narrow neural networks in the complex-valued case. Before stating the main result of that paper, we recall the universal approximation theorem for deep CVNNs from [90].

Theorem 3.5.1 ([90, Theorem 1.4]). *Let $\phi \in C(\mathbb{C}; \mathbb{C})$, $d \in \mathbb{N}$, and $L \geq 2$. Then the following are equivalent.*

- (1) *ϕ is neither holomorphic, nor anti-holomorphic, nor a polynomial in z and $\bar{z}$.*
- (2) *The set of CVNNs with L hidden layers, d input neurons, and fixed activation function ϕ is universal.*

Similarly to the case of RVNNs, one expects the set of “good” activation functions to grow if one moves from the classical setting to the deep narrow regime. However, the following two aspects quickly become apparent.

- (1) In case of a *holomorphic* activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$, the set of CVNNs with an arbitrary number of hidden layers with arbitrary width using activation function ϕ only consists of holomorphic functions too and can thus not be universal.
- (2) Similarly, in case of an *anti-holomorphic* activation function, the set of CVNNs with an arbitrary number of hidden layers with arbitrary width using that activation function only consists of functions that are either holomorphic or anti-holomorphic and can thus not be universal either.

Hence, the condition that ϕ is neither holomorphic, nor anti-holomorphic appearing in Theorem 3.5.1 can certainly not be weakened for the case of deep narrow CVNNs. The only condition where there is hope for a weakening is that ϕ must not be a polynomial in z and $\bar{z}$. The main result from [34] reads as follows.

Theorem 3.5.2 ([34, Theorem 1.1]). *Let $\phi \in C(\mathbb{C}; \mathbb{C})$ be real differentiable at $z_0 \in \mathbb{C}$ with $\phi'(z_0) \neq 0$ and $n, m \in \mathbb{N}$. Then the following are equivalent.*

- (1) *ϕ is neither holomorphic, nor anti-holomorphic, nor $\mathbb{R}$ -affine.*
- (2) *The set of CVNNs with activation function ϕ , with n input neurons, m output neurons, with an unrestricted number of hidden layers of width $2n + 2m + 5$ is universal.*

Theorem 3.5.2 shows that the restriction that ϕ must not be a polynomial in z and $\bar{z}$ can indeed be weakened to ϕ not being $\mathbb{R}$ -affine. As in the case of RVNNs, we hence obtain a significantly richer class of good activation functions than in the classical setting of the UAT. However, the additional assumption that there exists a point $z_0 \in \mathbb{C}$ at which ϕ is real differentiable with non-vanishing derivative is required.

The sufficient width of $2n + 2m + 5$ stated in Theorem 3.5.2 is significantly worse than the sufficient width in the case of RVNNs. Indeed, upon identifying $\mathbb{R}^2$ with $\mathbb{C}$, a function $\mathbb{C}^n \rightarrow \mathbb{C}^m$ may be viewed as a function $\mathbb{R}^{2n} \rightarrow \mathbb{R}^{2m}$ and can, according to [46, Theorem 3.2], be approximated locally uniformly by deep narrow RVNNs of width $2n + 2m + 2$, provided that the activation function is non-affine and satisfies the differentiability condition. The width of $2n + 2m + 5$ obtained in Theorem 3.5.2 is measured in the sense of *complex* neurons and hence corresponds to a width of $4n + 4m + 10$ in the real-valued sense. However, the occurrence of the factor 2 in Theorem 3.5.2 for the width is *not* a shortcoming of the proof techniques in [34]: As is shown in [34, Theorem 6.4], in

the considered generality of arbitrary, continuous complex activation functions, at least a width of $\max\{2n, 2m\}$ is necessary. Indeed, if the complex activation function is of the form $\phi(z) = \sigma(\operatorname{Re}(z))$ for $\sigma \in C(\mathbb{R}; \mathbb{R})$, each hidden neuron in a CVNN that uses ϕ as the activation function only stores a real number and hence corresponds to only a *single* real-valued neuron. For RVNNs with $2n$ (real) input and $2m$ (real) output neurons, it is well-known that, in order to obtain universal networks, the width needs to be at least $\max\{2n, 2m\}$; see [13, Lemma 1].

While a width of $\max\{2n, 2m\}$ is necessary in the generality of the considered setting, it is shown in [34] that for a large subclass of admissible activation functions a width of $n + m + 3$ is in fact sufficient. More precisely, this holds for each activation function $\phi \in C(\mathbb{C}; \mathbb{C})$ that is not $\mathbb{R}$ -affine and for which there exists $z_0 \in \mathbb{C}$ for which

$$(\partial_{\text{wirt}}\phi(z_0), \bar{\partial}_{\text{wirt}}\phi(z_0)) \in [\mathbb{C}^* \times \{0\}] \cup [\{0\} \times \mathbb{C}^*];$$

see [34, Theorems 4.2 & 4.9]. Here, $\mathbb{C}^* := \mathbb{C} \setminus \{0\}$ and ∂_{wirt} and $\bar{\partial}_{\text{wirt}}$ denote the Wirtinger derivatives as in (3.7).

The proof of Theorem 3.5.2 relies on a generalized Taylor expansion for functions from $\mathbb{C}$ to $\mathbb{C}$. More precisely, if $\phi \in C(\mathbb{C}; \mathbb{C})$ is real differentiable at $z_0 \in \mathbb{C}$, one gets

$$\phi(z + z_0) = \phi(z_0) + \partial_{\text{wirt}}\phi(z_0)z + \bar{\partial}_{\text{wirt}}\phi(z_0)\bar{z} + \Theta(z) \quad \text{for all } z \in \mathbb{C}, \quad (3.12)$$

for a function $\Theta : \mathbb{C} \rightarrow \mathbb{C}$ satisfying $\lim_{z \rightarrow 0} \frac{\Theta(z)}{z} = 0$. Under the assumption that $\phi'(z_0) \neq 0$, one can then show that the complex identity and/or the complex conjugation can be approximated locally uniformly by shallow CVNNs of width at most 2 and with activation function ϕ ; see [34, Proposition 3.1].

The analysis is then split into the case of non-polyharmonic and polyharmonic activation functions. In the case of a *non-polyharmonic* activation function, the universal approximation theorem for shallow CVNNs [90, Theorem 1.3] is applicable. Using a construction called the *register model* that was already used in [46], one can then “flip” the shallow CVNNs to obtain deep and narrow networks with the universal approximation property. If, on the other hand, ϕ is *polyharmonic*, ϕ is in particular smooth and using a second order Taylor expansion similar to (3.12) it can be shown that (variants) of the complex multiplication can be approximated locally uniformly by CVNNs with activation function ϕ . By gradually performing multiplications and conjugations, it can then be shown that multivariate complex polynomials in z and $\bar{z}$ can be approximated locally uniformly by CVNNs with activation function ϕ . The set of these polynomials is universal as follows from the Stone-Weierstrass theorem [30, Theorem 4.51].

We note that in the real-valued case, the techniques from [46] can also be used to obtain *quantitative* approximation results in terms of the number of hidden layers of the considered deep narrow networks [47]. Also in the case of CVNNs, the proof techniques from Theorem 3.5.2 can be used to prove such quantitative approximation statements. In [34], this is explicitly done for the case of non-polyharmonic activation functions; see [34, Theorem 5.3]. More precisely, the proof of this result combines the fact that shallow CVNNs with non-polyharmonic activation function can approximate polynomials (see Theorem 3.3.3) and the register model approach which is used to prove Theorem 3.5.2 for non-polyharmonic activation functions.

Bibliography

- [1] Z. Allen-Zhu, Y. Li, and Z. Song. “A Convergence Theory for Deep Learning via Over-Parameterization.” In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, June 9–15, 2019, pp. 242–252. URL: <https://proceedings.mlr.press/v97/allen-zhu19a.html>.
- [2] M. Arjovsky, A. Shah, and Y. Bengio. “Unitary Evolution Recurrent Neural Networks.” In: *Proceedings of the 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June 20–22, 2016, pp. 1120–1128. URL: <https://proceedings.mlr.press/v48/arjovsky16.html>.
- [3] A. Athalye, N. Carlini, and D. Wagner. “Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples.” In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholm, Sweden: PMLR, July 10–15, 2018, pp. 274–283. URL: <https://proceedings.mlr.press/v80/athalye18a.html>.
- [4] J. A. Barrachina, C. Ren, C. Morisseau, G. Vieillard, and J.-P. Ovarlez. “Comparison Between Equivalent Architectures of Complex-valued and Real-valued Neural Networks - Application on Polarimetric SAR Image Segmentation.” In: *Journal of Signal Processing Systems* 95.1 (Jan. 2023), pp. 57–66. DOI: 10.1007/s11265-022-01793-0.
- [5] A. Barron. “Universal approximation bounds for superpositions of a sigmoidal function.” In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945. DOI: 10.1109/18.256500.
- [6] P. Bartlett, S. Bubeck, and Y. Cherapanamjeri. “Adversarial Examples in Multi-Layer Random ReLU Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 9241–9252. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/4d19b37a2c399deace9082d464930022-Paper.pdf.
- [7] S. Bombari, S. Kiyani, and M. Mondelli. “Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels.” In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR, July 23–29, 2023, pp. 2738–2776. URL: <https://proceedings.mlr.press/v202/bombari23a.html>.
- [8] W. Brendel, J. Rauber, and M. Bethge. “Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models.” In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=SyZIOGWCZ>.
- [9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. “Language Models are Few-Shot Learners.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [10] S. Bubeck, Y. Cherapanamjeri, G. Gidel, and R. Tachet des Combes. “A single gradient step finds adversarial examples on random two-layers neural networks.” In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 10081–10091. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/536eecee295b92db6b32194e269541f8-Paper.pdf.

- [11] S. Bubeck and M. Sellke. “A Universal Law of Robustness via Isoperimetry.” In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 28811–28822. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/f197002b9a0853eca5e046d9ca4663d5-Paper.pdf.
- [12] S. Buchanan, D. Gilboa, and J. Wright. “Deep Networks and the Multiple Manifold Problem.” In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=0-6Pm_d_Q-.
- [13] Y. Cai. “Achieve the Minimum Width of Neural Networks for Universal Approximation.” In: *International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=hfUJ4ShyDEU>.
- [14] N. Carlini and D. Wagner. “Towards Evaluating the Robustness of Neural Networks.” In: *2017 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2017, pp. 39–57. DOI: 10.1109/SP.2017.49.
- [15] L. Chizat and F. Bach. “Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss.” In: *Proceedings of the 33rd Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. Online: PMLR, July 9–12, 2020, pp. 1305–1338. URL: <https://proceedings.mlr.press/v125/chizat20a.html>.
- [16] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. “Parseval Networks: Improving Robustness to Adversarial Examples.” In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 6–11, 2017, pp. 854–863. URL: <https://proceedings.mlr.press/v70/cisse17a.html>.
- [17] E. Cole, J. Cheng, J. Pauly, and S. Vasanawala. “Analysis of deep complex-valued convolutional neural networks for MRI reconstruction and phase-focused applications.” In: *Magnetic resonance in Medicine* 86.2 (2021), pp. 1093–1109. DOI: 10.1002/mrm.28733.
- [18] F. Croce and M. Hein. “Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks.” In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. Daumé III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. Virtual: PMLR, July 2020, pp. 2206–2216. URL: <https://proceedings.mlr.press/v119/croce20b.html>.
- [19] G. Cybenko. “Approximation by superpositions of a sigmoidal function.” In: *Mathematics of Control, Signals and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/BF02551274.
- [20] A. Daniely. *Existence of Adversarial Examples for Random Convolutional Networks via Isoperimetric Inequalities on $\mathbb{SO}(d)$* . 2025. arXiv: 2506.12613.
- [21] A. Daniely and H. Shacham. “Most ReLU Networks Suffer from ℓ^2 Adversarial Perturbations.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6629–6636. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/497476fe61816251905e8baafdf54c23-Paper.pdf.
- [22] R. A. DeVore, R. Howard, and C. Micchelli. “Optimal nonlinear approximation.” In: *manuscripta mathematica* 63.4 (Dec. 1989), pp. 469–478. DOI: 10.1007/BF01171759.
- [23] S. Dirksen, P. Finke, P. Geuchen, D. Stöger, and F. Voigtlaender. *Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks*. 2025. arXiv: 2506.19695.
- [24] S. Dirksen and S. Mendelson. “Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing.” In: *Journal of the European Mathematical Society* 23.9 (2021), pp. 2913–2947. DOI: 10.4171/JEMS/1066.
- [25] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su. “Adversarial Distributional Training for Robust Deep Learning.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 8270–8283. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/5de8a36008b04a6167761fa19b61aa6c-Paper.pdf.
- [26] S. S. Du, X. Zhai, B. Póczos, and A. Singh. “Gradient Descent Provably Optimizes Overparameterized Neural Networks.” In: *International Conference on Learning Representations*. 2019. URL: <https://openreview.net/forum?id=S1eK3i09YQ>.

-
- [27] R. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. ISBN: 9780521007542.
- [28] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. CRC Press, 1992. ISBN: 9780849371578.
- [29] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. “Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/95e1533eb1b20a9777749fb94fdb944-Paper.pdf.
- [30] G. B. Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999. ISBN: 9780471317166.
- [31] S. Frei, G. Vardi, P. Bartlett, and N. Srebro. “The Double-Edged Sword of Implicit Bias: Generalization vs. Robustness in ReLU Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 8885–8897. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/1c26c389d60ec419fd24b5fee5b35796-Paper-Conference.pdf.
- [32] Gemini Team. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. arXiv: 2312.11805.
- [33] P. Geuchen. “Approximation properties of complex-valued neural networks: An overview.” In: *15th International Conference on Sampling Theory and Applications*. 2025. URL: <https://openreview.net/forum?id=Kx844Vue41>.
- [34] P. Geuchen, T. Jahn, and H. Matt. “Universal approximation with complex-valued deep narrow neural networks.” In: *Constructive Approximation* (June 2025). DOI: 10.1007/s00365-025-09713-8.
- [35] P. Geuchen, P. Salanevich, O. Schavemaker, and F. Voigtlaender. “On best approximation by multivariate ridge functions with applications to generalized translation networks.” In: *Mathematical Foundations of Machine Learning 1.1* (July 2025), p. 1. DOI: 10.1007/s44439-025-00002-7.
- [36] P. Geuchen, D. Stöger, T. Telaar, and F. Voigtlaender. “Upper and lower bounds for the Lipschitz constant of random neural networks.” In: *Information and Inference: A Journal of the IMA* 14.2 (Apr. 2025), iaaf009. DOI: 10.1093/imaiai/iaaf009.
- [37] P. Geuchen and F. Voigtlaender. “Optimal approximation using complex-valued neural networks.” In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 1681–1737. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/05b69cc4c8ff6e24c5de1ecd27223d37-Paper-Conference.pdf.
- [38] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.
- [39] I. J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and Harnessing Adversarial Examples.” In: *International Conference on Learning Representations*. 2015. arXiv: 1412.6572.
- [40] Y. Gordon, V. E. Maiorov, M. Meyer, and S. Reisner. “On the Best Approximation by Ridge Functions in the Uniform Norm.” In: *Constructive Approximation* 18.1 (Jan. 2001), pp. 61–85. DOI: 10.1007/s00365-001-0009-5.
- [41] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. “Regularisation of neural networks by enforcing Lipschitz continuity.” In: *Machine Learning* 110.2 (Feb. 2021), pp. 393–416. DOI: 10.1007/s10994-020-05929-w.
- [42] A. Graves, A.-r. Mohamed, and G. Hinton. “Speech recognition with deep recurrent neural networks.” In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.
- [43] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. “Implicit Bias of Gradient Descent on Linear Convolutional Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.

- [44] K. He, X. Zhang, S. Ren, and J. Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.
- [45] A. Jacot, F. Gabriel, and C. Hongler. “Neural Tangent Kernel: Convergence and Generalization in Neural Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.
- [46] P. Kidger and T. Lyons. “Universal Approximation with Deep Narrow Networks.” In: *Proceedings of the 33rd Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. Virtual: PMLR, July 2020, pp. 2306–2327. URL: <https://proceedings.mlr.press/v125/kidger20a.html>.
- [47] A. Kratsios and L. Papon. “Universal Approximation Theorems for Differentiable Geometric Deep Learning.” In: *Journal of Machine Learning Research* 23.196 (2022), pp. 1–73. URL: <http://jmlr.org/papers/v23/21-0716.html>.
- [48] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [49] A. Kurakin, I. Goodfellow, and S. Bengio. “Adversarial examples in the physical world.” In: *International Conference on Learning Representations*. 2017. URL: <https://openreview.net/pdf?id=S10ufn1lx>.
- [50] Y. LeCun and Y. Bengio. “Convolutional networks for images, speech, and time series.” In: *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258. ISBN: 9780262511025.
- [51] Z. Lei, S. Gao, H. Hasegawa, Z. Zhang, M. Zhou, and K. Sedraoui. “Fully Complex-Valued Gated Recurrent Neural Network for Ultrasound Imaging.” In: *IEEE Transactions on Neural Networks and Learning Systems* 35.10 (2024), pp. 14918–14931. DOI: 10.1109/TNNLS.2023.3282231.
- [52] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function.” In: *Neural Networks* 6.6 (1993), pp. 861–867. DOI: [https://doi.org/10.1016/S0893-6080\(05\)80131-5](https://doi.org/10.1016/S0893-6080(05)80131-5).
- [53] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. “A Simple Tool for Bounding the Deviation of Random Matrices on Geometric Sets.” In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*. Ed. by B. Klartag and E. Milman. Cham: Springer International Publishing, 2017, pp. 277–299. ISBN: 9783319452821. DOI: 10.1007/978-3-319-45282-1_18.
- [54] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards Deep Learning Models Resistant to Adversarial Attacks.” In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=rJzIBfZAb>.
- [55] V. E. Maiorov. “Best approximation by ridge functions in L_p -spaces.” In: *Ukrainian Mathematical Journal* 62.3 (Oct. 2010), pp. 452–466. DOI: 10.1007/s11253-010-0364-0.
- [56] V. E. Maiorov. “On Best Approximation by Ridge Functions.” In: *Journal of Approximation Theory* 99.1 (1999), pp. 68–94. DOI: <https://doi.org/10.1006/jath.1998.3304>.
- [57] W. S. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity.” In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1943), pp. 115–133. DOI: 10.1007/BF02478259.
- [58] H. N. Mhaskar. “Neural Networks for Optimal Approximation of Smooth and Analytic Functions.” In: *Neural Computation* 8.1 (1996), pp. 164–177. DOI: 10.1162/neco.1996.8.1.164.
- [59] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. “Human-level control through deep reinforcement learning.” In: *Nature* 518.7540 (Feb. 2015), pp. 529–533. DOI: 10.1038/nature14236.
- [60] A. Montanari and Y. Wu. “Adversarial Examples in Random Neural Networks with General Activations.” In: *Mathematical Statistics and Learning* 6.1/2 (2023), pp. 143–200. DOI: 10.4171/MSL/41.

-
- [61] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. “DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks.” In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2574–2582. DOI: 10.1109/CVPR.2016.282.
- [62] Q. Nguyen, M. Mondelli, and G. F. Montúfar. “Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks.” In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 8119–8129. URL: <https://proceedings.mlr.press/v139/nguyen21g.html>.
- [63] S. Oymak and B. Recht. *Near-Optimal Bounds for Binary Embeddings of Arbitrary Sets*. 2015. arXiv: 1512.04433.
- [64] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. “Practical Black-Box Attacks against Machine Learning.” In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ASIA CCS ’17. Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2017, pp. 506–519. DOI: 10.1145/3052973.3053009.
- [65] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. “The Limitations of Deep Learning in Adversarial Settings.” In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2016, pp. 372–387. DOI: 10.1109/EuroSP.2016.36.
- [66] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. “Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks.” In: *2016 IEEE Symposium on Security and Privacy (SP)*. 2016, pp. 582–597. DOI: 10.1109/SP.2016.41.
- [67] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgöwer. “Training Robust Neural Networks Using Lipschitz Bounds.” In: *IEEE Control Systems Letters* 6 (2022), pp. 121–126. DOI: 10.1109/LCSYS.2021.3050444.
- [68] P. Petersen and F. Voigtlaender. “Optimal approximation of piecewise smooth functions using deep ReLU neural networks.” In: *Neural Networks* 108 (2018), pp. 296–330. DOI: 10.1016/j.neunet.2018.08.019.
- [69] A. Pinkus. *Ridge Functions*. Cambridge tracts in mathematics. Cambridge University Press, 2015. ISBN: 9781316436844.
- [70] Y.-D. Qu, R.-Q. Zhang, S.-Q. Shen, J. Yu, and M. Li. “Entanglement Detection with Complex-Valued Neural Networks.” In: *International Journal of Theoretical Physics* 62.9 (Sept. 2023), p. 206. DOI: 10.1007/s10773-023-05460-3.
- [71] H. El-Rewaidy, U. Neisius, J. Mancio, S. Kucukseymen, J. Rodriguez, A. Paskavitz, B. Menze, and R. Nezafat. “Deep complex convolutional network for fast reconstruction of 3D late gadolinium enhancement cardiac MRI.” In: *NMR in Biomedicine* 33.7 (2020). DOI: 10.1002/nbm.4312.
- [72] F. Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519.
- [73] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors.” In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. DOI: 10.1038/323533a0.
- [74] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN: 9781107057135.
- [75] J. W. Siegel and J. Xu. “Approximation rates for neural networks with general activation functions.” In: *Neural Networks* 128 (2020), pp. 313–321. DOI: 10.1016/j.neunet.2020.05.019.
- [76] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. “The Implicit Bias of Gradient Descent on Separable Data.” In: *Journal of Machine Learning Research* 19.70 (2018), pp. 1–57. URL: <http://jmlr.org/papers/v19/18-188.html>.
- [77] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. “Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models.” In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII*. Munich, Germany: Springer-Verlag, 2018, pp. 644–661. DOI: 10.1007/978-3-030-01258-8_39.
- [78] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions.” In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

- [79] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. “Intriguing properties of neural networks.” In: *International Conference on Learning Representations*. 2014. arXiv: 1312.6199.
- [80] X. Tan, M. Li, P. Zhang, Y. Wu, and W. Song. “Complex-Valued 3-D Convolutional Neural Network for PolSAR Image Classification.” In: *IEEE Geoscience and Remote Sensing Letters* 17.6 (2020), pp. 1022–1026. DOI: 10.1109/LGRS.2019.2940387.
- [81] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal. “Deep Complex Networks.” In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=H1T2hmZAb>.
- [82] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. “Robustness May Be at Odds with Accuracy.” In: *International Conference on Learning Representations*. 2019. URL: <https://openreview.net/forum?id=SyxAb30cY7>.
- [83] R. Van Handel. *Probability in high dimension*. Lecture Notes (Princeton University). 2014. URL: <https://web.math.princeton.edu/~rvan/APC550.pdf>.
- [84] G. Vardi, G. Yehudai, and O. Shamir. “Gradient Methods Provably Converge to Non-Robust Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 20921–20932. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/83e6913572ba09b0ab53c64c016c7d1a-Paper-Conference.pdf.
- [85] B. Vasudeva, P. Deora, S. Bhattacharya, and P. M. Pradhan. “Compressed Sensing MRI Reconstruction with Co-VeGAN: Complex-Valued Generative Adversarial Network.” In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 1779–1788. DOI: 10.1109/WACV51458.2022.00184.
- [86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need.” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [87] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN: 9781108415194.
- [88] A. Virmaux and K. Scaman. “Lipschitz regularity of deep neural networks: analysis and efficient estimation.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf.
- [89] P. Virtue, S. X. Yu, and M. Lustig. “Better than real: Complex-valued neural nets for MRI fingerprinting.” In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 3953–3957. DOI: 10.1109/ICIP.2017.8297024.
- [90] F. Voigtlaender. “The universal approximation theorem for complex-valued neural networks.” In: *Applied and Computational Harmonic Analysis* 64 (2023), pp. 33–61. DOI: 10.1016/j.acha.2022.12.002.
- [91] Y. Wang, E. Ullah, P. Mianjy, and R. Arora. “Adversarial Robustness is at Odds with Lazy Training.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 6505–6516. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/2aab664e0d1656e8b56c74f868e1ea69-Paper-Conference.pdf.
- [92] S. Wisdom, T. Powers, J. Hershey, J. Le Roux, and L. Atlas. “Full-Capacity Unitary Recurrent Neural Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/d9ff90f4000eacd3a6c9cb27f78994cf-Paper.pdf.

-
- [93] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri. “A Closer Look at Accuracy vs. Robustness.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 8588–8601. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/61d77652c97ef636343742fc3dcf3ba9-Paper.pdf.
- [94] D. Yarotsky. “Error bounds for approximations with deep ReLU networks.” In: *Neural Networks* 94 (2017), pp. 103–114. DOI: 10.1016/j.neunet.2017.07.002.
- [95] D. Yarotsky and A. Zhevnerchuk. “The phase diagram of approximation rates for deep neural networks.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 13005–13015. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/979a3f14bae523dc5101c52120c535e9-Paper.pdf.
- [96] A. Zhang, Z. Lipton, M. Li, and A. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. ISBN: 9781009389433.
- [97] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu. “An optical neural chip for implementing complex-valued neural network.” In: *Nature Communications* 12.1 (Jan. 2021), p. 457. DOI: 10.1038/s41467-020-20719-7.
- [98] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. “Theoretically Principled Trade-off between Robustness and Accuracy.” In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 9–15, 2019, pp. 7472–7482. URL: <https://proceedings.mlr.press/v97/zhang19p.html>.
- [99] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin. “Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 55.12 (2017), pp. 7177–7188. DOI: 10.1109/TGRS.2017.2743222.

A. Additional work regarding Chapter 2

In this appendix, we provide additional results regarding the study of the Lipschitz constant of ReLU neural networks with random parameters that was discussed in Chapter 2. Specifically, in Section A.1, we prove Theorem 2.3.2 which states that the ℓ^2 -Lipschitz constant of a random ReLU network as in Assumption 2.2.2 can depend exponentially on the number of hidden layers of the network if no assumptions on the distribution of the biases are imposed. Afterwards, in Section A.2, we generalize the upper and lower bounds from [36] to networks with weights following arbitrary sub-gaussian distributions.

A.1. Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. The idea of the proof is to let the entries of the bias vectors be sufficiently large such that the ReLU is active in every neuron of Φ when passing x_0 through the network.

For $\ell \in \{0, \dots, L-1\}$, we set

$$b^{(\ell)} := \left(2N^{\ell/2}, \dots, 2N^{\ell/2}\right)^T \in \mathbb{R}^N.$$

We claim that for every $\ell \in \{0, \dots, L-1\}$,

$$\left(W^{(\ell)}\Phi^{(\ell-1)}(x_0) + b^{(\ell)}\right)_i \in \left[N^{\ell/2}, 3 \cdot N^{\ell/2}\right] \quad \text{for all } i \in \{1, \dots, N\} \quad (\text{A.1})$$

with probability at least $1 - 2N(\ell+1)\exp(-c' \cdot N)$ for an absolute constant $c' > 0$. We prove this via induction over ℓ .

In the case $\ell = 0$, we observe that for every $i \in \{1, \dots, N\}$,

$$(W^{(0)}\Phi^{(-1)}(x_0))_i = (W^{(0)}x_0)_i \sim \mathcal{N}(0, 2\|x_0\|_2^2/N) = \mathcal{N}(0, 2/N).$$

Hence, since a Gaussian distribution is sub-gaussian, there exists an absolute constant $c_1 > 0$ such that

$$\left|(W^{(0)}\Phi^{(-1)}(x_0))_i\right| \leq 1$$

with probability at least $1 - 2\exp(-c_1 \cdot N)$. We perform a union bound over $i \in \{1, \dots, N\}$ and get

$$\max_{i \in \{1, \dots, N\}} \left|(W^{(0)}\Phi^{(-1)}(x_0))_i\right| \leq 1$$

with probability at least $1 - 2N\exp(-c_1 \cdot N)$. On this event, we get by definition of $b^{(0)}$ that

$$\left(W^{(0)}\Phi^{(-1)}(x_0) + b^{(0)}\right)_i \in [1, 3]$$

for every $i \in \{1, \dots, N\}$, proving the case $\ell = 0$.

For the induction step, we take $\ell \in \{1, \dots, L-1\}$ and assume that $\ell-1$ satisfies the claim. We condition on $W^{(0)}, \dots, W^{(\ell-1)}$ and assume that

$$\left(W^{(\ell-1)}\Phi^{(\ell-2)}(x_0) + b^{(\ell-1)}\right)_i \in \left[N^{(\ell-1)/2}, 3 \cdot N^{(\ell-1)/2}\right] \quad \text{for all } i \in \{1, \dots, N\}.$$

Note that this in particular implies

$$\Phi^{(\ell-1)}(x_0) = W^{(\ell-1)}\Phi^{(\ell-2)}(x_0) + b^{(\ell-1)}$$

and hence

$$\left\|\Phi^{(\ell-1)}(x_0)\right\|_2 \in \left[N^{\ell/2}, 3 \cdot N^{\ell/2}\right].$$

For every $i \in \{1, \dots, N\}$ we then have

$$\left(W^{(\ell)}\Phi^{(\ell-1)}(x_0)\right)_i \sim \mathcal{N}\left(0, 2\left\|\Phi^{(\ell-1)}(x_0)\right\|_2^2/N\right)$$

and hence, there exists an absolute constant $c_2 > 0$ such that for every $t \geq 0$,

$$\left|\left(W^{(\ell)}\Phi^{(\ell-1)}(x_0)\right)_i\right| \leq \frac{\sqrt{2}\left\|\Phi^{(\ell-1)}(x_0)\right\|_2 \cdot t}{\sqrt{N}} \leq 3\sqrt{2}N^{(\ell-1)/2} \cdot t$$

with probability at least $1 - \exp(-c_2 \cdot t^2)$. Letting $t := \frac{\sqrt{N}}{3\sqrt{2}}$, we get

$$\left|\left(W^{(\ell)}\Phi^{(\ell-1)}(x_0)\right)_i\right| \leq N^{\ell/2}$$

with probability at least $1 - 2\exp(-c' \cdot N)$, taking $c' \leq \frac{c_2}{18}$. Applying a union bound and recalling the choice of $b^{(\ell)}$, we thus get

$$\left(W^{(\ell)}\Phi^{(\ell-1)}(x_0) + b^{(\ell)}\right)_i \in \left[N^{\ell/2}, 3 \cdot N^{\ell/2}\right] \quad \text{for all } i \in \{1, \dots, N\}$$

with probability at least $1 - 2N\exp(-c' \cdot N)$. Recall that we conditioned on $W^{(0)}, \dots, W^{(\ell-1)}$ up to now. Reintroducing the randomness over $W^{(\ell)}$ gives us the induction step.

Overall, by induction we have shown (A.1). Note that this implies that

$$\bar{\nabla}\Phi(x_0) = \left(W^{(L)}W^{(L-1)} \dots W^{(1)}W^{(0)}\right)^T$$

with probability at least $1 - 2NL\exp(-c' \cdot N)$. Using Theorem 2.2.3, we hence get

$$\text{Lip}_2(\Phi) \geq \left\|\bar{\nabla}\Phi(x_0)\right\|_2 \geq \left\|W^{(L)}W^{(L-1)} \dots W^{(1)}W^{(0)}\right\|_2 \quad (\text{A.2})$$

with probability at least $1 - 2NL\exp(-c' \cdot N)$, where, with a slight abuse of notation, we write $\|x\|_2 := \|x^T\|_2$ for a vector $x \in \mathbb{R}^{1 \times N}$.

We claim that

$$\left\|W^{(L)}W^{(L-1)} \dots W^{(1)}W^{(0)}\right\|_2 \geq (1.4)^{L-1} \cdot \sqrt{d}$$

with high probability. To this end, we first show via induction over $\ell \in \{0, \dots, L-1\}$ that

$$\left\|W^{(L)} \dots W^{(L-\ell)}\right\|_2 \geq (1.4)^{\ell-1} \cdot \sqrt{N}$$

with probability at least $1 - 2(\ell+1)\exp(-c'' \cdot N)$ for an absolute constant $c'' > 0$.

We start with the case $\ell = 0$. Recall that $W^{(L)} \sim \mathcal{N}(0, I_N)$ and hence, according to [87, Theorem 3.1.1],

$$\left\| W^{(L)} \right\|_2 \geq \frac{\sqrt{N}}{1.4}$$

with probability at least $1 - 2 \exp(-c_3 \cdot N)$ for an absolute constant $c_3 > 0$. This gives us the case $\ell = 0$ by taking $c'' \leq c_3$. Now, let $\ell \in \{1, \dots, L-1\}$. Via induction, we may condition on $W^{(L)}, \dots, W^{(L-\ell+1)}$ and assume that

$$\left\| W^{(L)} \dots W^{(L-\ell+1)} \right\|_2 \geq (1.4)^{\ell-2} \cdot \sqrt{N},$$

which occurs with probability at least $1 - 2\ell \exp(-c'' \cdot N)$. We set $Z := W^{(L)} \dots W^{(L-\ell+1)} \in \mathbb{R}^{1 \times N}$. Then $(ZW^{(L-\ell)})^T \sim \mathcal{N}(0, 2\|Z\|_2^2 I_N/N)$ where we note that the randomness is only with respect to $W^{(L-\ell)}$. Hence, by again using [87, Theorem 3.1.1], we infer the existence of an absolute constant $c_4 > 0$ such that for every $t \geq 0$,

$$\left\| ZW^{(L-\ell)} \right\|_2 \geq \frac{\|Z\|_2 \sqrt{2}}{\sqrt{N}} (\sqrt{N} - t)$$

with probability at least $1 - 2 \exp(-c_4 \cdot t^2)$. We set $t := \left(1 - \frac{1.4}{\sqrt{2}}\right) \cdot \sqrt{N} > 0$ and get

$$\left\| ZW^{(L-\ell)} \right\|_2 \geq 1.4 \cdot \|Z\|_2 \geq (1.4)^{\ell-1} \cdot \sqrt{N}$$

with probability at least $1 - 2 \exp(-c'' \cdot N)$, taking $c'' \leq c_4 \cdot \left(1 - \frac{1.4}{\sqrt{2}}\right)^2$. Up to now we have conditioned on $W^{(L)}, \dots, W^{(L-\ell+1)}$. Reintroducing the randomness over these variables yields

$$\left\| W^{(L)} \dots W^{(L-\ell)} \right\|_2 \geq (1.4)^{\ell-1} \cdot \sqrt{N}$$

with probability at least $1 - 2(\ell+1) \exp(-c'' \cdot N)$. Via induction, we get

$$\left\| W^{(L)} \dots W^{(1)} \right\|_2 \geq (1.4)^{L-1} \cdot \sqrt{N}$$

with probability at least $1 - 2L \exp(-c'' \cdot N)$.

It remains to incorporate the first layer weights $W^{(0)}$. Again, we condition on $W^{(L)}, \dots, W^{(1)}$ and assume that

$$\left\| W^{(L)} \dots W^{(1)} \right\|_2 \geq (1.4)^{L-1} \cdot \sqrt{N}.$$

We set $Z := W^{(L)} \dots W^{(1)} \in \mathbb{R}^{1 \times N}$. Then $(ZW^{(0)})^T \sim \mathcal{N}(0, 2\|Z\|_2^2 I_d/N)$ and hence, again due to [87, Theorem 3.1.1]

$$\left\| ZW^{(0)} \right\|_2 \geq \frac{\|Z\|_2 \sqrt{2}}{\sqrt{N}} (\sqrt{d} - t)$$

with probability at least $1 - 2 \exp(-c_4 \cdot t^2)$ for every $t \geq 0$. We set $t := \left(1 - \frac{1.4}{\sqrt{2}}\right) \cdot \sqrt{d} > 0$ and get

$$\left\| ZW^{(0)} \right\|_2 \geq 1.4 \cdot \|Z\|_2 \cdot \frac{\sqrt{d}}{\sqrt{N}} \geq (1.4)^L \cdot \sqrt{d}$$

with probability at least $1 - 2 \exp(-c'' \cdot d)$ when the randomness is only with respect to $W^{(0)}$. Reintroducing the randomness over $W^{(L)}, \dots, W^{(1)}$, we get

$$\left\| W^{(L)} \dots W^{(0)} \right\|_2 \geq (1.4)^L \cdot \sqrt{d}$$

with probability at least $1 - 2L \exp(-c'' \cdot N) - 2 \exp(-c'' \cdot d)$.

Overall, recalling (A.2), we obtain

$$\text{Lip}_2(\Phi) \geq (1.4)^L \cdot \sqrt{d}$$

with probability at least

$$1 - 2NL \exp(-c' \cdot N) - 2L \exp(-c'' \cdot N) - 2 \exp(-c'' \cdot d).$$

Since $N \geq C \cdot \ln(e \cdot NL)$ and $\min\{N, d\} \geq C$,

$$1 - 2NL \exp(-c' \cdot N) - 2L \exp(-c'' \cdot N) - 2 \exp(-c'' \cdot d) \geq 1 - \exp(-c \cdot \min\{N, d\})$$

for a suitable constant $c > 0$, as claimed. $\square$

A.2. Generalization of [36] to other initializations

In this section, we investigate in what way the techniques from [36] can be used to prove bounds for random ReLU networks with initializations other than the He initialization. More precisely, we investigate the case where the weights follow *arbitrary* sub-gaussian distributions. The results and proofs of the present section are not contained in [36] and originate from additional work that was done by the author after [36] was completed. Again, the analysis is split into the proof of an upper bound and that of a lower bound.

A.2.1. Upper bound

In this section, we prove upper bounds for the ℓ^2 -Lipschitz constant of random ReLU networks, where the weights follow arbitrary sub-gaussian distributions. The assumption on the random initialization is formalized in the following.

Assumption A.2.1. *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (random) ReLU network with d input neurons, a single output neuron, with L hidden layers, and N neurons per hidden layer. We let $W^{(0)} \in \mathbb{R}^{N \times d}$, $W^{(1)}, \dots, W^{(L-1)} \in \mathbb{R}^{N \times N}$, and $W^{(L)} \in \mathbb{R}^{1 \times N}$ denote the weight matrices associated to Φ and $b^{(0)}, \dots, b^{(L-1)} \in \mathbb{R}^N$, $b^{(L)} \in \mathbb{R}$ denote the bias vectors of Φ . We assume that the matrices $W^{(\ell)}$ as well as the biases $b^{(\ell)}$ are randomly chosen in the following way:*

- *The entries of $W^{(0)}, b^{(0)}, \dots, W^{(L)}, b^{(L)}$ are jointly independent.*
- *All the entries of $W_{i,j}^{(\ell)}$ for $0 \leq \ell \leq L$ are drawn from (possibly different) mean-zero sub-gaussian distributions. For $\ell \in \{0, \dots, L\}$, we let $K^{(\ell)} := \max_{i,j} \|W_{i,j}^{(\ell)}\|_{\psi_2}$, where the maximum is formed over all suitable indices i, j .*
- *At this point, we do not impose any restrictions on the distribution of the biases (apart from independence).*

As in Section 2.3.1, the proof of the upper bound relies on an application of Dudley's inequality. More precisely, we get the following statement, which is a generalization of [36, Proposition 5.1] to the setting of general weight distributions.

Proposition A.2.2. *There exists an absolute constant $C > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption A.2.1 with L hidden layers of width N . Let the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ be fixed and set*

$$\Lambda := \left\| W^{(L-1)} \right\|_{2 \rightarrow 2} \cdots \left\| W^{(0)} \right\|_{2 \rightarrow 2}.$$

For $x \in \mathbb{R}^d$ and $z \in B^d$ we define

$$Y_{z,x} := D^{(L-1)}(x)W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}z \in \mathbb{R}^N$$

and further

$$\mathcal{L} = \mathcal{L}(d, N, L, W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}) := \{Y_{z,x} : x \in \mathbb{R}^d, z \in B^d\} \subseteq \mathbb{R}^N.$$

Then, given any $u \geq 0$,

$$\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \leq C \cdot K^{(L)} \cdot \left(\int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} \, d\varepsilon + u\Lambda \right)$$

with probability at least $1 - 2\exp(-u^2)$ (with respect to the choice of $W^{(L)}$). Moreover,

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] \leq C \cdot K^{(L)} \cdot \int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} \, d\varepsilon.$$

Proof. For $y \in \mathcal{L}$ there exist $x \in \mathbb{R}^d$ and $z \in B^d$ satisfying

$$y = Y_{z,x} = D^{(L-1)}(x)W^{(L-1)} \cdots D^{(0)}(x)W^{(0)}z.$$

We compute

$$\begin{aligned} \|y\|_2 &\leq \underbrace{\left\| D^{(L-1)}(x) \right\|_{2 \rightarrow 2}}_{\leq 1} \cdot \left\| W^{(L-1)} \right\|_{2 \rightarrow 2} \cdots \underbrace{\left\| D^{(0)}(x) \right\|_{2 \rightarrow 2}}_{\leq 1} \cdot \left\| W^{(0)} \right\|_{2 \rightarrow 2} \cdot \underbrace{\|z\|_2}_{\leq 1} \\ &\leq \left\| W^{(L-1)} \right\|_{2 \rightarrow 2} \cdots \left\| W^{(0)} \right\|_{2 \rightarrow 2} = \Lambda. \end{aligned}$$

Hence, since $0 = D^{(L-1)}(0)W^{(L-1)} \cdots D^{(0)}(0)W^{(0)}0 \in \mathcal{L}$,

$$\mathcal{N}(\mathcal{L}, \varepsilon) = 1 \quad \text{for } \varepsilon \geq \Lambda. \tag{A.3}$$

We rewrite

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 &= \sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) W^{(0)}\|_2 \\ &= \sup_{x \in \mathbb{R}^d, z \in B^d} \left\langle \left(W^{(L)} \right)^T, Y_{z,x} \right\rangle = \sup_{Y \in \mathcal{L}} \left\langle \left(W^{(L)} \right)^T, Y \right\rangle. \end{aligned}$$

Note that for $Y_1, Y_2 \in \mathcal{L}$ with $Y_1 \neq Y_2$,

$$\begin{aligned} \left\| \left\langle \left(W^{(L)} \right)^T, Y_1 \right\rangle - \left\langle \left(W^{(L)} \right)^T, Y_2 \right\rangle \right\|_{\psi_2} &= \left\| \left\langle \left(W^{(L)} \right)^T, \frac{Y_1 - Y_2}{\|Y_1 - Y_2\|_2} \right\rangle \right\|_{\psi_2} \cdot \|Y_1 - Y_2\|_2 \\ &\leq \left\| W^{(L)} \right\|_{\psi_2} \cdot \|Y_1 - Y_2\|_2 \\ &\leq C_1 \cdot K^{(L)} \cdot \|Y_1 - Y_2\|_2 \end{aligned}$$

with an absolute constant $C_1 > 0$. Here, the first inequality uses the definition of the sub-gaussian norm of a random vector (see (1.1)) and the second inequality is due to [87, Lemma 3.4.2]. Hence, the random process $\left(\left\langle (W^{(L)})^T, Y \right\rangle\right)_{Y \in \mathcal{L}}$ has sub-gaussian increments with respect to the ℓ^2 -norm. Moreover, since all the entries of $W^{(L)}$ are mean-zero, we have

$$\mathbb{E} \left[\left\langle (W^{(L)})^T, Y \right\rangle \right] = \sum_{j=1}^N \mathbb{E} \left[W_j^{(L)} Y_j \right] = 0$$

for every $Y \in \mathcal{L}$. Since $0 \in \mathcal{L}$ and $\text{diam}(\mathcal{L}) \leq 2\Lambda$, the claim then follows from Dudley's inequality and its high probability version [83, Corollary 5.25 & Theorem 5.29]. $\square$

Shallow networks

As in [36], we start by considering the case of shallow networks. Note that the statement of [36, Lemma 5.3] holds deterministically and thus remains true if initializations other than the He initialization are considered. We then get the following statement—the generalization of [36, Theorem 3.1] to the setting of more general weight initializations.

Theorem A.2.3. *There exists an absolute constant $C > 0$ such that the following holds. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network with N hidden neurons satisfying Assumption A.2.1, then*

$$(1) \quad \text{Lip}_2(\Phi) \leq \sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \leq C \cdot K^{(0)} \cdot K^{(1)} \cdot \sqrt{N} \cdot \sqrt{d}$$

with probability at least $(1 - 2\exp(-d))(1 - 2\exp(-N))$ and

$$(2) \quad \mathbb{E} [\text{Lip}_2(\Phi)] \leq \mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] \leq C \cdot K^{(0)} \cdot K^{(1)} \cdot \sqrt{N} \cdot \sqrt{d}.$$

Proof. We set $k := \min\{d, N\}$ and $\ell := \max\{d, N\}$. Note that the quantity Λ in Proposition A.2.2 satisfies $\Lambda = \|W^{(0)}\|_{2 \rightarrow 2}$ in the present setting. Conditioned on $W^{(0)}$ and $b^{(0)}$, we compute, with $\mathcal{L}$ as in Proposition A.2.2,

$$\begin{aligned} \int_0^{\|W^{(0)}\|_{2 \rightarrow 2}} \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} \, d\varepsilon &\stackrel{[36, \text{Lemma 5.3}]}{\leq} \int_0^{\|W^{(0)}\|_{2 \rightarrow 2}} \sqrt{C_1 \cdot k} \cdot \sqrt{\ln \left(\frac{9 \|W^{(0)}\|_{2 \rightarrow 2}}{\varepsilon} \right)} \, d\varepsilon \\ &= \sqrt{C_1} \cdot \sqrt{k} \cdot 9 \left\| W^{(0)} \right\|_{2 \rightarrow 2} \cdot \int_0^{\frac{1}{9}} \sqrt{\ln(1/\sigma)} \, d\sigma \\ &\leq C_2 \cdot \sqrt{k} \cdot \left\| W^{(0)} \right\|_{2 \rightarrow 2} \end{aligned}$$

with absolute constants $C_1, C_2 > 0$. At the equality, we applied the substitution $\frac{1}{\sigma} = \frac{9 \|W^{(0)}\|_{2 \rightarrow 2}}{\varepsilon}$. Taking the randomness only with respect to $W^{(1)}$, we combine this estimate with Proposition A.2.2 and get that

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \left\| W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)} \right\|_2 &\leq C_3 \cdot K^{(1)} \cdot \left(C_2 \cdot \sqrt{k} \cdot \left\| W^{(0)} \right\|_{2 \rightarrow 2} + \sqrt{k} \cdot \left\| W^{(0)} \right\|_{2 \rightarrow 2} \right) \\ &\leq C_3 \cdot (C_2 + 1) \cdot K^{(1)} \cdot \left\| W^{(0)} \right\|_{2 \rightarrow 2} \cdot \sqrt{k} \end{aligned}$$

with probability at least $1 - 2 \exp(-k)$, as well as

$$\mathbb{E}_{W^{(1)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(1)} D^{(0)}(x) W^{(0)}\|_2 \right] \leq C_3 \cdot K^{(1)} \cdot C_2 \cdot \sqrt{k} \cdot \|W^{(0)}\|_{2 \rightarrow 2},$$

where $C_3 > 0$ is the absolute constant from Proposition A.2.2. We then reintroduce the randomness over $W^{(0)}$ and $b^{(0)}$ and note that

$$\|W^{(0)}\|_{2 \rightarrow 2} \leq C_4 \cdot K^{(0)} \cdot \sqrt{\ell}$$

with probability at least $1 - 2 \exp(-\ell)$ as follows from [87, Theorem 4.4.5] and

$$\mathbb{E} \left[\|W^{(0)}\|_{2 \rightarrow 2} \right] \leq C_5 \cdot K^{(0)} \cdot \sqrt{\ell}$$

according to [87, Exercise 4.4.6], with absolute constants $C_4, C_5 > 0$. This implies the claim. $\square$

Note that in the case of the initialization studied in [36], we have $K^{(0)} \asymp \frac{1}{\sqrt{N}}$ and $K^{(1)} \asymp 1$. Theorem A.2.3 is thus a generalization of Theorem 2.3.3 to the case of general sub-gaussian distributions for the weights.

Deep networks

The upper bound in the case of deep networks follows by similar adjustments like in the case of shallow networks. Here, it is important that [36, Lemma 5.8] is not influenced by the fact that different initializations are considered. We then get the following result.

Theorem A.2.4. *There exists an absolute constant $C > 0$ such that the following holds. If $N > d + 2$ and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network with L hidden layers of width N and with random weights and biases following Assumption A.2.1, then*

$$(1) \text{ Lip}_2(\Phi) \leq \sup_{x \in \mathbb{R}^d} \|\nabla \Phi(x)\|_2 \leq C^L \cdot \left(\prod_{j=0}^L K^{(j)} \right) \cdot N^{L/2} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}$$

with probability at least $(1 - 2 \exp(-d)) (1 - 2 \exp(-N))^L$ and

$$(2) \mathbb{E} [\text{Lip}_2(\Phi)] \leq \mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|\nabla \Phi(x)\|_2 \right] \leq C^L \cdot \left(\prod_{j=0}^L K^{(j)} \right) \cdot N^{L/2} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}.$$

Proof. Conditioned on $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$, we get, with $\mathcal{L}$ and Λ as in Proposition A.2.2,

$$\begin{aligned} \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} &\leq \sqrt{\ln \left(\left(\frac{3\Lambda}{\varepsilon} \right)^d \cdot \left(\frac{eN}{d+1} \right)^{L(d+1)} \right)} \\ &\leq \sqrt{d} \cdot \sqrt{\ln \left(\frac{3\Lambda}{\varepsilon} \right)} + \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \end{aligned}$$

for every $\varepsilon \in (0, \Lambda)$, using [36, Lemma 5.8] and the elementary estimate $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$. This yields

$$\int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} \, d\varepsilon \leq \sqrt{d} \cdot \int_0^\Lambda \sqrt{\ln \left(\frac{3\Lambda}{\varepsilon} \right)} \, d\varepsilon + \Lambda \cdot \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)}. \quad (\text{A.4})$$

Using the substitution $\sigma = \frac{\varepsilon}{3\Lambda}$ we get

$$\int_0^\Lambda \sqrt{\ln\left(\frac{3\Lambda}{\varepsilon}\right)} d\varepsilon = 3\Lambda \cdot \int_0^{\frac{1}{3}} \sqrt{\ln(1/\sigma)} d\sigma \leq C_1 \cdot \Lambda \quad (\text{A.5})$$

with $C_1 := 3 \cdot \int_0^{1/3} \sqrt{\ln(1/\sigma)} d\sigma$. Overall, we thus see

$$\begin{aligned} \int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \varepsilon))} d\varepsilon &\stackrel{(\text{A.4}), (\text{A.5})}{\leq} \Lambda \cdot \left(C_1 \cdot \sqrt{d} + \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \right) \\ &\stackrel{d < d+1 \leq 2d}{\leq} \sqrt{2} \cdot C_1 \cdot \Lambda \cdot \left(\sqrt{d} + \sqrt{L} \cdot \sqrt{d} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \right) \\ &\stackrel{L \geq 1, N \geq d+1}{\leq} 2\sqrt{2} \cdot C_1 \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}. \end{aligned} \quad (\text{A.6})$$

Taking the randomness only with respect to $W^{(L)}$, we may apply Proposition A.2.2 and obtain the existence of an absolute constant $C_2 > 0$ such that

$$\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \leq C_2 \cdot K^{(L)} \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}$$

with probability at least $1 - 2\exp(-d)$ and

$$\mathbb{E}_{W^{(1)}} \left[\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \right] \leq C_2 \cdot K^{(L)} \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}.$$

We then reintroduce the randomness over $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ and note that

$$\Lambda \leq C_4^L \cdot \left(\prod_{j=0}^{L-1} K^{(j)} \right) \cdot N^{L/2}$$

with probability at least $(1 - 2\exp(-N))^L$ as follows from [87, Theorem 4.4.5], and

$$\mathbb{E}[\Lambda] \leq C_4^L \cdot \left(\prod_{j=0}^{L-1} K^{(j)} \right) \cdot N^{L/2}$$

as follows from [87, Exercise 4.4.6]. This implies the claim. $\square$

Theorem A.2.4 may be seen as a generalization of Theorem 2.3.1 to the setting of arbitrary sub-gaussian distributions, since in the setting of [36] we have $K^{(L)} \asymp 1$ and $K^{(j)} \asymp N^{-1/2}$ for $j \in \{0, \dots, L-1\}$. A significant difference, however, is the fact that it is more difficult to control the exact exponential dependence on L in the case of arbitrary sub-gaussian weights than in the setting of a Gaussian initialization. Exactly determining the exponential dependence in Theorem A.2.4 requires a careful bookkeeping of the absolute constants occurring in [87]. For a random matrix $A \in \mathbb{R}^{m \times n}$ with independent standard Gaussian entries, on the other hand, it is known that $\mathbb{E}[\|A\|_{2 \rightarrow 2}] \leq \sqrt{m} + \sqrt{n}$ without any occurrence of absolute constants; see [87, Theorem 7.3.1]. This makes it possible to quantify the exponential dependence on L more precisely in case of a Gaussian initialization; see Theorem 2.3.1.

A.2.2. Lower bound

We continue by proving lower bounds for random ReLU networks with arbitrary sub-gaussian distributions for the weights. This requires the following additional assumption.

Assumption A.2.5. *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (random) ReLU network as in Assumption A.2.1. We additionally assume that for every $\ell \in \{0, \dots, L\}$ all the entries of $W^{(\ell)}$ have the same standard deviation $\mathcal{V}^{(\ell)} > 0$. We set $\mathcal{R}^{(\ell)} := \mathcal{V}^{(\ell)} / K^{(\ell)}$.*

Shallow networks

Again, we start with the case of shallow networks. Here, as in [36], the proof of the lower bound is based on the fact that the ℓ^2 -Lipschitz constant of such a network can be lower bounded by the ℓ^2 -Lipschitz constant of the corresponding linear network; see [36, Proposition 4.8].

Theorem A.2.6. *There exists an absolute constant $c > 0$ such that the following holds. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network with width N and with random weights and biases following Assumption A.2.5, then*

$$(1) \quad \text{Lip}_2(\Phi) \geq \frac{1}{8} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N}$$

with probability at least $(1 - 2 \exp(-c \cdot N \cdot (\mathcal{R}^{(1)})^4))_+ (1 - 2 \exp(-c \cdot d \cdot (\mathcal{R}^{(0)})^4))_+$ and

$$(2) \quad \mathbb{E}[\text{Lip}_2(\Phi)] \geq \frac{1}{32} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N} \quad \text{if } N \geq \frac{\ln(4)}{c \cdot (\mathcal{R}^{(1)})^4} \text{ and } d \geq \frac{\ln(4)}{c \cdot (\mathcal{R}^{(0)})^4}.$$

Proof. From [36, Proposition 4.8] we know that

$$\text{Lip}_2(\Phi) \geq \frac{1}{2} \cdot \left\| W^{(1)} W^{(0)} \right\|_2.$$

We set $U := (W^{(0)})^T \in \mathbb{R}^{d \times N}$ and $V := (W^{(1)})^T \in \mathbb{R}^N$, so the goal is to establish a lower bound for $\|UV\|_2$. Let $V' := V/\mathcal{V}^{(1)}$. Then the entries of V' have unit variance and

$$\max_{i=1, \dots, N} \|V'_i\|_{\psi_2} = K^{(1)} / \mathcal{V}^{(1)} = (\mathcal{R}^{(1)})^{-1}.$$

By applying [87, Theorem 3.1.1 & Equation (2.14)] to V' , we infer the existence of an absolute constant $c' > 0$ such that for every $t \geq 0$,

$$\mathbb{P} \left(\|V'\|_2 \geq \sqrt{N} - t \right) \geq 1 - 2 \exp \left(-c' \cdot (\mathcal{R}^{(1)})^4 \cdot t^2 \right).$$

Choosing $t := \frac{\sqrt{N}}{2}$, we obtain

$$\|V\|_2 \geq \mathcal{V}^{(1)} \cdot \frac{\sqrt{N}}{2}$$

with probability at least $1 - 2 \exp(-c_1 \cdot N \cdot (\mathcal{R}^{(1)})^4)$ with an absolute constant $c_1 > 0$.

In the following, we condition on V and only consider randomness in U . We assume that $\|V\|_2 \geq \mathcal{V}^{(1)} \cdot \frac{\sqrt{N}}{2}$. For arbitrary $i \in \{1, \dots, d\}$ we then get

$$\|(UV)_i\|_{\psi_2}^2 = \left\| \sum_{k=1}^N U_{i,k} V_k \right\|_{\psi_2}^2 \leq C_1 \sum_{k=1}^N \|U_{i,k} V_k\|_{\psi_2}^2 \leq C_1 \cdot (K^{(0)})^2 \cdot \|V\|_2^2,$$

where $C_1 > 0$ is an absolute constant according to [87, Proposition 2.6.1] and where we used that the entries of U are jointly independent. Moreover, we obtain

$$\mathbb{E}_U [(UV)_i^2] = \sum_{k=1}^N \mathbb{E}_U [U_{i,k}^2 V_k^2] = (\mathcal{V}^{(0)})^2 \cdot \|V\|_2^2,$$

where at the first equality we used that the entries of U are independent and mean-zero. We then set $U' := UV/(\mathcal{V}^{(0)} \cdot \|V\|_2)$ and note that the entries of U' have unit variance and that

$$\max_{i=1,\dots,N} \|U'_i\|_{\psi_2} \leq \sqrt{C_1} \cdot (\mathcal{R}^{(0)})^{-1}.$$

We can therefore again apply [87, Theorem 3.1.1 & Equation (2.14)] to U' and get

$$\|UV\|_2 \geq \frac{1}{2} \cdot \mathcal{V}^{(0)} \cdot \|V\|_2 \cdot \sqrt{d}$$

with probability at least $1 - 2\exp(-c_2 \cdot d \cdot (\mathcal{R}^{(0)})^4)$ with an absolute constant $c_2 > 0$. By reintroducing the randomness over V and letting $c := \min\{c_1, c_2\}$ we get (1).

For (2), we assume $N \geq \frac{\ln(4)}{c \cdot (\mathcal{R}^{(1)})^4}$ and $d \geq \frac{\ln(4)}{c \cdot (\mathcal{R}^{(0)})^4}$ and apply Markov's inequality to get

$$\begin{aligned} \mathbb{E}[\text{Lip}_2(\Phi)] &\geq \mathbb{P}\left(\text{Lip}_2(\Phi) \geq \frac{1}{8} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N}\right) \cdot \frac{1}{8} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N} \\ &\geq \underbrace{(1 - 2\exp(-c \cdot N \cdot (\mathcal{R}^{(1)})^4))}_{\geq 1/2} \underbrace{(1 - 2\exp(-c \cdot d \cdot (\mathcal{R}^{(0)})^4))}_{\geq 1/2} \cdot \frac{1}{8} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N} \\ &\geq \frac{1}{32} \cdot \mathcal{V}^{(0)} \cdot \mathcal{V}^{(1)} \cdot \sqrt{d} \cdot \sqrt{N}. \end{aligned} \quad \square$$

In the case of the initialization studied in [36], we have $\mathcal{V}^{(0)} = \sqrt{2/N}$, $\mathcal{V}^{(1)} = 1$, $K^{(0)} \asymp \sqrt{2/N}$, and $K^{(1)} \asymp 1$. In particular, we have $\mathcal{R}^{(0)} \asymp \mathcal{R}^{(1)} \asymp 1$. Thus, Theorem A.2.6 is a generalization of Theorem 2.3.4 to the case of more general distributions.

Deep networks

Proving a lower bound in the case of deep networks is the part of this section that requires the most additional work, as several results from [36] have to be carefully generalized. We need the following additional assumptions.

Assumption A.2.7. *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network as in Assumption A.2.5. We additionally assume that all the weights follow continuous and symmetric distributions and that the entries of the bias vectors follow symmetric distributions (but not necessarily continuous and sub-gaussian). Lastly, we assume that the distribution of $W^{(L)}$ is rotation invariant, i.e., $W^{(L)}U \sim W^{(L)}$ for every orthogonal matrix $U \in \mathbb{R}^{N \times N}$.*

We recall from Section 2.3.2 that the proof of the lower bound for deep networks in [36] works by lower bounding the Euclidean norm of the gradient at a single fixed non-zero input $x_0 \in \mathbb{R}^d$ by an iterative application of the matrix deviation inequality [53, Theorem 3]. For the matrix deviation inequality to be applicable, one needs to show that the rows of each matrix $\sqrt{N}D^{(\ell)}(x_0)W^{(\ell)}$ are independent, isotropic, and sub-gaussian random vectors. The proof of the isotropy in [36] makes heavy use of the rotation invariance of the Gaussian distribution; see [36, Lemma 6.3]. However, using a simplified approach, one can prove the same statement for much more general distributions.

Lemma A.2.8. *Let $W \in \mathbb{R}^{N \times k}$ be a random matrix and $b \in \mathbb{R}^N$ a random vector where all the entries of W and b are jointly independent. We assume that the entries $W_{i,j}$ follow symmetric, continuous, and sub-gaussian distributions with $\mathbb{E}[W_{i,j}^2] = \frac{2}{N}$ for all suitable indices i, j . We set*

$$K := \max_{i,j} \|W_{i,j}\|_{\psi_2}.$$

Regarding the entries of b , we assume that they follow symmetric distributions (but not necessarily continuous and sub-gaussian). Let $0 \neq x \in \mathbb{R}^k$ be fixed. Then the matrix

$$V := \sqrt{N} \cdot \text{diag}(\mathbb{1}_{(Wx)_1 + b_1 > 0}, \dots, \mathbb{1}_{(Wx)_N + b_N > 0}) \cdot W \in \mathbb{R}^{N \times k}$$

has jointly independent, isotropic, and sub-gaussian rows with

$$\|V_{i,:}\|_{\psi_2} \leq C \cdot \sqrt{N} \cdot K \quad \text{for all } i \in \{1, \dots, N\},$$

where $C > 0$ is an absolute constant.

Proof. By definition, the i th row of V only depends on $W_{i,:}$ and b_i which implies that the rows are jointly independent.

For the isotropy, we fix $i \in \{1, \dots, N\}$ and two indices $j_1, j_2 \in \{1, \dots, k\}$ and observe by definition

$$\mathbb{E}[V_{i,j_1} \cdot V_{i,j_2}] = N \cdot \mathbb{E}[\mathbb{1}_{(Wx)_i + b_i > 0} \cdot W_{i,j_1} \cdot W_{i,j_2}]. \quad (\text{A.7})$$

Since $(W, b) \stackrel{d}{=} (-W, -b)$ by symmetry, we get

$$\mathbb{E}[\mathbb{1}_{(Wx)_i + b_i > 0} \cdot W_{i,j_1} \cdot W_{i,j_2}] = \mathbb{E}[\mathbb{1}_{(Wx)_i + b_i < 0} \cdot W_{i,j_1} \cdot W_{i,j_2}],$$

which gives us

$$\mathbb{E}[\mathbb{1}_{(Wx)_i + b_i > 0} \cdot W_{i,j_1} \cdot W_{i,j_2}] = \frac{1}{2} \cdot \mathbb{E}[\mathbb{1}_{(Wx)_i + b_i \neq 0} \cdot W_{i,j_1} \cdot W_{i,j_2}]. \quad (\text{A.8})$$

Note that

$$(Wx)_i + b_i = \left(\sum_{j=1}^k W_{i,j} x_j \right) + b_i$$

is a sum of independent random variables, where at least one summand follows a continuous distribution (since $x \neq 0$ and since the weights have continuous distributions). Thus, $(Wx)_i + b_i$ has a continuous distribution too.¹ Therefore,

$$\mathbb{E}[\mathbb{1}_{(Wx)_i + b_i \neq 0} \cdot W_{i,j_1} \cdot W_{i,j_2}] = \mathbb{E}[W_{i,j_1} \cdot W_{i,j_2}] = \mathbb{1}_{j_1=j_2} \cdot \frac{2}{N}.$$

Here, at the last step we used the independence of W_{i,j_1} and W_{i,j_2} if $j_1 \neq j_2$ and the fact that by symmetry we have $\mathbb{E}[W_{i,j_1}] = \mathbb{E}[W_{i,j_2}] = 0$. In view of (A.7) and (A.8), this proves the isotropy; see [87, Definition 3.2.1].

¹If X and Y are independent real-valued random variables and if, e.g., X has a continuous distribution, then we have $\mathbb{P}(X + Y = x) = \mathbb{E}[\mathbb{1}_{X+Y=x}] = \mathbb{E}_Y[\mathbb{E}_X[\mathbb{1}_{X=x-Y}]] = 0$ for every $x \in \mathbb{R}$.

Lastly, for every vector $y \in \mathbb{R}^k$ and every $i \in \{1, \dots, N\}$ we see

$$\begin{aligned} |V_{i,:} \cdot y| &= \sqrt{N} \cdot \left| \sum_{j=1}^k \mathbb{1}_{(Wx+b)_i > 0} \cdot W_{i,j} y_j \right| \\ &= \sqrt{N} \cdot \mathbb{1}_{(Wx+b)_i > 0} \cdot \left| \sum_{j=1}^k W_{i,j} y_j \right| \leq \sqrt{N} \cdot \left| \sum_{j=1}^k W_{i,j} y_j \right|. \end{aligned}$$

Note that the $W_{i,j} y_j$ ($j \in \{1, \dots, k\}$) are independent, whence

$$\left\| \sqrt{N} \cdot \sum_{j=1}^k W_{i,j} y_j \right\|_{\psi_2}^2 \leq C_1 \cdot N \cdot \sum_{j=1}^k \|W_{i,j} y_j\|_{\psi_2}^2 \leq C_1 \cdot N \cdot K^2 \cdot \|y\|_2^2,$$

with an absolute constant $C_1 > 0$, as follows from [87, Theorem 2.6.1]. Hence, by definition (see [87, Section 2.5.1 and Definition 3.4.1]) it follows that the random variable $V_{i,:} \cdot y$ is sub-gaussian and, since $y \in \mathbb{R}^k$ has been chosen arbitrarily, we deduce that $V_{i,:}$ is sub-gaussian with

$$\|V_{i,:}\|_{\psi_2} \leq \sup_{y \in \mathbb{S}^{k-1}} \left\| \sum_{j=1}^k \sqrt{N} W_{i,j} y_j \right\|_{\psi_2} \leq \sqrt{C_1} \cdot \sqrt{N} \cdot K,$$

which is as desired if we set $C := \sqrt{C_1}$. $\square$

Via an iterative application of the matrix deviation inequality, we can now prove a statement which generalizes [36, Proposition 6.5] to the setting of more general distributions.

Proposition A.2.9. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption A.2.7 with L hidden layers of width N . Set $K := \max_{\ell=0, \dots, L-1} K^{(\ell)}$ and assume $\mathcal{V}^{(\ell)} = \sqrt{2/N}$ for every $\ell \in \{0, \dots, L-1\}$. We further assume that $C \cdot K^2 \cdot L \cdot \sqrt{Nd} \leq 1$ and fix $\ell \in \{0, \dots, L-1\}$, $x \in \mathbb{R}^d \setminus \{0\}$. Then, with probability at least*

$$\prod_{j=0}^{\ell} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (K^{(j)})^{-4} \cdot N^{-1} \cdot L^{-2}) \right)$$

the following properties are simultaneously satisfied:

- (1) $\left(W^{(\ell')} \Phi^{(\ell'-1)}(x) + b^{(\ell')} \right)_i \neq 0$ for all $i \in \{1, \dots, N\}$ and $\ell' \in \{0, \dots, \ell\}$,
- (2) $\Phi^{(\ell')}(x) \neq 0$ for all $\ell' \in \{0, \dots, \ell\}$,
- (3) $\|D^{(\ell)}(x) W^{(\ell)} \dots D^{(0)}(x) W^{(0)} y\|_2 \geq \left(1 - \frac{1}{2L}\right)^{\ell+1} \|y\|_2$ for all $y \in \mathbb{R}^d$.

Proof. The proof is via induction over ℓ . We start with the case $\ell = 0$. Note that $\Phi^{(-1)}(x) = x$ per definition and we hence get

$$\left(W^{(0)} \Phi^{(-1)}(x) + b^{(0)} \right)_i = \left(W^{(0)} x + b^{(0)} \right)_i$$

for every $i \in \{1, \dots, N\}$, where the latter random variable follows a continuous and symmetric distribution, which follows with the same reasoning as in the proof of Lemma A.2.8. From the

continuity, we infer that (1) (with $\ell' = 0$) occurs even almost surely. Moreover, since

$$\Phi^{(0)}(x) = 0 \iff \left(W^{(0)}x + b^{(0)}\right)_i \leq 0 \text{ for all } i \in \{1, \dots, N\},$$

and due to symmetry, independence, and continuity, we obtain

$$\mathbb{P}\left(\Phi^{(0)}(x) \neq 0\right) \geq 1 - \frac{1}{2^N}. \quad (\text{A.9})$$

Further, note that $\sqrt{N}D^{(0)}(x)W^{(0)}$ is a matrix with independent isotropic sub-gaussian rows according to Lemma A.2.8. Therefore, the high probability version of the matrix deviation inequality (see [53, Theorem 3]) yields for every $u \geq 0$,

$$\sup_{y \in B^d} \left| \left\| \sqrt{N}D^{(0)}(x)W^{(0)}y \right\|_2 - \sqrt{N} \|y\|_2 \right| \leq C_1 \cdot (K^{(0)})^2 \cdot N \cdot (\sqrt{d} + u)$$

with probability at least $1 - \exp(-u^2)$, where we employed [36, Proposition 4.6]. $C_1 > 0$ is an absolute constant according to Lemma A.2.8. Hence, we get that for every $u \geq 0$,

$$\|D^{(0)}(x)W^{(0)}y\|_2 \geq \left(1 - C_1(K^{(0)})^2\sqrt{N}(\sqrt{d} + u)\right)_+ \|y\|_2 \text{ for all } y \in \mathbb{R}^d$$

with probability at least $1 - \exp(-u^2)$. By taking $C > 0$ large enough, we have

$$C_1 \cdot (K^{(0)})^2 \cdot \sqrt{Nd} \leq \frac{1}{4L}$$

and by picking $u := \frac{1}{4C_1 \cdot (K^{(0)})^2 \cdot \sqrt{N} \cdot L}$, we obtain

$$\|D^{(0)}(x)W^{(0)}y\|_2 \geq \left(1 - \frac{1}{2L}\right) \cdot \|y\|_2 \text{ for all } y \in \mathbb{R}^d$$

with probability at least

$$1 - \exp(-c \cdot (K^{(0)})^{-4} \cdot N^{-1} \cdot L^{-2})$$

with an absolute constant $c > 0$. Combining this with (A.9), we obtain the claim in the case $\ell = 0$.

For the induction step, we assume that for some fixed $\ell \in \{0, \dots, L-2\}$ the properties (1)-(3) are satisfied with probability at least

$$\prod_{j=0}^{\ell} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (K^{(j)})^{-4} \cdot N^{-1} \cdot L^{-2})\right).$$

Note that these properties only depend on $W^{(0)}, \dots, W^{(\ell)}$ and $b^{(0)}, \dots, b^{(\ell)}$. We can thus condition on these random variables and assume that properties (1)-(3) are satisfied for ℓ . In particular, we then have $\Phi^{(\ell)}(x) \neq 0$ and thus, each random variable

$$\left(W^{(\ell+1)}\Phi^{(\ell)}(x) + b^{(\ell+1)}\right)_i$$

for $i \in \{1, \dots, N\}$ follows a symmetric and continuous distribution (if the randomness is only with respect to $W^{(\ell+1)}$ and $b^{(\ell+1)}$). This again follows with the same reasoning as in the proof of Lemma A.2.8. Due to continuity, we in particular have

$$\left(W^{(\ell+1)}\Phi^{(\ell)}(x) + b^{(\ell+1)}\right)_i \neq 0 \text{ for all } i \in \{1, \dots, N\}$$

with probability 1, if the randomness is only with respect to $W^{(\ell+1)}$ and $b^{(\ell+1)}$. Moreover,

$$\Phi^{(\ell+1)}(x) = 0 \iff \left(W^{(\ell+1)} \Phi^{(\ell)}(x) + b^{(\ell+1)} \right)_i \leq 0 \quad \text{for all } i \in \{1, \dots, N\}.$$

Due to symmetry, independence, and continuity, we get

$$\mathbb{P} \left(\Phi^{(\ell+1)}(x) \neq 0 \right) \geq 1 - \frac{1}{2^N} \quad (\text{A.10})$$

where the randomness is only taken with respect to $W^{(\ell+1)}$ and $b^{(\ell+1)}$. For property (3), we set $V' := D^{(\ell)}(x)W^{(\ell)} \dots D^{(0)}(x)W^{(0)}$. According to Lemma A.2.8, the matrix $\sqrt{N}D^{(\ell+1)}(x)W^{(\ell+1)}$ has independent, sub-gaussian, isotropic rows. Hence, we may again apply the high-probability version of the matrix deviation inequality ([53, Theorem 3]) and obtain

$$\sup_{y \in \text{Im}(V') \cap \overline{B}_N(0,1)} \left| \sqrt{N} \left\| D^{(\ell+1)}(x)W^{(\ell+1)}y \right\|_2 - \sqrt{N} \|y\|_2 \right| \leq C_1 (K^{(\ell+1)})^2 N (\sqrt{d} + u)$$

with probability at least $1 - \exp(-u^2)$ for every $u \geq 0$, where we again used [36, Proposition 4.6], noting that the subspace $\text{Im}(V')$ is at most d -dimensional since $V' \in \mathbb{R}^{N \times d}$. But this directly implies

$$\|D^{(\ell+1)}(x)W^{(\ell+1)}y\|_2 \geq \left(1 - C_1 (K^{(\ell+1)})^2 \sqrt{N}(\sqrt{d} + u) \right)_+ \|y\|_2 \quad \text{for all } y \in \text{Im}(V')$$

with probability at least $1 - \exp(-u^2)$. As in the case $\ell = 0$, this implies

$$\|D^{(\ell+1)}(x)W^{(\ell+1)}y\|_2 \geq \left(1 - \frac{1}{2L} \right) \|y\|_2 \quad \text{for all } y \in \text{Im}(V')$$

with probability at least $1 - \exp(-c \cdot (K^{(\ell+1)})^{-4} \cdot N^{-1} \cdot L^{-2})$, which implies

$$\begin{aligned} \|D^{(\ell+1)}(x)W^{(\ell+1)} \dots D^{(0)}(x)W^{(0)}y\|_2 &\geq \left(1 - \frac{1}{2L} \right) \|D^{(\ell)}(x)W^{(\ell)} \dots D^{(0)}(x)W^{(0)}y\|_2 \\ &\geq \left(1 - \frac{1}{2L} \right)^{\ell+1} \cdot \|y\|_2 \quad \text{for all } y \in \mathbb{R}^d. \end{aligned}$$

Overall, we thus get that properties (1)-(3) are satisfied with probability at least

$$1 - \frac{1}{2^N} - \exp(-c \cdot (K^{(\ell+1)})^{-4} \cdot N^{-1} \cdot L^{-2})$$

if the randomness is only with respect to $W^{(\ell+1)}$ and $b^{(\ell+1)}$. Reintroducing the randomness over $W^{(0)}, \dots, W^{(\ell)}$ and $b^{(0)}, \dots, b^{(\ell)}$ yields the claim. $\square$

Using the rotation invariance of $W^{(L)}$ we can now incorporate the final layer to get a lower bound on the Euclidean norm of the gradient at x .

Proposition A.2.10. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption A.2.7 with L hidden layers of width N . Set $K := \max_{\ell=0, \dots, L-1} K^{(\ell)}$ and assume $\mathcal{V}^{(\ell)} = \sqrt{2/N}$ for every $\ell \in \{0, \dots, L-1\}$. We assume that $C \cdot K^2 \cdot L \cdot \sqrt{Nd} \leq 1$ and fix $x \in \mathbb{R}^d \setminus \{0\}$. We set $k := \min\{d, N\}$. Then, with probability at*

least

$$\left[\prod_{j=0}^{L-1} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (K^{(j)})^{-4} \cdot N^{-1} \cdot L^{-2}) \right) \right] \left(1 - 2 \exp(-c \cdot k \cdot (\mathcal{R}^{(L)})^4) \right)$$

the following properties are simultaneously satisfied:

- (1) $(W^{(\ell)} \Phi^{(\ell-1)}(x) + b^{(\ell-1)})_i \neq 0$ for all $i \in \{1, \dots, N\}$ and $\ell \in \{0, \dots, L-1\}$,
- (2) $\|\bar{\nabla} \Phi(x)\|_2 \geq \frac{1}{4} \cdot \mathcal{V}^{(L)} \cdot \sqrt{k}$.

Proof. We condition on $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ and assume that the properties from Proposition A.2.9 with $\ell = L-1$ are satisfied. Note that this in particular implies (1).

We set $V' := D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}$ and let $V' = U\Sigma V^T$ be the singular value decomposition of V' . Here, $U \in \mathbb{R}^{N \times N}$ and $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{N \times d}$ is of the form

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_d & \\ 0 & \dots & 0 & \\ \vdots & \ddots & \vdots & \\ 0 & \dots & 0 & \end{pmatrix} \quad \text{if } N \geq d \quad \text{or} \quad \Sigma = \begin{pmatrix} \sigma_1 & & 0 & \dots & 0 \\ & \ddots & \vdots & \ddots & \vdots \\ & & \sigma_N & 0 & \dots & 0 \end{pmatrix} \quad \text{if } N \leq d$$

with $\sigma_1 \geq \dots \geq \sigma_k > 0$. Recall that we assume the properties from Proposition A.2.9 to be satisfied, which implies, according to [87, Equation (4.5)],

$$\sigma_k \geq \left(1 - \frac{1}{2L} \right)^L \geq \frac{1}{2},$$

where the last step uses Bernoulli's inequality. Using the rotation invariance of $W^{(L)}$, we obtain

$$\|W^{(L)}V'\|_2 = \|W^{(L)}U\Sigma V^T\|_2 = \|W^{(L)}U\Sigma\|_2 \stackrel{d}{=} \|W^{(L)}\Sigma\|_2,$$

where the distributional equivalence is to be understood with respect to the randomness in $W^{(L)}$. Note that

$$\|W^{(L)}\Sigma\|_2 = \sqrt{\sum_{j=1}^k (\sigma_j \cdot W_{1,j}^{(L)})^2} \geq \sigma_k \cdot \|W'\|_2,$$

where $W' \in \mathbb{R}^{1 \times k}$ arises from $W^{(L)}$ by only picking the first k entries. Applying [87, Theorem 3.1.1] to the rescaled vector $W'/\mathcal{V}^{(L)}$, we get

$$\|W'\|_2 \geq \mathcal{V}^{(L)} \cdot \frac{\sqrt{k}}{2}$$

with probability at least $1 - 2 \exp(-c_1 \cdot k \cdot (\mathcal{R}^{(L)})^4)$ with an absolute constant $c_1 > 0$. Reintroducing the randomness over $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ yields the claim. $\square$

Up to now, we have assumed that $\mathcal{V}^{(\ell)} = \sqrt{2/N}$ for every $\ell \in \{0, \dots, L-1\}$. Via a rescaling argument, we can generalize the statement to arbitrary choices of $\mathcal{V}^{(\ell)}$.

Corollary A.2.11. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption A.2.7 with L hidden layers of width N . Set $\mathcal{R} := \min_{\ell=0, \dots, L-1} \mathcal{R}^{(\ell)}$ and $k := \min\{d, N\}$. We assume that $C \cdot \mathcal{R}^{-4} \cdot L^2 \cdot d \leq N$ and fix $x \in \mathbb{R}^d \setminus \{0\}$. Then, with probability at least*

$$\left[\prod_{j=0}^{L-1} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (\mathcal{R}^{(j)})^4 \cdot N/L^2) \right) \right] \left(1 - 2 \exp(-c \cdot k \cdot (\mathcal{R}^{(L)})^4) \right)$$

the following properties are simultaneously satisfied:

(1) Φ is differentiable at x with $\nabla \Phi(x) = \bar{\nabla} \Phi(x)$,

(2) $\|\bar{\nabla} \Phi(x)\|_2 \geq \frac{1}{4} \cdot \left(\prod_{j=0}^L \mathcal{V}^{(j)} \right) \cdot \left(\frac{\sqrt{N}}{\sqrt{2}} \right)^L \cdot \sqrt{k}$.

Proof. For a random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfying Assumption A.2.7, we define the rescaled network $\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}$ to have the weight matrices and bias vectors

$$\tilde{W}^{(\ell)} := W^{(\ell)} \cdot \frac{\sqrt{2}}{\mathcal{V}^{(\ell)} \cdot \sqrt{N}}, \quad \tilde{b}^{(\ell)} := b^{(\ell)} \cdot \prod_{j=0}^{\ell} \left[\frac{\sqrt{2}}{\mathcal{V}^{(j)} \cdot \sqrt{N}} \right], \quad \ell \in \{0, \dots, L-1\}$$

and $\tilde{W}^{(L)} := W^{(L)}$. Then, via induction, it is easy to see that

$$\left(\tilde{W}^{(\ell)} \tilde{\Phi}^{(\ell-1)}(x) + \tilde{b}^{(\ell)} \right) = \left[\prod_{j=0}^{\ell} \left(\frac{\sqrt{2}}{\mathcal{V}^{(j)} \cdot \sqrt{N}} \right) \right] \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) + b^{(\ell)} \right) \quad (\text{A.11})$$

for all $x \in \mathbb{R}^d$ and $\ell \in \{-1, \dots, L-1\}$. This implies $\tilde{D}^{(\ell)}(x) = D^{(\ell)}(x)$ for every $\ell \in \{0, \dots, L-1\}$, where $\tilde{D}^{(\ell)}(x)$ denotes the ℓ th D -matrix of the rescaled network $\tilde{\Phi}$ at x . Lastly, we note that

$$\tilde{K}^{(\ell)} = (K^{(\ell)} / \mathcal{V}^{(\ell)}) \cdot \sqrt{2/N} = \frac{1}{\mathcal{R}^{(\ell)}} \cdot \sqrt{2/N}, \quad \ell \in \{0, \dots, L-1\},$$

where $\tilde{K}^{(\ell)}$ refers to the maximal sub-gaussian norm of the entries of $\tilde{W}^{(\ell)}$. With $\tilde{K} := \max_{\ell=0, \dots, L-1} \tilde{K}^{(\ell)}$ it is clear that $\tilde{K} = \frac{1}{\mathcal{R}} \cdot \sqrt{2/N}$. Hence, if C is chosen sufficiently large, the assumption $C \cdot \mathcal{R}^{-4} \cdot L^2 \cdot d \leq N$ implies $C' \cdot \tilde{K}^2 \cdot L \cdot \sqrt{Nd} \leq 1$, where $C' > 0$ is the constant given by Proposition A.2.10. Consequently, $\tilde{\Phi}$ satisfies the conditions of Proposition A.2.10 and we obtain

$$\|\bar{\nabla} \Phi(x)\|_2 = \left[\prod_{j=0}^{L-1} \mathcal{V}^{(j)} \right] \cdot \left(\frac{\sqrt{N}}{\sqrt{2}} \right)^L \cdot \|\bar{\nabla} \tilde{\Phi}(x)\|_2 \geq \frac{1}{4} \cdot \left[\prod_{j=0}^L \mathcal{V}^{(j)} \right] \cdot \left(\frac{\sqrt{N}}{\sqrt{2}} \right)^L \cdot \sqrt{k}$$

with probability at least

$$\left[\prod_{j=0}^{L-1} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (\mathcal{R}^{(j)})^4 \cdot N/L^2) \right) \right] \left(1 - 2 \exp(-c \cdot k \cdot (\mathcal{R}^{(L)})^4) \right)$$

for an absolute constant $c > 0$ as follows from Proposition A.2.10. This proves (2).

On the same event, as stated in Proposition A.2.10, we have

$$\left(\widetilde{W}^{(\ell)}\widetilde{\Phi}^{(\ell-1)}(x) + \widetilde{b}^{(\ell)}\right)_i \neq 0$$

for every $i \in \{1, \dots, N\}$ and $\ell \in \{0, \dots, L-1\}$. By (A.11), we then also have

$$\left(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}\right)_i \neq 0$$

for every $i \in \{1, \dots, N\}$ and $\ell \in \{0, \dots, L-1\}$. But since the ReLU is differentiable on $\mathbb{R} \setminus \{0\}$ with $\text{ReLU}'(y) = 0$ for $y < 0$ and $\text{ReLU}'(y) = 1$ for $y > 0$, we get (1) via an iterative application of the chain rule. $\square$

Corollary A.2.11 directly implies a lower bound for the Lipschitz constant of Φ .

Theorem A.2.12. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption A.2.7 with L hidden layers of width N . We set $\mathcal{R} := \min_{\ell=0, \dots, L-1} \mathcal{R}^{(\ell)}$ and $k := \min\{d, N\}$. We assume that $C \cdot \mathcal{R}^{-4} \cdot L^2 \cdot d \leq N$. Then with probability at least*

$$\left[\prod_{j=0}^{L-1} \left(1 - \frac{1}{2^N} - \exp(-c \cdot (\mathcal{R}^{(j)})^4 \cdot N/L^2) \right) \right] \left(1 - 2 \exp(-c \cdot k \cdot (\mathcal{R}^{(L)})^4) \right)$$

we have

$$\text{Lip}_2(\Phi) \geq \frac{1}{4} \cdot \left(\prod_{j=0}^L \mathcal{V}^{(j)} \right) \cdot \left(\frac{\sqrt{N}}{\sqrt{2}} \right)^L \cdot \sqrt{k}.$$

Proof. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (deterministic) ReLU network and $x \in \mathbb{R}^d$ with the property that Φ is differentiable at x . If $\nabla \Phi(x) \neq 0$, we set $\nu := \nabla \Phi(x) / \|\nabla \Phi(x)\|_2$ and get

$$\|\nabla \Phi(x)\|_2 = \langle \nabla \Phi(x), \nu \rangle = \lim_{t \rightarrow 0} \frac{|\Phi(x + t\nu) - \Phi(x)|}{t} = \lim_{t \rightarrow 0} \frac{|\Phi(x + t\nu) - \Phi(x)|}{\|(x + t\nu) - x\|_2} \leq \text{Lip}_2(\Phi).$$

This inequality trivially remains true in the case $\nabla \Phi(x) = 0$. The claim thus follows immediately from Corollary A.2.11. $\square$

We note that in the setting studied in [36] we have $\mathcal{R}^{(\ell)} \asymp 1$ for every $\ell \in \{0, \dots, L\}$, $\mathcal{V}^{(\ell)} = \frac{\sqrt{2}}{\sqrt{N}}$ for $\ell \in \{0, \dots, L-1\}$ and $\mathcal{V}^{(L)} = 1$. Thus, Theorem A.2.12 generalizes Theorem 2.3.5 to the setting of more general initializations.

B. Full texts of the papers

This appendix contains the full texts of the five papers [23, 34, 35, 36, 37] that this thesis is based on.

B.1. Upper and lower bounds for the Lipschitz constant of random neural networks

Upper and lower bounds for the Lipschitz constant of random neural networks

PAUL GEUCHEN 

Mathematical Institute for Machine Learning and Data Science (MIDS), KU Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany

*Corresponding author. Email: paul.geuchen@ku.de

DOMINIK STÖGER 

MIDS, KU Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany

THOMAS TELAAR

Chair of Mathematics—Statistics, KU Eichstätt-Ingolstadt, Ostenstraße 28, 85072 Eichstätt, Germany

AND

FELIX VOIGTLAENDER

MIDS, KU Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany

[Received on 25 March 2024; accepted on 24 February 2025]

Empirical studies have widely demonstrated that neural networks are highly sensitive to small, adversarial perturbations of the input. The worst-case robustness against these so-called adversarial examples can be quantified by the Lipschitz constant of the neural network. In this paper, we study upper and lower bounds for the Lipschitz constant of *random ReLU neural networks*. Specifically, we assume that the weights and biases follow a generalization of the He initialization, where general symmetric distributions for the biases are permitted. For deep networks of fixed depth and sufficiently large width, our established upper bound is larger than the lower bound by a factor that is logarithmic in the width. In contrast, for shallow neural networks we characterize the Lipschitz constant up to an absolute numerical constant that is independent of all parameters.

Keywords: Lipschitz constant; random neural networks; adversarial robustness.

1. Introduction

Deep neural networks have achieved remarkable success across a diverse range of applications in the past decade. However, it is well-known that small, adversarially chosen perturbations of the input can drastically alter the output of the neural network. These perturbations, which often can be chosen to be invisible to the human eye, are known in the literature as *adversarial examples*. One of the first papers to empirically examine this phenomenon was [40]. This observation has led to an extensive body of research, focusing both on devising increasingly sophisticated adversarial attacks (see, for example, [3, 12, 26, 32, 35]) and on constructing robust defenses against such attacks (see, for example, [11, 17, 28, 34, 36]). Arguably, this has resulted in an ‘arms race’ between adversarial attacks and defenses. The theoretical understanding of adversarial examples, however, remains limited.

A natural measure for the worst-case robustness against adversarial examples is given by the Lipschitz constant of a neural network function $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, which is defined via

$$\text{Lip}(\Phi) := \sup_{x, y \in \mathbb{R}^d, x \neq y} \frac{|\Phi(x) - \Phi(y)|}{\|x - y\|_2},$$

where by $\|\cdot\|_2$ we denote the ℓ_2 -norm of a vector. Indeed, the Lipschitz constant measures how much a small change of the input can perturb the output. A small Lipschitz constant may imply that no adversarial examples exist, whereas a large Lipschitz constant is indicative of the existence of adversarial examples. For these reasons, several works have proposed algorithms to determine upper bounds for the Lipschitz constant of a trained deep neural network—see, e.g., [19, 21, 39]—in order to quantify the worst-case robustness against adversarial examples.

Random neural networks, i.e., neural networks whose weights and biases are chosen at random, are often used as a first step in theoretically understanding empirically observed phenomena in deep learning; see for instance [4, 15, 16]. In this spirit, in this paper, we establish upper and lower bounds for the Lipschitz constant of random ReLU neural networks. Our paper makes the following two main contributions:

1. In the case of shallow neural networks with ReLU activation function and random weights following a variant of the He initialization [25], we prove high-probability upper and lower bounds for the Lipschitz constant. These bounds match up to an absolute numerical constant.
2. In the case of deep neural networks with ReLU activation function, we again derive high-probability upper and lower bounds for the Lipschitz constant. For this, we assume that our networks are wide enough, i.e., they satisfy $N \gtrsim L^2 d$, where N is the width of the neural network, d is the dimension of the input and L is the depth of the neural network. If the depth L is assumed to be a constant, our upper and lower bounds match up to a factor that is logarithmic in the width.

Regarding the biases, in our results we allow arbitrary symmetric distributions, whereas earlier works on the Lipschitz constant of neural networks [10, 33] require all biases to be initialized to zero, as in the classical He initialization.

The paper is structured as follows: Section 2 provides technical background regarding random ReLU networks. In Section 3, we present our main theoretical results and in Section 4 we review some technical tools needed for our proofs. The proofs regarding the upper bound of the Lipschitz constant are given in Section 5. The corresponding lower bound is derived in Section 6.

1.1 Related work

In [8], it has been conjectured that any shallow neural network (with one hidden layer) that interpolates the training data must have large Lipschitz constant, unless the neural network is extremely overparameterized. In [9], this conjecture was proven and generalized to a much broader class of models beyond (shallow) neural networks. However, while this result states that neural networks interpolating the data can only be adversarially robust if they are extremely overparameterized, it does not imply that overparameterized networks are inherently adversarially robust; this is because only a *lower* bound for the Lipschitz constant (which tends to zero as the number of network parameters goes to infinity) is presented. Moreover, when the data is no longer *perfectly* interpolated, the lower bound for the Lipschitz constant in [9] no longer applies. This leaves open the possibility of training adversarially robust neural networks without huge overparameterization under such conditions. For these reasons, a more detailed analysis of the Lipschitz constant of neural networks is essential.

Upper bounds for the Lipschitz constants of random neural networks were derived in [10, 33] in the context of the lazy training regime [13]. In contrast to our work, these two results do not contain any lower bounds. We refer to Remark 3.5 for a detailed comparison of the results in [10, 33] with our bounds.

A *lower* bound on the gradient of random ReLU networks at a fixed point was derived in [4] in the context of proving the existence of an adversarial example at a fixed point. This result immediately implies a lower bound on the Lipschitz constant as well, since random ReLU networks are almost surely differentiable at a fixed point (cf. Appendix E). We refer to Remark 3.6 for a comparison to the lower bound that we derive.

Our analysis of the Lipschitz constant is based on bounding the *supremum* of the norm of the gradients of a random ReLU network. Regarding the gradient at a fixed point, [23] even characterized the asymptotic distribution of the norm of the directional derivative in a fixed direction, which is shown to be a log normal distribution. However, similar to [33] this bound only considers the feature maps, i.e., the neural network excluding the output layer.

A line of research related to studying the Lipschitz constant of neural networks is concerned with proving the existence of adversarial perturbations in *random* neural networks. The first result was obtained in [38], where it was shown that by only perturbing a few entries of the input vector one can construct an adversarial example. Realizing that these perturbations may be very large, this result was further improved in [15], where it was shown that adversarial examples can be constructed with much smaller perturbations. However, this result applies only to random ReLU networks for which the width of each layer is small relative to the width of the previous layer and for which the penultimate layer still needs to be large enough (depending on the input dimension). In the case of shallow networks, [7] improved the latter result for networks with ReLU activation or smooth activation whose width depends subexponentially on the input dimension. The follow-up work [4] extended this result to the multi-layer setting. Recently, this has been further improved in [29], which presented a result that requires no restriction on the width of the neural network and which applies to general non-constant locally Lipschitz continuous activations. All of the above results, however, are of a somewhat different nature compared to the results of the paper at hand. The above results show, for an arbitrarily *fixed* input, the existence, with high probability, of a small adversarial perturbation. In contrast, our paper considers the adversarial robustness over all possible inputs and, thereby, characterizes worst-case adversarial robustness.

Finally, let us note that going beyond random neural networks, in [44] the existence of adversarial examples was recently shown for trained neural networks in the lazy training regime [13]. Moreover, in [22, 42] it was shown that in neural networks with one hidden layer, the bias of gradient flow towards margin maximization induces the existence of adversarial examples. In [6], adversarial robustness of (trained) random features and neural tangent kernel regression models was studied. Here, adversarial robustness was measured with respect to a sensitivity measure, which is the ℓ_2 -norm of the gradient at a data point scaled by the ℓ_2 -norm of this data point. This sensitivity measure was analyzed for one fixed vector from the input data distribution, i.e., the *average-case* adversarial robustness was analyzed.

1.2 Notation

For any mathematical statement A , we set $\mathbb{1}_A = 1$ if A is true and $\mathbb{1}_A = 0$ otherwise. The notation $\mathbb{1}_A$ will also be used for the indicator function of a set A . We write $a_+ := \max\{0, a\}$ for $a \in \mathbb{R}$. For two quantities α and β we write $\alpha \lesssim \beta$ if there exists an absolute constant $C > 0$ with $\alpha \leq C \cdot \beta$. Moreover, we write $\alpha \gtrsim \beta$ if $\beta \lesssim \alpha$ and $\alpha \asymp \beta$ if both $\alpha \lesssim \beta$ and $\beta \lesssim \alpha$.

For two vectors $x, y \in \mathbb{R}^k$, we write $\langle x, y \rangle := \sum_{i=1}^k x_i y_i$ for the inner product of x and y . Moreover, we denote by $\|x\|_2 := \sqrt{\langle x, x \rangle}$ the *Euclidean norm* of x . Given a metric space T with metric ϱ , for $x \in T$ and $r > 0$ we write

$$B_T^\varrho(x, r) := \{y \in T : \varrho(x, y) < r\}, \quad \overline{B}_T^\varrho(x, r) := \{y \in T : \varrho(x, y) \leq r\}.$$

When considering the Euclidean norm on $\mathbb{R}^k$, we denote the open and closed ball by $B_k(x, r)$ and $\bar{B}_k(x, r)$, respectively. Furthermore, we denote $\mathbb{S}^{k-1} := \{x \in \mathbb{R}^k : \|x\|_2 = 1\}$. For a subset $T \subseteq \mathbb{R}^k$, we denote the diameter of T as $\text{diam}(T) := \sup_{x, y \in T} \|x - y\|_2 \in [0, \infty]$.

For a matrix $A \in \mathbb{R}^{k_1 \times k_2}$, we define

$$\|A\|_2 := \sup_{x \in \bar{B}_{k_2}(0,1)} \|Ax\|_2.$$

By $A_{i,-}$ and $A_{-,j}$ we denote the i -th row and j -th column of A , respectively. We write $v \odot w \in \mathbb{R}^k$ for the *Hadamard product* of two vectors $v, w \in \mathbb{R}^k$, which is defined as $(v \odot w)_j := v_j \cdot w_j$ for $j \in \{1, \dots, k\}$.

For any set A and an element $a \in A$, we denote by δ_a the *Dirac-measure* at a . For $\sigma > 0$, we write $\mathcal{N}(0, \sigma^2)$ for the standard normal distribution with expectation 0 and variance σ^2 . We write $\mathcal{N}(0, I_k)$ for the distribution of a k -dimensional random vector with independent $\mathcal{N}(0, 1)$ -entries. Moreover, we write $\mathbb{E}[X]$ for the expectation of a random variable X . By $\mathbb{P}^X$ we denote the probability measure induced by some random variable X , i.e., $\mathbb{P}^X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$. Furthermore, we denote A^c for the complement of any event A .

2. Background

In the present paper, we derive bounds on the Lipschitz constant of random ReLU networks at initialization. We consider a variant of the so-called *He initialization* as introduced in [25]. In [25], the entries of the weight matrices $W^{(\ell)} \in \mathbb{R}^{N^{(\ell+1)} \times N^{(\ell)}}$ are identically and independently drawn from a Gaussian distribution with zero mean and standard deviation $\sqrt{2/N^{(\ell)}}$ and the biases are initialized to zero. We consider a standard deviation of $\sqrt{2/N^{(\ell+1)}}$, which is commonly used in theoretical studies of random ReLU networks (cf. [2, 10, 16]). This choice of the random initialization slightly simplifies the proofs and is natural since with this initialization, each layer is *isometric in expectation*, meaning that $\mathbb{E}[\|\text{ReLU}(W^{(\ell)}x)\|_2^2] = \|x\|_2^2$ for every vector $x \in \mathbb{R}^{N^{(\ell)}}$ that is fed into the ℓ th layer of the network. Since the ReLU is positively homogeneous, our results readily imply corresponding bounds for other initialization choices where the network weights are normally distributed, possibly even with a different variance in different layers. In particular, for the usual He initialization, one has to multiply all our upper/lower bounds for Lipschitz constants by $1/\sqrt{d}$.

Moreover, we allow more general biases than in [25]. Specifically, we require that the biases are also drawn independently from certain probability distributions. In order to derive upper bounds for the Lipschitz constant (see Section 5) and lower bounds in the case of shallow networks (see Section 6.1), we do not need to impose *any* assumptions on these probability distributions. To derive lower bounds in the case of deep networks (see Section 6.2), however, we require these probability distributions to be symmetric. Note that assuming symmetry of the distributions of the biases is very natural and in particular covers the zero initialization that is used in [25]. In the following assumption, we formally introduce the considered random initialization.

ASSUMPTION 2.1. We consider ReLU networks with d input neurons, 1 output neuron, a width of N and L hidden layers. Precisely, we consider maps

$$\Phi : \mathbb{R}^d \rightarrow \mathbb{R}, \quad \Phi(x) := \left(V^{(L)} \circ \text{ReLU} \circ V^{(L-1)} \circ \dots \circ \text{ReLU} \circ V^{(0)} \right)(x). \quad (2.1)$$

Here, ReLU denotes the function defined as

$$\text{ReLU}(x) := \max\{0, x\}, \quad x \in \mathbb{R},$$

and the application of ReLU in (2.1) is to be understood componentwise, i.e.,

$$\text{ReLU}((x_1, \dots, x_N)) = (\text{ReLU}(x_1), \dots, \text{ReLU}(x_N)).$$

Furthermore, the maps $V^{(\ell)}$ for $0 \leq \ell \leq L$ are affine-linear maps, which in detail means the following: There exist matrices $W^{(0)} \in \mathbb{R}^{N \times d}$, $W^{(\ell)} \in \mathbb{R}^{N \times N}$ for $1 \leq \ell \leq L-1$ and $W^{(L)} \in \mathbb{R}^{1 \times N}$ as well as *biases* $b^{(\ell)} \in \mathbb{R}^N$ for $0 \leq \ell < L$ and $b^{(L)} \in \mathbb{R}$ such that

$$V^{(\ell)}(x) = W^{(\ell)}x + b^{(\ell)} \quad \text{for every } 0 \leq \ell \leq L.$$

We assume that the matrices $W^{(\ell)}$ as well as the biases $b^{(\ell)}$ are randomly chosen in the following way: For $0 \leq \ell < L$, we have

$$\left(W^{(\ell)}\right)_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N), \quad \left(W^{(L)}\right)_{1,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

Concerning the biases, we assume

$$\left(b^{(\ell)}\right)_i \sim \mathcal{D}_i^{(\ell)}, \quad 0 \leq \ell \leq L.$$

Here, each $\mathcal{D}_i^{(\ell)}$ is an arbitrary probability distribution over $\mathbb{R}$. Furthermore, the entries of the random variables $W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)}$ are assumed to be jointly independent.

The above assumption, which suffices for deriving the upper bound on the Lipschitz constant for networks of arbitrary depth and the lower bound in the case of shallow networks, imposes almost no restrictions on the distribution of the biases. However, for deriving lower bounds for deep networks we will use the following more restrictive assumption on the initialization of the biases.

ASSUMPTION 2.2. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network satisfying Assumption 2.1. Moreover, we assume that each $\mathcal{D}_i^{(\ell)}$ is a symmetric probability distribution over $\mathbb{R}$ (e.g., $\mathcal{N}(0, \sigma^2)$ for some parameter $\sigma > 0$, uniform on some interval $[-m, m] \subseteq \mathbb{R}$, or initialized to zero).

We restrict ourselves to square matrices (i.e., we assume that each layer of the network has the same amount of neurons). We believe, however, that our proofs still work in the case of layers with varying widths that are uniformly bounded from below by $\Omega(dL^2)$.

3. Main results

In this section, we discuss the main results of the paper. In the case of shallow networks, we present upper and lower bounds on the Lipschitz constant that hold almost without any restrictions on the input dimension d and the network width N . Moreover, the upper and lower bounds match up to an absolute

multiplicative constant. In the case of deep networks, we have to impose the additional assumption $N \gtrsim dL^2$. Furthermore, there is a gap between the upper and lower bounds; this gap grows exponentially with the depth L of the network and logarithmically with the network width N .

3.1 Shallow networks

We provide a high-probability version and an expectation version for the upper and lower bounds. The result for the upper bound reads as follows:

THEOREM 3.1. There exist absolute constants $C, c_1 > 0$ satisfying the following: If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network of arbitrary width $N \in \mathbb{N}$ with random weights and random biases following Assumption 2.1, then the following hold:

1. We have

$$\text{Lip}(\Phi) \leq C \cdot \sqrt{d}$$

with probability at least $(1 - 2 \exp(-\min\{d, N\})) \cdot (1 - 2 \exp(-c_1 \cdot \max\{d, N\}))_+$.

2. $\mathbb{E}[\text{Lip}(\Phi)] \leq C \cdot \sqrt{d}$.

The proof of Theorem 3.1 can be found in Section 5.1.

The upper bound from above is complemented by the following lower bound, which indeed shows that the bounds are tight up to multiplicative constants.

THEOREM 3.2. There exist absolute constants $c, c_1 > 0$ satisfying the following: If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network with width N and with random weights and biases following Assumption 2.1, then the following hold:

1. The inequality

$$\text{Lip}(\Phi) \geq \frac{1}{4\sqrt{2}} \cdot \sqrt{d}$$

holds with probability at least $(1 - 2 \exp(-c_1 N))_+ (1 - 2 \exp(-c_1 d))_+$.

2. If $d, N > \frac{\ln(2)}{c_1}$, it holds

$$\mathbb{E}[\text{Lip}(\Phi)] \geq c \cdot \sqrt{d}.$$

The proof of Theorem 3.2 can be found in Section 6.1.

The two results combined show that in the case of shallow networks it holds

$$\text{Lip}(\Phi) \asymp \sqrt{d},$$

where the involved constants are absolute. This holds with high probability and in expectation and moreover without almost any restrictions on the width N or the input dimension d .

3.2 Deep networks

In the case of deep networks, our upper bound for the Lipschitz constant reads as follows:

THEOREM 3.3. There exist absolute constants $C, c_1 > 0$ such that, if $N > d + 2$ and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network of width N and with L hidden layers with random weights and biases following Assumption 2.1, the following hold:

1. The inequality

$$\text{Lip}(\Phi) \leq C \cdot (3\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}$$

holds with probability at least $(1 - 2\exp(-d)) \left((1 - 2\exp(-c_1 N))_+\right)^L$.

2. $\mathbb{E}[\text{Lip}(\Phi)] \leq C \cdot (2\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}$.

The proof of Theorem 3.3 can be found in Section 5.2. We remark that for the upper bounds in the case of deep networks above, we impose the assumption $N > d + 2$ on the network width, which is not present in our results for shallow networks.

Again, we complement the upper bound by a lower bound, which shows—at least for fixed depth L —that the bounds are tight up to a factor that is logarithmic in the width of the network.

THEOREM 3.4. There exist absolute constants $C, c, c_1 > 0$ with the following property: If $N \geq CdL^2$ and if $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network of width N , with L hidden layers with random weights and biases following Assumption 2.2, then the following two properties are satisfied:

1. It holds

$$\text{Lip}(\Phi) \geq \frac{1}{4}\sqrt{d}$$

with probability at least $\left(1 - \frac{1}{2^N} - \exp(-N/(CL^2))\right)^L \cdot (1 - 2\exp(-c_1 d))_+$.

2. If we additionally assume $N \geq CL^2 \ln(4L)$ and $d > \frac{\ln(2)}{c_1}$, it holds

$$\mathbb{E}[\text{Lip}(\Phi)] \geq c \cdot \sqrt{d}.$$

Theorem 3.4 is proven in Section 6.2. At first glance, it might be irritating that the probability in (1) depends on $N/(CL^2)$. Here, it should be noted that, by rearranging, the condition $N \geq CdL^2$ is equivalent to $N/(CL^2) \geq d$. Therefore, the bound on the Lipschitz constant in particular holds with high probability depending on the input dimension d . We remark that in contrast to the upper bound, where we assume that $N > d + 2$, for the lower bound we assume that $N \gtrsim dL^2$. Moreover, we assume that the biases are drawn from symmetric distributions, whereas this assumption is not needed in Theorems 3.1 to 3.3. However, we emphasize once more that the assumption that the biases are drawn from symmetric distributions is natural and includes in particular the initialization to zero (as originally introduced in [25]), a Gaussian initialization, a symmetric uniform initialization as well as a Rademacher initialization. Moreover, we

note that the fact that the numerical constant in Theorem 3.4(1) is better than the one in Theorem 3.2(1) is merely a proof artifact since we did not try to optimize numerical constants.

The above results provide bounds on the Lipschitz constant of deep ReLU networks. The upper and lower bounds are tight up to a factor that depends exponentially on the depth L , but is only logarithmic in N/d .

Note that for both the upper and lower bound for *deep* networks, we impose some additional assumptions on the input dimension d , the width N and the depth L , whereas no such assumption is needed in the case of shallow networks. Specifically, for the upper bound we require $N > d + 2$ and for the lower bound $N \gtrsim dL^2$.

REMARK 3.5. In this remark, we compare the upper bounds for the Lipschitz constant of random ReLU networks that we derive in this work (see Theorems 3.1 and 3.3) to the bounds that were shown in [10, 33] in the context of the lazy training regime [13].

We first note that neither of the papers [10, 33] provides *lower* bounds for the Lipschitz constant.

The analysis in [33] concentrates on bounding the Lipschitz constant of the feature maps and does not include the final layer. In our setting (L hidden layers, each with N neurons and input dimension d with weights following the variant of the He initialization), the bound in [33, Theorem 6.2] shows that the output of the penultimate layer has Lipschitz constant at most $\mathcal{O}(2^L \max\{1, \sqrt{d/N}\})$ up to log factors, with high probability. If one naively combines this bound with the (sharp) bound of $\sqrt{N}$ for the Lipschitz constant of the function mapping the output of the penultimate layer to the output of the network, one achieves a bound of $\mathcal{O}(2^L \max\{\sqrt{N}, \sqrt{d}\})$, which is (for fixed depth L and $N \gg d$) significantly weaker than the bound of $\mathcal{O}(C^L \sqrt{d})$ (up to log factors) that we derive. Moreover, the analysis in [33] heavily uses the fact that the biases are all zero, whereas our techniques can handle quite general distributions of the biases.

A sharper version of this bound has been obtained as an auxiliary result in [10, Theorem B.5]: Under the additional assumption $N \gtrsim d^4 L \ln^4(N)$, this result implies that the Lipschitz constant is bounded by $\mathcal{O}(\sqrt{d})$ up to log factors, with high probability. In particular, it is remarkable that the bound is independent of the network depth L , whereas our bound depends *exponentially* on L . Removing the exponential dependence on L in our setting is interesting but beyond the scope of the present work. Adapting the techniques from [10] to our setting with quite general distributions of the biases is not straightforward, since [10] heavily uses that all biases are zero in order to reduce the problem of bounding the global Lipschitz constant to bounding the Lipschitz constant on the sphere. Finally, we point out that the upper bound in [10] implies that our lower bound of $\Omega(\sqrt{d})$ is sharp, at least in the case of all biases being zero and $N \gtrsim d^4 L \ln^4(N) + dL^2$.

REMARK 3.6. This remark is dedicated to comparing the lower bounds that we derive in this work (see Theorems 3.2 and 3.4) with a lower bound on the norm of the gradient of random ReLU networks at a fixed point that was derived in [4, Lemma 2.2]. Noting that the initialization scheme considered in [4] is slightly different from the setting that we consider, transferred to our setting the lower bound from [4] states that the norm of the gradient at a single fixed point is lower bounded by $\Omega(\sqrt{d} \cdot \sqrt{2}^{-L})$ with high probability. This lower bound immediately implies the same lower bound for the Lipschitz constant of random ReLU networks, since such networks are almost surely differentiable at a fixed point (see Appendix E). While this lower bound is for fixed depth L at par with the lower bound of $\Omega(\sqrt{d})$ that we derive, it is important to note that our bound does *not* depend on the depth L of the network, whereas the bound from [4] has *exponential* decay in L . Moreover, in [4] it is also assumed that all the biases are equal to zero, whereas we allow more general distributions for the biases.

In principle, it might be possible to adapt the techniques from [4] to derive a lower bound that is similar to ours but this is not done in [4]. As far as we are aware, our paper is the first to state and derive a sharp lower bound matching the upper bound from [10].

4. Technical preliminaries

In this section, we discuss the necessary technical preliminaries for our proofs.

4.1 The gradient of ReLU networks

To obtain bounds on the Lipschitz constant of a neural network, we will use the gradient of the network. To conveniently express this gradient, we introduce the following notation: For a vector $x \in \mathbb{R}^N$ we define the diagonal matrix $\Delta(x) \in \mathbb{R}^{N \times N}$ via

$$(\Delta(x))_{ij} := \begin{cases} \mathbb{1}_{x_i > 0}, & i = j, \\ 0, & i \neq j. \end{cases}$$

This leads to the following recursive representation of a ReLU network Φ with the same notation as in Assumption 2.1: Let $x =: x^{(0)} \in \mathbb{R}^d$ and define recursively

$$\begin{aligned} D^{(\ell)}(x) &:= \Delta(W^{(\ell)}x^{(\ell)} + b^{(\ell)}), \\ x^{(\ell+1)} &:= D^{(\ell)}(x) \cdot (W^{(\ell)}x^{(\ell)} + b^{(\ell)}) = \text{ReLU}(W^{(\ell)}x^{(\ell)} + b^{(\ell)}), \quad 0 \leq \ell < L. \end{aligned} \quad (4.1)$$

Then it holds $\Phi(x) = W^{(L)}x^{(L)} + b^{(L)}$.

A ReLU network is not necessarily everywhere differentiable, since the ReLU itself is not differentiable at 0. Nevertheless, the following proposition states how the gradient of a ReLU network can be computed almost everywhere.

PROPOSITION 4.1 (cf. [5, Theorem III.1]). Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU network. Then it holds for almost every $x \in \mathbb{R}^d$ that Φ is differentiable at x with

$$(\nabla \Phi(x))^T = W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}.$$

Here, the matrices $D^{(\ell)}(x)$ are defined as in (4.1) and Φ is any realization of a random ReLU network as in Assumption 2.1.

4.2 The Lipschitz constant of ReLU networks

The following well-known proposition establishes a relation between the Lipschitz constant and the gradient of a function.

PROPOSITION 4.2. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be Lipschitz continuous and $M \subseteq \mathbb{R}^d$ any measurable subset of $\mathbb{R}^d$ with Lebesgue measure $\lambda^d(\mathbb{R}^d \setminus M) = 0$ such that f is differentiable in every $x \in M$. Then it holds that

$$\text{Lip}(f) = \sup_{x \in M} \|\nabla f(x)\|_2.$$

For the sake of completeness, we provide a proof of Proposition 4.2 in Appendix A. It should be noted that the existence of a set M as in the proposition follows from the fact that f is Lipschitz continuous; this is known as Rademacher's theorem (cf. [20, Section 3.1.2]).

Every ReLU network Φ is Lipschitz continuous as the composition of Lipschitz continuous functions. Hence, from Propositions 4.1 and 4.2, we infer

$$\text{Lip}(\Phi) \leq \sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2. \quad (4.2)$$

We remark that in general, one does *not* necessarily have equality in (4.2), since the supremum is taken over all of $\mathbb{R}^d$, including points at which Φ might not be differentiable.

As an example, consider the shallow ReLU network $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}$ with three hidden neurons built from the matrices $W^{(0)} := \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & -1 \end{pmatrix}$ and $W^{(1)} := \begin{pmatrix} -1 & 1 & 1 \end{pmatrix}$ and all biases equal to zero. Then a direct computation shows that for each vector $(x, y) \in \mathbb{R}^2$ it holds

$$\left(W^{(1)} D^{(0)}(x, y) W^{(0)}\right)^T = \begin{pmatrix} -\mathbb{1}_{x>y} - \mathbb{1}_{y>x} + 2\mathbb{1}_{2x>y} \\ \mathbb{1}_{x>y} + \mathbb{1}_{y>x} - \mathbb{1}_{2x>y} \end{pmatrix} = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & x \neq y, 2x > y, \\ \begin{pmatrix} -1 \\ 1 \end{pmatrix}, & x \neq y, 2x \leq y, \\ \begin{pmatrix} 2 \\ -1 \end{pmatrix}, & x = y, 2x > y, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & x = y, 2x \leq y. \end{cases}$$

Since Φ is Lipschitz continuous, differentiable on $\{(x, y) \in \mathbb{R}^2 : x \neq y \text{ and } y \neq 2x\}$ and this set has full measure, we conclude using Proposition 4.2 that $\text{Lip}(\Phi) = \sqrt{2}$. On the other hand, the above computation shows

$$\sup_{(x,y) \in \mathbb{R}^2} \|W^{(1)} D^{(0)}(x, y) W^{(0)}\|_2 = \sqrt{5} > \text{Lip}(\Phi).$$

The estimate (4.2) is essential for our derivation of *upper* bounds for Lipschitz constants of random ReLU networks. On the other hand, for any $x_0 \in \mathbb{R}^d$ with the property that Φ is differentiable at x_0 it holds

$$\text{Lip}(\Phi) \geq \|\nabla \Phi(x_0)\|_2, \quad (4.3)$$

which also follows directly from Propositions 4.1 and 4.2. This will be useful to derive *lower* bounds for Lipschitz constants of random ReLU networks.

For technical reasons, since we form expressions such as $\mathbb{E}[\text{Lip}(\Phi)]$, it is important to observe that the map

$$(W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)}) \mapsto \text{Lip}(\Phi),$$

where Φ is the ReLU network built from $(W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)})$, is measurable. This follows from the continuity of the map

$$(x, y) \mapsto \frac{|\Phi(x) - \Phi(y)|}{\|x - y\|_2}$$

on $\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : x \neq y\}$ for a fixed network Φ , combined with the separability of the latter set.

4.3 Covering numbers and VC-dimension

Let (T, ϱ) be a metric space. For $\varepsilon > 0$, we define the ε -covering number of T as

$$\mathcal{N}(T, \varrho, \varepsilon) := \inf \left\{ |K| : K \subseteq T, \bigcup_{k \in K} \bar{B}_T^\varrho(k, \varepsilon) = T \right\} \in \mathbb{N} \cup \{\infty\},$$

where $|K|$ denotes the cardinality of a set K . Any set $K \subseteq T$ with $\bigcup_{k \in K} \bar{B}_T^\varrho(k, \varepsilon) = T$ is called an ε -net of T .

It is well-known that the ε -covering number of the unit ball in k -dimensional Euclidean space with respect to the $\|\cdot\|_2$ -norm can be upper bounded by $\left(1 + \frac{2}{\varepsilon}\right)^k$ (see, e.g., [43, Corollary 4.2.13]). We, however, need a slightly modified version of this result, which we prove in Appendix A.

PROPOSITION 4.3. Let $\varepsilon > 0$ and $V \subseteq \mathbb{R}^k$ be a linear subspace of $\mathbb{R}^k$. Then it holds

$$\mathcal{N}(\bar{B}_k(0, 1) \cap V, \|\cdot\|_2, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^{\dim(V)}.$$

For a set $\mathcal{F}$ of Boolean functions defined on a set Ω (i.e., $\mathcal{F} \subseteq \{f : \Omega \rightarrow \{0, 1\}\}$), we denote by $\text{VC}(\mathcal{F})$ its VC-dimension, i.e.,

$$\text{VC}(\mathcal{F}) := \sup \left\{ |\Lambda| : \Lambda \subseteq \Omega, |\{f|_\Lambda : f \in \mathcal{F}\}| = 2^{|\Lambda|} \right\}.$$

We refer to [37, Chapter 6] and [43, Chapter 8.3] for details on the concept of the VC-dimension.

It is well-known that the VC-dimension of the class of homogeneous halfspaces in k -dimensional Euclidean space equals k (see, e.g., [37, Theorem 9.2]). As in the case of the covering number of Euclidean balls, we need a slightly modified version of this result. The proof is also deferred to Appendix A.

PROPOSITION 4.4. Let $k \in \mathbb{N}$ and $V \subseteq \mathbb{R}^k$ be a linear subspace. For $\alpha \in \mathbb{R}^k$, we define

$$f_\alpha : \mathbb{R}^k \rightarrow \{0, 1\}, \quad x \mapsto \mathbb{1}_{\alpha^T x > 0}.$$

Let $\mathcal{F} := \{f_\alpha|_V : \alpha \in V\}$. Then it holds $\text{VC}(\mathcal{F}) = \dim(V)$.

We will further make use of the following estimate, which enables us to bound the covering number of a class of Boolean functions with respect to $L^2(\mu)$ for some probability measure μ using the VC-dimension of this set.

PROPOSITION 4.5 (cf. [43, Theorem 8.3.18]). There exists an absolute constant $C > 0$ with the following property: For any class of measurable Boolean functions $\mathcal{F} \neq \emptyset$ on some probability space $(\Omega, \mathcal{A}, \mu)$ and for every $\varepsilon \in (0, 1)$, we have

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{C \cdot \text{VC}(\mathcal{F})}.$$

4.4 Sub-gaussian random variables and Gaussian width

Sub-gaussian random variables occur frequently over the course of our paper; we therefore briefly recall their definition in this subsection. We refer to [43, Chapters 2.5 and 3.4] for a detailed introduction to this topic.

A real-valued random variable X is called *sub-gaussian* if

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t^2/C_1^2)$$

holds with a constant $C_1 > 0$ and all $t > 0$. This is equivalent (see [43, Proposition 2.5.2]) to the existence of a constant $C_2 > 0$ satisfying

$$\mathbb{E} \left[\exp(X^2/C_2^2) \right] \leq 2. \quad (4.4)$$

The *sub-gaussian* norm of X (see [43, Equation (2.13)]) is defined as the infimum of all numbers satisfying (4.4), i.e.,

$$\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E} \left[\exp(X^2/t^2) \right] \leq 2\}.$$

Following [43, Definition 3.4.1], a k -dimensional *random vector* X is called sub-gaussian if and only if $\langle X, x \rangle$ is sub-gaussian for every $x \in \mathbb{R}^k$. The sub-gaussian norm of X is defined as

$$\|X\|_{\psi_2} := \sup_{x \in \mathbb{S}^{k-1}} \|\langle X, x \rangle\|_{\psi_2}.$$

In Section 6, we will use the notion of Gaussian width since this quantity appears in the high-probability version of the matrix deviation inequality [27, Theorem 3]. Specifically, we will need to bound the Gaussian width of low-dimensional balls in a higher-dimensional space. Also here, we defer the proof to Appendix A.

PROPOSITION 4.6. Let $k \in \mathbb{N}$ and $\emptyset \neq T \subseteq \mathbb{R}^k$. Following [43, Definition 7.5.1], we then define

$$w(T) := \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in T} \langle g, x \rangle \right].$$

$w(T)$ is called the *Gaussian width* of T .

Let $V \subseteq \mathbb{R}^k$ be a linear subspace of $\mathbb{R}^k$. Then it holds

$$w(\bar{B}_k(0, 1) \cap V) \leq \sqrt{\dim(V)}.$$

4.5 Omitting the ReLU activation

It is natural to ask how the Lipschitz constant of a ReLU network is related to the Lipschitz constant of the corresponding linear network, i.e., using the identity function instead of the ReLU as the activation function. Precisely, for a given random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ that is defined as in Section 2, we define the corresponding linear network via

$$\tilde{\Phi} := V^{(L)} \circ V^{(L-1)} \circ \dots \circ V^{(1)} \circ V^{(0)}. \quad (4.5)$$

We first note that $\tilde{\Phi}$ is an affine map with

$$\text{Lip}(\tilde{\Phi}) = \|W^{(L)} \cdot W^{(L-1)} \dots W^{(1)} \cdot W^{(0)}\|_2.$$

At first glance, one might be tempted to think that the ReLU activation reduces the Lipschitz constant at most, since the ReLU itself is Lipschitz continuous with Lipschitz constant 1. The following proposition, however, demonstrates that this is in general *not* the case, even for shallow networks. We defer the proof to Appendix A.

PROPOSITION 4.7. Let $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a random shallow ReLU network with width $N = 3$ satisfying Assumption 2.1. Moreover, let $C > 0$ be arbitrary. Then with positive probability it holds for $\tilde{\Phi}$ as in (4.5) that

$$\text{Lip}(\Phi) > C \cdot \text{Lip}(\tilde{\Phi}).$$

Hence, we see that in general it does *not* hold

$$\text{Lip}(\Phi) \lesssim \text{Lip}(\tilde{\Phi}) \quad \text{almost surely.}$$

Nevertheless, Theorems 3.1 and 3.2 show that at least for shallow networks, we indeed have that

$$\text{Lip}(\Phi) \asymp \text{Lip}(\tilde{\Phi}) \quad \text{with high probability}$$

but proving this is nontrivial.

Remarkably, the converse estimate to what is considered in Proposition 4.7 even holds deterministically, at least for shallow networks.

PROPOSITION 4.8. Let $d, N \in \mathbb{N}$ and let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a fixed (deterministic) shallow ReLU network with N hidden neurons. Then it holds for $\tilde{\Phi}$ as in (4.5) that

$$\text{Lip}(\Phi) \geq \frac{1}{2} \text{Lip}(\tilde{\Phi}).$$

Proof. In this proof, we use a different notation for shallow ReLU networks than the one introduced in Section 2. Specifically, we write

$$\Phi(x) := \sum_{i=1}^N [c_i \text{ReLU}(\langle a_i, x \rangle + b_i)] + \beta \quad \text{and} \quad \tilde{\Phi}(x) := \sum_{i=1}^N [c_i \cdot (\langle a_i, x \rangle + b_i)] + \beta$$

with $a_1, \dots, a_N \in \mathbb{R}^d$, $b_1, \dots, b_N, c_1, \dots, c_N, \beta \in \mathbb{R}$.

Let

$$v_0 := \sum_{i=1}^N c_i a_i \in \mathbb{R}^d.$$

If $v_0 = 0$, there is nothing to show, since in that case we have $\text{Lip}(\tilde{\Phi}) = \|v_0\|_2 = 0$. Hence, we assume $v_0 \neq 0$ and define $v := \frac{v_0}{\|v_0\|_2}$. We denote $\alpha_i := \langle a_i, v \rangle$ for $i \in \{1, \dots, N\}$ and define the sets

$$\begin{aligned} I_+ &:= \{i \in \{1, \dots, N\} : \alpha_i > 0\}, \\ I_- &:= \{i \in \{1, \dots, N\} : \alpha_i < 0\}, \\ I_0 &:= \{i \in \{1, \dots, N\} : \alpha_i = 0\} \quad \text{and} \\ I &:= I_+ \cup I_-. \end{aligned}$$

We then see for every $t \in \mathbb{R}$ that

$$\begin{aligned} \Phi(tv) &= \sum_{i \in I_+} [c_i \text{ReLU}(t\alpha_i + b_i)] + \sum_{i \in I_-} [c_i \text{ReLU}(t\alpha_i + b_i)] + \sum_{i \in I_0} [c_i \text{ReLU}(t\alpha_i + b_i)] + \beta \\ &= \sum_{i \in I_+} [c_i \text{ReLU}(t\alpha_i + b_i)] + \sum_{i \in I_-} [c_i \text{ReLU}(t\alpha_i + b_i)] + \sum_{i \in I_0} [c_i \text{ReLU}(b_i)] + \beta. \end{aligned}$$

Fix $t_0 \in \mathbb{R}$ with

$$t_0 > \max \left\{ -\frac{b_i}{\alpha_i} : i \in I \right\}.$$

For $i \in I_+$, this implies

$$t_0 \alpha_i + b_i > \left(-\frac{b_i}{\alpha_i} \right) \alpha_i + b_i = 0 \tag{4.6}$$

and for $i \in I_-$, we see

$$t_0 \alpha_i + b_i < \left(-\frac{b_i}{\alpha_i} \right) \alpha_i + b_i = 0. \tag{4.7}$$

By continuity, there exists $\delta > 0$ such that the inequalities (4.6) and (4.7) hold for any $t \in (t_0 - \delta, t_0 + \delta)$. In particular, it holds

$$\Phi(tv) = \sum_{i \in I_+} [c_i \cdot (t\alpha_i + b_i)] + \sum_{i \in I_0} [c_i \text{ReLU}(b_i)] + \beta$$

for every $t \in (t_0 - \delta, t_0 + \delta)$. This means that $t \mapsto \Phi(tv)$ is differentiable on $(t_0 - \delta, t_0 + \delta)$ with

$$\left. \frac{d}{dt} \right|_{t=t_0} [\Phi(tv)] = \sum_{i \in I_+} c_i \alpha_i.$$

We then get

$$\left| \sum_{i \in I_+} c_i \alpha_i \right| = \left| \left. \frac{d}{dt} \right|_{t=t_0} [\Phi(tv)] \right| = \lim_{t \rightarrow t_0} \frac{|\Phi(tv) - \Phi(t_0 v)|}{|t - t_0|} = \lim_{t \rightarrow t_0} \frac{|\Phi(tv) - \Phi(t_0 v)|}{\|tv - t_0 v\|_2} \leq \text{Lip}(\Phi).$$

Similarly, by picking $t_0 < \min \left\{ -\frac{b_i}{\alpha_i} : i \in I \right\}$, we get

$$\text{Lip}(\Phi) \geq \left| \sum_{i \in I_-} c_i \alpha_i \right|.$$

Hence, combining these two estimates, we arrive at

$$\text{Lip}(\Phi) \geq \frac{1}{2} \left[\left| \sum_{i \in I_+} c_i \alpha_i \right| + \left| \sum_{i \in I_-} c_i \alpha_i \right| \right] \geq \frac{1}{2} \left[\left| \sum_{i \in I} c_i \alpha_i \right| \right] = \frac{1}{2} \left[\left| \sum_{i=1}^N c_i \alpha_i \right| \right],$$

where the last equality follows from $\alpha_i = 0$ for every $i \in I_0$. To get the final claim, note that

$$\left| \sum_{i=1}^N c_i \alpha_i \right| = \left| \sum_{i=1}^N c_i \langle a_i, v \rangle \right| = \left| \left\langle \sum_{i=1}^N c_i a_i, v \right\rangle \right| = \langle v_0, v \rangle = \|v_0\|_2 = \text{Lip}(\tilde{\Phi}),$$

as was to be shown. $\square$

This bound will be useful in order to derive a lower bound for the Lipschitz constant of *shallow* ReLU networks. Specifically, in order to get a lower bound on the Lipschitz constant of shallow random ReLU networks, it suffices to establish a lower bound for $\|W^{(1)} W^{(0)}\|_2$, where $W^{(0)} \in \mathbb{R}^{N \times d}$ is a Gaussian matrix and $W^{(1)} \in \mathbb{R}^{1 \times N}$ is a Gaussian row vector. We refer to Section 6.1 for the details.

Unfortunately, an analogous bound does not hold for *deep* networks anymore, even for depth $L = 2$, as is stated in the following proposition, the proof of which can also be found in Appendix A.

PROPOSITION 4.9. We consider a ReLU network $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ with depth $L = 2$ and width $N = 1$ satisfying Assumption 2.1. Moreover, we assume that $\mathbb{P}(b^{(1)} \leq 0) > 0$. Then, it holds with positive probability for

$\tilde{\Phi}$ as in Equation (4.5) that

$$\text{Lip}(\Phi) = 0 < \text{Lip}(\tilde{\Phi}).$$

Proposition 4.9 does not rule out the possibility that the inequality $\text{Lip}(\Phi) \gtrsim \text{Lip}(\tilde{\Phi})$ holds with high probability as $d \rightarrow \infty$. However, the proposition shows that the case of deep networks is qualitatively different from the case of shallow networks, where $\text{Lip}(\Phi) \gtrsim \text{Lip}(\tilde{\Phi})$ holds even deterministically (see Proposition 4.8). Hence, one must carry out a more detailed analysis of the gradients in the case of deep networks (see Section 6.2).

5. Proof of the upper bound

Referring to (4.2), the goal of this section is to establish upper bounds for

$$\sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2$$

that hold with high probability and in expectation, where the randomness is over the random matrices $W^{(0)}, \dots, W^{(L)}$ and the random bias vectors $b^{(0)}, \dots, b^{(L-1)}$. For this to make sense one first needs to know that

$$\sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2 \quad (5.1)$$

is indeed measurable; we refer to Appendix B for a proof of this fact.

Throughout the entire section, we assume that Assumption 2.1 is satisfied. Since the random matrices $W^{(\ell)}$ and the random biases $b^{(\ell)}$ are jointly independent it is possible to calculate the expectation iteratively by first assuming that the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ are fixed and deriving an upper bound when only $W^{(L)}$ is assumed to be random. Then as the final step we are also going to allow randomness in $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ to get the desired result. In other words, we are conditioning on $W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}$.

The central tool for deriving these bounds is *Dudley's inequality*, which can be found for example in [41, Theorems 5.25 and 5.29]. We refer to Appendix D for details on Dudley's inequality. The key idea of this section is contained in the following proposition.

PROPOSITION 5.1. Let the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ be fixed and define $\Lambda := \|W^{(L-1)}\|_2 \cdots \|W^{(0)}\|_2$. For $x \in \mathbb{R}^d$ and $z \in \bar{B}_d(0, 1)$ we set

$$Y_{z,x} := D^{(L-1)}(x) W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)} z \in \mathbb{R}^N$$

and further

$$\mathcal{L} = \mathcal{L}(d, N, L, W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}) := \left\{ Y_{z,x} : x \in \mathbb{R}^d, z \in \bar{B}_d(0, 1) \right\} \subseteq \mathbb{R}^N.$$

Then there exists an absolute constant $C > 0$ such that the following holds: Given any $u \geq 0$, we have

$$\sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \dots D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C \cdot \left(\int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} \, d\varepsilon + u\Lambda \right)$$

with probability at least $(1 - 2 \exp(-u^2))$ (with respect to the choice of $W^{(L)}$). Moreover,

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \dots D^{(0)}(x) \cdot W^{(0)}\|_2 \right] \leq C \cdot \int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} \, d\varepsilon.$$

Proof. For $y \in \mathcal{L}$ it holds that there are $x \in \mathbb{R}^d$ and $z \in \bar{B}_d(0, 1)$ satisfying

$$y = Y_{z,x} = D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}z.$$

We compute

$$\begin{aligned} \|y\|_2 &\leq \underbrace{\|D^{(L-1)}(x)\|_2}_{\leq 1} \cdot \|W^{(L-1)}\|_2 \dots \underbrace{\|D^{(0)}(x)\|_2}_{\leq 1} \cdot \|W^{(0)}\|_2 \cdot \underbrace{\|z\|_2}_{\leq 1} \\ &\leq \|W^{(L-1)}\|_2 \dots \|W^{(0)}\|_2 = \Lambda. \end{aligned}$$

Hence, since $0 = D^{(L-1)}(0)W^{(L-1)} \dots D^{(0)}(0)W^{(0)}0 \in \mathcal{L}$ it follows

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) = 1 \quad \text{for } \varepsilon \geq \Lambda. \quad (5.2)$$

To get the final result we rewrite

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) \cdot W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 &= \sup_{x \in \mathbb{R}^d, z \in \bar{B}_d(0,1)} \left\langle \left(W^{(L)}\right)^T, Y_{z,x} \right\rangle \\ &= \sup_{Y \in \mathcal{L}} \left\langle \left(W^{(L)}\right)^T, Y \right\rangle. \end{aligned}$$

From the observation $\mathcal{L} \subseteq \bar{B}_N(0, \Lambda)$ we infer $\text{diam}(\mathcal{L}) \leq 2\Lambda$. Proposition D.4 and (5.2) then yield the claim by noticing once again that $0 \in \mathcal{L}$. $\square$

Given the above proposition, the problem of bounding the Lipschitz constant of random ReLU networks has been transferred to bounding the covering numbers of the set $\mathcal{L}$. Finding upper bounds for these covering numbers is the essential task in the following two subsections. In fact, in the following we show that

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{9\|W^{(L-1)}\|_2 \dots \|W^{(0)}\|_2}{\varepsilon} \right)^{C(d+1)} \cdot \left(\frac{eN}{d+1} \right)^{L(d+1)},$$

at least if $N > d + 2$. Here, $C > 0$ is an absolute constant. However, in the case of shallow networks, i.e., $L = 1$, it is possible to show that the above inequality holds without the additional factor $\left(\frac{eN}{d+1}\right)^{L(d+1)}$, without the assumption $N > d + 2$ and d can be replaced by $\min\{N, d\}$, which in the end leads to a sharper bound on the Lipschitz constant. Therefore, the cases of shallow and deep networks are treated separately in Sections 5.1 and 5.2, respectively. The final bounds on the covering numbers can be found in Lemmas 5.3 and 5.8.

5.1 The shallow case

Firstly, we consider shallow neural networks, i.e., networks that only have a single hidden layer and can hence be written as

$$\left(x \mapsto (W^{(1)} \cdot x + b^{(1)})\right) \circ \text{ReLU} \circ \left(x \mapsto (W^{(0)} \cdot x + b^{(0)})\right).$$

As already explained above, we are from now on going to assume that the matrix $W^{(0)}$ and the vector $b^{(0)}$ are fixed and only assume randomness in $W^{(1)}$ and $b^{(1)}$.

LEMMA 5.2. Let $W^{(0)} \in \mathbb{R}^{N \times d}$ and $b^{(0)} \in \mathbb{R}^N$ be fixed. We recall that for $\alpha \in \mathbb{R}^{d+1}$ we define

$$f_\alpha : \mathbb{R}^{d+1} \rightarrow \{0, 1\}, \quad x \mapsto \mathbb{1}_{\alpha^T x > 0}.$$

Furthermore, for any vector $z \in \mathbb{R}^d$, let

$$\tau_{\alpha, z} := (W^{(0)} z) \odot \left(f_\alpha \left((W_{1,-}^{(0)}, b_1^{(0)})^T\right), \dots, f_\alpha \left((W_{N,-}^{(0)}, b_N^{(0)})^T\right)\right) \in \mathbb{R}^N.$$

Then the following two statements hold:

1. $Y_{z,x} = \tau_{(x,1)^T, z}$ for every $x \in \mathbb{R}^d$ and $z \in \bar{B}_d(0, 1)$,
2. $\|\tau_{\alpha, z}\|_2 \leq \|W^{(0)}\|_2 \cdot \|z\|_2$ for all $\alpha \in \mathbb{R}^{d+1}$, $z \in \mathbb{R}^d$.

Here, $Y_{z,x}$ is as introduced in Proposition 5.1.

Proof.

1. Let $x \in \mathbb{R}^d$ and $z \in \bar{B}_d(0, 1)$. For every $i \in \{1, \dots, N\}$ we calculate

$$\begin{aligned} (D^{(0)}(x) \cdot W^{(0)} \cdot z)_i &= \mathbb{1}_{W_{i,-}^{(0)} x + b_i^{(0)} > 0} \cdot (W^{(0)} z)_i \\ &= (W^{(0)} z)_i \cdot f_{(x,1)^T} \left((W_{i,-}^{(0)}, b_i^{(0)})^T \right) = (\tau_{(x,1)^T, z})_i, \end{aligned}$$

which yields the claim.

2. It is immediate that

$$\|\tau_{\alpha, z}\|_2 \leq \|W^{(0)} z\|_2 \cdot \underbrace{\max_{i=1, \dots, N} \left| f_\alpha \left((W_{i,-}^{(0)}, b_i^{(0)})^T \right) \right|}_{\leq 1} \leq \|W^{(0)}\|_2 \cdot \|z\|_2.$$

□

The desired bound for the covering number of $\mathcal{L}$ in the case of shallow networks is contained in the following lemma.

LEMMA 5.3. Assume that $W^{(0)} \in \mathbb{R}^{N \times d}$ and $b^{(0)} \in \mathbb{R}^N$ are fixed. There exists an absolute constant $C > 0$ such that, writing $k := \text{rank}(W^{(0)})$, for every $\varepsilon \in (0, \|W^{(0)}\|_2)$ it holds

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{9\|W^{(0)}\|_2}{\varepsilon} \right)^{C \cdot k}.$$

Here, $\mathcal{L}$ is as introduced in Proposition 5.1.

Proof. Without loss of generality, we assume that $W^{(0)} \neq 0$, since otherwise $(0, \|W^{(0)}\|_2) = \emptyset$. Note that according to Lemma 5.2 we can write

$$\mathcal{L} = \left\{ \tau_{(x,1)^T, v} : x \in \mathbb{R}^d, v \in \bar{B}_d(0, 1) \right\}.$$

We can even weaken this identity and infer

$$\mathcal{L} = \left\{ \tau_{(x,1)^T, v} : x \in \ker(W^{(0)})^\perp, v \in \bar{B}_d(0, 1) \cap \ker(W^{(0)})^\perp \right\}.$$

Here, $\ker(W^{(0)})^\perp$ denotes the orthogonal complement of $\ker(W^{(0)})$. We further note

$$\dim(\ker(W^{(0)})^\perp) = k.$$

Let $\varepsilon \in (0, \|W^{(0)}\|_2)$. From Proposition 4.3 we infer the existence of a natural number $M \in \mathbb{N}$ with $M \leq \left(\frac{8\|W^{(0)}\|_2}{\varepsilon} + 1 \right)^k$ and $v_1, \dots, v_M \in \bar{B}_d(0, 1) \cap \ker(W^{(0)})^\perp$ such that

$$\bar{B}_d(0, 1) \cap \ker(W^{(0)})^\perp \subseteq \bigcup_{i=1}^M \bar{B}_d\left(v_i, \frac{\varepsilon}{4\|W^{(0)}\|_2}\right) \cap \ker(W^{(0)})^\perp.$$

For $i \in \{1, \dots, M\}$ let $w_i := W^{(0)}v_i$. Fix $i \in \{1, \dots, M\}$ and first assume $w_i \neq 0$. Define a probability measure μ_i on $\ker(W^{(0)})^\perp \times \mathbb{R}$ by

$$\mu_i := \frac{1}{\|w_i\|_2^2} \cdot \sum_{\ell=1}^N (w_i)_\ell^2 \cdot \delta_{(w_{\ell,-}^{(0)}, b_\ell^{(0)})^T}.$$

Here, we note that by definition it holds $(w_{\ell,-}^{(0)})^T \in \ker(W^{(0)})^\perp$ for every $\ell \in \{1, \dots, N\}$.

Let $\mathcal{F} := \{f_\alpha|_{\ker(W^{(0)})^\perp \times \mathbb{R}} : \alpha \in \ker(W^{(0)})^\perp \times \mathbb{R}\}$, where f_α is as introduced in the previous Lemma 5.2. From Proposition 4.4 we infer that

$$\text{VC}(\mathcal{F}) = k + 1.$$

Further, Proposition 4.5 shows for every $\delta \in (0, 1)$ that

$$\mathcal{M}(\mathcal{F}, L^2(\mu_i), \delta) \leq \left(\frac{2}{\delta}\right)^{C'(k+1)}$$

with an absolute constant $C' > 0$. Thus, there exists $K_i \in \mathbb{N}$ with $K_i \leq \left(\frac{8\|W^{(0)}\|_2}{\varepsilon}\right)^{C'(k+1)}$ and vectors $\alpha_1^{(i)}, \dots, \alpha_{K_i}^{(i)} \in \ker(W^{(0)})^\perp \times \mathbb{R}$ such that

$$\mathcal{F} \subseteq \bigcup_{j=1}^{K_i} \overline{B}_{\mathcal{F}}^{L^2(\mu_i)} \left(f_{\alpha_j^{(i)}}|_{\ker(W^{(0)})^\perp \times \mathbb{R}}, \frac{\varepsilon}{4\|W^{(0)}\|_2} \right).$$

If $w_i = 0$ let $\alpha_j^{(i)} := 0 \in \ker(W^{(0)})^\perp \times \mathbb{R}$ for $1 \leq j \leq K_i := 1$.

Now, let $v \in \overline{B}_d(0, 1) \cap \ker(W^{(0)})^\perp$ and $x \in \ker(W^{(0)})^\perp$ be arbitrary. Then there exists $i \in \{1, \dots, M\}$ such that

$$\|v - v_i\|_2 \leq \frac{\varepsilon}{4\|W^{(0)}\|_2}. \quad (5.3)$$

Let us first consider the case $w_i = W^{(0)}v_i \neq 0$. Then there exists $j \in \{1, \dots, K_i\}$ such that

$$\left\| f_{(x,1)^T}|_{\ker(W^{(0)})^\perp \times \mathbb{R}} - f_{\alpha_j^{(i)}}|_{\ker(W^{(0)})^\perp \times \mathbb{R}} \right\|_{L^2(\mu_i)} \leq \frac{\varepsilon}{4\|W^{(0)}\|_2}.$$

We compute

$$\begin{aligned} \left\| \tau_{(x,1)^T, v} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2 &\leq \left\| \tau_{(x,1)^T, v} - \tau_{(x,1)^T, v_i} \right\|_2 + \left\| \tau_{(x,1)^T, v_i} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2 \\ &= \left\| \tau_{(x,1)^T, v-v_i} \right\|_2 + \left\| \tau_{(x,1)^T, v_i} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2 \\ &\stackrel{\text{Lemma 5.2 (2)}}{\leq} \|W^{(0)}\|_2 \cdot \|v - v_i\|_2 + \left\| \tau_{(x,1)^T, v_i} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2 \\ &\stackrel{(5.3)}{\leq} \frac{\varepsilon}{4} + \left\| \tau_{(x,1)^T, v_i} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2. \end{aligned} \quad (5.4)$$

Finally, we note because of $w_i = W^{(0)}v_i$ and by definition of μ_i that

$$\begin{aligned} \left\| \tau_{(x,1)^T, v_i} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2^2 &= \sum_{\ell=1}^N (W^{(0)}v_i)_\ell^2 \cdot \left(f_{(x,1)^T} \left((W_{\ell,-}^{(0)}, b_\ell^{(0)})^T \right) - f_{\alpha_j^{(i)}} \left((W_{\ell,-}^{(0)}, b_\ell^{(0)})^T \right) \right)^2 \\ &= \|w_i\|_2^2 \cdot \left\| f_{(x,1)^T} \Big|_{\ker(W^{(0)})^\perp \times \mathbb{R}} - f_{\alpha_j^{(i)}} \Big|_{\ker(W^{(0)})^\perp \times \mathbb{R}} \right\|_{L^2(\mu_i)}^2 \\ &\leq \|w_i\|_2^2 \cdot \left(\frac{\varepsilon}{4\|W^{(0)}\|_2} \right)^2 \leq \|W^{(0)}\|_2^2 \cdot \|v_i\|_2^2 \cdot \left(\frac{\varepsilon}{4\|W^{(0)}\|_2} \right)^2 \leq \left(\frac{\varepsilon}{4} \right)^2, \end{aligned} \quad (5.5)$$

and this trivially remains true in the case $w_i = 0$ if we choose $j = 1$.

Overall, (5.4) and (5.5) together imply in any case that

$$\left\| \tau_{(x,1)^T, v} - \tau_{\alpha_j^{(i)}, v_i} \right\|_2 \leq \frac{\varepsilon}{2}.$$

Hence, the set

$$\left\{ \tau_{\alpha_j^{(i)}, v_i} : 1 \leq i \leq M, 1 \leq j \leq K_i \right\}$$

is an $\frac{\varepsilon}{2}$ -net of $\mathcal{L}$ with respect to $\|\cdot\|_2$. However, this set does not necessarily have to be a subset of $\mathcal{L}$. Yet, using [43, Exercise 4.2.9] for the first inequality, we get

$$\begin{aligned} \mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) &\leq \sum_{i=1}^M K_i \leq \left(\frac{8\|W^{(0)}\|_2}{\varepsilon} + 1 \right)^k \cdot \left(\frac{8\|W^{(0)}\|_2}{\varepsilon} \right)^{C'(k+1)} \\ &\stackrel{\varepsilon < \|W^{(0)}\|_2}{\leq} \left(\frac{9\|W^{(0)}\|_2}{\varepsilon} \right)^{C'(k+1)+k} \leq \left(\frac{9\|W^{(0)}\|_2}{\varepsilon} \right)^{(2C'+1)k}, \end{aligned}$$

so the claim follows choosing $C = 2C' + 1$. $\square$

The derived bound for the covering number of $\mathcal{L}$ leads to the following bound when we only assume randomness in $W^{(1)}$.

PROPOSITION 5.4. There exists an absolute constant $C > 0$ such that for fixed $W^{(0)} \in \mathbb{R}^{N \times d}$ and fixed $b^{(0)} \in \mathbb{R}^N$, writing $k = \text{rank}(W^{(0)})$, the following hold:

1. For any $u \geq 0$, we have

$$\sup_{x \in \mathbb{R}^d} \left\| W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)} \right\|_2 \leq C \cdot \|W^{(0)}\|_2 \cdot (\sqrt{k} + u)$$

with probability at least $1 - 2 \exp(-u^2)$ (with respect to the choice of $W^{(1)}$).

$$2. \quad \mathbb{E}_{W^{(1)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \right] \leq C \cdot \sqrt{k} \cdot \|W^{(0)}\|_2.$$

Proof. Without loss of generality we assume $k \geq 1$. We observe

$$\begin{aligned} \int_0^{\|W^{(0)}\|_2} \sqrt{\ln(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} \, d\varepsilon &\stackrel{\text{Lemma 5.3}}{\leq} \int_0^{\|W^{(0)}\|_2} \sqrt{C_1 \cdot k} \cdot \sqrt{\ln\left(\frac{9\|W^{(0)}\|_2}{\varepsilon}\right)} \, d\varepsilon \\ &= \sqrt{C_1} \cdot \sqrt{k} \cdot 9\|W^{(0)}\|_2 \cdot \int_0^{\frac{1}{9}} \sqrt{\ln(1/\sigma)} \, d\sigma \\ &\leq C_2 \cdot \sqrt{k} \cdot \|W^{(0)}\|_2. \end{aligned}$$

Here, $C_1 > 0$ is the absolute constant from Lemma 5.3 and $C_2 := 9 \cdot \sqrt{C_1} \cdot \int_0^{1/9} \sqrt{\ln(1/\sigma)} \, d\sigma$. At the equality, we applied the substitution $\frac{1}{\sigma} = \frac{9\|W^{(0)}\|_2}{\varepsilon}$. We combine this estimate with Proposition 5.1 and get for any $u \geq 0$ that

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 &\leq C_3 \cdot \left(C_2 \cdot \sqrt{k} \cdot \|W^{(0)}\|_2 + u \cdot \|W^{(0)}\|_2 \right) \\ &\leq C_3 \cdot \max\{1, C_2\} \cdot \|W^{(0)}\|_2 \cdot (\sqrt{k} + u) \end{aligned}$$

with probability at least $1 - 2 \exp(-u^2)$, as well as

$$\mathbb{E}_{W^{(1)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(1)} D^{(0)}(x) W^{(0)}\|_2 \right] \leq C_3 \cdot C_2 \cdot \sqrt{k} \cdot \|W^{(0)}\|_2,$$

where $C_3 > 0$ is the absolute constant from Proposition 5.1. Hence, the claim follows in the end by letting $C := C_3 \max\{C_2, 1\}$. $\square$

Until now, we have conditioned on $W^{(0)}, b^{(0)}$. Reintroducing the randomness with respect to $W^{(0)}, b^{(0)}$ leads to the following statement.

PROPOSITION 5.5. There exist absolute constants $C, c_1 > 0$ such that, writing $k := \min\{d, N\}$, the following hold:

1. For any $u, t \geq 0$ it holds

$$\sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C \cdot \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right) (\sqrt{k} + u)$$

with probability at least $(1 - 2 \exp(-u^2))_+ \cdot (1 - 2 \exp(-c_1 t^2))_+$ with respect to the choice of $W^{(0)}, W^{(1)}, b^{(0)}, b^{(1)}$.

$$2. \quad \mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \right] \leq C \cdot \sqrt{d}.$$

Proof. We first note that for any matrix $W^{(0)}$ it holds $\text{rank}(W^{(0)}) \leq k$.

Let us first deal with Part (1). Let $C_2 > 0$ be the (absolute) constant from Proposition 5.4 and $C_3 := \sqrt{2}C_2$. For fixed $u, t \geq 0$ let

$$A := \left\{ (W^{(1)}, W^{(0)}, b^{(0)}) : \sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C_3 \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}}\right) (\sqrt{k} + u) \right\}.$$

Furthermore let

$$A_1 := \left\{ (W^{(0)}, b^{(0)}) : \|W^{(0)}\|_2 \leq \sqrt{2} \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}}\right) \right\}$$

and for fixed $(W^{(0)}, b^{(0)})$, let

$$A_2(W^{(0)}, b^{(0)}) := \left\{ W^{(1)} : \sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C_2 \cdot \|W^{(0)}\|_2 \cdot (\sqrt{k} + u) \right\}.$$

Note that then

$$(W^{(0)}, b^{(0)}) \in A_1, W^{(1)} \in A_2(W^{(0)}, b^{(0)}) \implies (W^{(1)}, W^{(0)}, b^{(0)}) \in A.$$

From Proposition 5.4 and since probabilities are always non-negative we infer

$$\mathbb{P}^{W^{(1)}}(A_2(W^{(0)}, b^{(0)})) \geq (1 - 2\exp(-u^2))_+$$

for any $(W^{(0)}, b^{(0)})$. Furthermore, it holds

$$\|W^{(0)}\|_2 = \sqrt{\frac{2}{N}} \cdot \left\| \sqrt{\frac{N}{2}} \cdot W^{(0)} \right\|_2 \leq \sqrt{\frac{2}{N}} \cdot (\sqrt{N} + \sqrt{d} + t) = \sqrt{2} \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}}\right)$$

with probability at least $(1 - 2\exp(-c_1 t^2))_+$ for some absolute constant $c_1 > 0$, as follows from [43, Corollary 7.3.3] by noting that the matrix $\sqrt{\frac{N}{2}} W^{(0)}$ has independent $\mathcal{N}(0, 1)$ -entries. The claim of Part (1) then follows from Proposition C1.

Let us now deal with Part (2). Using Proposition 5.4 we derive

$$\mathbb{E}_{W^{(0)}, b^{(0)}, W^{(1)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)}\|_2 \right] \leq C_2 \cdot \sqrt{k} \cdot \mathbb{E}_{W^{(0)}} \|W^{(0)}\|_2.$$

From [43, Theorem 7.3.1] we get

$$\mathbb{E}_{W^{(0)}} \|W^{(0)}\|_2 = \sqrt{\frac{2}{N}} \cdot \mathbb{E}_{W^{(0)}} \left\| \sqrt{\frac{N}{2}} \cdot W^{(0)} \right\|_2 \leq \sqrt{\frac{2}{N}} (\sqrt{N} + \sqrt{d}) = \sqrt{2} \left(1 + \sqrt{\frac{d}{N}}\right).$$

Hence, we obtain

$$\mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \left\| W^{(1)} \cdot D^{(0)}(x) \cdot W^{(0)} \right\|_2 \right] \leq C_3 \cdot \left(\sqrt{k} + \sqrt{\frac{kd}{N}} \right).$$

If $k = N$ (so that $N \leq d$), we get

$$\sqrt{k} + \sqrt{\frac{kd}{N}} = \sqrt{N} + \sqrt{d} \stackrel{N \leq d}{\leq} 2 \cdot \sqrt{d}$$

and if $k = d$ (so that $d \leq N$), we have

$$\sqrt{k} + \sqrt{\frac{kd}{N}} = \sqrt{d} + \frac{d}{\sqrt{N}} \stackrel{\sqrt{d} \leq \sqrt{N}}{\leq} 2 \cdot \sqrt{d}.$$

Thus, the final claim follows letting $C := 2C_3$. $\square$

The transfer of Proposition 5.5 to obtain an upper bound of the Lipschitz constant of shallow ReLU networks follows directly from (4.2).

THEOREM 5.6. There exist absolute constants $C, c_1 > 0$ such that if $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network with width N satisfying Assumption 2.1 and writing $k := \min\{d, N\}$, the following hold:

1. For any $u, t \geq 0$, we have

$$\text{Lip}(\Phi) \leq C \cdot \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right) (\sqrt{k} + u)$$

with probability at least $(1 - 2 \exp(-u^2))_+ \cdot (1 - 2 \exp(-c_1 t^2))_+$.

2. $\mathbb{E} [\text{Lip}(\Phi)] \leq C \cdot \sqrt{d}$.

By plugging in certain values for u and t , we can now prove Theorem 3.1.

Proof of Theorem 3.1. Let $\tilde{C}$ and $\tilde{c}_1$ be the relabeled constants from Theorem 5.6 and $k := \min\{d, N\}$, as well as $\ell := \max\{d, N\}$. To show Part (1), we set $u = \sqrt{k}$ and $t = \sqrt{\ell}$. Then, since the inequality $\sqrt{\ln(2)} < \sqrt{\ln(e)} = 1$ holds, we get $u \geq \sqrt{\ln(2)}$ and thus $1 - 2 \exp(-u^2) \geq 0$. Theorem 5.6 shows

$$\text{Lip}(\Phi) \leq 2\tilde{C} \cdot \left(1 + \frac{\sqrt{d} + \sqrt{\ell}}{\sqrt{N}} \right) \sqrt{k}$$

with probability at least $(1 - 2 \exp(-k)) \cdot (1 - 2 \exp(-\tilde{c}_1 \ell))_+$. If $d \leq N$ we get

$$2\tilde{C} \cdot \left(1 + \frac{\sqrt{d} + \sqrt{\ell}}{\sqrt{N}} \right) \sqrt{k} \leq 2\tilde{C} \cdot \left(1 + \frac{\sqrt{N} + \sqrt{N}}{\sqrt{N}} \right) \cdot \sqrt{d} = 6\tilde{C} \cdot \sqrt{d}$$

and if $d \geq N$ we infer

$$2\tilde{C} \cdot \left(1 + \frac{\sqrt{d} + \sqrt{\ell}}{\sqrt{N}}\right) \sqrt{k} = 2\tilde{C} \cdot \left(1 + \frac{\sqrt{d} + \sqrt{d}}{\sqrt{N}}\right) \sqrt{N} = 2\tilde{C} \cdot (\sqrt{N} + \sqrt{d} + \sqrt{d}) \leq 6\tilde{C} \cdot \sqrt{d}.$$

Part (2) follows directly from Theorem 5.6(2). $\square$

5.2 The deep case

In the following, we treat the case of deep networks, meaning $L > 1$. The proofs of this section also apply in the case of shallow networks, but are only relevant in the case of deep networks, since in the case of shallow networks better bounds have been derived in the preceding subsection. Again, we first assume that the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ are fixed and the matrix $W^{(L)}$ is initialized randomly according to Assumption 2.1. In this setting, we denote

$$\Lambda := \|W^{(L-1)}\|_2 \cdots \|W^{(0)}\|_2.$$

As the first step, we state a bound on the cardinality of the set of all possible combinations $(D^{(L-1)}(x), \dots, D^{(0)}(x))$ of the D -matrices occurring in the formula for $\nabla \Phi(x)$ from Proposition 4.1. This cardinality is known as *the number of activation patterns* in the literature [24] and is related but not necessarily identical to the number of linear regions, which has been studied in, e.g., [31, 45]. Naively, one can bound the number of activation patterns by 2^{LN} . A much sharper bound that, however, only holds on bounded input sets has been derived in [24]. Moreover, a bound sharper than 2^{LN} that holds on the entire input space has been derived in [30, Proposition 3]. The following inequality is similar to the one presented in [30]. In order to keep the paper self-contained we decided to include the short proof below.

LEMMA 5.7. Assume $d + 2 < N$. For fixed $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ we define

$$\mathcal{D} := \left\{ \left(D^{(L-1)}(x), \dots, D^{(0)}(x) \right) : x \in \mathbb{R}^d \right\},$$

with $D^{(0)}(x), \dots, D^{(L-1)}(x)$ as defined in Section 4.1. Then it holds that

$$|\mathcal{D}| \leq \left(\frac{eN}{d+1} \right)^{L(d+1)}.$$

Proof. For any $0 \leq \ell \leq L-1$ we define

$$\mathcal{D}^{(\ell)} := \left\{ \left(D^{(\ell)}(x), \dots, D^{(0)}(x) \right) : x \in \mathbb{R}^d \right\}$$

and claim that it holds

$$|\mathcal{D}^{(\ell)}| \leq \left(\frac{eN}{d+1} \right)^{(\ell+1)(d+1)}, \quad (5.6)$$

which we will show by induction over ℓ .

To this end, we start with the case $\ell = 0$. Using the notation introduced in proposition 4.4, we see for any $i \in \{1, \dots, N\}$ and $x \in \mathbb{R}^d$ that

$$\left(D^{(0)}(x)\right)_{i,i} = f_{(x,1)^T} \left(W_{i,-}^{(0)}, b_i^{(0)}\right)^T.$$

By definition of $D^{(0)}(x)$, this yields

$$\begin{aligned} |\mathcal{D}^{(0)}| &= \left| \left\{ \left(f_{(x,1)^T} \left(W_{1,-}^{(0)}, b_1^{(0)} \right)^T, \dots, f_{(x,1)^T} \left(W_{N,-}^{(0)}, b_N^{(0)} \right)^T \right) : x \in \mathbb{R}^d \right\} \right| \\ &\leq \left| \left\{ \left(f_{\alpha} \left(W_{1,-}^{(0)}, b_1^{(0)} \right)^T, \dots, f_{\alpha} \left(W_{N,-}^{(0)}, b_N^{(0)} \right)^T \right) : \alpha \in \mathbb{R}^{d+1} \right\} \right| \leq \left(\frac{eN}{d+1} \right)^{d+1}. \end{aligned}$$

To obtain the last inequality we employed Sauer's lemma (cf. [37, Lemma 6.10]) and the estimate for the VC-dimension of halfspaces (cf. Proposition 4.4) and the assumption $d+2 < N$.

We now assume that the claim holds for some fixed $0 \leq \ell \leq L-2$. We then see

$$\begin{aligned} \mathcal{D}^{(\ell+1)} &= \left\{ \left(D^{(\ell+1)}(x), \dots, D^{(0)}(x) \right) : x \in \mathbb{R}^d \right\} \\ &= \biguplus_{(C^{(\ell)}, \dots, C^{(0)}) \in \mathcal{D}^{(\ell)}} \left\{ \left(D^{(\ell+1)}(x), C^{(\ell)}, \dots, C^{(0)} \right) : x \in \mathbb{R}^d \text{ with } D^{(j)}(x) = C^{(j)} \text{ for all } j \in \{0, \dots, \ell\} \right\} \end{aligned}$$

and hence

$$\left| \mathcal{D}^{(\ell+1)} \right| = \sum_{(C^{(\ell)}, \dots, C^{(0)}) \in \mathcal{D}^{(\ell)}} \left| \left\{ D^{(\ell+1)}(x) : x \in \mathbb{R}^d \text{ with } D^{(j)}(x) = C^{(j)} \text{ for all } j \in \{0, \dots, \ell\} \right\} \right|. \quad (5.7)$$

We thus fix $(C^{(\ell)}, \dots, C^{(0)}) \in \mathcal{D}^{(\ell)}$ and seek to bound $|\mathcal{A}|$ where

$$\mathcal{A} := \left\{ D^{(\ell+1)}(x) : x \in \mathbb{R}^d, D^{(j)}(x) = C^{(j)} \text{ for all } j \in \{0, \dots, \ell\} \right\}.$$

With $x^{(\ell)}$ as in Section 4.1 (for $1 \leq \ell \leq L$), it is immediate that

$$\begin{aligned} |\mathcal{A}| &= \left| \left\{ \left(\mathbb{1}_{W_{1,-}^{(\ell+1)} x^{(\ell+1)} + b_1^{(\ell+1)} > 0}, \dots, \mathbb{1}_{W_{N,-}^{(\ell+1)} x^{(\ell+1)} + b_N^{(\ell+1)} > 0} \right) : x \in \mathbb{R}^d \text{ with } \forall 0 \leq j \leq \ell : D^{(j)}(x) = C^{(j)} \right\} \right|. \end{aligned}$$

Fix $x \in \mathbb{R}^d$ with $D^{(j)}(x) = C^{(j)}$ for every $j \in \{0, \dots, \ell\}$. From the definition of $x^{(\ell+1)}$ (see Section 4.1) we may write

$$x^{(\ell+1)} = C^{(\ell)} W^{(\ell)} \dots C^{(0)} W^{(0)} x + \bar{c},$$

where $\bar{c} = \bar{c}(C^{(0)}, \dots, C^{(\ell)}, W^{(0)}, \dots, W^{(\ell)}, b^{(0)}, \dots, b^{(\ell)}) \in \mathbb{R}^N$ is a fixed vector. Thus, for any $i \in \{1, \dots, N\}$ we get that

$$W_{i,-}^{(\ell+1)} x^{(\ell+1)} + b_i^{(\ell+1)} = W_{i,-}^{(\ell+1)} C^{(\ell)} W^{(\ell)} \dots C^{(0)} W^{(0)} x + c_i,$$

where $c = c(C^{(0)}, \dots, C^{(\ell)}, W^{(0)}, \dots, W^{(\ell+1)}, b^{(0)}, \dots, b^{(\ell+1)}) \in \mathbb{R}^N$ is fixed. Writing

$$V := W^{(\ell+1)} C^{(\ell)} W^{(\ell)} \dots C^{(0)} W^{(0)} \in \mathbb{R}^{N \times d}$$

we infer

$$\begin{aligned} |\mathcal{A}| &= \left| \left\{ \left(f_{(x,1)^T} \left((V_{1,-}, c_1)^T \right), \dots, f_{(x,1)^T} \left((V_{N,-}, c_N)^T \right) \right) : x \in \mathbb{R}^d \right\} \right| \\ &\leq \left| \left\{ \left(f_{\alpha} \left((V_{1,-}, c_1)^T \right), \dots, f_{\alpha} \left((V_{N,-}, c_N)^T \right) \right) : \alpha \in \mathbb{R}^{d+1} \right\} \right| \leq \left(\frac{eN}{d+1} \right)^{d+1} \end{aligned}$$

where we again used [37, Lemma 6.10] and Proposition 4.4 for the last inequality, again noting that we assumed $d+2 < N$.

Combining this result with (5.7) and the induction hypothesis, we see

$$|\mathcal{D}^{(\ell+1)}| \leq |\mathcal{D}^{(\ell)}| \cdot \left(\frac{eN}{d+1} \right)^{d+1} \leq \left(\frac{eN}{d+1} \right)^{(\ell+2)(d+1)}.$$

By induction this shows that (5.6) holds for all $\ell \in \{0, \dots, L-1\}$. The statement of the lemma then follows by noting that $\mathcal{D} = \mathcal{D}^{(L-1)}$. $\square$

Having established a bound for the cardinality of the set $\mathcal{D}$, we can now bound the covering numbers of $\mathcal{L}$ in the case of deep networks.

LEMMA 5.8. Let $N > d+2$ and $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ be fixed. Moreover, let

$$\Lambda := \|W^{(L-1)}\|_2 \dots \|W^{(0)}\|_2.$$

Then for every arbitrary $\varepsilon \in (0, \Lambda)$ it holds that

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{3\Lambda}{\varepsilon} \right)^d \cdot \left(\frac{eN}{d+1} \right)^{L(d+1)}.$$

Here, $\mathcal{L}$ is as defined in Proposition 5.1.

Proof. We can assume that $\Lambda > 0$ since otherwise $(0, \Lambda) = \emptyset$. Using the notation $\mathcal{D}$ as introduced previously in lemma 5.7, we see immediately that

$$\mathcal{L} = \bigcup_{(C^{(L-1)}, \dots, C^{(0)}) \in \mathcal{D}} \left[C^{(L-1)} W^{(L-1)} \dots C^{(0)} W^{(0)} \bar{B}_d(0, 1) \right]$$

and hence it holds for every $\varepsilon > 0$ that

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \varepsilon) \leq \sum_{(C^{(L-1)}, \dots, C^{(0)}) \in \mathcal{D}} \mathcal{N}\left(C^{(L-1)}W^{(L-1)} \dots C^{(0)}W^{(0)}\bar{B}_d(0, 1), \|\cdot\|_2, \varepsilon\right). \quad (5.8)$$

This can be seen from the fact that the union of ε -nets of $C^{(L-1)}W^{(L-1)} \dots C^{(0)}W^{(0)}\bar{B}_d(0, 1)$, where $(C^{(L-1)}, \dots, C^{(0)})$ runs through all elements of $\mathcal{D}$, is an ε -net of $\mathcal{L}$.

Fix $(C^{(L-1)}, \dots, C^{(0)}) \in \mathcal{D}$ and define $V := C^{(L-1)}W^{(L-1)} \dots C^{(0)}W^{(0)}$. From Proposition 4.3 we infer that there are $w_1, \dots, w_M \in \bar{B}_d(0, 1)$ such that

$$\bar{B}_d(0, 1) \subseteq \bigcup_{i=1}^M \bar{B}_d\left(w_i, \frac{\varepsilon}{\Lambda}\right)$$

with $M \leq \left(\frac{2\Lambda}{\varepsilon} + 1\right)^d$. Let $v_i := Vw_i$ for every $i \in \{1, \dots, M\}$. Hence, it holds $v_i \in V\bar{B}_d(0, 1)$ for every $i \in \{1, \dots, M\}$. Let $u \in V\bar{B}_d(0, 1)$ be arbitrary and choose $u' \in \bar{B}_d(0, 1)$ satisfying $u = Vu'$. By choice of $w_1, \dots, w_M$ there exists $i \in \{1, \dots, M\}$ with

$$\|u' - w_i\|_2 \leq \frac{\varepsilon}{\Lambda}.$$

But then it holds

$$\begin{aligned} \|u - v_i\|_2 &= \|u - Vw_i\|_2 \leq \|V\|_2 \cdot \|u' - w_i\|_2 \\ &\leq \underbrace{\|C^{(L-1)}\|_2}_{\leq 1} \cdot \|W^{(L-1)}\|_2 \dots \underbrace{\|C^{(0)}\|_2}_{\leq 1} \cdot \|W^{(0)}\|_2 \cdot \frac{\varepsilon}{\|W^{(L-1)}\|_2 \dots \|W^{(0)}\|_2} \\ &\leq \varepsilon. \end{aligned}$$

Hence, we conclude

$$\mathcal{N}(V\bar{B}_d(0, 1), \|\cdot\|_2, \varepsilon) \leq \left(\frac{2\Lambda}{\varepsilon} + 1\right)^d \leq \left(\frac{3\Lambda}{\varepsilon}\right)^d,$$

since $\varepsilon \leq \Lambda$. Combining this estimate with Equation (5.8) and lemma 5.7 yields the claim. $\square$

The following proposition establishes the final bound when the expectation is calculated only with respect to $W^{(L)}$, i.e., conditioning on $W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}$.

PROPOSITION 5.9. Let the matrices $W^{(L-1)}, \dots, W^{(0)}$ and the biases $b^{(L-1)}, \dots, b^{(0)}$ be fixed and define $\Lambda := \|W^{(L-1)}\|_2 \dots \|W^{(0)}\|_2$. Moreover, assume that $d + 2 < N$. Then the following hold, with an absolute constant $C > 0$:

1. For any $u \geq 0$, we have

$$\sup_{x \in \mathbb{R}^d} \|W^{(L)}D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)}\|_2 \leq C \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot (\sqrt{d} + u)$$

with probability at least $(1 - 2 \exp(-u^2))$ (with respect to the choice of $W^{(L)}$).

$$2. \quad \mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 \right] \leq C \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}.$$

Proof. Using Lemma 5.8 and the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$ we infer

$$\begin{aligned} \sqrt{\ln(\mathcal{M}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} &\leq \sqrt{\ln \left(\left(\frac{3\Lambda}{\varepsilon} \right)^d \cdot \left(\frac{eN}{d+1} \right)^{L(d+1)} \right)} \\ &\leq \sqrt{d} \cdot \sqrt{\ln \left(\frac{3\Lambda}{\varepsilon} \right)} + \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \end{aligned}$$

for any $\varepsilon \in (0, \Lambda)$ where $\mathcal{L}$ is as in Proposition 5.1. This yields

$$\int_0^\Lambda \sqrt{\ln(\mathcal{M}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} \, d\varepsilon \leq \sqrt{d} \cdot \int_0^\Lambda \sqrt{\ln \left(\frac{3\Lambda}{\varepsilon} \right)} \, d\varepsilon + \Lambda \cdot \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)}. \quad (5.9)$$

Using the substitution $\sigma = \frac{\varepsilon}{3\Lambda}$ we get

$$\int_0^\Lambda \sqrt{\ln \left(\frac{3\Lambda}{\varepsilon} \right)} \, d\varepsilon = 3\Lambda \cdot \int_0^{\frac{1}{3}} \sqrt{\ln(1/\sigma)} \, d\sigma \leq C_1 \cdot \Lambda \quad (5.10)$$

with $C_1 := 3 \cdot \int_0^{1/3} \sqrt{\ln(1/\sigma)} \, d\sigma$. Overall, we thus see

$$\begin{aligned} \int_0^\Lambda \sqrt{\ln(\mathcal{M}(\mathcal{L}, \|\cdot\|_2, \varepsilon))} \, d\varepsilon &\stackrel{(5.9), (5.10)}{\leq} \Lambda \cdot \left(C_1 \cdot \sqrt{d} + \sqrt{L} \cdot \sqrt{d+1} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \right) \\ &\stackrel{d < d+1 \leq 2d}{\leq} \sqrt{2} \cdot \max\{1, C_1\} \cdot \Lambda \cdot \left(\sqrt{d} + \sqrt{L} \cdot \sqrt{d} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \right) \\ &\stackrel{L \geq 1, N \geq d+1}{\leq} \underbrace{2\sqrt{2} \cdot \max\{1, C_1\}}_{=: C_2} \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}. \quad (5.11) \end{aligned}$$

We can now prove (1). From Proposition 5.1 we obtain an absolute constant $C_3 > 0$ such that for any $u \geq 0$ the estimate

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 &\leq C_3 \left(C_2 \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d} + u \cdot \Lambda \right) \\ &\leq C_3 C_2 \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot (\sqrt{d} + u) \end{aligned}$$

holds with probability at least $(1 - 2 \exp(-u^2))$ with respect to $W^{(L)}$.

For the expectation bound, simply note that

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathbb{R}^d} \|W^{(L)} \cdot D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 \right] \leq C_3 C_2 \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}$$

follows directly from Proposition 5.1. Hence, the claim follows letting $C := C_2 C_3$. $\square$

Incorporating randomness in $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$ leads to the following theorem.

THEOREM 5.10. There exist absolute constants $C, c_1 > 0$ such that for $N > d+2$, random weight matrices $W^{(0)}, \dots, W^{(L)}$ and random bias vectors $b^{(0)}, \dots, b^{(L)}$ as in Assumption 2.1 the following hold:

1. For every $u, t \geq 0$ we have

$$\begin{aligned} &\sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 \\ &\leq C \cdot \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right) \left(2\sqrt{2} + \frac{\sqrt{2}t}{\sqrt{N}} \right)^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot (\sqrt{d} + u) \end{aligned}$$

with probability at least $(1 - 2 \exp(-u^2))_+ ((1 - 2 \exp(-c_1 t^2))_+)^L$ with respect to $W^{(0)}, \dots, W^{(L)}$ and $b^{(0)}, \dots, b^{(L)}$.

$$\begin{aligned} 2. \quad &\mathbb{E} \left[\sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2 \right] \\ &\leq C \cdot \left(1 + \frac{\sqrt{d}}{\sqrt{N}} \right) \cdot (2\sqrt{2})^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}. \end{aligned}$$

Proof. Let $\tilde{C}$ be the relabeled constant from Proposition 5.9 and define $C := \sqrt{2}\tilde{C}$. We start with (1): In view of Propositions 5.9 and C.1 it suffices to show that for every $t \geq 0$ the estimate

$$\Lambda \leq \sqrt{2} \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right) \left(2\sqrt{2} + \frac{\sqrt{2}t}{\sqrt{N}} \right)^{L-1}$$

holds with probability at least $((1 - 2 \exp(-c_1 t^2))_+)^L$, where

$$\Lambda := \|W^{(L-1)}\|_2 \cdots \|W^{(0)}\|_2.$$

To show this, note that [43, Corollary 7.3.3] yields

$$\|W^{(0)}\|_2 = \sqrt{\frac{2}{N}} \left\| \sqrt{\frac{N}{2}} W^{(0)} \right\|_2 \leq \sqrt{2} \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right)$$

with probability at least $(1 - 2 \exp(-c_1 t^2))_+$, as well as

$$\|W^{(\ell)}\|_2 = \sqrt{\frac{2}{N}} \left\| \sqrt{\frac{N}{2}} W^{(\ell)} \right\|_2 \leq \sqrt{\frac{2}{N}} (2\sqrt{N} + t) = 2\sqrt{2} + \frac{\sqrt{2}t}{\sqrt{N}}$$

with probability at least $(1 - 2 \exp(-c_1 t^2))_+$ for any $1 \leq \ell \leq L - 1$. Hence, the claim follows from an iterative application of Proposition C1.

To conclude (2) we apply [43, Theorem 7.3.1] to the matrices $\sqrt{\frac{N}{2}} W^{(\ell)}$ for $0 \leq \ell \leq L - 1$. This yields $\mathbb{E}_{W^{(\ell)}} \|W^{(\ell)}\|_2 \leq 2\sqrt{2}$ for every $1 \leq \ell \leq L - 1$ and further $\mathbb{E}_{W^{(0)}} \|W^{(0)}\|_2 \leq \sqrt{2} \left(1 + \frac{\sqrt{d}}{\sqrt{N}} \right)$. The independence of the matrices $W^{(\ell)}$ combined with Proposition 5.9 then yields the claim. $\square$

With a simple modification of the proof of Theorem 5.10 one can obtain the same high-probability bound even without assuming that the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ are jointly independent, at the cost of changing the quantity $((1 - 2 \exp(-c_1 t^2))_+)^L$ to $1 - 2L \exp(-c_1 t^2)$. Note that one still needs to assume that $W^{(L)}$ and $b^{(L)}$ are independent of $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$.

Now we obtain an upper bound on the Lipschitz constant directly using Theorem 5.10 and (4.2):

THEOREM 5.11. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network of width N and with L hidden layers satisfying Assumption 2.1. Moreover, let $d + 2 < N$. Then the following hold, for certain absolute constants $C, c_1 > 0$:

1. For every $u, t \geq 0$, we have

$$\text{Lip}(\Phi) \leq C \cdot \left(1 + \frac{\sqrt{d} + t}{\sqrt{N}} \right) \left(2\sqrt{2} + \frac{\sqrt{2}t}{\sqrt{N}} \right)^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot (\sqrt{d} + u)$$

with probability at least $(1 - 2 \exp(-u^2))_+ ((1 - 2 \exp(-c_1 t^2))_+)^L$.

2. $\mathbb{E} [\text{Lip}(\Phi)] \leq C \cdot \left(1 + \frac{\sqrt{d}}{\sqrt{N}} \right) \cdot (2\sqrt{2})^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln \left(\frac{eN}{d+1} \right)} \cdot \sqrt{d}.$

By plugging in special values for t and u and using $d \leq N$ we can now prove Theorem 3.3.

Proof of Theorem 3.3. Let $\tilde{C}$ and $\tilde{c}_1$ be the relabeled constants from Theorem 5.11 and let $C := 6\tilde{C}$ and $c_1 := \tilde{c}_1$. Part (1) follows from Theorem 5.11 by plugging in $u = \sqrt{d}$ and $t = \sqrt{N}$,

which yields

$$\text{Lip}(\Phi) \leq \tilde{C} \cdot 3 \cdot (3\sqrt{2})^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot 2 \cdot \sqrt{d} = 6\tilde{C} \cdot (3\sqrt{2})^{L-1} \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d}$$

with probability at least $(1 - 2\exp(-d)) \left((1 - 2\exp(-c_1 N))_+\right)^L$, where we also used that $d \leq N$ and $1 - 2\exp(-u^2) = 1 - 2\exp(-d) \geq 0$. Part (2) follows immediately from $d \leq N$ and the second part of Theorem 5.11. $\square$

6. Proof of the lower bound

In this section, we establish *lower* bounds on the Lipschitz constant of randomly initialized ReLU networks. The strategy for deriving these lower bounds for shallow networks differs significantly from the approach for deep networks. The lower bound in the case of shallow networks (see Section 6.1) follows from the fact that the Lipschitz constant of a shallow ReLU network can be lower bounded by the Lipschitz constant of the corresponding *linear* network (see Proposition 4.8), combined with concentration properties of Gaussian matrices and vectors. In the case of deep networks (see Section 6.2), we follow the approach that was already described in (4.3): Our strategy is to fix a point $x^{(0)} \in \mathbb{R}^d \setminus \{0\}$ and derive lower bounds for the expression

$$\|W^{(L)} D^{(L-1)}(x_0) W^{(L-1)} \dots D^{(0)}(x_0) W^{(0)}\|_2.$$

6.1 The shallow case

In this subsection, we deal with the case of shallow networks, i.e., $L = 1$. We make heavy use of Proposition 4.8, which states that we only have to consider the corresponding *linear* network in that case, i.e., the network that arises from a ReLU network by omitting the ReLU activation. This reduces the problem to bounding the norm of the product of a Gaussian matrix with a Gaussian vector (from below).

THEOREM 6.1. There exists a constant $c > 0$ with the following property: If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random shallow ReLU network with width N satisfying Assumption 2.1, then for every $t, u \geq 0$ it holds

$$\text{Lip}(\Phi) \geq \frac{1}{\sqrt{2}} \cdot \left(1 - \frac{u}{\sqrt{N}}\right)_+ \cdot (\sqrt{d} - t)_+$$

with probability at least $(1 - 2\exp(-ct^2))_+ (1 - 2\exp(-cu^2))_+$. Recall that we write $a_+ = \max\{0, a\}$ for any number $a \in \mathbb{R}$.

Proof. From Proposition 4.8 we infer that it holds

$$\text{Lip}(\Phi) \geq \frac{1}{2} \text{Lip}(\tilde{\Phi}) = \frac{1}{2} \cdot \|W^{(1)} \cdot W^{(0)}\|_2.$$

Therefore, in the following we only consider the expression $\|W^{(1)} \cdot W^{(0)}\|_2$.

To this end, we introduce the notations $U^{(0)} := (W^{(0)})^T \in \mathbb{R}^{d \times N}$ and $U^{(1)} := (W^{(1)})^T \in \mathbb{R}^N$. Moreover, we fix $t, u \geq 0$. Consider the set

$$A_1 := \{W^{(1)} : \|W^{(1)}\|_2 = \|U^{(1)}\|_2 \geq (\sqrt{N} - u)_+\}.$$

Since the entries of $W^{(1)}$ are $\mathcal{N}(0, 1)$ -distributed, and since the norm of a Gaussian random vector concentrates around the square root of its size (see, e.g., [43, Theorem 3.1.1 & Equation (2.14)], we infer

$$\mathbb{P}^{W^{(1)}}(A_1) \geq (1 - 2 \exp(-cu^2))_+$$

with a suitably chosen constant $c > 0$. We now fix the vector $W^{(1)}$ and consider the set

$$A_2(W^{(1)}) := \left\{ W^{(0)} : \|U^{(0)} U^{(1)}\|_2 \geq \frac{\sqrt{2} \|U^{(1)}\|_2}{\sqrt{N}} \cdot (\sqrt{d} - t)_+ \right\}.$$

Firstly, assume that $W^{(1)} \neq 0$. Note then that $\frac{\sqrt{N}}{\sqrt{2} \|U^{(1)}\|_2} \cdot U^{(0)} U^{(1)} \sim \mathcal{N}(0, I_d)$, as follows from the independence of the rows of $U^{(0)}$ and [43, Exercise 3.3.3 (a)]. Therefore, using again [43, Theorem 3.1.1 & Equation (2.14)] we get

$$\begin{aligned} \frac{\sqrt{N}}{\sqrt{2} \|U^{(1)}\|_2} \cdot \|U^{(0)} U^{(1)}\|_2 &\geq (\sqrt{d} - t)_+ \\ \Leftrightarrow \|U^{(0)} U^{(1)}\|_2 &\geq \frac{\sqrt{2} \|U^{(1)}\|_2}{\sqrt{N}} \cdot (\sqrt{d} - t)_+ \end{aligned}$$

with probability at least $1 - 2 \exp(-ct^2)$ and the last inequality remains true in the case $U^{(1)} = 0$. Hence, we see

$$\mathbb{P}^{W^{(0)}}(A_2(W^{(1)})) \geq (1 - 2 \exp(-ct^2))_+.$$

For any tuple $(W^{(0)}, W^{(1)})$ with $W^{(1)} \in A_1$ and $W^{(0)} \in A_2(W^{(1)})$ we get

$$\begin{aligned} \|U^{(0)} U^{(1)}\|_2 &\geq \frac{\sqrt{2} \|U^{(1)}\|_2}{\sqrt{N}} \cdot (\sqrt{d} - t)_+ \geq \frac{\sqrt{2}(\sqrt{N} - u)_+}{\sqrt{N}} \cdot (\sqrt{d} - t)_+ \\ &= \sqrt{2} \cdot \left(1 - \frac{u}{\sqrt{N}}\right)_+ \cdot (\sqrt{d} - t)_+. \end{aligned}$$

Therefore, Proposition C1 yields the claim. $\square$

Again, we plug in special values for u and t to derive the main result.

Proof of Theorem 3.2. Let $c_2 > 0$ be the constant appearing in Theorem 6.1. We then pick $u = \frac{\sqrt{N}}{2}$ and $t = \frac{\sqrt{d}}{2}$ and directly get

$$\text{Lip}(\Phi) \geq \frac{1}{\sqrt{2}} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \sqrt{d} = \frac{1}{4\sqrt{2}} \cdot \sqrt{d}$$

with probability at least $(1 - 2\exp(-c_2N/4))_+(1 - 2\exp(-c_2d/4))_+$. Hence, the first claim follows picking $c_1 := c_2/4$.

For the expectation bound, we assume $d, N > \frac{\ln(2)}{c_1}$ and use Markov's inequality to get

$$\begin{aligned} \mathbb{E}[\text{Lip}(\Phi)] &\geq \mathbb{P}\left(\text{Lip}(\Phi) \geq \frac{1}{4\sqrt{2}} \cdot \sqrt{d}\right) \cdot \frac{1}{4\sqrt{2}} \cdot \sqrt{d} \\ &\geq (1 - 2\exp(-c_1N))(1 - 2\exp(-c_1d)) \cdot \frac{1}{4\sqrt{2}} \cdot \sqrt{d} \\ &\geq \underbrace{(1 - 2\exp(-c_1 \cdot (\lfloor \ln(2)/c_1 \rfloor + 1)))(1 - 2\exp(-c_1 \cdot (\lfloor \ln(2)/c_1 \rfloor + 1)))}_{=:c} \cdot \frac{1}{4\sqrt{2}} \cdot \sqrt{d} \\ &= c \cdot \sqrt{d}. \end{aligned}$$

□

6.2 The deep case

In this subsection, we deal with the case of deep networks, i.e., $L \geq 2$. We note that the condition $L \geq 2$ is not needed in order for our proofs to work and in particular that our proofs also work in the case of shallow networks. However, for what follows we will need an additional assumption on the distribution of the biases (see Assumption 2.2) and the additional condition $N \gtrsim dL^2$ to prove our final lower bound on the Lipschitz constant, whereas these assumptions are *not* needed in the case of shallow networks. That is why we presented a different proof for shallow networks in Section 6.1.

The basic observation is that, if we fix a point $x_0 \in \mathbb{R}^d \setminus \{0\}$, the weights $W^{(0)}, \dots, W^{(\ell-1)}$ and biases $b^{(0)}, \dots, b^{(\ell-1)}$ and assume that the output of the ℓ -th layer is non-zero (i.e., $x^{(\ell)} \neq 0$ with $x^{(\ell)}$ as in Equation (4.1)), then the matrix $\sqrt{N}D^{(\ell)}(x_0)W^{(\ell)}$ has *isotropic, independent and sub-gaussian rows* (with respect to the randomness in $W^{(\ell)}$ and $b^{(\ell)}$), which is shown in Lemmas 6.3 and 6.4. Here, a random vector $X \in \mathbb{R}^k$ is called isotropic iff

$$\mathbb{E}[XX^T] = I_{k \times k}$$

with $I_{k \times k}$ denoting the k -dimensional identity matrix.

Afterwards, using the *matrix deviation inequality* (see [27, Theorem 3]) we show that the product $D^{(L-1)}(x_0)W^{(L-1)} \dots D^{(0)}(x_0)W^{(0)}$ is almost isometric with high probability, which then implies the claim.

We start by showing that for some fixed $x_0 \in \mathbb{R}^d \setminus \{0\}$ the matrices

$$\sqrt{N}D^{(\ell)}(x_0)W^{(\ell)}$$

have isotropic, independent and sub-gaussian rows when conditioning on the previous weights and biases $W^{(0)}, \dots, W^{(\ell-1)}, b^{(0)}, \dots, b^{(\ell-1)}$ and assuming $x^{(\ell)} \neq 0$. Since $W^{(0)} \in \mathbb{R}^{N \times d}$ whereas in contrast $W^{(\ell)} \in \mathbb{R}^{N \times N}$ for $1 \leq \ell \leq L-1$, we are going to prove the result generally for matrices $W \in \mathbb{R}^{N \times k}$. Let us therefore first introduce the basic assumptions of what follows.

ASSUMPTION 6.2. Let $W \in \mathbb{R}^{N \times k}$ be a random matrix and $b \in \mathbb{R}^N$ a random vector with

$$W_{ij} \sim \mathcal{N}(0, 2/N), \quad b_i \sim \mathcal{D}_i, \quad \text{for } 1 \leq i \leq N \quad \text{and} \quad 1 \leq j \leq k,$$

where each $\mathcal{D}_i$ is a symmetric probability distribution on $\mathbb{R}$. Furthermore, we assume that all the entries of W and b are jointly independent.

First, we show that the rows of the matrix are indeed isotropic, which is done in the following lemma.

LEMMA 6.3. Let Assumption 6.2 be satisfied and fix any vector $x \in \mathbb{R}^k \setminus \{0\}$. Then each row of

$$\sqrt{N} \cdot \Delta(Wx + b) \cdot W \in \mathbb{R}^{N \times k}$$

is an isotropic random vector. Here, Δ is defined as in Section 4.1.

Proof. We first consider the case $x = (\alpha, 0, \dots, 0)^T \in \mathbb{R}^k$ with $\alpha \in \mathbb{R} \setminus \{0\}$. Let $i \in \{1, \dots, N\}$ and define $V := \sqrt{N} \cdot \Delta(Wx + b) \cdot W$. It is well-known that $V_{i,-}$ is an isotropic random vector if and only if

$$\mathbb{E} \langle (V_{i,-})^T, y \rangle^2 = \|y\|_2^2$$

for every $y \in \mathbb{R}^k$; see [43, Lemma 3.2.3]. Therefore, take any arbitrary vector $y \in \mathbb{R}^k$. A direct calculation yields

$$\begin{aligned} \mathbb{E} \left[\langle (V_{i,-})^T, y \rangle^2 \right] &= \mathbb{E} \left[\left(\sum_{\ell=1}^k V_{i,\ell} y_\ell \right)^2 \right] = \mathbb{E} \left[\sum_{j,\ell=1}^k V_{i,\ell} y_\ell V_{i,j} y_j \right] = \sum_{j,\ell=1}^k y_j y_\ell \mathbb{E} [V_{i,j} V_{i,\ell}] \\ &= N \cdot \sum_{j,\ell=1}^k y_j y_\ell \mathbb{E} \left(\mathbb{1}_{(Wx+b)_i > 0} \cdot W_{i,j} \cdot W_{i,\ell} \right) \\ &= N \cdot \sum_{j,\ell=1}^k y_j y_\ell \mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,j} \cdot W_{i,\ell} \right). \end{aligned} \tag{6.1}$$

If $j \neq \ell$ and w.l.o.g. $j \neq 1$ (since $j \neq 1$ or $\ell \neq 1$, because otherwise $j = 1 = \ell$), it follows by independence that

$$\mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,j} \cdot W_{i,\ell} \right) = \underbrace{\mathbb{E} (W_{i,j})}_{=0} \cdot \mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,\ell} \right) = 0. \tag{6.2}$$

If $j = \ell \neq 1$, we see

$$\mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,j} \cdot W_{i,\ell} \right) = \mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \right) \cdot \mathbb{E} \left(W_{i,j}^2 \right) = \frac{1}{2} \cdot \frac{2}{N} = \frac{1}{N}, \quad (6.3)$$

using that $\alpha W_{i,1} + b_i$ has a symmetric and continuous probability distribution. Here, the continuity follows from $\alpha \neq 0$ and the fact that the random variables $W_{i,1}$ and b_i are independent with $W_{i,1}$ having an absolutely continuous distribution (see, e.g., [18, Proposition 9.1.6]). If $j = \ell = 1$, we have

$$\mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,j} \cdot W_{i,\ell} \right) = \mathbb{E} \left(\mathbb{1}_{\alpha W_{i,1} + b_i > 0} \cdot W_{i,1}^2 \right).$$

For simplicity, we write $X = W_{i,1}$ and $Y = b_i$ and note since $(X, Y) \sim (-X, -Y)$ (since X and Y are independent and both symmetrically distributed) that

$$\mathbb{E} \left(\mathbb{1}_{\alpha X + Y > 0} \cdot X^2 \right) = \mathbb{E} \left(\mathbb{1}_{-\alpha X - Y > 0} \cdot (-X)^2 \right) = \mathbb{E} \left(\mathbb{1}_{\alpha X + Y < 0} \cdot X^2 \right).$$

This yields

$$\begin{aligned} \mathbb{E} \left(\mathbb{1}_{\alpha X + Y > 0} \cdot X^2 \right) &= \frac{1}{2} \left(\mathbb{E} \left(\mathbb{1}_{\alpha X + Y > 0} \cdot X^2 \right) + \mathbb{E} \left(\mathbb{1}_{\alpha X + Y < 0} \cdot X^2 \right) \right) \\ &= \frac{1}{2} \mathbb{E} \left((\mathbb{1}_{\alpha X + Y > 0} + \mathbb{1}_{\alpha X + Y < 0}) \cdot X^2 \right) = \frac{1}{2} \mathbb{E}(X^2) = 1/N. \end{aligned} \quad (6.4)$$

Here, we used that $\mathbb{E}(\mathbb{1}_{\alpha X + Y = 0} \cdot X^2) = 0$, since αX has an absolutely continuous distribution (note $\alpha \neq 0$) and X and Y are independent so that $\mathbb{P}(\alpha X + Y = 0) = 0$. Inserting (6.2), (6.3) and (6.4) into (6.1) yields

$$\mathbb{E} \langle (V_{i,-})^T, y \rangle^2 = \|y\|_2^2,$$

hence the rows of V are isotropic in the case $x = (\alpha, 0, \dots, 0)^T \in \mathbb{R}^k$ with $\alpha \in \mathbb{R} \setminus \{0\}$.

Now consider an arbitrary vector $x \in \mathbb{R}^k \setminus \{0\}$. Taking an orthogonal deterministic matrix $U \in \mathbb{R}^{k \times k}$ with $x = U\alpha e_1$ for some $\alpha \in \mathbb{R} \setminus \{0\}$ yields

$$\Delta(Wx + b) \cdot W = \Delta(WU\alpha e_1 + b) \cdot WU \cdot U^{-1}.$$

For every $j \in \{1, \dots, N\}$ we see

$$\left((WU)^T \right)_{-j} = \left(U^T W^T \right)_{-j} = U^T \left(W^T \right)_{-j} \sim \mathcal{N} \left(0, \frac{2}{N} I_k \right),$$

where in the last step we used the rotation invariance of the Gaussian distribution [43, Proposition 3.3.2]. This yields

$$(WU)_{j,-} \sim \mathcal{N} \left(0, \frac{2}{N} I_k \right)$$

for every $j \in \{1, \dots, N\}$. Since $(WU)_{j,-}$ only depends on $W_{j,-}$, this implies

$$(WU)_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N) \quad \text{for all } i \in \{1, \dots, N\}, j \in \{1, \dots, k\}.$$

Hence, the matrix $V := \Delta((WU)\alpha e_1 + b) \cdot (WU)$ has isotropic rows by what has just been shown (the case $x = (\alpha, 0, \dots, 0)^T$). But then $V \cdot U^{-1}$ has isotropic rows too. Indeed, let $i \in \{1, \dots, N\}$. Then we have $((V \cdot U^{-1})_{i,-})^T = U \cdot (V_{i,-})^T$ where $(V_{i,-})^T$ is an isotropic random vector. Writing $Z := (V_{i,-})^T$, we see for any vector $y \in \mathbb{R}^k$ that

$$\mathbb{E}\langle UZ, y \rangle^2 = \mathbb{E}\langle Z, U^T y \rangle^2 = \|U^T y\|_2^2 = \|y\|_2^2,$$

so UZ is isotropic as was to be shown. $\square$

Our next lemma collects all of the properties of the matrices $\sqrt{N}D^{(\ell)}(x_0)W^{(\ell)}$ that we will need.

LEMMA 6.4. Let Assumption 6.2 be satisfied and fix $x \in \mathbb{R}^k \setminus \{0\}$. Let $\Delta(v)$, $v \in \mathbb{R}^N$, be as defined in Section 4.1. Then the rows of $\sqrt{N} \Delta(Wx + b) \cdot W$ are jointly independent, isotropic random vectors, and sub-gaussian with

$$\left\| \left(\sqrt{N} \cdot \Delta(Wx + b) \cdot W \right)_{i,-} \right\|_{\psi_2} \leq C \quad \text{for all } i \in \{1, \dots, N\},$$

where $C > 0$ is an absolute constant.

Proof. Note that

$$\left(\sqrt{N} \Delta(Wx + b) \cdot W \right)_{i,-} = \sqrt{N} \cdot \mathbb{1}_{(Wx+b)_i > 0} \cdot W_{i,-},$$

which only depends on b_i and the i -th row of W , which implies that the rows of this matrix are jointly independent. For every vector $y \in \mathbb{R}^k$ and every $i \in \{1, \dots, N\}$ we see

$$\begin{aligned} \left| \left(\sqrt{N} \Delta(Wx + b) \cdot W \right)_{i,-} \cdot y \right| &= \left| \sum_{j=1}^k \left(\Delta(Wx + b) \cdot \sqrt{N}W \right)_{ij} y_j \right| = \left| \sum_{j=1}^k \mathbb{1}_{(Wx+b)_i > 0} \cdot \sqrt{N}W_{ij} y_j \right| \\ &= \mathbb{1}_{(Wx+b)_i > 0} \cdot \left| \sum_{j=1}^k \sqrt{N}W_{ij} y_j \right| \leq \left| \sum_{j=1}^k \sqrt{N}W_{ij} y_j \right|. \end{aligned}$$

Note that the $W_{ij}y_j$ ($j \in \{1, \dots, k\}$) are independent with $\sqrt{N}W_{ij}y_j \sim \mathcal{N}(0, 2y_j^2)$ and hence

$$\sum_{j=1}^k \sqrt{N}W_{ij}y_j \sim \mathcal{N}(0, 2 \cdot \|y\|_2^2).$$

Hence, by definition (see [43, Section 2.5.2 and Definition 3.4.1]) it follows that the random variable $\left(\sqrt{N} \Delta(Wx + b) \cdot W\right)_{i,-} \cdot y$ is sub-gaussian and since $y \in \mathbb{R}^k$ has been chosen arbitrarily we deduce that $\left(\sqrt{N} \Delta(Wx + b) \cdot W\right)_{i,-}$ is sub-gaussian with

$$\left\| \left(\sqrt{N} \Delta(Wx + b) \cdot W\right)_{i,-} \right\|_{\psi_2} \leq \sup_{y \in \mathbb{S}^{k-1}} \left\| \sum_{j=1}^k \sqrt{N} W_{i,j} y_j \right\|_{\psi_2} \leq \sqrt{2} \cdot C_1 =: C,$$

where C_1 is an absolute constant according to [43, Example 2.5.8 (i)].

The isotropy has already been shown in Lemma 6.3. $\square$

We now turn to the proof of the lower bound in the case of deep networks. For what follows, we assume that the considered ReLU network satisfies Assumption 2.2. We take a *fixed* vector

$$x_0 := x^{(0)} \in \mathbb{R}^d \setminus \{0\}$$

and define the matrices $D^{(0)}(x_0), \dots, D^{(L-1)}(x_0)$ as introduced in Section 4.1. Since x_0 is a fixed vector, we omit the argument and just write $D^{(\ell)}$ instead of $D^{(\ell)}(x_0)$.

First, we prove that the product matrix $D^{(L-1)} W^{(L-1)} \dots D^{(0)} W^{(0)}$ is almost isometric with high probability. This will be based on the fact that the rows of $\sqrt{N} D^{(\ell)} W^{(\ell)}$ are independent, isotropic random vectors (see Lemma 6.4). However, in order to guarantee these properties, we have to make sure that the output of the previous layer $x^{(\ell)}$ is *not* zero. Hence, in the following proposition we carefully keep track of this event as well. Moreover, since the ultimate goal is to apply Equation (4.3), we have to ensure that the network Φ is differentiable at x_0 with

$$(\nabla \Phi(x_0))^T = W^{(L)} \cdot D^{(L-1)} \cdot W^{(L-1)} \dots D^{(0)} \cdot W^{(0)},$$

which is satisfied if all pre-activations throughout the network are non-zero. This is why we also consider this event in the following proposition.

PROPOSITION 6.5. Let $W^{(0)}, \dots, W^{(L)}$ and $b^{(0)}, \dots, b^{(L)}$ as in Assumption 2.2, and let $x^{(0)} \in \mathbb{R}^d \setminus \{0\}$ be fixed and let $D^{(0)}, \dots, D^{(L-1)}$ and $x^{(1)}, \dots, x^{(L)}$ as in Section 4.1. For every $C > 0, u \geq 0$ and $\ell \in \{1, \dots, L\}$, we write $\mathcal{G}^{(\ell)} = \mathcal{G}^{(\ell)}(u, C)$ for the event defined via the following three properties:

1. $(W^{(\ell')} x^{(\ell')} + b^{(\ell')})_i \neq 0$ for all $\ell' \in \{0, \dots, \ell - 1\}$ and $i \in \{1, \dots, N\}$,
2. $x^{(\ell')} \neq 0$ for all $\ell' \in \{0, \dots, \ell\}$,
3. $\left(\left(1 - \frac{C(\sqrt{d}+u)}{\sqrt{N}} \right)_+ \right)^\ell \|y\|_2 \leq \|D^{(\ell-1)} W^{(\ell-1)} \dots D^{(0)} W^{(0)} y\|_2 \leq \left(1 + \frac{C(\sqrt{d}+u)}{\sqrt{N}} \right)^\ell \|y\|_2$

holds uniformly over all $y \in \mathbb{R}^d$.

Recall that $a_+ = \max\{a, 0\}$ for any $a \in \mathbb{R}$. Then, there exists an absolute constant $C > 0$ with

$$\mathbb{P}(\mathcal{G}^{(\ell)}) = \mathbb{P}(\mathcal{G}^{(\ell)}(u, C)) \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^\ell$$

for every $u \geq 0$.

Proof. The proof is via induction over ℓ , where the constant C is determined later. We note that for fixed ℓ , the defining conditions of $\mathcal{G}^{(\ell)}$ only depend on the weights $W^{(0)}, \dots, W^{(\ell-1)}$ and the biases $b^{(0)}, \dots, b^{(\ell-1)}$.

We start with the case $\ell = 1$. We denote

$$\begin{aligned} A(1) &:= \left\{ (W^{(0)}, \dots, W^{(\ell-1)}, b^{(0)}, \dots, b^{(\ell-1)}) : (1) \text{ is satisfied} \right\} \\ B(1) &:= \left\{ (W^{(0)}, \dots, W^{(\ell-1)}, b^{(0)}, \dots, b^{(\ell-1)}) : (2) \text{ is satisfied} \right\} \\ C(1) &:= \left\{ (W^{(0)}, \dots, W^{(\ell-1)}, b^{(0)}, \dots, b^{(\ell-1)}) : (3) \text{ is satisfied} \right\}. \end{aligned}$$

For every $i \in \{1, \dots, N\}$, we have, using $*$ to denote the convolution of two measures,

$$\left(W^{(0)} x^{(0)} + b^{(0)} \right)_i \sim \mathcal{N}(0, 2/N \cdot \|x^{(0)}\|_2^2) * \mathcal{D}_i^{(0)},$$

where the latter is an absolutely continuous and symmetric probability distribution. Note that we assume $x^{(0)} \neq 0$. Hence, we first conclude from the joint independence of the above random variables that

$$\mathbb{P}^{(W^{(0)}, b^{(0)})}(A(1)) = 1.$$

Moreover, we have

$$(W^{(0)}, b^{(0)}) \notin B(1) \iff \left(W^{(0)} x^{(0)} + b^{(0)} \right)_i \leq 0 \quad \text{for all } i \in \{1, \dots, N\}.$$

From the joint independence, the symmetry and the fact that the random variables $(W^{(0)} x^{(0)} + b^{(0)})_i$ follow an absolutely continuous distribution for every i , we infer

$$\mathbb{P}^{(W^{(0)}, b^{(0)})}(B(1)^c) = \frac{1}{2^N}.$$

Moreover, note that $\sqrt{N} D^{(0)} W^{(0)} = \sqrt{N} \Delta(W^{(0)} x^{(0)} + b^{(0)}) W^{(0)}$ is a matrix with independent isotropic sub-gaussian rows according to Lemma 6.4. Therefore, the high probability version of the matrix deviation inequality (see [27, Theorem 3]) yields

$$\sup_{y \in \bar{B}_d(0,1)} \left| \|\sqrt{N} D^{(0)} W^{(0)} y\|_2 - \sqrt{N} \|y\|_2 \right| \leq C(\sqrt{d} + u)$$

with probability at least $1 - \exp(-u^2)$, where we employed Proposition 4.6. $C > 0$ is an absolute constant according to Lemma 6.4. This yields

$$\left(1 - \frac{C(\sqrt{d} + u)}{\sqrt{N}} \right)_+ \|y\|_2 \leq \|D^{(0)} W^{(0)} y\|_2 \leq \left(1 + \frac{C(\sqrt{d} + u)}{\sqrt{N}} \right) \|y\|_2 \quad \text{for all } y \in \mathbb{R}^d$$

with probability at least $1 - \exp(-u^2)$. This gives us $\mathbb{P}(C(1)^c) \leq \exp(-u^2)$. This gives us in total

$$\mathbb{P}^{(W^{(0)}, b^{(0)})}(\mathcal{G}^{(1)}) \geq 1 - \mathbb{P}(A(1)^c) - \mathbb{P}(B(1)^c) - \mathbb{P}(C(1)^c) \geq 1 - 0 - \frac{1}{2^N} - \exp(-u^2).$$

Fix $1 \leq \ell < L$, set $V' := D^{(\ell-1)}W^{(\ell-1)} \dots D^{(0)}W^{(0)}$ and write $\vec{W} := (W^{(0)}, \dots, W^{(\ell-1)})$ for the tuple of the first ℓ weight matrices and $\vec{b} := (b^{(0)}, \dots, b^{(\ell-1)})$ for the tuple of the first ℓ bias vectors and assume by induction that

$$\mathbb{P}\left(\vec{W}, \vec{b} \in \mathcal{G}^{(\ell)}\right) \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2)\right)_+\right)^\ell. \quad (6.5)$$

Furthermore, we write

$$A(\vec{W}, \vec{b}) := \left\{ (W^{(\ell)}, b^{(\ell)}) : \left(W^{(\ell)}x^{(\ell)} + b^{(\ell)} \right)_i \neq 0 \text{ for all } i \in \{1, \dots, N\} \right\}$$

$$B(\vec{W}, \vec{b}) := \left\{ (W^{(\ell)}, b^{(\ell)}) : x^{(\ell+1)} = \text{ReLU}(W^{(\ell)}x^{(\ell)} + b^{(\ell)}) \neq 0 \right\}$$

$$C(\vec{W}, \vec{b}) := \left\{ (W^{(\ell)}, b^{(\ell)}) : \left(1 - \frac{C(\sqrt{d}+u)}{\sqrt{N}} \right)_+ \|V'y\|_2 \leq \|D^{(\ell)}W^{(\ell)}V'y\|_2 \leq \left(1 + \frac{C(\sqrt{d}+u)}{\sqrt{N}} \right) \|V'y\|_2 \text{ for all } y \in \mathbb{R}^d \right\}$$

for each $\vec{W}, \vec{b}$, where we note that V' is a function of only $\vec{W}$ and $\vec{b}$ (and $x^{(0)}$, which is fixed). Then we see that

$$(\vec{W}, \vec{b}) \in \mathcal{G}^{(\ell)}, (W^{(\ell)}, b^{(\ell)}) \in A(\vec{W}, \vec{b}) \cap B(\vec{W}, \vec{b}) \cap C(\vec{W}, \vec{b}) \implies (\vec{W}, W^{(\ell)}, \vec{b}, b^{(\ell)}) \in \mathcal{G}^{(\ell+1)}.$$

In view of Proposition C1, we thus seek to bound

$$\mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(A(\vec{W}, \vec{b}) \cap B(\vec{W}, \vec{b}) \cap C(\vec{W}, \vec{b}) \right)$$

from below, where $(\vec{W}, \vec{b}) \in \mathcal{G}^{(\ell)}$ is fixed.

To this end, we consider the sets $A(\vec{W}, \vec{b}), B(\vec{W}, \vec{b}), C(\vec{W}, \vec{b})$ individually. Note that the vector $x^{(\ell)}$ is fixed and from $(\vec{W}, \vec{b}) \in \mathcal{G}^{(\ell)}$ we infer $x^{(\ell)} \neq 0$. Hence,

$$\left(W^{(\ell)}x^{(\ell)} + b^{(\ell)} \right)_i \sim \mathcal{N}\left(0, 2/N \cdot \|x^{(\ell)}\|_2^2\right) * \mathcal{D}_i^{(\ell)}$$

for every $i \in \{1, \dots, N\}$, where the latter is again a symmetric and (absolutely) continuous probability distribution. Similar to the case $\ell = 1$, this gives us

$$\mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(A(\vec{W}, \vec{b}) \right) = 1.$$

Moreover, as in the case $\ell = 1$, we get

$$(W^{(\ell)}, b^{(\ell)}) \notin B(\vec{W}, \vec{b}) \iff \left(W^{(\ell)} x^{(\ell)} + b^{(\ell)} \right)_i \leq 0 \quad \text{for all } i \in \{1, \dots, N\}.$$

Again, from the joint independence, the symmetry and the absolute continuity of the distribution of the random variables

$$\left(W^{(\ell)} x^{(\ell)} + b^{(\ell)} \right)_i,$$

we get

$$\mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(B(\vec{W}, \vec{b})^c \right) = \frac{1}{2^N}.$$

Lastly, according to Lemma 6.4, the matrix

$$\sqrt{N} D^{(\ell)} W^{(\ell)} = \sqrt{N} \Delta(W^{(\ell)} x^{(\ell)} + b^{(\ell)}) W^{(\ell)}$$

has independent, sub-gaussian, isotropic rows. Hence, we may again apply the high-probability version of the matrix deviation inequality ([27, Theorem 3]) and obtain

$$\sup_{y \in \text{Im}(V') \cap \bar{B}_N(0,1)} \left| \sqrt{N} \|D^{(\ell)} W^{(\ell)} y\|_2 - \sqrt{N} \|y\|_2 \right| \leq C(\sqrt{d} + u)$$

with probability at least $1 - \exp(-u^2)$, where we again used Proposition 4.6, noting that the subspace $\text{Im}(V')$ is at most d -dimensional since

$$V' = D^{(\ell-1)} W^{(\ell-1)} \dots D^{(0)} W^{(0)} \in \mathbb{R}^{N \times d}.$$

But this directly implies

$$\left(1 - \frac{C(\sqrt{d} + u)}{\sqrt{N}} \right)_+ \|y\|_2 \leq \|D^{(\ell)} W^{(\ell)} y\|_2 \leq \left(1 + \frac{C(\sqrt{d} + u)}{\sqrt{N}} \right) \|y\|_2 \quad \forall y \in \text{Im}(V')$$

with probability at least $1 - \exp(-u^2)$, which means

$$\mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} (C(\vec{W}, \vec{b})) \geq 1 - \exp(-u^2).$$

Hence, we get

$$\begin{aligned}
& \mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(A(\vec{W}, \vec{b}) \cap B(\vec{W}, \vec{b}) \cap C(\vec{W}, \vec{b}) \right) \\
& \geq 1 - \mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(A(\vec{W}, \vec{b})^c \right) - \mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(B(\vec{W}, \vec{b})^c \right) - \mathbb{P}^{(W^{(\ell)}, b^{(\ell)})} \left(C(\vec{W}, \vec{b})^c \right) \\
& \geq 1 - 0 - \frac{1}{2^N} - \exp(-u^2) = 1 - \frac{1}{2^N} - \exp(-u^2)
\end{aligned}$$

for every fixed $(\vec{W}, \vec{b}) \in \mathcal{G}^{(\ell)}$. Using Proposition C.1, we obtain

$$\mathbb{P}(\mathcal{G}^{(\ell+1)}) \geq \left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \cdot \mathbb{P}^{(\vec{W}, \vec{b})}(\mathcal{G}^{(\ell)}) \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^{\ell+1},$$

as was to be shown. Here, we applied the $+$ -operator since probabilities are non-negative. $\square$

To finalize our result, we introduce randomness in $W^{(L)}$ as well and use an argument based on the singular value decomposition of the product $D^{(L-1)} W^{(L-1)} \dots D^{(0)} W^{(0)}$.

PROPOSITION 6.6. Let Φ and $W^{(0)}, \dots, W^{(L)}$ as well as $b^{(0)}, \dots, b^{(L)}$ as in Assumption 2.2 and let the input vector $x_0 := x^{(0)} \in \mathbb{R}^d \setminus \{0\}$ be fixed. Let $D^{(\ell)} := D^{(\ell)}(x_0)$ for $0 \leq \ell \leq L-1$ with $D^{(\ell)}(x_0)$ as in Section 4.1. We let $k := \min\{N, d\}$, $u \geq 0$, $C > 0$ and $t \geq 0$ and define the event $\mathcal{A} = \mathcal{A}(u, t, C)$ via the properties

1.

$$\begin{aligned}
\left(\left(1 - \frac{C \cdot (\sqrt{d} + u)}{\sqrt{N}} \right)_+ \right)^L \cdot (\sqrt{k} - t) & \leq \|W^{(L)} D^{(L-1)} W^{(L-1)} \dots D^{(0)} W^{(0)}\|_2 \\
& \leq \left(1 + \frac{C \cdot (\sqrt{d} + u)}{\sqrt{N}} \right)^L \cdot (\sqrt{k} + t),
\end{aligned}$$

2. Φ is differentiable at x_0 with

$$(\Phi(x_0))^T = W^{(L)} \cdot D^{(L-1)} \cdot W^{(L-1)} \dots D^{(0)} \cdot W^{(0)}.$$

Then there exist absolute constants $C, c_1 > 0$ such that

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A}(u, t, C)) \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^L \cdot (1 - 2 \exp(-c_1 t^2))_+$$

for every $u, t \geq 0$.

Proof. We again denote $V' := D^{(L-1)} W^{(L-1)} \dots D^{(0)} W^{(0)}$. We decompose $V' = U \Sigma Q^T$ with orthogonal matrices $U \in \mathbb{R}^{N \times N}$ and $Q \in \mathbb{R}^{d \times d}$ and a matrix $\Sigma \in \mathbb{R}^{N \times d}$ of the form

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_d & \\ 0 & \dots & 0 & \\ \vdots & \ddots & \vdots & \\ 0 & \dots & 0 & \end{pmatrix} \quad \text{if } N \geq d \quad \text{or} \quad \Sigma = \begin{pmatrix} \sigma_1 & & 0 & \dots & 0 \\ & \ddots & \vdots & \ddots & \vdots \\ & & \sigma_N & 0 & \dots & 0 \end{pmatrix} \quad \text{if } N \leq d$$

with $\sigma_1 \geq \dots \geq \sigma_k \geq 0$ (singular value decomposition). Hence, recalling that $W^{(L)} \in \mathbb{R}^{1 \times N}$, we get

$$\begin{aligned} \|W^{(L)} V'\|_2 &= \|W^{(L)} U \Sigma Q^T\|_2 = \|W^{(L)} U \Sigma\|_2 = \left\| \left(W^{(L)} U \Sigma \right)^T \right\|_2 \\ &= \sqrt{\sum_{i=1}^k (\sigma_i \cdot (W^{(L)} U)_i)^2} \begin{cases} \leq \sigma_1 \cdot \|W'\|_2, \\ \geq \sigma_k \cdot \|W'\|_2 \end{cases} \end{aligned} \quad (6.6)$$

with $W' \in \mathbb{R}^{1 \times k}$ denoting the vector of the first k entries of $W^{(L)} U$.

Let $C > 0$ be the absolute constant from Proposition 6.5. We denote $\vec{W} := (W^{(0)}, \dots, W^{(L-1)})$ and $\vec{b} := (b^{(0)}, \dots, b^{(L-1)})$ and fix $u, t \geq 0$. Furthermore, let

$$\mathcal{B}(\vec{W}, \vec{b}) := \left\{ W^{(L)} : \sqrt{d} + t \geq \|W'(\vec{W}, \vec{b}, W^{(L)})\|_2 \geq \sqrt{d} - t \right\},$$

where we wrote $W'(\vec{W}, \vec{b}, W^{(L)})$ to emphasize the dependence of W' on $\vec{W}, \vec{b}$ and $W^{(L)}$.

Note that for $(\vec{W}, \vec{b}) \in \mathcal{G}^{(L)}$ (where $\mathcal{G}^{(L)}$ is as in Proposition 6.5), property (2) follows directly and independent of the choice of $W^{(L)}$ and $b^{(L)}$. This is due to defining property (1) in Proposition 6.5 and the fact that the ReLU is differentiable on $\mathbb{R} \setminus \{0\}$. Hence, from (6.6) we infer

$$(\vec{W}, \vec{b}) \in \mathcal{G}^{(L)}, W^{(L)} \in \mathcal{B}(\vec{W}, \vec{b}) \implies (\vec{W}, \vec{b}, W^{(L)}) \in \mathcal{A}, \quad (6.7)$$

where we also applied [43, Equation (4.5)]. From Proposition 6.5 we deduce

$$\mathbb{P}^{(\vec{W}, \vec{b})}(\mathcal{G}^{(L)}) \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^L. \quad (6.8)$$

Furthermore, for fixed $\vec{W}$ and $\vec{b}$, the rotation invariance of the Gaussian distribution [43, Proposition 3.3.2] implies

$$W^{(L)} U \sim W^{(L)}.$$

Therefore, W' is a k -dimensional random vector with $(W')^T \sim \mathcal{N}(0, I_k)$. Thus, [43, Theorem 3.1.1] yields

$$\left\| \|W'\|_2 - \sqrt{k} \right\|_{\psi_2} \leq C_2 \quad (\text{conditioned on } \vec{W}, \vec{b})$$

with an absolute constant $C_2 > 0$. From [43, Equation (2.14)] we get

$$\mathbb{P}^{W^{(L)}} \left(\left| \|W'\|_2 - \sqrt{k} \right| \geq t \right) \leq 2 \exp(-c_1 t^2) \quad \text{for fixed } \vec{W}, \vec{b}$$

with an absolute constant $c_1 > 0$, and hence

$$\mathbb{P}^{W^{(L)}} \left(\sqrt{k} + t \geq \|W'\|_2 \geq (\sqrt{k} - t)_+ \right) \geq 1 - 2 \exp(-c_1 t^2) \quad \text{for fixed } \vec{W}, \vec{b}.$$

From Proposition C1 and (6.7), we see

$$\begin{aligned} \mathbb{P}^{(\vec{W}, \vec{b}, W^{(L)})}(\mathcal{A}) &\geq (1 - 2 \exp(-c_1 t^2))_+ \cdot \mathbb{P}^{(\vec{W}, \vec{b})}(\mathcal{G}^{(L)}) \\ &\stackrel{(6.8)}{\geq} (1 - 2 \exp(-c_1 t^2))_+ \cdot \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^L, \end{aligned}$$

as was to be shown. $\square$

We remark that the previous result, in addition to providing a lower bound on the Lipschitz constant of random neural networks, is of independent interest on its own, since it provides a lower and *upper* bound on the gradient of a random ReLU network at a fixed point $x_0 \neq 0$. It is an interesting question whether this *pointwise* estimate can be used to get a uniform estimate that holds for every point $x^{(0)}$, thus yielding an *upper* bound on the Lipschitz constant as well.

Moreover, we note that one can even show that a random ReLU network Φ is *almost surely* differentiable with

$$\nabla \Phi(x_0)^T = W^{(L)} \cdot D^{(L-1)} \dots D^{(0)} \cdot W^{(0)}$$

at any fixed point $x_0 \neq 0$ (and not only with high probability as stated in Proposition 6.6). A proof of this fact (which we expect to be of independent interest) is contained in Appendix E.

Proposition 6.6 and Equation (4.3) directly give us the following lower bound on the Lipschitz constant of random ReLU networks.

THEOREM 6.7. There exist absolute constants $C, c_1 > 0$ with the following property: If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random ReLU network of width N and with L hidden layers according to the random initialization as described in Assumption 2.2, then for any $u, t \geq 0$, writing $k := \min\{d, N\}$, it holds

$$\text{Lip}(\Phi) \geq \left(\left(1 - \frac{C \cdot (\sqrt{d} + u)}{\sqrt{N}} \right)_+ \right)^L \cdot (\sqrt{k} - t)$$

with probability at least $\left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^L \cdot (1 - 2 \exp(-c_1 t^2))_+.$

By plugging in special values for t and u and assuming $N \gtrsim dL^2$, we can now prove the main result.

Proof of Theorem 3.4. Let $\tilde{C}$ and $\tilde{c}_1$ be the relabeled constants from Theorem 6.7. We can clearly assume $\tilde{C} \geq 1$. We define the new constants $C := (4\tilde{C})^2$ and $c_1 := \tilde{c}_1/4$. We assume $N \geq C \cdot d \cdot L^2$ and let $u = \frac{\sqrt{N}}{4\tilde{C}L}$ and $t = \sqrt{d}/2$. Note that $N \geq CdL^2$ is equivalent to

$$\sqrt{d} \leq \frac{\sqrt{N}}{4\tilde{C}L}.$$

We get

$$\tilde{C} \cdot (\sqrt{d} + u) = \tilde{C} \cdot \left(\sqrt{d} + \frac{\sqrt{N}}{4\tilde{C}L} \right) \leq \tilde{C} \cdot \left(\frac{\sqrt{N}}{4\tilde{C}L} + \frac{\sqrt{N}}{4\tilde{C}L} \right) = \frac{\sqrt{N}}{2L} \leq \sqrt{N}$$

and hence

$$1 - \frac{\tilde{C} \cdot (\sqrt{d} + u)}{\sqrt{N}} \geq 0.$$

Moreover, $N \geq C \cdot d \cdot L^2 \geq d$ and thus $k = \min\{N, d\} = d$. Therefore, Theorem 6.7 yields

$$\text{Lip}(\Phi) \geq \left(1 - \frac{\tilde{C}(\sqrt{d} + u)}{\sqrt{N}} \right)^L \cdot \frac{1}{2} \cdot \sqrt{d}$$

with probability at least

$$\begin{aligned} & \left(\left(1 - \frac{1}{2^N} - \exp(-u^2) \right)_+ \right)^L \cdot (1 - 2 \exp(-\tilde{c}_1 d/4))_+ \\ &= \left(1 - \frac{1}{2^N} - \exp(-N/(CL^2)) \right)_+^L \cdot (1 - 2 \exp(-c_1 d))_+. \end{aligned}$$

Here, we implicitly used that

$$1 - \frac{1}{2^N} - \exp(-N/(CL^2)) \geq 1 - \frac{1}{2^N} - \exp(-d) \geq 0,$$

where we employed $N \geq CdL^2$. Moreover, note that

$$\frac{\tilde{C} \cdot (\sqrt{d} + u)}{\sqrt{N}} = \tilde{C} \cdot \left(\frac{\sqrt{d}}{\sqrt{N}} + \frac{1}{4\tilde{C}L} \right) \leq \tilde{C} \cdot \left(\frac{1}{4\tilde{C}L} + \frac{1}{4\tilde{C}L} \right) = \frac{1}{2L},$$

which yields

$$\left(1 - \frac{\tilde{C} \cdot (\sqrt{d} + u)}{\sqrt{N}}\right)^L \geq \left(1 - \frac{1}{2L}\right)^L \geq 1/2$$

for every $L \in \mathbb{N}$, which follows from Bernoulli's inequality. This gives us (1).

For the expectation bound, note that by Markov's inequality we have

$$\mathbb{E}[\text{Lip}(\Phi)] \geq \mathbb{P}(\text{Lip}(\Phi) \geq \sqrt{d}/4) \cdot \frac{\sqrt{d}}{4} \geq \left(\left(1 - \frac{1}{2^N} - \exp(-u^2)\right)^L \cdot (1 - 2\exp(-c_1 d)) \right) \cdot \frac{\sqrt{d}}{4}.$$

First note that there is an absolute constant $c_3 > 0$ with

$$1 - 2\exp(-c_1 d) \geq c_3$$

for every $d \geq \left\lfloor \frac{\ln(2)}{c_1} \right\rfloor + 1$. Therefore, it remains to find a uniform bound for

$$\left(1 - \frac{1}{2^N} - \exp(-u)\right)^L.$$

We apply Bernoulli's inequality and $L \leq N$ to obtain

$$\begin{aligned} \left(1 - \frac{1}{2^N} - \exp(-u^2)\right)^L &\geq 1 - \frac{L}{2^N} - L\exp(-u^2) \geq 1 - \frac{N}{2^N} - L\exp(-N/(CL^2)) \\ &\geq \frac{1}{2} - L\exp(-N/(CL^2)). \end{aligned}$$

If we now assume $N \geq CL^2 \ln(4L)$, we get

$$\exp(-N/(CL^2)) \leq \exp(-CL^2 \ln(4L)/(CL^2)) = \exp(-\ln(4L)) = \frac{1}{4L}.$$

Hence, the claim follows letting $c := \frac{c_3}{16}$. □

Acknowledgements

The authors thank Felix Krahmer, Afonso Bandeira and Sam Buchanan for fruitful discussions and helpful comments. The authors thank the anonymous reviewers for suggesting several improvements to the presentation of the results.

Funding

This work was supported by German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1 and by the Hightech Agenda Bavaria.

Data availability statement

Does not apply.

REFERENCES

1. ALIPRANTIS, C. D. & BORDER, K. C. (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd edn. Berlin: Springer.
2. ALLEN-ZHU, Z., LI, Y. & SONG, Z. (2019) A convergence theory for deep learning via over-parameterization. *International Conference on Machine Learning*. (K. CHAUDHURI & R. SALAKHUTDINOV eds). Long Beach: PMLR, pp. 242–252.
3. ATHALYE, A., CARLINI, N. & WAGNER, D. (2018) Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. *International Conference on Machine Learning*. (D. JENNIFER & K. ANDREAS eds). Stockholm: PMLR, pp. 274–283.
4. BARTLETT, P., BUBECK, S. & CHERAPANAMJERI, Y. (2021) Adversarial examples in multi-layer random relu networks. *Adv. Neural Inf. Process. Syst.*, **34**, 9241–9252.
5. BERNER, J., ELBRÄCHTER, D., GROHS, P. & JENTZEN, A. (2019) Towards a regularity theory for relu networks—chain rule and global error estimates. In *2019 13th International Conference on Sampling Theory and Applications (SampTA)*. Bordeaux: IEEE, pp. 1–5.
6. BOMBARI, S., KIYANI, S. & MONDELLI, M. (2023) Beyond the universal law of robustness: sharper laws for random features and neural tangent kernels. arXiv:2302.01629.
7. BUBECK, S., CHERAPANAMJERI, Y., GIDEL, G. & COMBES, R. T. D. (2021) A single gradient step finds adversarial examples on random two-layers neural networks. *Adv. Neural Inf. Process. Syst.*, **34**, 10081–10091.
8. BUBECK, S., LI, Y. & NAGARAJ, D. M. (2021) A law of robustness for two-layers neural networks. *Conference on Learning Theory*. (M. BELKIN & S. KPOUFE eds). Boulder (Colorado): PMLR, pp. 804–820.
9. BUBECK, S. & SELKE, M. (2021) A universal law of robustness via isoperimetry. *Adv. Neural Inf. Process. Syst.*, **34**, 28811–28822.
10. BUCHANAN, S., GILBOA, D. & WRIGHT, J. (2021) Deep networks and the multiple manifold problem. In *International Conference on Learning Representations*. Vienna.
11. BUI, A. T., LE, T., TRAN, Q. H., ZHAO, H. & PHUNG, D. (2021) A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*. Vienna.
12. CARLINI, N. & WAGNER, D. (2018) Audio adversarial examples: targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. San Francisco: IEEE, pp. 1–7.
13. CHIZAT, L., OYALLON, E. & BACH, F. (2019) On lazy training in differentiable programming. *Adv. Neural Inf. Process. Syst.*, **32**.
14. COHN, D. L. (2013) *Measure Theory*, 2nd edn. New York: Springer New York.
15. DANIELY, A. & SHACHAM, H. (2020) Most relu networks suffer from ℓ^2 -adversarial perturbations. *Adv. Neural Inf. Process. Syst.*, **33**, 6629–6636.
16. DIRKSEN, S., GENZEL, M., JACQUES, L. & STOLLENWERK, A. (2022) The separation capacity of random neural networks. *J. Mach. Learn. Res.*, **23**, 13924–13970.
17. DONG, Y., DENG, Z., PANG, T., ZHU, J. & HANG, S. (2020) Adversarial distributional training for robust deep learning. *Adv. Neural Inf. Process. Syst.*, **33**, 8270–8283.
18. DUDLEY, R. M. (2002) *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.
19. EBIHARA, Y., DAI, X., MAGRON, V., PEAUCELLE, D. & TARBOURIECH, S. (2023) Local lipschitz constant computation of relu-fnns: upper bound computation with exactness verification. arXiv:2310.11104.
20. EVANS, L. C. & GARIEPY, R. F. (1992) *Measure Theory and Fine Properties of Functions*. *Studies in Advanced Mathematics*. Boca Raton, Florida: CRC Press.
21. FAZLYAB, M., ROBAY, A., HASSANI, H., MORARI, M. & PAPPAS, G. (2019) Efficient and accurate estimation of lipschitz constants for deep neural networks. *Adv. Neural Inf. Process. Syst.*, **32**.

22. FREI, S., VARDI, G., BARTLETT, P. L. & SREBRO, N. (2023) The double-edged sword of implicit bias: generalization vs. robustness in relu networks. arXiv:2303.01456.
23. HANIN, B. & NICA, M. (2020) Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.*, **376**, 287–322.
24. HANIN, B. & ROLNICK, D. (2019) Deep relu networks have surprisingly few activation patterns. *Adv. Neural Inf. Process. Syst.*, **32**.
25. HE, K., ZHANG, X., REN, S. & SUN, J. (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*. Santiago: IEEE, pp. 1026–1034.
26. KURAKIN, A., GOODFELLOW, I. J. & BENGIO, S. (2018) Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, 1st edn. Boca Raton, FL: CRC Press/Taylor & Francis Group, pp. 99–112. Chapman and Hall/CRC, 2018.
27. LIAW, C., MEHRABIAN, A., PLAN, Y. & VERSHYNIN, R. (2017) *A Simple Tool for Bounding the Deviation of Random Matrices on Geometric Sets*. Cham: Springer International Publishing, pp. 277–299.
28. MADRY, A., MAKELOV, A., SCHMIDT, L., TSIPRAS, D. & VLADU, A. (2018) Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*. Vancouver.
29. MONTANARI, A. & YUCHEN, W. (2023) Adversarial examples in random neural networks with general activations. *Math. Stat. Learn.*, **6**, 143–200.
30. MONTUFAR, G. (2017) *Notes on the Number of Linear Regions of Deep Neural Networks*. SampTA. Tallinn.
31. MONTUFAR, G. F., PASCANU, R., CHO, K. & BENGIO, Y. (2014) On the number of linear regions of deep neural networks. *Adv. Neural Inf. Process. Syst.*, **27**.
32. MOOSAVI-DEZFOOLI, S.-M., FAWZI, A. & FROSSARD, P. (2016) Deepfool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, pp. 2574–2582.
33. NGUYEN, Q., MONDELLI, M. & MONTUFAR, G. F. (2021) Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*. (M. MARINA & Z. TONG eds). Vienna: PMLR, pp. 8119–8129.
34. PANG, T., KUN, X., CHAO, D., CHEN, N. & ZHU, J. (2019) Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*. (K. CHAUDHURI & R. SALAKHUTDINOV eds). Long Beach: PMLR, pp. 4970–4979.
35. PAPERNOT, N., MCDANIEL, P., JHA, S., MATT FREDRIKSON, Z., CELIK, B. & SWAMI, A. (2016) The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. Saarbruecken: IEEE, pp. 372–387.
36. SCHOTT, L., RAUBER, J., BETHGE, M. & BRENDL, W. (2018) Towards the first adversarially robust neural network model on mnist. In *International Conference on Learning Representations*. Vancouver.
37. SHALEV-SHWARTZ, S. & BEN-DAVID, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.
38. SHAMIR, A., SAFRAN, I., RONEN, E. & DUNKELMAN, O. (2019) A simple explanation for the existence of adversarial examples with small hamming distance. arXiv:1901.10861.
39. SHI, Z., WANG, Y., ZHANG, H., KOLTER, J. Z. & HSIEH, C.-J. (2022) Efficiently computing local Lipschitz constants of neural networks via bound propagation. *Adv. Neural Inf. Process. Syst.*, **35**, 2350–2364.
40. SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I. & FERGUS, R. (2014) Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*. Banff: ICLR 2014.
41. VAN HANDEL (2014) *Probability in High Dimension*. Princeton: Lecture Notes (Princeton University).
42. VARDI, G., YEHUDAI, G. & SHAMIR, O. (2022) Gradient methods provably converge to non-robust networks. *Adv. Neural Inf. Process. Syst.*, **35**, 20921–20932.
43. VERSHYNIN, R. (2018) *High-Dimensional Probability: An Introduction With Applications in Data Science. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

44. WANG, Y., ULLAH, E., MIANY, P. & ARORA, R. (2022) Adversarial robustness is at odds with lazy training. *Adv. Neural Inf. Process. Syst.*, **35**, 6505–6516.
45. ZASLAVSKY, T. (1975) *Facing Up to Arrangements: Face-Count Formulas for Partitions of Space by Hyperplanes*. Mem. Am. Math. Soc, vol. **1**. Providence, RI: American Mathematical Society (AMS).

A. Postponed proofs for the preliminary results

In this appendix, we provide the postponed proofs for the preliminary results (see Section 4). Specifically, we prove Propositions 4.2 to 4.4, 4.6, 4.7 and 4.9.

We start with the proof of Proposition 4.2. This proposition is similar to other statements in the literature and is probably folklore. However, for the sake of completeness, we include a proof, as we could not locate a convenient reference.

Proof of Proposition 4.2. By [20, Section 4.2.3] it holds $f \in W_{\text{loc}}^{1,\infty}(\mathbb{R}^d)$. Here, $f \in W_{\text{loc}}^{1,\infty}(\mathbb{R}^d)$ denotes the set of functions that are locally (on every open bounded set $U \subseteq \mathbb{R}^d$) in the Sobolev space $W^{1,\infty}$. Then, [20, Theorem 1 in Section 6.2] implies that there exists a nullset $N \subseteq \mathbb{R}^d$ such that for every $x \in \mathbb{R}^d \setminus N$, f is differentiable at x , and the partial derivatives agree with the weak partial derivatives on $\mathbb{R}^d \setminus N$. For $i \in \{1, \dots, d\}$ let $\tilde{\partial}_i f$ be an explicit weak i -th partial derivative of f satisfying

$$\tilde{\partial}_i f = \partial_i f \quad \text{on } \mathbb{R}^d \setminus N$$

where ∂_i denotes the usual i -th partial derivative. We further write

$$\tilde{\nabla} f := (\tilde{\partial}_1 f, \dots, \tilde{\partial}_d f)^T.$$

Now, let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be a smooth function with compact support satisfying $\varphi \geq 0$ and furthermore $\int_{\mathbb{R}^d} \varphi(x) \, dx = 1$. See, e.g., [20, Section 4.2.1] for an explicit example of such a function. For $\varepsilon > 0$ let $\varphi_\varepsilon(x) := \varepsilon^{-d} \varphi(x/\varepsilon)$. Since f is continuous, the convolution $f_\varepsilon := f * \varphi_\varepsilon$ converges pointwise (even locally uniformly) to f as $\varepsilon \rightarrow 0$; see [20, Theorem 1 in Section 4.2]. The same theorem also shows that $f_\varepsilon \in C^\infty(\mathbb{R}^d)$ with

$$\partial_i f_\varepsilon = (\tilde{\partial}_i f) * \varphi_\varepsilon.$$

Hence, since $M \subseteq \mathbb{R}^d$ is a set of full measure, and N is a null-set,

$$\begin{aligned} \|\nabla f_\varepsilon(x)\|_2 &= \|(\tilde{\nabla} f * \varphi_\varepsilon)(x)\|_2 = \left\| \int_{\mathbb{R}^d} \tilde{\nabla} f(y) \varphi_\varepsilon(x-y) \, dy \right\|_2 = \left\| \int_{M \setminus N} \tilde{\nabla} f(y) \varphi_\varepsilon(x-y) \, dy \right\|_2 \\ &\leq \int_{M \setminus N} \|\nabla f(y)\|_2 |\varphi_\varepsilon(x-y)| \, dy \leq \sup_{y \in M} \|\nabla f(y)\|_2 =: L \end{aligned}$$

for every $x \in \mathbb{R}^d$. This implies

$$|f_\varepsilon(x) - f_\varepsilon(y)| \leq L \cdot \|x - y\|_2$$

for all $x, y \in \mathbb{R}^d$ and $\varepsilon > 0$. But then it also follows

$$|f(x) - f(y)| = \lim_{\varepsilon \rightarrow 0} |f_\varepsilon(x) - f_\varepsilon(y)| \leq L \cdot \|x - y\|_2$$

and thus $\text{Lip}(f) \leq L$.

To prove the inequality in the other direction, let $x \in M$. Assume without loss of generality that $\nabla f(x) \neq 0$ and let $v := \frac{\nabla f(x)}{\|\nabla f(x)\|_2}$. Then it holds

$$\begin{aligned}\|\nabla f(x)\|_2 &= \langle \nabla f(x), v \rangle = \lim_{t \rightarrow 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t \rightarrow 0} \frac{|f(x + tv) - f(x)|}{|t|} \\ &= \lim_{t \rightarrow 0} \frac{|f(x + tv) - f(x)|}{\|(x + tv) - x\|_2} \leq \text{Lip}(f),\end{aligned}$$

as was to be shown. Note that we applied the absolute value at the third equality, since the limit is positive. $\square$

We now move to the proofs of Propositions 4.3, 4.4 and 4.6. To this end, we first formulate a well-known lemma, which is needed for the proofs of these propositions. For the sake of completeness, we decided to include a proof of this lemma in this appendix.

LEMMA A1. Let $V \subseteq \mathbb{R}^k$ be a linear subspace and let $\ell = \dim(V) > 0$. Then there exists a matrix $U \in \mathbb{R}^{k \times \ell}$ with $U^T U = I_{\ell \times \ell}$, $\text{Im}(U) = V$ and $\langle Ux_1, Ux_2 \rangle = \langle x_1, x_2 \rangle$ for each $x_1, x_2 \in \mathbb{R}^\ell$ as well as $\langle U^T y_1, U^T y_2 \rangle = \langle y_1, y_2 \rangle$ for every $y_1, y_2 \in V$. Here, $I_{\ell \times \ell}$ denotes the ℓ -dimensional identity matrix.

Proof. We pick an orthonormal basis $u_1, \dots, u_\ell$ of V , arrange it columnwise in a matrix and denote the resulting matrix by $U = (u_1 | \dots | u_\ell)$. From the definition of an orthonormal basis, it follows $U^T U = I_{\ell \times \ell}$ and $\text{Im}(U) = V$. Moreover, for $x_1, x_2 \in \mathbb{R}^\ell$ we get by definition

$$\langle Ux_1, Ux_2 \rangle = \langle U^T Ux_1, x_2 \rangle \stackrel{U^T U = I_{\ell \times \ell}}{=} \langle x_1, x_2 \rangle.$$

For $y_1, y_2 \in V$ we can pick $a_1, a_2 \in \mathbb{R}^\ell$ with $Ua_1 = y_1$ and $Ua_2 = y_2$. This then gives us

$$\langle U^T y_1, U^T y_2 \rangle = \langle U^T Ua_1, U^T Ua_2 \rangle = \langle a_1, a_2 \rangle = \langle Ua_1, Ua_2 \rangle = \langle y_1, y_2 \rangle,$$

as was to be shown. $\square$

Note that Lemma A1 in particular implies that both U and U^T preserve the Euclidean norm, i.e., it holds $\|Ux\|_2 = \|x\|_2$ for every $x \in \mathbb{R}^\ell$ and $\|U^T y\|_2 = \|y\|_2$ for every $y \in V$ (but not in general for $y \in \mathbb{R}^k$).

We can now show Propositions 4.3, 4.4 and 4.6. The proof of Proposition 4.3 relies on the well-known fact that the ε -covering-number of the Euclidean unit ball in $\mathbb{R}^k$ can be bounded from above by $(1 + \frac{2}{\varepsilon})^k$, combined with Lemma A1.

Proof of Proposition 4.3. Let $\ell := \dim(V) > 0$. If $\ell = 0$, the claim is trivial since in this case we have $\bar{B}_k(0, 1) \cap V = \{0\}$. Hence, we can assume $\ell > 0$. Choose $U \in \mathbb{R}^{k \times \ell}$ according to Lemma A1. Pick $x_1, \dots, x_M \in \bar{B}_\ell(0, 1)$ with $M \leq \left(\frac{2}{\varepsilon} + 1\right)^\ell$ and

$$\bar{B}_\ell(0, 1) \subseteq \bigcup_{i=1}^M \bar{B}_\ell(x_i, \varepsilon),$$

which is possible due to [43, Corollary 4.2.13]. We then define $y_i := Ux_i \in V$ for $i = 1, \dots, M$. Then it holds $\|y_i\|_2 = \|x_i\|_2 \leq 1$ and hence, $y_i \in V \cap \bar{B}_k(0, 1)$ for every $i = 1, \dots, M$. Let $y \in V \cap \bar{B}_k(0, 1)$ be arbitrary and note that $U^T y \in \bar{B}_\ell(0, 1)$. Hence, there exists $i \in \{1, \dots, M\}$ with $\|x_i - U^T y\|_2 \leq \varepsilon$. But then

we get

$$\|y_i - y\|_2 = \|Ux_i - y\|_2 = \|U^T(Ux_i - y)\|_2 = \|x_i - U^T y\|_2 \leq \varepsilon,$$

where the second equality follows since U^T is norm-preserving on V and $Ux_i - y \in V$. Since y was arbitrary, we see

$$\bar{B}_k(0, 1) \cap V \subseteq \bigcup_{i=1}^M \bar{B}_k(y_i, \varepsilon) \cap V. \quad (\text{A.1})$$

This yields the claim. $\square$

For the proof of Proposition 4.4, we use the well-known fact that the VC-dimension of the set of homogeneous half spaces in $\mathbb{R}^k$ equals k , combined with Lemma A1.

Proof of Proposition 4.4. Let $\ell := \dim(V)$. In case of $\ell = 0$, we have $\mathcal{F} = \{0\}$ and hence $\text{VC}(\mathcal{F}) = 0$, so that the claim is trivial. Hence, we can assume $\ell > 0$. Pick $U \in \mathbb{R}^{k \times \ell}$ according to Lemma A1. For $\alpha \in V$ we define

$$\tilde{f}_\alpha : \mathbb{R}^\ell \rightarrow \mathbb{R}, \quad x \mapsto \mathbb{1}_{(U^T \alpha)^T x > 0} \quad \text{and} \quad \tilde{\mathcal{F}} := \{\tilde{f}_\alpha : \alpha \in V\}.$$

Since U^T induces a surjective mapping from V to $\mathbb{R}^\ell$ (because U^T is isometric on V and $\dim(V) = \ell$) we infer

$$\tilde{\mathcal{F}} = \{f_\alpha : \alpha \in \mathbb{R}^\ell\}.$$

From [37, Theorem 9.2] we infer $\text{VC}(\tilde{\mathcal{F}}) = \ell$. Therefore, it suffices to show that

$$\text{VC}(\mathcal{F}) = \text{VC}(\tilde{\mathcal{F}}).$$

To this end, let $t \leq \text{VC}(\mathcal{F})$ and let $\{v_1, \dots, v_t\} \subseteq V$ be a subset of V with t elements that is shattered by $\mathcal{F}$. Let $\eta \in \{0, 1\}^t$ be arbitrary and $\alpha \in V$ such that

$$(f_\alpha(v_1), \dots, f_\alpha(v_t)) = \eta.$$

We then consider $\{U^T v_1, \dots, U^T v_t\} \subseteq \mathbb{R}^\ell$ and note that this set also consists of t elements since U^T induces an injective mapping $V \rightarrow \mathbb{R}^\ell$. Fix $j \in \{1, \dots, t\}$. Then it holds

$$\tilde{f}_\alpha(U^T v_j) = \mathbb{1}_{\langle U^T \alpha, U^T v_j \rangle > 0} = \mathbb{1}_{\langle \alpha, v_j \rangle > 0} = f_\alpha(v_j) = \eta_j.$$

But, since j was picked arbitrarily, this shows

$$(\tilde{f}_\alpha(U^T v_1), \dots, \tilde{f}_\alpha(U^T v_t)) = \eta$$

and thus, the set $\{U^T v_1, \dots, U^T v_t\}$ is shattered by $\tilde{\mathcal{F}}$. Therefore, we conclude $\text{VC}(\mathcal{F}) \leq \text{VC}(\tilde{\mathcal{F}})$. Similarly, we can derive the bound in the other direction: Let $t \leq \text{VC}(\tilde{\mathcal{F}})$ and $\{w_1, \dots, w_t\} \subseteq \mathbb{R}^\ell$ a subset of $\mathbb{R}^\ell$ with t elements that is shattered by $\tilde{\mathcal{F}}$. Let $\eta \in \{0, 1\}^t$ be arbitrary and $\alpha \in V$ such that

$$(\tilde{f}_\alpha(w_1), \dots, \tilde{f}_\alpha(w_t)) = \eta.$$

We then consider $\{Uw_1, \dots, Uw_t\} \subseteq V$ and note that this set also consists of t elements since U induces an injective mapping $\mathbb{R}^\ell \rightarrow V$. Moreover, for every $j \in \{1, \dots, t\}$ we have

$$f_\alpha(Uw_j) = \mathbb{1}_{\langle \alpha, Uw_j \rangle > 0} = \mathbb{1}_{\langle U^T \alpha, U^T Uw_j \rangle > 0} = \tilde{f}_\alpha(w_j) = \eta_j,$$

which gives us

$$(f_\alpha(U_{w_1}), \dots, f_\alpha(U_{w_t})) = \eta$$

and thus, the set $\{U_{w_1}, \dots, U_{w_t}\}$ is shattered by $\mathcal{F}$. This yields $\text{VC}(\mathcal{F}) \geq \text{VC}(\tilde{\mathcal{F}})$ and overall, we get the claim. $\square$

The proof of Proposition 4.6 makes heavy use of the fact that the Gaussian width of the Euclidean unit ball in $\mathbb{R}^k$ can be bounded from above by $\sqrt{k}$. Again, we apply Lemma A1 to transfer the problem to some lower-dimensional space $\mathbb{R}^\ell$ and use the rotation invariance of the Gaussian distribution.

Proof of Proposition 4.6. Let $\ell := \dim(V)$. If $\ell = 0$, then $\bar{B}_k(0, 1) \cap V = \{0\}$, and we trivially have $w(\{0\}) = 0$. Hence, we can assume $\ell > 0$. Choose $U \in \mathbb{R}^{k \times \ell}$ according to Lemma A1. From the fact that U and U^T preserve the Euclidean norm (on $\mathbb{R}^\ell$ and V , respectively), we get

$$\bar{B}_k(0, 1) \cap V = U\bar{B}_\ell(0, 1).$$

This yields

$$\begin{aligned} w(\bar{B}_k(0, 1) \cap V) &= \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in \bar{B}_k(0, 1) \cap V} \langle g, x \rangle \right] = \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{v \in U\bar{B}_\ell(0, 1)} \langle g, v \rangle \right] \\ &= \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in \bar{B}_\ell(0, 1)} \langle g, Ux \rangle \right] = \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in \bar{B}_\ell(0, 1)} \langle U^T g, x \rangle \right]. \end{aligned}$$

By basis completion, pick a matrix $U' \in \mathbb{R}^{k \times (k-\ell)}$ such that

$$U_0 := (U \mid U') \in \mathbb{R}^{k \times k}$$

is orthogonal. Then $(U_0)^T$ is also orthogonal and from the rotation invariance of $\mathcal{N}(0, I_k)$ (see, e.g., [43, Proposition 3.3.2]) we see that, if we take $g \sim \mathcal{N}(0, I_k)$, it holds

$$(U_0)^T g = \left(\frac{U^T g}{(U')^T g} \right) \sim \mathcal{N}(0, I_k),$$

which implies $U^T g \sim \mathcal{N}(0, I_\ell)$. Therefore, we get

$$w(\bar{B}_k(0, 1) \cap V) = \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in \bar{B}_\ell(0, 1)} \langle U^T g, x \rangle \right] = \mathbb{E}_{g \sim \mathcal{N}(0, I_\ell)} \left[\sup_{x \in \bar{B}_\ell(0, 1)} \langle g, x \rangle \right] = w(\bar{B}_\ell(0, 1)).$$

The claim now follows using [43, Proposition 7.5.2], noting that $\text{diam}(\bar{B}_\ell(0, 1)) = 2$. $\square$

We now move the proofs of Propositions 4.7 and 4.9. The idea of the proof of Proposition 4.7 is to find one specific choice of weight matrices for which the inequality

$$\text{Lip}(\Phi) = 0 < \text{Lip}(\tilde{\Phi})$$

holds for arbitrary biases. Then, using a continuity argument, we infer that the desired inequality even holds with positive probability.

Proof of Proposition 4.7. Let $b^{(0)} \in \mathbb{R}^3$ and $b^{(1)} \in \mathbb{R}$ be arbitrary but from now on fixed bias vectors.

We consider the shallow network Φ_0 given via the biases $b^{(0)}$ and $b^{(1)}$ and the weights $W_*^{(0)} := \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $W_*^{(1)} := (1 \ -1 \ -1)$. This implies

$$\Phi_0(x) = \text{ReLU}(x_1 + x_2 + b_1^{(0)}) - \text{ReLU}(x_1 + b_2^{(0)}) - \text{ReLU}(x_2 + b_3^{(0)}) + b^{(1)}$$

for all $x \in \mathbb{R}^2$. Moreover, $\widetilde{\Phi}_0(x) = b_1^{(0)} - b_2^{(0)} - b_3^{(0)} + b^{(1)}$ for all $x \in \mathbb{R}^2$ and hence, $\text{Lip}(\widetilde{\Phi}_0) = 0$. On the other hand, we have

$$\Phi_0(x) = x_1 + x_2 + b_1^{(0)} - x_2 - b_3^{(0)} + b^{(1)} = x_1 + b_1^{(0)} - b_3^{(0)} + b^{(1)},$$

whenever $x_1 \leq -b_2^{(0)}$, and $x_2 \geq \max\{-b_3^{(0)}, -x_1 - b_1^{(0)}\}$, whence $\text{Lip}(\Phi_0) \geq 1$. Therefore, it holds

$$\text{Lip}(\Phi_0) > C \cdot \text{Lip}(\widetilde{\Phi}_0).$$

It remains to show that this even holds with positive probability. To see this, we still let $b^{(0)}$ and $b^{(1)}$ be fixed and note that $\text{Lip}(\widetilde{\Phi}) = \|W^{(1)}W^{(0)}\|_2$ depends continuously on the network weights $W^{(0)}$ and $W^{(1)}$. Moreover, we fix the point $x_2 := \max\{-b_3^{(0)}, -b_1^{(0)} + b_2^{(0)} + 1\}$ and note that the expression $|\Phi((-b_2^{(0)}, x_2)) - \Phi(-b_2^{(0)} - 1, x_2)|$ also depends continuously on the network weights. Therefore, for $(W^{(0)}, W^{(1)})$ on a sufficiently small open neighborhood of $(W_*^{(0)}, W_*^{(1)})$, we have

$$\frac{\text{Lip}(\Phi)}{C} \geq \frac{|\Phi((-b_2^{(0)}, x_2)) - \Phi(-b_2^{(0)} - 1, x_2)|}{C} > 1/(2C) > \text{Lip}(\widetilde{\Phi}).$$

Hence, for fixed biases $b^{(0)}$ and $b^{(1)}$ we note

$$\mathbb{P}^{(W^{(0)}, W^{(1)})}(\text{Lip}(\Phi) > C \cdot \text{Lip}(\widetilde{\Phi})) > 0.$$

In total, allowing randomness in $b^{(0)}$ and $b^{(1)}$ as well gives us

$$\mathbb{P}(\text{Lip}(\Phi) > C \cdot \text{Lip}(\widetilde{\Phi})) = \mathbb{E}_{b^{(0)}, b^{(1)}} \left[\underbrace{\mathbb{P}^{(W^{(0)}, W^{(1)})}(\text{Lip}(\Phi) > C \cdot \text{Lip}(\widetilde{\Phi}))}_{>0} \right] > 0. \quad \square$$

The proof of Proposition 4.9 is based on identity

$$\text{ReLU}(-\text{ReLU}(x)) = 0 \quad \text{for all } x \in \mathbb{R},$$

which follows directly from the definition of the ReLU.

Proof of Proposition 4.9. Whenever $b^{(1)} \leq 0$, $W^{(1)} < 0$ and $W^{(0)}, W^{(2)} \neq 0$ (which happens with positive probability), we get

$$\underbrace{\text{ReLU}(W^{(1)} \cdot \underbrace{\text{ReLU}(W^{(0)}x + b^{(0)})}_{\geq 0})}_{\leq 0} + b^{(1)} = 0 \quad \text{for all } x \in \mathbb{R},$$

which implies $\Phi \equiv 0$ and in particular $\text{Lip}(\Phi) = 0$ in that case. On the other hand, we get

$$\text{Lip}(\tilde{\Phi}) = |W^{(2)} \cdot W^{(1)} \cdot W^{(0)}| > 0.$$

□

B. A note on measurability

In this appendix we elaborate on the measurability of the function

$$(W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)}) \mapsto \sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2. \quad (\text{B.1})$$

Note that we computed (upper bounds for) the expectation of this function with respect to the weights $W^{(0)}, \dots, W^{(L)}$ and the biases $b^{(0)}, \dots, b^{(L)}$ in Section 5 in order to establish upper bounds for the Lipschitz constant of ReLU networks. In order for the expectation to make sense, the function needs to be measurable, which we verify (to a sufficient extent) in this appendix.

As a preparation, we recall from [14, Section 1.5] the notion of the *completion* of a measure. Given a measure μ on a measurable space $(\Omega, \mathcal{A})$, we define

$$\mathcal{A}_\mu := \{A \subseteq \Omega : \exists E, F \in \mathcal{A} : E \subseteq A \subseteq F \text{ and } \mu(F \setminus E) = 0\}.$$

Then, $\mathcal{A}_\mu$ is a σ -algebra with $\mathcal{A} \subseteq \mathcal{A}_\mu$, and we can extend μ to a well-defined measure $\bar{\mu}$ on $\mathcal{A}_\mu$ by setting

$$\bar{\mu}(A) := \mu(E) \quad \text{for } E, F \in \mathcal{A} \text{ with } E \subseteq A \subseteq F \text{ and } \mu(F \setminus E) = 0.$$

Moreover, the measure space $(X, \mathcal{A}_\mu, \bar{\mu})$ is then *complete*, meaning that $A \in \mathcal{A}_\mu$, $\bar{\mu}(A) = 0$ and $B \subseteq A$ together imply $B \in \mathcal{A}_\mu$ for all subsets $A, B \subseteq \Omega$.

Next, given a measurable space $(\Omega, \mathcal{A})$, we say that $A \subseteq \Omega$ is *universally measurable* if $A \in \mathcal{A}_\mu$ for every finite measure $\mu : \mathcal{A} \rightarrow [0, \infty]$. It is easy to see that the universally measurable sets over $(\Omega, \mathcal{A})$ form a σ -algebra, and that every $A \in \mathcal{A}$ is universally measurable.

We show that the map considered in (B1) is universally measurable, i.e., that the preimage of every Borel set is universally measurable.

PROPOSITION B1. Let $\vec{W} := (W^{(0)}, \dots, W^{(L)}) \in \mathbb{R}^{N \times d} \times \mathbb{R}^{N \times N} \times \dots \times \mathbb{R}^{N \times N} \times \mathbb{R}^{1 \times N} \cong \mathbb{R}^{D_1}$ and further $\vec{b} := (b^{(0)}, \dots, b^{(L)}) \in \mathbb{R}^N \times \dots \times \mathbb{R}^N \times \mathbb{R} \cong \mathbb{R}^{D_2}$. Then the map

$$\xi : \mathbb{R}^{D_1} \times \mathbb{R}^{D_2} \rightarrow \mathbb{R}, \quad (\vec{W}, \vec{b}) \mapsto \sup_{x \in \mathbb{R}^d} \|W^{(L)} D^{(L-1)}(x) W^{(L-1)} \dots D^{(0)}(x) W^{(0)}\|_2$$

is universally measurable. That is, $\xi^{-1}(U)$ is universally measurable over $(\mathbb{R}^{D_1} \times \mathbb{R}^{D_2}, \mathcal{B})$ (with $\mathcal{B}$ denoting the Borel σ -algebra over $\mathbb{R}^{D_1} \times \mathbb{R}^{D_2}$) for any Borel measurable set $U \subseteq \mathbb{R}$. Here, the matrices $D^{(L-1)}(x), \dots, D^{(0)}(x)$ are as defined in Section 4.1.

Proof. According to [1, Corollary 4.25] it suffices to show that $\xi^{-1}((a, \infty))$ is universally measurable for every $a \in \mathbb{R}$ in order to show that ξ is universally measurable.

Note that

$$\begin{aligned}\xi^{-1}((a, \infty)) &= \left\{ (\vec{W}, \vec{b}) \in \mathbb{R}^{D_1+D_2} : \exists x \in \mathbb{R}^d \text{ with } \|W^{(L)}D^{(L-1)}(x) \cdots D^{(0)}(x)W^{(0)}\|_2 > a \right\} \\ &= \pi \left(\left\{ (\vec{W}, \vec{b}, x) \in \mathbb{R}^{D_1} \times \mathbb{R}^{D_2} \times \mathbb{R}^d : \|W^{(L)}D^{(L-1)}(x) \cdots D^{(0)}(x)W^{(0)}\|_2 > a \right\} \right).\end{aligned}$$

Here,

$$\pi : \mathbb{R}^{D_1+D_2+d} \rightarrow \mathbb{R}^{D_1+D_2}, \quad (\vec{W}, \vec{b}, x) \mapsto (\vec{W}, \vec{b})$$

denotes the projection from $\mathbb{R}^{D_1+D_2+d}$ onto $\mathbb{R}^{D_1+D_2}$. It is easy to see that the map

$$\mathbb{R}^{D_1+D_2+d} \rightarrow \mathbb{R}, \quad (\vec{W}, \vec{b}, x) \mapsto \|W^{(L)}D^{(L-1)}(x)W^{(L-1)} \cdots D^{(0)}(x)W^{(0)}\|_2$$

is Borel measurable, whence the set

$$\left\{ (\vec{W}, \vec{b}, x) \in \mathbb{R}^{D_1} \times \mathbb{R}^{D_2} \times \mathbb{R}^d : \|W^{(L)}D^{(L-1)}(x) \cdots D^{(0)}(x)W^{(0)}\|_2 > a \right\}$$

is Borel measurable. Since π is trivially Borel measurable, it follows from [1, Theorem 12.24] that $\xi^{-1}((a, \infty))$ is an analytic set and hence, $\xi^{-1}((a, \infty))$ is universally measurable, as follows from [1, Theorem 12.41]. $\square$

It remains unclear, whether ξ is not merely universally measurable, but in fact Borel measurable.

In order for the expectation in Section 5 to be well-defined we have to require that the map

$$\mathcal{W} : \Omega \rightarrow \mathbb{R}^{D_1+D_2}, \quad \omega \mapsto (\vec{W}(\omega), \vec{b}(\omega)),$$

is measurable with respect to $\mathcal{A}$ and the σ -algebra of universally measurable sets on $\mathbb{R}^{D_1+D_2}$. Here, $(\Omega, \mathcal{A}, \mathbb{P})$ denotes the underlying ‘master probability space’. This can be achieved by defining the probability space to be $\Omega := \mathbb{R}^{D_1+D_2}$ together with the σ -algebra of universally measurable subsets, $\mathcal{W} = \text{id}$ and $\mathbb{P}$ as the (completion of the) probability measure representing the distribution of $(\vec{W}, \vec{b})$ according to Assumption 2.1.

C. A proposition for deriving high-probability results

This appendix contains a folklore proposition, which is needed for proving the high-probability results derived in the present paper. For the sake of completeness, we decided to include a proof of this proposition in the present paper. It states the following:

PROPOSITION C1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ two measurable spaces and $X : \Omega \rightarrow \Omega_1, Y : \Omega \rightarrow \Omega_2$ two stochastically independent random variables. Moreover, let the set $A \subseteq \Omega_1 \times \Omega_2$ be measurable (with respect to the product-sigma-algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$), for every $x \in \Omega_1$ let some $A_2(x) \in \mathcal{A}_2$ be given and let $A_1 \in \mathcal{A}_1$. Finally, assume that

$$\forall (x, y) \in \Omega_1 \times \Omega_2 : x \in A_1, y \in A_2(x) \implies (x, y) \in A. \quad (\text{C.1})$$

Then it holds

$$\mathbb{P}((X, Y) \in A) \geq \mathbb{P}^X(A_1) \cdot \inf_{x \in A_1} \mathbb{P}^Y(A_2(x)).$$

Proof. The condition (C.1) can be translated to

$$\mathbb{1}_A(x, y) \geq \mathbb{1}_{A_1}(x) \cdot \mathbb{1}_{A_2(x)}(y)$$

for every $(x, y) \in \Omega_1 \times \Omega_2$. Note that due to the stochastic independence of X and Y it holds that $\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes \mathbb{P}^Y$. Applying Tonelli's theorem, we get

$$\mathbb{P}((X, Y) \in A) = \int_{\Omega_1 \times \Omega_2} \mathbb{1}_A(x, y) d\mathbb{P}^{(X,Y)}(x, y) = \int_{\Omega_1} \int_{\Omega_2} \mathbb{1}_A(x, y) d\mathbb{P}^Y(y) d\mathbb{P}^X(x). \quad (\text{C.2})$$

For fixed $x \in \Omega_1$ we get

$$\begin{aligned} \int_{\Omega_2} \mathbb{1}_A(x, y) d\mathbb{P}^Y(y) &\geq \mathbb{1}_{A_1}(x) \int_{\Omega_2} \mathbb{1}_{A_2(x)}(y) d\mathbb{P}^Y(y) = \begin{cases} \mathbb{P}^Y(A_2(x)), & x \in A_1, \\ 0, & x \notin A_1, \end{cases} \\ &\geq \mathbb{1}_{A_1}(x) \cdot \inf_{x \in A_1} \mathbb{P}^Y(A_2(x)). \end{aligned}$$

Inserting this into (C.2) yields

$$\mathbb{P}((X, Y) \in A) \geq \int_{\Omega_1} \mathbb{1}_{A_1}(x) \cdot \inf_{x \in A_1} \mathbb{P}^Y(A_2(x)) d\mathbb{P}^X(x) = \mathbb{P}^X(A_1) \cdot \inf_{x \in A_1} \mathbb{P}^Y(A_2(x)),$$

as was to be shown. $\square$

D. On Dudley's inequality

In this section we add important notes on Dudley's inequality, which is crucial for deriving the upper bounds in Section 5.

We first define the notion of separability of a random process.

DEFINITION D1 (cf. [41, Definition 5.22]). A random process $(X_t)_{t \in T}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ indexed by a metric space (T, ϱ) is called *separable* if there exists a countable set $T_0 \subseteq T$ and a $\mathbb{P}$ -null set $N \subseteq \Omega$ such that for all $\omega \in \Omega \setminus N$ and $t \in T$ there exists a sequence $(s_n)_{n \in \mathbb{N}} \in T_0^{\mathbb{N}}$ with $s_n \rightarrow t$ and

$$\lim_{n \rightarrow \infty} X_{s_n}(\omega) = X_t(\omega).$$

PROPOSITION D2. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, (T, ϱ) a separable metric space and $(X_t)_{t \in T}$ a random process with $X_t : \Omega \rightarrow \mathbb{R}$ for every $t \in T$. Moreover, we assume that for every fixed $\omega \in \Omega$ the map

$$T \rightarrow \mathbb{R}, \quad t \mapsto X_t(\omega)$$

is continuous. Then $(X_t)_{t \in T}$ is a separable random process.

Proof. Let $T_0 \subseteq T$ be a countable dense subset of T and $\omega \in \Omega$, $t \in T$. Since T_0 is dense there is a sequence $(s_n)_{n \in \mathbb{N}} \in T_0^{\mathbb{N}}$ with $s_n \rightarrow t$. Then, using the continuity assumption, it holds

$$\lim_{n \rightarrow \infty} X_{s_n}(\omega) = X_t(\omega),$$

which yields the claim. $\square$

We can now formulate and prove a slight extension of Dudley's inequality from [41].

THEOREM D3 (cf. [41, Theorems 5.25 and 5.29]). There exists a constant $C > 0$ with the following property: If $(X_t)_{t \in T}$ is a mean-zero separable random process indexed by a metric space (T, ϱ) satisfying

$$\|X_t - X_s\|_{\psi_2} \leq K \varrho(t, s)$$

for every $t, s \in T$ with a constant $K > 0$ that does not depend on s and t , the following hold:

1. For all $t_0 \in T$ and $u \geq 0$ it holds

$$\sup_{t \in T} (X_t - X_{t_0}) \leq CK \left(\int_0^\infty \sqrt{\ln(\mathcal{M}(T, \varrho, \varepsilon))} \, d\varepsilon + u \operatorname{diam}(T) \right)$$

with probability at least $1 - 2 \exp(-u^2)$.

2. $\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq CK \int_0^\infty \sqrt{\ln(\mathcal{M}(T, \varrho, \varepsilon))} \, d\varepsilon$.

Proof. Firstly, note that the separability assumption implies that $\sup_{t \in T} X_t = \sup_{t \in T_0} X_t$ almost surely for some countable set $T_0 \subseteq T$ so that no measurability issues arise (at least if we are willing to pass to the completion of the underlying probability space).

Let $C_1 > 0$ be the absolute constant appearing in [43, Equation (2.16)] and define $Y_t := \frac{X_t}{K\sqrt{2C_1}}$ for $t \in T$. Then it holds

$$\|Y_t - Y_s\|_{\psi_2} = \frac{\|X_t - X_s\|_{\psi_2}}{K\sqrt{2C_1}} \leq \frac{\varrho(t, s)}{\sqrt{2C_1}}$$

for every $t, s \in T$ and hence, using [43, Equation (2.16)], we get

$$\mathbb{E} [\exp(\lambda(Y_t - Y_s))] \leq \exp(C_1 \lambda^2 \|Y_t - Y_s\|_{\psi_2}^2) \leq \exp(\lambda^2 \varrho(t, s)^2 / 2)$$

for all $\lambda \in \mathbb{R}$. Hence, $(Y_t)_{t \in T}$ is a *sub-gaussian* process (see [41, Definition 5.20]) and trivially inherits the property of separability from $(X_t)_{t \in T}$. Hence, [41, Theorems 5.25 and 5.29] yield the existence of a constant $C_2 > 0$ such that the following hold:

1. For all $u \geq 0$ and $t_0 \in T$ it holds

$$\sup_{t \in T} (Y_t - Y_{t_0}) \leq C_2 \left(\int_0^\infty \sqrt{\ln(\mathcal{M}(T, \varrho, \varepsilon))} \, d\varepsilon + u \operatorname{diam}(T) \right)$$

with probability at least $1 - C_2 \exp(-u^2)$.

2. $\mathbb{E} \left[\sup_{t \in T} Y_t \right] \leq C_2 \int_0^\infty \sqrt{\ln(\mathcal{M}(T, \varrho, \varepsilon))} \, d\varepsilon$.

Note that in (1) we can bound the probability even by $1 - 2 \exp(-u^2)$, which is due to the first inequality on p.137 in [41]. Specifically, we have

$$\sum_{k=1}^{\infty} e^{-k/2} \leq \sum_{k=1}^{\infty} \left(\frac{9}{4}\right)^{-k/2} = \sum_{k=1}^{\infty} \left(\frac{3}{2}\right)^{-k} = \frac{2/3}{1 - 2/3} = 2.$$

Using this fact and the exact definition of $(Y_t)_{t \in T}$ yields the claim by letting $C := \sqrt{2C_1} C_2$. $\square$

Now we apply Dudley's inequality to certain finite-dimensional Gaussian processes.

PROPOSITION D4 (special version of Dudley's inequality). There exists an absolute constant $C > 0$ with the following property: For any $T \subseteq \mathbb{R}^N$ with $0 \in T$ and any random vector $X \in \mathbb{R}^N$ with

$$X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for } 1 \leq i \leq N,$$

the following hold:

1. For any $u \geq 0$, we have

$$\sup_{t \in T} \langle X, t \rangle \leq C \left(\int_0^\infty \sqrt{\ln(\mathcal{N}(T, \|\cdot\|_2, \varepsilon))} \, d\varepsilon + u \cdot \text{diam}(T) \right)$$

with probability at least $1 - 2 \exp(-u^2)$.

2. $\mathbb{E} \left[\sup_{t \in T} \langle X, t \rangle \right] \leq C \int_0^\infty \sqrt{\ln(\mathcal{N}(T, \|\cdot\|_2, \varepsilon))} \, d\varepsilon.$

Proof. For $t, s \in T$ we have

$$\|\langle X, t \rangle - \langle X, s \rangle\|_{\psi_2} = \|\langle X, t - s \rangle\|_{\psi_2} \leq C_1 \|t - s\|_2$$

with an absolute constant $C_1 > 0$. Here we used that $\langle X, t - s \rangle \sim \mathcal{N}(0, \|t - s\|_2^2)$ (see for instance [43, Exercise 3.3.3]) and [43, Example 2.5.8 (i)]. Furthermore, it holds

$$\mathbb{E}[\langle X, t \rangle] = 0$$

for fixed $t \in T$ since $\langle X, t \rangle \sim \mathcal{N}(0, \|t\|_2^2)$. For a fixed realization of X it is trivial that $t \mapsto \langle X, t \rangle$ is continuous and hence, the process $(\langle X, t \rangle)_{t \in T}$ is separable by Proposition D2. Dudley's inequality and its high-probability version (cf. Theorem D3) then show that:

1. For any $u \geq 0$, we have

$$\sup_{t \in T} \langle X, t \rangle = \sup_{t \in T} (\langle X, t \rangle - \langle X, 0 \rangle) \stackrel{0 \in T}{\leq} C_1 C_2 \left(\int_0^\infty \sqrt{\ln(\mathcal{N}(T, \|\cdot\|_2, \varepsilon))} \, d\varepsilon + u \cdot \text{diam}(T) \right)$$

with probability at least $1 - 2 \exp(-u^2)$.

2. $\mathbb{E} \left[\sup_{t \in T} \langle X, t \rangle \right] \leq C_1 C_2 \int_0^\infty \sqrt{\ln(\mathcal{N}(T, \|\cdot\|_2, \varepsilon))} \, d\varepsilon$

with an absolute constant $C_2 > 0$. Hence, the claim follows letting $C := C_1 C_2$. $\square$

E. Differentiability of a random ReLU network at a fixed point

In this appendix, we prove that, if we fix a point $x_0 \in \mathbb{R}^d \setminus \{0\}$, a random ReLU network that follows Assumption 2.1 is differentiable at x_0 with

$$(\nabla \Phi(x_0))^T = W^{(L)} \cdot D^{(L-1)}(x_0) \cdot W^{(L-1)} \cdots D^{(0)}(x_0) \cdot W^{(0)}.$$

Since the vector x_0 is fixed, we often omit the argument and write $D^{(\ell)}$ instead of $D^{(\ell)}(x_0)$ for each $\ell \in \{0, \dots, L-1\}$. Let us now formulate and prove the statement.

THEOREM E1. Let $x_0 \in \mathbb{R}^d \setminus \{0\}$ be fixed and consider a random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ of width N and depth L , satisfying the random initialization as described in Assumption 2.1. Then Φ is almost

surely differentiable at x_0 with

$$(\nabla \Phi(x_0))^T = W^{(L)} \cdot D^{(L-1)} \cdot W^{(L-1)} \dots D^{(0)} \cdot W^{(0)}. \quad (\text{E.1})$$

Proof. We denote by $\mathcal{A}$ the event that Φ is differentiable at x_0 and (E.1) holds. Let $x^{(\ell)}$, $\ell \in \{0, \dots, L-1\}$ be defined as in (4.1). We denote

$$\begin{aligned} \eta &= \eta(W^{(0)}, \dots, W^{(L-1)}, b^{(0)}, \dots, b^{(L-1)}) \\ &:= \min \left\{ \ell \in \{0, \dots, L-1\} : \exists i \in \{1, \dots, N\} \text{ with } (W^{(\ell)} x^{(\ell)} + b^{(\ell)})_i = 0 \right\} \end{aligned}$$

and $\eta := L$ if the minimum above is not well-defined. Since the events $\{\eta = \ell\}$ for $\ell \in \{0, \dots, L\}$ are pairwise disjoint and their union is the entire probability space, we get

$$\mathbb{P}(\mathcal{A}) = \sum_{\ell=0}^L \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\}).$$

The goal is now to show that

$$\mathbb{P}(\mathcal{A} \cap \{\eta = \ell\}) = \mathbb{P}(\{\eta = \ell\}) \quad \text{for every } \ell \in \{0, \dots, L\}, \quad (\text{E.2})$$

which will directly give us $\mathbb{P}(\mathcal{A}) = 1$.

To this end, we start with the case $\ell = L$. Note that

$$\eta = L \iff (W^{(\ell)} x^{(\ell)} + b^{(\ell)})_i \neq 0 \quad \text{for all } \ell \in \{0, \dots, L-1\}, i \in \{1, \dots, N\}.$$

But since the ReLU is differentiable on $\mathbb{R} \setminus \{0\}$, this directly implies the differentiability of Φ at x_0 and that (E1) holds. Hence, it even holds

$$\mathcal{A} \cap \{\eta = L\} = \{\eta = L\}.$$

We proceed with the case $\ell \in \{1, \dots, L-1\}$. Note that it holds

$$\mathbb{P}(\mathcal{A} \cap \{\eta = \ell\}) = \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\} \cap \{x^{(\ell)} \neq 0\}) + \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\} \cap \{x^{(\ell)} = 0\}).$$

We get

$$\begin{aligned} \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\} \cap \{x^{(\ell)} \neq 0\}) &\leq \mathbb{P}(\{\eta = \ell\} \cap \{x^{(\ell)} \neq 0\}) \\ &= \mathbb{E}_{\substack{W^{(0)}, \dots, W^{(\ell-1)} \\ b^{(0)}, \dots, b^{(\ell-1)}}} \left[\mathbb{1}_{x^{(\ell)} \neq 0} \cdot \mathbb{E}_{W^{(\ell)}, b^{(\ell)}} \mathbb{1}_{\eta=\ell} \right] \\ &= 0. \end{aligned} \quad (\text{E.3})$$

Here, the last equality stems from the fact that, if $x^{(\ell)} \neq 0$, it follows that, conditioning on the weights $W^{(0)}, \dots, W^{(\ell-1)}$ and the biases $b^{(0)}, \dots, b^{(\ell-1)}$, it holds

$$(W^{(\ell)} x^{(\ell)} + b^{(\ell)})_i \sim \mathcal{N}\left(0, \frac{2}{N} \|x^{(\ell)}\|_2^2\right) * \mathcal{D}_i^{(\ell)} \quad \text{for each } i \in \{1, \dots, N\},$$

the fact that the distribution on the right-hand side is an absolutely continuous distributions (as follows from [18, Proposition 9.16]) and the fact that all the random variables $(W^{(\ell)} x^{(\ell)} + b^{(\ell)})_i$, $i \in \{1, \dots, N\}$, are jointly independent. Here, $*$ denotes the convolution of two (probability) measures.

We now show that it holds

$$\mathcal{A} \cap \{\eta = \ell\} \cap \{x^{(\ell)} = 0\} = \{\eta = \ell\} \cap \{x^{(\ell)} = 0\}. \quad (\text{E.4})$$

To this end, assume that $\eta = \ell$ and $x^{(\ell)} = 0$ hold. This implies

$$\left(W^{(\ell-1)}x^{(\ell-1)} + b^{(\ell-1)}\right)_i \neq 0 \quad \text{and} \quad \left(W^{(\ell-1)}x^{(\ell-1)} + b^{(\ell-1)}\right)_i \leq 0 \quad \text{for all } i \in \{1, \dots, N\},$$

or equivalently

$$\left(W^{(\ell-1)}x^{(\ell-1)} + b^{(\ell-1)}\right)_i < 0 \quad \text{for all } i \in \{1, \dots, N\}.$$

But by continuity, these inequalities also hold in an open neighborhood of x_0 . This implies that Φ is constant on a neighborhood of x_0 and hence, Φ is differentiable at x_0 with

$$(\nabla \Phi(x_0))^T = 0 = W^{(L)} \cdot D^{(L-1)} \cdot W^{(L-1)} \dots D^{(0)} \cdot W^{(0)}.$$

Here, the last equality stems from $D^{(\ell-1)} = \Delta(W^{(\ell-1)}x^{(\ell-1)} + b^{(\ell-1)}) = 0$. This proves (E.4). Overall, we get

$$\begin{aligned} \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\}) &\stackrel{(\text{E.3})}{=} \mathbb{P}(\mathcal{A} \cap \{\eta = \ell\} \cap \{x^{(\ell)} = 0\}) \stackrel{(\text{E.4})}{=} \mathbb{P}(\{\eta = \ell\} \cap \{x^{(\ell)} = 0\}) \\ &\stackrel{(\text{E.3})}{=} \mathbb{P}(\{\eta = \ell\} \cap \{x^{(\ell)} = 0\}) + \mathbb{P}(\{\eta = \ell\} \cap \{x^{(\ell)} \neq 0\}) \\ &= \mathbb{P}(\{\eta = \ell\}). \end{aligned}$$

It remains to deal with the case $\ell = 0$. Since $x^{(0)} = x_0 \neq 0$, we get with the same reasoning as above (Equation (E.3)) that

$$0 \leq \mathbb{P}(\mathcal{A} \cap \{\eta = 0\}) \leq \mathbb{P}(\eta = 0) = 0,$$

such that equality has to hold everywhere.

We thus have shown (E.2), which implies the claim. $\square$

B.2. Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks

Near-optimal estimates for the ℓ^p -Lipschitz constants of deep random ReLU neural networks

SJOERD DIRKSEN, PATRICK FINKE, PAUL GEUCHEN, DOMINIK STÖGER, AND FELIX VOIGTLAENDER

ABSTRACT. This paper studies the ℓ^p -Lipschitz constants of ReLU neural networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with random parameters for $p \in [1, \infty]$. The distribution of the weights follows a variant of the He initialization and the biases are drawn from symmetric distributions. We derive high probability upper and lower bounds for wide networks that differ at most by a factor that is logarithmic in the network's width and linear in its depth. In the special case of shallow networks, we obtain matching bounds. Remarkably, the behavior of the ℓ^p -Lipschitz constant varies significantly between the regimes $p \in [1, 2)$ and $p \in [2, \infty]$. For $p \in [2, \infty]$, the ℓ^p -Lipschitz constant behaves similarly to $\|g\|_{p'}$, where $g \in \mathbb{R}^d$ is a d -dimensional standard Gaussian vector and $1/p + 1/p' = 1$. In contrast, for $p \in [1, 2)$, the ℓ^p -Lipschitz constant aligns more closely to $\|g\|_2$.

1 Introduction

From empirical research it is well known that neural networks are vulnerable to adversarial attacks [19]. By introducing small perturbations to the input, the output of the neural network can be significantly altered. One way to assess the worst-case robustness of a neural network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ to adversarial attacks is by considering its ℓ^p -Lipschitz constants

$$\text{Lip}_p(\Phi) := \sup_{\substack{x, y \in \mathbb{R}^d \\ x \neq y}} \frac{|\Phi(x) - \Phi(y)|}{\|x - y\|_p}, \quad p \in [1, \infty].$$

For example, in image classification the case $p = \infty$ is especially relevant. This scenario allows for perturbations of each pixel which are inceptible to the human eye. Consequently, several algorithms were proposed to estimate bounds for Lipschitz constants of a trained deep neural network, see, e.g., [9, 11].

In this paper, we aim to understand the Lipschitz constants of neural networks with *random parameters*, i.e., the biases and weights are chosen at random. These networks appear naturally as the random initialization of any trained neural network. Understanding their properties is hence an important component in building a rigorous theory for deep learning. For instance, the ℓ^2 -Lipschitz constant of a random neural network has been used in the study of neural networks in the Neural Tangent Kernel (NTK) regime, see, e.g., [4]. More generally, random neural networks are often initially considered when trying to rigorously understand empirical phenomena in deep learning, see, e.g., [2, 5].

In this paper, we consider fully-connected neural networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ of the form

$$\Phi(x) := \left(V^{(L)} \circ \text{ReLU} \circ V^{(L-1)} \circ \dots \circ \text{ReLU} \circ V^{(0)} \right)(x),$$

where $\text{ReLU}(x) := \max\{0, x\}$ is applied componentwise and the affine maps $V^{(\ell)}$ are defined by

$$V^{(\ell)}(x) = W^{(\ell)}x + b^{(\ell)} \quad \text{for every } 0 \leq \ell \leq L.$$

Date: June 25, 2025.

2020 Mathematics Subject Classification. 68T07, 26A16, 60B20, 60G15.

Key words and phrases. Lipschitz constant, Random neural networks, ℓ^p -norms, Biases, Adversarial robustness.

We assume that both the weight matrices $W^{(0)} \in \mathbb{R}^{d \times N}$, $(W^{(\ell)})_{\ell=1}^{L-1} \subset \mathbb{R}^{N \times N}$, $W^{(L)} \in \mathbb{R}^{1 \times N}$ and the bias vectors $(b^{(\ell)})_{\ell=0}^{L-1} \subset \mathbb{R}^N$ and $b^{(L)} \in \mathbb{R}$ are initialized at random according to a variant of the popular He initialization [14]. Namely, we assume that the entries of $W^{(\ell)}$ are i.i.d. Gaussian with zero mean and standard deviation $\sqrt{2/N}$ for $0 \leq \ell \leq L-1$ and i.i.d. standard Gaussian for $\ell = L$. This variant of the He initialization is a natural choice for our study: in case of zero biases, each layer acts as an isometry in expectation, i.e., $\mathbb{E} [\|\text{ReLU}(W^{(\ell)}x)\|_2^2] = \|x\|_2^2$. Due to the homogeneity of the ReLU-function our results can be easily generalized to other initialization schemes. For the biases $(b^{(\ell)})_{\ell=0}^L$, we allow for more general distributions.

In this paper, we are interested in upper and lower bounds for the Lipschitz constant $\text{Lip}_p(\Phi)$ for all $p \in [1, \infty]$. Previous works made notable progress on this question in the case $p = 2$. It was shown in [13] that for shallow nets (i.e., $L = 1$) with arbitrary width, we have $\text{Lip}_2(\Phi) \asymp \sqrt{d}$ with high probability. For deeper networks, i.e., $L \geq 2$, they showed that

$$\sqrt{d} \lesssim \text{Lip}_2(\Phi) \lesssim C^L \sqrt{Ld \ln(N/d)}$$

for some absolute constant $C > 0$ if $N \gtrsim d$. Hence, the gap between the lower bound and the upper bound grows exponentially in L and logarithmically in N . An improved upper bound in the special case of sufficiently wide *zero-bias* networks was derived in [4]: in this case it was shown that $\text{Lip}_2(\Phi) \lesssim \sqrt{d \log(N)}$ if $N \gtrsim d^4 L$ (up to logarithmic factors). In this work we derive near-matching upper and lower bounds for $\text{Lip}_p(\Phi)$ for all $p \in [1, \infty]$. In the process, we also obtain (partial) improvements of the above results in the case $p = 2$. In the summary of our main findings we distinguish between the zero-bias case and the more general case.

Zero-bias case: We first consider the zero-bias case, that is, we assume that $b^{(\ell)} = 0$ for all $0 \leq \ell \leq L$. In this scenario, we establish that, with high-probability, the Lipschitz constants satisfy

$$\begin{cases} \sqrt{d/L}, \\ d^{1-1/p} \end{cases} \lesssim \text{Lip}_p(\Phi) \lesssim \begin{cases} \sqrt{d \cdot \ln(N/d)} & \text{for } p \in [1, 2), \\ d^{1-1/p} \cdot \sqrt{\ln(N/d)} & \text{for } p \in [2, \infty]. \end{cases}$$

In the case $p = 2$ and $L \geq 2$, we recover an upper bound that is slightly better than the estimate from [4] described above. Whereas [4, Theorem B.5] requires $N \gtrsim d^4 L$ (up to logarithmic factors), our result holds in the regime $N \gtrsim dL^3$ (again up to logarithmic factors). Thus, under the (practical) assumption $L \ll d$ our result holds for a wider range of parameters N . In the shallow case we do not completely recover the optimal bound $\text{Lip}_2(\Phi) \asymp \sqrt{d}$ from [13], suggesting that the logarithmic factor may be an artifact of our proof also in the case $L \geq 2$. In this special case, our work can also be considered complementary to [13]: indeed, our analysis in the case $p \neq 2$ together with the optimal bound in the case $p = 2$ together imply the matching bounds

$$\text{Lip}_p(\Phi) \asymp \begin{cases} \sqrt{d} & \text{for } p \in [1, 2), \\ d^{1-1/p} & \text{for } p \in [2, \infty], \end{cases} \quad (1.1)$$

provided that $L = 1$ and $N \gtrsim d$ (up to logarithmic factors). Our proof of the upper bounds relies on pointwise estimates of the gradient combined with an ε -net covering argument. The main technical innovation lies in adapting recent methods developed for analyzing random hyperplane tessellations [6, 7, 15] to control the deviation between gradients at points within ε -distance of each other.

To establish the lower bounds, we obtain from a straightforward duality argument that $\text{Lip}_p(\Phi)$ can be bounded from below by $\|\nabla \Phi(x)\|_{p'}$ for any fixed $x \in \mathbb{R}^d$, where p' is the Hölder conjugate of p , i.e., $1/p + 1/p' = 1$. By estimating $\|\nabla \Phi(x)\|_{p'}$ at a single point $x \in \mathbb{R}^d$, we obtain a lower bound for $\text{Lip}_p(\Phi)$ when $p \in [2, \infty]$ that match the upper bounds up to logarithmic factors. Interestingly, this approach does *not* yield optimal lower bounds when $p \in [1, 2)$. To address this, instead of considering a fixed point x we consider $\sup_x \|\nabla \Phi(x)\|_{p'}$, where the supremum is taken over all points $x \in \mathbb{R}^d$ in which Φ is differentiable. To obtain a lower bound for $p \in [1, 2]$ for this supremum we combine Sudakov minoration, decoupling arguments, and methods for analyzing random hyperplane tessellations.

p range	Type	Bound	Reference
$p \in [1, 2)$	$\gtrsim$	$\sqrt{d/L}$	Theorem 5.1
	$\lesssim$	$\sqrt{d \cdot \ln(N/d)}$	Corollary 4.2
$p \in [2, \infty]$	$\gtrsim$	$d^{1-1/p}$	Corollary 3.6
	$\lesssim$	$d^{1-1/p} \sqrt{\ln(N/d)}$	Corollary 4.2

FIGURE 1. Our results for neural networks with zero bias

A summary of our main results in the zero-bias case is presented in Figure 1, where we also indicate the corresponding locations of the bounds within the paper.

General case (non-zero bias): For networks with non-zero biases, we derive near-matching bounds on the Lipschitz constants under suitable assumptions on the biases. For instance, these may be drawn independently from a Gaussian distribution with a zero mean and a fixed variance, or a symmetric uniform distribution on an interval of a fixed size. Under these assumptions, we show that

$$\begin{cases} \sqrt{d/L}, \\ d^{1-1/p} \end{cases} \lesssim \text{Lip}_p(\Phi) \lesssim \begin{cases} \sqrt{d} \left(\sqrt{L} + \sqrt{\ln(N/d)} \right) & \text{for } p \in [1, 2), \\ d^{1-1/p} \left(\sqrt{L} + \sqrt{\ln(N/d)} \right) & \text{for } p \in [2, \infty]. \end{cases}$$

These bounds essentially remain valid for distributions other than the Gaussian or uniform distribution. Specifically, the lower bounds hold *for any* choice of the bias distribution. For the upper bounds, we only need to make the mild assumption that the biases follow absolutely continuous distributions, with bounded probability density function. For an overview, we refer to Figure 2, which lists our bounds along with the corresponding assumptions and locations of the bounds within the manuscript. Compared to [13], in the case $p = 2$ we obtain an upper bound with linear rather than exponential dependence on the depth, under slightly more stringent assumptions on the biases. Moreover, we establish near-matching bounds for all $p \neq 2$. Similar to the zero-bias case, for shallow networks our work complements [13] to yield matching bounds for all $p \in [1, \infty]$: the estimates (1.1) remain valid for $L = 1$ and $N \gtrsim d$ (up to logarithmic factors) for arbitrary distributions of the biases; see Corollary 6.8.

The lower bounds are derived via a reduction to the zero-bias case. We observe that for any signal $x \in M = \{x \in \mathbb{R}^d : \|x\|_2 \geq R\}$, for sufficiently large $R > 0$, as x propagates through the individual layers of the network, it remains relatively large compared to the biases. Consequently, the perturbation of the signal induced by the biases remains minimal. In particular, on the set M the network Φ can be approximated well by a zero-bias network $\tilde{\Phi}$ obtained from Φ by setting all its biases to zero.

In our proof of the upper bounds for $\text{Lip}_p(\Phi)$ we again aim to leverage the results from the zero-bias case. However, on M^c we can no longer expect that the zero-bias network $\tilde{\Phi}$ approximates Φ . We deal with this using a different approach that leads to an additional $\sqrt{L}$ -factor in the upper bound when compared to the zero-bias case.

Structure of the paper: In Section 2, we provide background regarding random neural networks and describe our main technical assumptions. Moreover, we describe several results on random hyperplane tessellations, which are a key component in our proofs. Section 3 focuses on proving both upper and lower bounds for the ℓ^p -norms of the gradient of the network at a fixed point $x \in \mathbb{R}^d$. This directly leads to near-optimal lower bounds for the Lipschitz constants of the network Φ when $p \in [2, \infty]$. In Section 4, we prove an upper bound for the ℓ^p -Lipschitz constant for any $p \in [1, \infty]$. In Section 5, we establish a lower bound for the Lipschitz constants of zero-bias networks in the regime $p \in [1, 2]$. In Section 6, we explain how to generalize our results from zero-bias networks to general networks with non-zero biases. Throughout, we provide proof sketches of our results which contain the key ideas. The lengthier proofs with full details can be found in the Appendix.

p range	Type	Bound	Assumption	Reference
$p \in [1, 2)$	$\succ$	$\sqrt{d/L}$	none	Theorem 6.6
	$\succsim$	$\sqrt{d} \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right)$	bounded pdf+symmetric	Corollary 6.4
$p \in [2, \infty]$	$\succ$	$d^{1-1/p}$	none	Theorem 6.7
	$\succsim$	$d^{1-1/p} \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right)$	bounded pdf+symmetric	Corollary 6.4

FIGURE 2. Our results for neural networks with non-zero bias. The constant C_τ is an upper bound for the pdf of the biases and λ specifies their decay rate.

1.1 Notation

For a statement A , we set $\mathbb{1}_A := 1$ if A is true and $\mathbb{1}_A := 0$ otherwise. We denote constants whose enlargement does not change a given statement using capital letters and constants whose shrinkage does not change a given statement using lowercase letters. Without loss of generality, we may assume that all constants of the first kind are greater or equal than 1, and all constants of the second kind are less or equal than 1. We write $\alpha \lesssim \beta$ if there exists an absolute constant $C > 0$ such that $\alpha \leq C \cdot \beta$, and we write $\alpha \asymp \beta$ if both $\alpha \lesssim \beta$ and $\beta \lesssim \alpha$.

Let $x, y \in \mathbb{R}^n$. We denote their inner product by $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$. For $p \in [1, \infty)$, we denote the ℓ^p -norm by $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ and the ℓ^∞ -norm by $\|x\|_\infty := \max_{i=1, \dots, n} |x_i|$. We denote the open and closed Euclidean balls with center x and radius $r > 0$ by

$$B_n(x, r) := \{y \in \mathbb{R}^n : \|x - y\|_2 < r\} \quad \text{and} \quad \overline{B}_n(x, r) := \{y \in \mathbb{R}^n : \|x - y\|_2 \leq r\},$$

and the unit sphere by $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. The d -dimensional Lebesgue measure is denoted by $\mathbb{X}^d$.

Let $K \subseteq \mathbb{R}^n$. The interior of K (with respect to the standard topology on $\mathbb{R}^n$) is denoted by $\text{int}(K)$. We define the cone generated by K and the Minkowski difference of K with itself by

$$\text{cone}(K) := \{sx : x \in K, s > 0\} \quad \text{and} \quad K - K := \{x - y : x, y \in K\},$$

respectively. Finally, we denote the diameter of K by $\text{diam}(K) := \sup_{x, y \in K} \|x - y\|_2 \in [0, \infty]$.

Let $A \in \mathbb{R}^{m \times n}$. We denote its i -th row by $A_{i,:}$ and its j -th column by $A_{:,j}$. For $p, q \in [1, \infty]$, we define the operator norm by

$$\|A\|_{p \rightarrow q} := \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p \leq 1}} \|Ax\|_q.$$

By $|A|$ we denote the matrix whose entries are the absolute values of the entries of A . We denote the trace by $\text{Tr}(A) := \sum_{i=1}^{\min\{m, n\}} A_{i,i}$. $I_{m \times n} \in \mathbb{R}^{m \times n}$ denotes the matrix with $(I_{m \times n})_{i,j} = \mathbb{1}_{i=j}$ for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$, and we use the shorthand $I_n := I_{n \times n}$. By $\text{diag}\{0, 1, -1\}^{m \times m}$ we denote the set of $m \times m$ diagonal matrices with diagonal entries in $\{0, 1, -1\}$. For $a_1, \dots, a_m \in \mathbb{R}$ we denote by $\text{diag}(a_1, \dots, a_m) \in \mathbb{R}^{m \times m}$ the $m \times m$ diagonal matrix with entries $a_1, \dots, a_m$ on the diagonal.

For $\ell_1, \ell_2 \in \mathbb{Z}$ with $\ell_1 \leq \ell_2$ and matrices $A_{\ell_1}, \dots, A_{\ell_2}$ with suitable dimensions, we let

$$\prod_{i=\ell_2}^{\ell_1} A_i := A_{\ell_2} \cdots A_{\ell_1} \tag{1.2}$$

denote their matrix product in reverse order. If $A_{\ell_1} \in \mathbb{R}^{m \times n}$ and $\ell_2 < \ell_1$, define the empty product as

$$\prod_{i=\ell_2}^{\ell_1} A_i := I_{n \times n}.$$

We write $\mathcal{N}(0, \sigma^2)$ for the normal distribution with expectation 0 and variance σ^2 , and $\mathcal{N}(0, I_k)$ for the distribution of a k -dimensional random vector with independent standard Gaussian entries. We use $\text{Unif}(K)$ to denote a uniform distribution on a set K .

Let X and Y be random variables. We write $X \stackrel{d}{=} Y$ if X and Y are identically distributed. If X is real-valued, we denote the expectation of X by $\mathbb{E}[X]$, the variance by $\mathbb{V}[X]$ and let $\|X\|_{\psi_1}$ and $\|X\|_{\psi_2}$ denote the sub-exponential and sub-gaussian norm of X , respectively; see [20, Chapters 2.5 & 2.7]. For a subset $K \subseteq \mathbb{R}^k$, we let

$$w(K) := \mathbb{E}_{g \sim \mathcal{N}(0, I_k)} \left[\sup_{x \in K} \langle x, g \rangle \right]$$

denote the *Gaussian width* of K ; see [20, Section 7.5].

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. We denote by $f|_K$ the restriction of f onto $K \subseteq \mathbb{R}^n$. If f is differentiable at $x_0 \in \mathbb{R}^n$, we denote by $\text{Df}(x_0) \in \mathbb{R}^{m \times n}$ the Jacobian of f at x_0 . If $m = 1$, we let $\nabla f(x_0) \in \mathbb{R}^n$ be the gradient of f at x_0 .

For given $\varepsilon > 0$, we call a set $\mathcal{N} \subseteq K$ an ε -net of K , if for every $x \in K$ there exists $x^* \in \mathcal{N}$ with $\|x - x^*\|_2 \leq \varepsilon$. We denote the ε -covering number (with respect to Euclidean distances) of K by

$$\mathcal{N}(K, \varepsilon) := \min \{ \#\Lambda : \Lambda \subseteq K \text{ is an } \varepsilon\text{-net of } K \},$$

that is, the smallest possible cardinality of an ε -net of K . Similarly, we call a set $\Lambda \subseteq K$ δ -separated if $\|x - y\|_2 > \delta$ for every $x, y \in \Lambda$ with $x \neq y$. The δ -packing number of K is defined as

$$\mathcal{P}(K, \delta) := \sup \{ \#\Lambda : \Lambda \subseteq K, \Lambda \text{ is } \delta\text{-separated} \}.$$

It is well-known that the covering numbers and packing numbers are equivalent in the sense that

$$\mathcal{P}(K, 2\delta) \leq \mathcal{N}(K, \delta) \leq \mathcal{P}(K, \delta);$$

see [20, Lemma 4.2.8].

2 Setting and preliminary results

2.1 Setting

We consider ReLU networks with d input neurons, a single output neuron and L hidden layers of width N . Formally, we consider maps

$$\Phi : \mathbb{R}^d \rightarrow \mathbb{R}, \quad \Phi(x) := \left(V^{(L)} \circ \text{ReLU} \circ V^{(L-1)} \circ \dots \circ \text{ReLU} \circ V^{(0)} \right)(x). \quad (2.1)$$

Here, ReLU denotes the function

$$\text{ReLU}(x) := \max\{0, x\}, \quad x \in \mathbb{R},$$

and its application in (2.1) is to be understood componentwise, i.e.,

$$\text{ReLU}((x_1, \dots, x_N)) = (\text{ReLU}(x_1), \dots, \text{ReLU}(x_N)).$$

The maps $V^{(\ell)}$ for $0 \leq \ell \leq L$ are affine: there exist *weights* $W^{(0)} \in \mathbb{R}^{N \times d}$, $W^{(1)}, \dots, W^{(L-1)} \in \mathbb{R}^{N \times N}$ and $W^{(L)} \in \mathbb{R}^{1 \times N}$ as well as *biases* $b^{(0)}, \dots, b^{(L-1)} \in \mathbb{R}^N$ and $b^{(L)} \in \mathbb{R}$ such that

$$V^{(\ell)}(x) = W^{(\ell)}x + b^{(\ell)} \quad \text{for every } 0 \leq \ell \leq L.$$

We refer to a network as *zero-bias*, if the biases are all set to zero.

In the present work, we specifically analyze the properties of *random* ReLU networks, i.e., ReLU networks where the weights and biases are drawn at random. We collect the different assumptions on this random initialization that are used over the course of the present paper in the following.

Assumption 2.1. We assume that the weights and biases are randomly chosen in the following way: For $0 \leq \ell < L$, we have

$$\left(W^{(\ell)}\right)_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N) \quad \text{and} \quad \left(W^{(L)}\right)_{1,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),$$

where all entries are jointly independent. The entries of the biases are drawn independently from (possibly different) real-valued random distributions and are jointly independent of the weights.

Assumption 2.2. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.1. We say that the biases are symmetric if the components of the bias vectors are drawn from symmetric distributions.

Assumption 2.3. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.2. We say that the biases satisfy a small-ball property with constant $C_\tau > 0$ if, for every $t \in \mathbb{R}$ and $\varepsilon > 0$,

$$\mathbb{P}\left(b_j^{(\ell)} \in (t - \varepsilon, t + \varepsilon)\right) \leq C_\tau \cdot \varepsilon$$

for any j and ℓ . This is equivalent to every $b_j^{(\ell)}$ having an absolutely continuous probability distribution with probability density function bounded by $C_\tau/2$ almost everywhere; see Theorem I.1.

2.2 The Lipschitz constant and the formal gradient of ReLU networks

For any $K \subseteq \mathbb{R}^d$, any function $f : K \rightarrow \mathbb{R}^k$ and any $p, q \in [1, \infty]$ we denote its *Lipschitz constant* with respect to $\|\cdot\|_p$ and $\|\cdot\|_q$ as

$$\text{Lip}_{p \rightarrow q}(f) := \sup_{\substack{x, y \in K \\ x \neq y}} \frac{\|f(x) - f(y)\|_q}{\|x - y\|_p}.$$

We call f *Lipschitz continuous* if its Lipschitz constant is finite.¹ Since we study ReLU networks Φ with an output dimension of 1, we have $\|\Phi(x) - \Phi(y)\|_q = |\Phi(x) - \Phi(y)|$ for any $q \in [1, \infty]$. Therefore, the value of $\text{Lip}_{p \rightarrow q}(\Phi)$ is actually independent of q and we simply write $\text{Lip}_p(\Phi)$ instead of $\text{Lip}_{p \rightarrow q}(\Phi)$.

The following well-known proposition establishes a relation between the Lipschitz constant and the Jacobian of a function.

Proposition 2.4. Let $K \subseteq \mathbb{R}^d$ be a convex set with non-empty interior. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ be Lipschitz continuous and $M \subseteq \mathbb{R}^d$ any measurable subset of $\mathbb{R}^d$ with Lebesgue measure $\lambda^d(\mathbb{R}^d \setminus M) = 0$ such that f is differentiable in every $x \in M$. Let $p, q \in [1, \infty]$. Then,

$$\text{Lip}_{p \rightarrow q}(f|_K) = \sup_{x \in M \cap \text{int}(K)} \|Df(x)\|_{p \rightarrow q}.$$

In particular, if $k = 1$,

$$\text{Lip}_p(f|_K) = \sup_{x \in M \cap \text{int}(K)} \|\nabla f(x)\|_{p'},$$

where $p' \in [1, \infty]$ satisfies $1/p + 1/p' = 1$.

Note that the existence of a set M with the required properties follows from the fact that f is Lipschitz continuous; this is known as Rademacher's theorem (cf. [10, Section 3.1.2]). For the sake of completeness, we provide a proof of Proposition 2.4 in Appendix A.

Inspired by the above proposition, we base our estimation of the Lipschitz constant of a ReLU network on the computation of its gradient. Note, however, that a ReLU network is *not* necessarily everywhere differentiable, since the ReLU itself is not differentiable at 0. To overcome this problem, in the following, we introduce a computable proxy of the gradient which coincides almost everywhere with the actual gradient.

¹This is independent of p and q as all norms are equivalent on $\mathbb{R}^n$.

For this purpose, we introduce the following notation: For a vector $x \in \mathbb{R}^N$ we define the diagonal matrix $\Delta(x) \in \mathbb{R}^{N \times N}$ via

$$(\Delta(x))_{i,j} := \begin{cases} \mathbb{1}_{x_i > 0} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

This leads to the following recursive representation of the ReLU network Φ . Using the same notation as in Section 2.1, let $x =: \Phi^{(-1)}(x) \in \mathbb{R}^d$ and define recursively

$$\begin{aligned} D^{(\ell)}(x) &:= \Delta(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}), \\ \Phi^{(\ell)}(x) &:= D^{(\ell)}(x) \cdot (W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}) = \text{ReLU}(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}), \quad 0 \leq \ell < L. \end{aligned} \quad (2.2)$$

Note that then $\Phi(x) = W^{(L)}\Phi^{(L-1)}(x) + b^{(L)}$. For convenience, we further set $D^{(L)}(x) := 1$ for every $x \in \mathbb{R}^d$.

Using this notation, we define the formal gradient of the network, and similarly the formal Jacobian of the intermediate layers as follows.

Definition 2.5. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU network with L hidden layers as in Section 2.1, $x \in \mathbb{R}^d$ and $\ell \in \{-1, \dots, L-1\}$. We define the *formal Jacobian* of $\Phi^{(\ell)}$ at x and the *formal gradient* of Φ at x by

$$\overline{D}\Phi^{(\ell)}(x) := \prod_{j=\ell}^0 D^{(j)}(x)W^{(j)} \quad \text{and} \quad \overline{\nabla}\Phi(x) := \left(W^{(L)} \left[\prod_{j=L-1}^0 D^{(j)}(x)W^{(j)} \right] \right)^T, \quad (2.3)$$

respectively. Here, the reverse order matrix product is as defined in (1.2). This in particular implies $\overline{D}\Phi^{(-1)}(x) = I_{d \times d}$. Furthermore, for $\ell_1 \in \{0, \dots, L\}$ and $\ell_2 \in \{-1, \dots, L-1\}$, we define

$$\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x) := \prod_{j=\ell_2}^{\ell_1} D^{(j)}(x)W^{(j)}, \quad x \in \mathbb{R}^d.$$

As already indicated above, the formal gradient and Jacobian serve as a proxy for their true counterparts. This is justified by the following proposition.

Proposition 2.6. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU neural network with L hidden layers as in Section 2.1 and fix $\ell \in \{-1, \dots, L-1\}$. Then, for almost every $x \in \mathbb{R}^d$, $\Phi^{(\ell)}$ is differentiable at x with

$$D\Phi^{(\ell)}(x) = \overline{D}\Phi^{(\ell)}(x).$$

Moreover, for almost every $x \in \mathbb{R}^d$ the entire network Φ is differentiable at x with

$$\nabla\Phi(x) = \overline{\nabla}\Phi(x).$$

The above is probably already well known to the community and the statement can (at least partially) be found in [3, Theorem III.1]. In fact, the proof in [3] already includes the statement for the intermediate layers but it is not stated explicitly in the theorem. For this reason, and in order to clarify the proof itself, we decided to include a detailed argument in Appendix A.

Remark 2.7. We emphasize that the formal gradient may in general *not* be viewed as an extension of the true gradient. In fact, there might exist points of differentiability where $\nabla\Phi(x) \neq \overline{\nabla}\Phi(x)$.

As an example, consider the shallow zero-bias ReLU network $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ with weight matrices

$$W^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad W^{(1)} = (1 \quad -1).$$

A simple computation shows $\Phi(x) = x$ for every $x \in \mathbb{R}$. Hence, Φ is everywhere differentiable with $\nabla\Phi(x) = 1$. On the other hand, $D^{(0)}(0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and therefore also $\overline{\nabla}\Phi(0) = 0$. $\diamond$

At a fixed non-zero input, a random ReLU network is almost surely differentiable with the actual gradient coinciding with the formal gradient; see [13, Theorem E.1]. Combining this fact with Propositions 2.4 and 2.6 we get the following result, which is the starting point in our analysis of the Lipschitz constant of random ReLU networks.

Theorem 2.8. *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers satisfying Assumption 2.1, and let $p, p', q \in [1, \infty]$ with $1/p + 1/p' = 1$. Let*

$$M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla \Phi(x) = \bar{\nabla} \Phi(x)\}.$$

Moreover, fix $x_0 \in \mathbb{R}^d \setminus \{0\}$. Then, almost surely with respect to the randomness in Φ ,

$$\|\bar{\nabla} \Phi(x_0)\|_{p'} \leq \text{Lip}_p(\Phi) = \sup_{x \in M_\Phi} \|\bar{\nabla} \Phi(x)\|_{p'} \leq \sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_{p'}$$

and, in case Φ is zero-bias, even

$$\|\bar{\nabla} \Phi(x_0)\|_{p'} \leq \text{Lip}_p(\Phi) = \sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x)\|_{p'} \leq \sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x)\|_{p'}.$$

Moreover, for every convex set $K \subseteq \mathbb{R}^d$ with non-empty interior and $\ell \in \{-1, \dots, L-1\}$,

$$\text{Lip}_{p \rightarrow q} \left(\Phi^{(\ell)} \Big|_K \right) \leq \sup_{x \in \text{int}(K)} \|\bar{D} \Phi^{(\ell)}(x)\|_{p \rightarrow q}$$

and, in case Φ is zero-bias,

$$\text{Lip}_{p \rightarrow q} \left(\Phi^{(\ell)} \right) \leq \sup_{x \in \mathbb{S}^{d-1}} \|\bar{D} \Phi^{(\ell)}(x)\|_{p \rightarrow q}.$$

Notably, in the special case of a zero-bias network, we may restrict ourselves to inputs from the sphere. This is mainly due to the fact that in this case the formal gradient is invariant under scaling by positive constants; see Appendix A for the details. The same is not possible if one considers networks with non-zero biases. The possibility to restrict to the sphere simplifies the analysis of the Lipschitz constant for zero-bias networks tremendously, since the compactness of the sphere allows for a covering-based approach. Hence, we consider the zero-bias case first in Sections 4 and 5 and handle the case of networks with general biases by reducing it to the zero-bias setting in Section 6.

2.3 The randomized gradient

A major challenge in analyzing the gradient of a random ReLU network at a fixed input x is that, while the weight matrices $W^{(\ell)}$ are stochastically independent, the matrices $D^{(\ell)}(x)$ are *not* independent, since each depends on all the weights and biases from the previous layers. In this section, we show that, for a fixed input x , in the case of symmetric biases, on a high probability event and *in distribution* we may effectively replace the $D^{(\ell)}(x)$ by matrices that are independent of the weights and biases, and of each other. This significantly simplifies the analysis of the gradient at a fixed input.

The motivation behind the specific definition of the matrices $D^{(\ell)}(x)$ constructed in (2.2) is that the ReLU is differentiable on $\mathbb{R} \setminus \{0\}$ with $\text{ReLU}' \equiv 1$ on $\mathbb{R}_{>0}$ and $\text{ReLU}' \equiv 0$ on $\mathbb{R}_{<0}$. However, it is not quite clear which value one should assign for a diagonal entry of a D -matrix, if the pre-activation, i.e., the input of the ReLU, of the corresponding neuron is exactly zero. Therefore, as was already noted in [2], to facilitate the analysis of the gradient at a fixed point when considering random networks, it is convenient to introduce an additional source of randomness: if the preactivation of a single neuron is exactly zero, then the corresponding entry on the diagonal of the D -matrix is either 0 or 1, each with probability 1/2. We formalize this idea, in the following definition.

Definition 2.9. Fix $x \in \mathbb{R}^d$ and let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N as in Assumption 2.1. For $\ell \in \{0, \dots, L-1\}$ we define the random matrices $\bar{D}^{(\ell)}(x)$ via

$$\bar{D}^{(\ell)}(x) := D^{(\ell)}(x) + \text{diag} \left(\mathbb{1}_{(W^{(\ell)} \Phi^{(\ell-1)}(x))_1 + b_1^{(\ell)} = 0} \cdot \varepsilon_1^{(\ell)}, \dots, \mathbb{1}_{(W^{(\ell)} \Phi^{(\ell-1)}(x))_N + b_N^{(\ell)} = 0} \cdot \varepsilon_N^{(\ell)} \right),$$

where $\varepsilon_i^{(\ell)} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{0, 1\})$, for $\ell \in \{0, \dots, L-1\}$ and $i \in \{1, \dots, N\}$, are jointly independent of $(W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)})$.

Notably, at a fixed input $x \in \mathbb{R}^d \setminus \{0\}$, the matrices $\overline{D}^{(\ell)}(x)$ agree with the matrices $D^{(\ell)}(x)$ with high probability. This is shown in the following result, whose proof is deferred to Appendix A.

Proposition 2.10. *Fix $x \in \mathbb{R}^d \setminus \{0\}$ and let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases and with L hidden layers of width N as in Assumption 2.2. Then, with probability at least $(1 - \frac{1}{2^N})^L$,*

$$D^{(\ell)}(x) = \overline{D}^{(\ell)}(x) \quad \text{for every } \ell \in \{0, \dots, L-1\}.$$

Having introduced the randomized matrices $\overline{D}^{(\ell)}(x)$, one can show that *in distribution* one can replace the matrices $\overline{D}^{(\ell)}(x)$ by *independent* diagonal matrices $\widehat{D}^{(\ell)}$ with i.i.d. $\text{Unif}\{0, 1\}$ -distributed entries on the diagonal. This distributional equivalence was originally stated and proven in [2, Lemma 2.1]. In the following, we prove a generalization of that statement, where the arguments of the $\overline{D}$ -matrices may vary from layer to layer; see below for a precise formulation. The proof can be found in Appendix B.

Proposition 2.11 (generalization of [2, Lemma 2.1]). *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases and with L hidden layers of width N as in Assumption 2.2. Fix $\ell_1 \in \{0, \dots, L-1\}$, $\ell_2 \in \{-1, \dots, L-1\}$ and vectors $z_0, \dots, z_{L-1} \in \mathbb{R}^d$. Set*

$$\mu := \begin{cases} d, & \text{if } \ell_1 = 0, \\ N, & \text{otherwise,} \end{cases}$$

let $\nu \in \mathbb{N}$ be arbitrary and let $A = A(W^{(0)}, \dots, W^{(\ell_1-1)}, b^{(0)}, \dots, b^{(\ell_1-1)}) \in \mathbb{R}^{\mu \times \nu}$ be a matrix which may solely depend on $W^{(0)}, \dots, W^{(\ell_1-1)}$ and $b^{(0)}, \dots, b^{(\ell_1-1)}$. Then,

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \left[\prod_{i=\ell_2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2} \quad \text{and}$$

$$W^{(L)} \left[\prod_{i=L-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] \stackrel{d}{=} W^{(L)} \left[\prod_{i=L-1}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right]$$

with

$$\widehat{D}^{(i)} := \text{diag}(\varepsilon_1^{(i)}, \dots, \varepsilon_N^{(i)}) \quad \text{for every } i \in \{0, \dots, L-1\}.$$

Here, the $\varepsilon_j^{(i)}$ are i.i.d. $\text{Unif}\{0, 1\}$ and jointly independent of $(W^{(0)}, \dots, W^{(L)}, b^{(0)}, \dots, b^{(L)})$.

Propositions 2.10 and 2.11 together show that, by passing to a high probability event, we may effectively replace the matrices $D^{(\ell)}$ by the randomized matrices $\overline{D}^{(\ell)}$ and afterwards replace these matrices by the stochastically independent diagonal matrices $\widehat{D}^{(\ell)}$. We use this technique throughout this paper, especially in Section 3.

2.4 Random hyperplane tessellations

It is of central importance in the present work to control the number of neurons at which the activation patterns differ for two inputs $x, y \in \mathbb{R}^d$. More precisely, one wants to control the expression $\text{Tr} |D^{(\ell)}(x) - D^{(\ell)}(y)|$.

To gain some intuition, let us for now focus on the case $\ell = 0$. Then the problem of estimating

$$\text{Tr} |D^{(0)}(x) - D^{(0)}(y)| = \# \left\{ i \in \{1, \dots, N\} : \text{sgn} \left((W^{(0)}x)_i + b_i^{(0)} \right) \neq \text{sgn} \left((W^{(0)}y)_i + b_i^{(0)} \right) \right\}$$

is intimately related to *random hyperplane tessellations* (cf. for instance [6, 17]). Indeed, each row of $W^{(0)}$ and the corresponding entry of the bias vector $b^{(0)}$ determine a random hyperplane in $\mathbb{R}^d$ and $\text{Tr} |D^{(0)}(x) - D^{(0)}(y)|$ corresponds to the number of hyperplanes separating x and y .

In this work, we make heavy use of the results presented in [6, 15]. These provide *local* uniform tessellation results for random hyperplane tessellations, meaning that with high probability the number of separating hyperplanes is large for all vectors x, y with $\|x - y\|_2$ sufficiently large and small for all vectors x, y with $\|x - y\|_2$ sufficiently small. These results were obtained in [15] without biases and in [6] with biases drawn from a uniform distribution. In fact, for many results in [6] one does not necessarily need to require the biases to be drawn from a uniform distribution; it suffices to assume that they satisfy a small-ball property, as we do in Proposition 2.15.

For two *fixed* inputs $x, y \in \mathbb{S}^{d-1}$, we get the following well-known result. It states that the fraction of separating random hyperplanes with zero biases between x and y is proportional to their Euclidean distance, with high probability. For the sake of completeness, we provide a proof in Appendix C.

Proposition 2.12. *There exist absolute constants $C, c > 0$ such that the following holds. For $\sigma > 0$, let $A \in \mathbb{R}^{m \times n}$ be a random matrix with entries $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, let $x, y \in \mathbb{S}^{n-1}$ and set $\delta := \|x - y\|_2$. Then, with probability at least $1 - 2 \exp(-c \cdot \delta m)$,*

$$c \cdot \delta m \leq \#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i) \neq \text{sgn}((Ay)_i)\} \leq C \cdot \delta m.$$

Next, we move on to the localized results. The relevant result from [15] reads as follows.

Proposition 2.13 ([15, Theorem 3.2]). *There exist absolute constants $C, c > 0$ such that the following holds. For $\sigma > 0$, let $A \in \mathbb{R}^{m \times n}$ be a random matrix with entries $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, and let $K \subseteq \mathbb{S}^{n-1}$ be non-empty and $\delta \in (0, e^{-1})$. Suppose that $0 < \varepsilon \leq c \cdot \delta \cdot \ln^{-1/2}(1/\delta)$ and*

$$m \geq C \cdot \max\{\delta^{-1} \ln(e \cdot \mathcal{N}(K, \varepsilon/2)), \delta^{-3} w^2((K - K) \cap \overline{B}_n(0, \varepsilon))\}.$$

Then, with probability at least $1 - \exp(-c \cdot \delta m)$,

$$\#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i) \neq \text{sgn}((Ay)_i)\} \leq \delta \cdot m \quad \text{for all } x, y \in K \text{ with } \|x - y\|_2 \leq \varepsilon/2.$$

Since [15] has to the best of our knowledge only appeared as a preprint so far and for the sake of completeness, we provide a proof for Proposition 2.13 in Appendix C.

Moreover, for the case of networks with possibly non-zero biases, we make use of the following two results, which can be proven with similar techniques as in [6]; see Appendix C for the details. The first one states that the bias terms do not play a significant role for inputs with large Euclidean norm.

Proposition 2.14. *There exist absolute constants $C, c > 0$ such that the following holds. Assume that $A \in \mathbb{R}^{m \times n}$ is a random matrix with entries $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and let $K \subseteq \mathbb{S}^{n-1}$ be non-empty and $\delta \in (0, e^{-1})$. Suppose that $0 < \varepsilon \leq c \cdot \delta \cdot \ln^{-1/2}(1/\delta)$ and*

$$m \geq C \cdot \max\{\delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon)), \delta^{-3} \cdot w^2((K - K) \cap \overline{B}_n(0, \varepsilon))\}.$$

Let further $\lambda > 0$ and $\tau_1, \dots, \tau_m \in \mathbb{R}$ such that $\max_i |\tau_i| \leq \lambda$. Then, with probability at least $1 - \exp(-c \cdot \delta m)$,

$$\#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i + \tau_i) \neq \text{sgn}((Ax)_i)\} \leq \delta m \quad \text{for all } x \in \text{cone}(K) \text{ with } \|x\|_2 \geq C \cdot \delta^{-1} \lambda.$$

We can further establish a bound similar to Proposition 2.13 for networks with non-zero biases, which generalizes the upper bound from [6, Theorem 2.9].

Proposition 2.15. *There exist absolute constants $C, c > 0$ such that the following holds. Assume that $A \in \mathbb{R}^{m \times n}$ is a random matrix with $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and let $b \in \mathbb{R}^m$ be a random vector with independent entries which is independent of A . Suppose that the entries of b satisfy a small-ball property with constant $C_\tau > 0$, that is, for any $i \in \{1, \dots, m\}$, and for all $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathbb{P}(b_i \in (t - \varepsilon, t + \varepsilon)) \leq C_\tau \cdot \varepsilon.$$

Let further $\delta \in (0, e^{-1})$ and $K \subseteq \mathbb{R}^n$ non-empty. Suppose that $\varepsilon \leq c \cdot C_\tau^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$ and

$$m \geq C \cdot \max\{\delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon/2)), \delta^{-3} \cdot w^2((K - K) \cap \overline{B}_n(0, \varepsilon)) \cdot C_\tau^2\}.$$

Then, with probability at least $1 - \exp(-c \cdot \delta m)$,

$$\#\{i \in \{1, \dots, m\} : \operatorname{sgn}((Ax)_i + b_i) \neq \operatorname{sgn}((Ay)_i + b_i)\} \leq \delta m \quad \text{for all } x, y \in K \text{ with } \|x - y\|_2 \leq \varepsilon/2.$$

3 Pointwise bounds on the Jacobian and gradient

As a first step, we provide *pointwise* estimates on the formal Jacobian and gradient of random ReLU networks, i.e., estimates that hold at fixed, non-zero inputs but not necessarily uniformly over the input space. The proofs in this section heavily rely on the randomized gradient introduced in Section 2.3 and in particular exploit the distributional equivalence from Proposition 2.11. This allows to condition on the $\widehat{D}$ -matrices and to then only consider the randomness over the weight matrices $W^{(\ell)}$, so that the results can be shown using standard bounds for Gaussian matrices and vectors. Note that, in order to be able to apply the techniques from Section 2.3, we need to assume that the biases are symmetric, which we do throughout this section. Note that this in particular includes the case of zero-bias networks. We refer to Appendix D for detailed proofs of what follows.

In particular, in Theorem 3.4 we provide upper and lower bounds for the ℓ^q -norm of the gradient of a random ReLU network with symmetric biases as in Assumption 2.2 at a fixed input point. As already discussed in Theorem 2.8, these bounds immediately yield a corresponding lower bound for the Lipschitz constant $\operatorname{Lip}_p(\Phi)$, where $1/p + 1/q = 1$. Quite surprisingly, in the case $p \in [2, \infty]$ these pointwise bounds on the gradient are up to a logarithmic factor of the same order as the upper bounds for the Lipschitz constant established in Theorem 4.1, where the supremum over the entire input space is taken. Roughly speaking, taking the supremum over the entire input space does not change anything (except for logarithmic factors) compared to considering just a single point. Note that this is not the case in the regime $p \in [1, 2)$; see Section 5.

Moreover, in Lemmas 3.1 to 3.3, we state three auxiliary results which are frequently used over the course of the paper and which we expect to be of independent interest.

We start with the following lemma, which states that matrices of the form $\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i)W^{(i)}$ satisfy a certain isometry property.

Lemma 3.1. *There exist constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases, with L hidden layers of width N as in Assumption 2.2 and with $N \geq C \cdot L^2 \cdot \ln(eL)$. Let $\ell_1 \in \{0, \dots, L-1\}$, $\ell_2 \in \{-1, \dots, L-1\}$, $z_0, \dots, z_{L-1} \in \mathbb{R}^d \setminus \{0\}$,*

$$\mu := \begin{cases} d, & \text{if } \ell_1 = d, \\ N, & \text{otherwise,} \end{cases}$$

and let $v = v(W^{(0)}, \dots, W^{(\ell_1-1)}, b^{(0)}, \dots, b^{(\ell_1-1)}) \in \mathbb{R}^\mu$ be a vector which may solely depend on $W^{(0)}, \dots, W^{(\ell_1-1)}$ and $b^{(0)}, \dots, b^{(\ell_1-1)}$. Then, with probability at least $1 - \exp(-c \cdot N/L^2)$,

$$\frac{1}{2} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i)W^{(i)} \right] v \right\|_2 \leq e \cdot \|v\|_2.$$

Note that in the special case of a *zero-bias* ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, we have

$$\Phi^{(\ell)}(x) = \left[\prod_{i=\ell}^0 D^{(i)}(x)W^{(i)} \right] x, \quad x \in \mathbb{R}^d.$$

Hence, Lemma 3.1 in particular states that in the case of a zero-bias random ReLU network following Assumption 2.1, we have

$$\|\Phi^{(\ell)}(x)\|_2 \asymp \|x\|_2$$

with high probability. Note that this statement already appears in a slightly stronger form in [1, Lemma 7.1]; see also Lemma F.1.

We continue with the following statement, which essentially states that

$$\left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2} \lesssim \|A\|_{2 \rightarrow 2}$$

for every matrix A with $\text{rank}(A) \leq N/L^2$. We even state a slightly more general version, where the arguments of the D -matrices may vary from layer to layer.

Lemma 3.2. *There exist absolute constants $C, c > 0$ such that the following holds. Let $x \in \mathbb{R}^d \setminus \{0\}$ be fixed and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases, with L hidden layers of width N as in Assumption 2.2 with $N \geq C \cdot L^2 \cdot \ln(eL)$. Moreover, fix $\ell_2 \in \{-1, \dots, L-1\}$, $\ell_1 \in \{0, \dots, L\}$ and $z_0, \dots, z_{L-1} \in \mathbb{R}^d \setminus \{0\}$. Let*

$$\mu := \begin{cases} d & \text{if } \ell_1 = 0, \\ N & \text{if } \ell_1 > 0, \end{cases}$$

pick $\nu \in \mathbb{N}$ and let $A \in \mathbb{R}^{\mu \times \nu}$ with $\text{rank}(A) \leq N/L^2$. Then,

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i)W^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq C \cdot \|A\|_{2 \rightarrow 2}$$

with probability at least $1 - \exp(-c \cdot N/L^2)$.

Note that, by putting $\ell_1 = 0$, $A = I_{d \times d}$ and $z_0 = z_{L-1} = x$, Lemma 3.2 provides a bound on the spectral norm of $\overline{D}\Phi^{(\ell_2)}(x)$, under the assumption $N \gtrsim L^2(d + \ln(eL))$.

Next, the following lemma provides an estimate similar to the one from Lemma 3.2 incorporating also the final layer $W^{(L)}$.

Lemma 3.3. *There exist absolute constants $C > 0$ and $c > 0$ such that the following holds. Let $z_0, \dots, z_{L-1} \in \mathbb{R}^d \setminus \{0\}$ be fixed and let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases, with L hidden layers of width N as in Assumption 2.2 with $N \geq C \cdot L^2 \cdot \ln(eL)$ and moreover, fix $\ell \in \{0, \dots, L\}$. Let*

$$\mu := \begin{cases} d & \text{if } \ell = 0, \\ N & \text{if } \ell > 0, \end{cases}$$

let $\nu \in \mathbb{N}$ and let $A \in \mathbb{R}^{\mu \times \nu}$ with $\text{rank}(A) \leq N/L^2$. Then, for every $C \leq t \leq \sqrt{N}/L$,

$$\left\| W^{(L)} \cdot \left[\prod_{i=L-1}^{\ell} D^{(i)}(z_i)W^{(i)} \right] \cdot A \right\|_{2 \rightarrow 2} \leq C \cdot \|A\|_{2 \rightarrow 2} \cdot (\sqrt{\text{rank}(A)} + t)$$

with probability at least $1 - \exp(-c \cdot t^2)$.

In particular, picking $\ell = 0$, $z_0 = \dots = z_{L-1} = x$ and $A = I_{d \times d}$, we get a pointwise upper bound on the ℓ^2 -norm of the gradient $\nabla \Phi(x)$ and hence in fact a special case of Theorem 3.4.

Last but not least, we provide bounds on the ℓ^p -norm of the gradient of a random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ at a fixed non-zero input $x_0 \in \mathbb{R}^d$ as already discussed in the beginning of this section.

Theorem 3.4. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases, with L hidden layers of width N as in Assumption 2.2 with $N \geq C \cdot L^2 \cdot \ln(eL)$ and $d \geq C$ and let $x_0 \in \mathbb{R}^d \setminus \{0\}$ be fixed.*

(1) *If $p \in [1, 2]$, we have*

$$\mathbb{P} \left(c \cdot d^{1/p} \leq \|\nabla \Phi(x_0)\|_p \leq C \cdot d^{1/p} \right) \geq 1 - \exp(-c \cdot \min\{d, N/L^2\}).$$

(2) *If $p \in (2, c \cdot \ln(d))$, we have*

$$\mathbb{P} \left(c \cdot \sqrt{p} \cdot d^{1/p} \leq \|\nabla \Phi(x_0)\|_p \leq C \cdot \sqrt{p} \cdot d^{1/p} \right) \geq 1 - \exp(-c \cdot \min\{p \cdot d^{2/p}, N/L^2\}).$$

(3) If $p \geq \ln(d)$, we have

$$\mathbb{P}\left(c \cdot \sqrt{\ln(d)} \leq \|\bar{\nabla}\Phi(x_0)\|_p \leq C \cdot \sqrt{\ln(d)}\right) \geq 1 - \exp(-c \cdot \min\{\ln(d), N/L^2\}).$$

The proof of Theorem 3.4 is based on a careful evaluation of the gradient $\bar{\nabla}\Phi(x_0)$ using the techniques from Section 2.3, combined with the following result regarding the concentration of the ℓ^p -norm of a standard Gaussian vector.

Lemma 3.5 (Consequence of [16, Proposition 2.4, Theorem 4.11]). *There exist absolute constants $C, c > 0$ such that the following holds for $d \geq C$:*

(1) For every $p \in [1, 2]$ we have

$$\mathbb{P}_{X \sim \mathcal{N}(0, I_d)}\left(c \cdot d^{1/p} \leq \|X\|_p \leq C \cdot d^{1/p}\right) \geq 1 - C \cdot \exp(-c \cdot d).$$

(2) For every $p \in (2, c \cdot \ln(d))$ we have

$$\mathbb{P}_{X \sim \mathcal{N}(0, I_d)}\left(c \cdot \sqrt{p} \cdot d^{1/p} \leq \|X\|_p \leq C \cdot \sqrt{p} \cdot d^{1/p}\right) \geq 1 - C \cdot \exp(-c \cdot p \cdot d^{2/p}).$$

(3) For every $p \geq \ln(d)$ we have

$$\mathbb{P}_{X \sim \mathcal{N}(0, I_d)}\left(c \cdot \sqrt{\ln(d)} \leq \|X\|_p \leq C \cdot \sqrt{\ln(d)}\right) \geq 1 - C \cdot \exp(-c \cdot \ln(d)).$$

Below, we provide a proof sketch for Theorem 3.4, illustrating the main ideas of the proof. The proofs of Lemmas 3.1 to 3.3 are based on a similar approach.

Proof sketch of Theorem 3.4. By definition of the formal gradient,

$$(\bar{\nabla}\Phi(x_0))^T = W^{(L)} \left[\prod_{\ell=L-1}^0 D^{(\ell)}(x_0) W^{(\ell)} \right].$$

Using Proposition 2.10, by passing to a high probability event and by using the distributional equivalence in Proposition 2.11, it suffices to study

$$W^{(L)} \left[\prod_{\ell=L-1}^0 \bar{D}^{(\ell)}(x_0) W^{(\ell)} \right] \stackrel{d}{=} W^{(L)} \left[\prod_{\ell=L-1}^0 \hat{D}^{(\ell)} W^{(\ell)} \right],$$

where the matrices $\bar{D}^{(\ell)}(x_0)$ and $\hat{D}^{(\ell)}$ are defined as in Section 2.3. Since the matrices $\hat{D}^{(\ell)}$ are independent of the weight matrices, we can condition on these matrices and assume that

$$\text{for all } \ell \in \{0, \dots, L-1\}: \quad \frac{N}{2} \left(1 - \frac{1}{4L}\right) \leq \text{Tr}(\hat{D}^{(\ell)}) \leq \frac{N}{2} \left(1 + \frac{1}{4L}\right), \quad (3.1)$$

which occurs with probability at least $1 - 2\exp(-c \cdot N/L^2)$. Because of $\text{Tr}(\hat{D}^{(\ell)}) = \sum_{j=1}^N \varepsilon_j^{(\ell)}$, this follows from Hoeffding's inequality for bounded random variables (see [20, Theorem 2.2.5]) and the assumption $N \geq C \cdot L^2 \cdot \ln(eL)$.

For fixed $\hat{D}^{(\ell)}$ -matrices, we set $\alpha_\ell := \text{Tr}(\hat{D}^{(\ell)})$. We obtain

$$\left\| W^{(L)} \left[\prod_{\ell=L-1}^0 \hat{D}^{(\ell)} W^{(\ell)} \right] \right\|_p \stackrel{d}{=} \left\| \prod_{\ell=L}^0 W_{\dagger}^{(\ell)} \right\|_p,$$

where $W_{\dagger}^{(L)} \in \mathbb{R}^{1 \times \alpha_{L-1}}$, $W_{\dagger}^{(\ell)} \in \mathbb{R}^{\alpha_\ell \times \alpha_{\ell-1}}$ for $\ell \in \{1, \dots, L-1\}$ and $W_{\dagger}^{(0)} \in \mathbb{R}^{\alpha_0 \times d}$. Further, we have $(W_{\dagger}^{(L)})_{1,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $(W_{\dagger}^{(\ell)})_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N)$ for $\ell \in \{0, \dots, L-1\}$ and the matrices $W_{\dagger}^{(0)}, \dots, W_{\dagger}^{(L)}$ are jointly independent.

We now condition on the matrices $W_{\dagger}^{(1)}, \dots, W_{\dagger}^{(L)}$ and set

$$Z := \prod_{\ell=L}^1 W_{\dagger}^{(\ell)} \in \mathbb{R}^{1 \times \alpha_0}.$$

Then,

$$(Z \cdot W_{\dagger}^{(0)})^T \sim \mathcal{N}\left(0, \frac{2\|Z\|_2^2}{N} \cdot I_d\right)$$

where the randomness is only with respect to $W_{\dagger}^{(0)}$. Hence, using Lemma 3.5, we get for instance in the case $p \in [1, 2]$ that

$$\|Z \cdot W_{\dagger}^{(0)}\|_p \asymp \frac{\|Z\|_2}{\sqrt{N}} \cdot d^{1/p}$$

with high probability. Hence, it remains to show that $\|Z\|_2 \asymp \sqrt{N}$ with high probability. This is achieved via an inductive reasoning under an iterative application of [20, Theorem 3.1.1] (concentration of the ℓ^2 -norm of a Gaussian random vector) and using (3.1) with

$$\frac{1}{2} \leq \left(1 - \frac{1}{2L}\right)^L \leq \left(1 + \frac{1}{2L}\right)^L \leq e;$$

see Appendix D for the details. $\square$

The following corollary records the consequences of Theorem 3.4 for the Lipschitz constants of random ReLU neural networks. It follows by applying Theorem 3.4 with the setting $p = 1$, using the standard inequality $\|\cdot\|_1 \leq d^{1-1/p} \cdot \|\cdot\|_p$ for $p \in [1, \infty]$ and by then invoking Theorem 2.8.

Corollary 3.6. *There exist absolute constants $C, c > 0$ such that the following holds. For a random ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ with symmetric biases, with L hidden layers of width N as in Assumption 2.2 with $d \geq C$ and $N \geq C \cdot L^2 \cdot \ln(eL)$,*

$$\text{Lip}_p(\Phi) \geq c \cdot d^{1-1/p} \quad \text{for all } p \in [2, \infty]$$

with probability at least $1 - \exp(-c \cdot \min\{N/L^2, d\})$.

4 The upper bound for zero-bias networks

In this section, we state the upper bound for the Lipschitz constant in the case of zero-bias networks and provide a proof sketch for it. We refer to Appendix E for the full proof. We note that the same upper bound was shown in [4, Theorem B.5] in the regime $N \gtrsim d^4 L \ln^4(N)$. The main result reads as follows.

Theorem 4.1. *There exist absolute constants $C, c > 0$ satisfying the following: If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random zero-bias ReLU network with L hidden layers of width N as in Assumption 2.1 and if we assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot (\ln(N/d) \cdot L^3 + \ln^2(N/d) L^2)$$

then, with probability at least $1 - \exp(-c \cdot d \ln(N/d))$, for all $\ell \in \{0, \dots, L-1\}$,

$$\text{Lip}_{2 \rightarrow 2}(\Phi^{(\ell)}) \leq \sup_{x \in \mathbb{S}^{d-1}} \|\overline{\text{D}}\Phi^{(\ell)}(x)\|_{2 \rightarrow 2} \leq C,$$

and, under the stronger assumption $N \geq C \cdot d \cdot (L^3 \ln^2(N/d) + L^2 \ln^3(N/d))$,

$$\text{Lip}_2(\Phi) \leq \sup_{x \in \mathbb{S}^{d-1}} \|\overline{\nabla}\Phi(x)\|_2 \leq C \cdot \sqrt{d \cdot \ln(N/d)}.$$

Using the standard inequalities

$$\|\cdot\|_2 \leq \|\cdot\|_p \quad \text{for } p \in [1, 2] \quad \text{and} \quad d^{1/p-1/2} \|\cdot\|_2 \leq \|\cdot\|_p \quad \text{for } p \in [2, \infty], \quad (4.1)$$

we obtain the following corollary.

Corollary 4.2. *For $p \in [1, \infty]$, we have on the same event as in Theorem 4.1, and under the stated additional assumption,*

$$\text{Lip}_p(\Phi) \leq C \cdot \begin{cases} \sqrt{d \cdot \ln(N/d)} & \text{if } p \in [1, 2], \\ d^{1-1/p} \cdot \sqrt{\ln(N/d)} & \text{if } p \in [2, \infty]. \end{cases}$$

Notice that we handle the two regimes $p \in [1, 2]$ and $p \in [2, \infty]$ by application of (4.1), which might appear to be loose at first sight. However, the resulting bound in fact matches the lower lower bounds from Corollary 3.6 and Theorem 5.1 up to a factor that is polynomial in the number of hidden layers L (this factor only appears for $p \in [1, 2]$) and logarithmic in the width N .

We now give a proof sketch for Theorem 4.1.

Proof sketch of Theorem 4.1. The basic strategy is to *inductively* bound the Lipschitz constants of the intermediate maps $\Phi^{(\ell)}$ and to then incorporate the final layer in the very last step. Following Theorem 2.8, this is achieved by bounding

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| \overline{D}\Phi^{(\ell)}(x) \right\|_{2 \rightarrow 2},$$

where $\overline{D}\Phi^{(\ell)}(x) = D^{(\ell)}(x)W^{(\ell)} \dots D^{(0)}(x)W^{(0)}$ is the formal Jacobian.

We pick $\delta \asymp \frac{d}{N} \cdot \ln(N/d)$ and $\varepsilon \asymp \delta \cdot \ln^{-1/2}(1/\delta)$ in the spirit of Proposition 2.13 and let $\mathcal{N}_\varepsilon \subseteq \mathbb{S}^{d-1}$ be an ε -net of $\mathbb{S}^{d-1}$ satisfying $\ln |\mathcal{N}_\varepsilon| \lesssim d \cdot \ln(1/\varepsilon) \asymp d \cdot \ln(N/d)$ (existence follows from [20, Corollary 4.2.13]). Combining Lemma 3.2 and a union bound, we get

$$\text{for all } x^* \in \mathcal{N}_\varepsilon : \quad \left\| \overline{D}\Phi^{(\ell)}(x^*) \right\|_{2 \rightarrow 2} \lesssim 1$$

with probability at least $1 - \exp(-c \cdot N/L^2)$ for an appropriate constant $c > 0$, where we used the assumption $N \gtrsim L^2 d \ln(N/d)$.

The remaining and more challenging part is to control the deviation $\overline{D}\Phi^{(\ell)}(x^*) - \overline{D}\Phi^{(\ell)}(y)$ for $x^* \in \mathcal{N}_\varepsilon$ and $y \in \mathbb{S}^{d-1}$ with $\|x^* - y\|_2 \leq \varepsilon$. In order to do so, we use a decomposition of this difference which already appeared in [2], namely,

$$\overline{D}\Phi^{(\ell)}(x^*) - \overline{D}\Phi^{(\ell)}(y) = \sum_{j=0}^{\ell} \left(\overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) \left(D^{(j)}(x^*) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right);$$

see Lemma E.1. To get a bound on the spectral norm of this difference, we use the triangle inequality and consider each summand individually. For a fixed $j \in \{0, \dots, \ell\}$ we use the estimate

$$\begin{aligned} & \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) \left(D^{(j)}(x^*) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right\|_{2 \rightarrow 2} \\ & \leq \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) \left(D^{(j)}(x^*) - D^{(j)}(y) \right) \right\|_{2 \rightarrow 2} \left\| \left(D^{(j)}(x^*) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right\|_{2 \rightarrow 2}. \end{aligned} \quad (4.2)$$

Note that $D := D^{(j)}(x^*) - D^{(j)}(y)$ is a diagonal matrix with values from $\{0, 1, -1\}$ on the diagonal and thus $D = D \cdot |D|$. By definition, we have

$$\begin{aligned} & \text{Tr} \left| D^{(j)}(x^*) - D^{(j)}(y) \right| \\ & = \# \left\{ i \in \{1, \dots, N\} : \text{sgn} \left(\left(W^{(j)} \Phi^{(j-1)}(x^*) \right)_i \right) \neq \text{sgn} \left(\left(W^{(j)} \Phi^{(j-1)}(y) \right)_i \right) \right\} \\ & = \# \left\{ i \in \{1, \dots, N\} : \text{sgn} \left(\left(W^{(j)} \frac{\Phi^{(j-1)}(x^*)}{\|\Phi^{(j-1)}(x^*)\|_2} \right)_i \right) \neq \text{sgn} \left(\left(W^{(j)} \frac{\Phi^{(j-1)}(y)}{\|\Phi^{(j-1)}(y)\|_2} \right)_i \right) \right\}, \end{aligned}$$

putting us in the situation of Proposition 2.13. According to Lemma 3.1 and a union bound, we have

$$\text{for all } x^* \in \mathcal{N}_\varepsilon : \quad \left\| \Phi^{(j-1)}(x^*) \right\|_2 \gtrsim 1, \quad (4.3)$$

with high probability. Further, by the induction hypothesis, we may assume that $\text{Lip}_2(\Phi^{(j-1)}) \lesssim 1$, which implies

$$\begin{aligned} \left\| \frac{\Phi^{(j-1)}(x^*)}{\left\| \Phi^{(j-1)}(x^*) \right\|_2} - \frac{\Phi^{(j-1)}(y)}{\left\| \Phi^{(j-1)}(y) \right\|_2} \right\|_2 &\stackrel{\text{Lem. I.2}}{\lesssim} \frac{\left\| \Phi^{(j-1)}(x^*) - \Phi^{(j-1)}(y) \right\|_2}{\left\| \Phi^{(j-1)}(x^*) \right\|_2} \\ &\stackrel{(4.3)}{\lesssim} \left\| \Phi^{(j-1)}(x^*) - \Phi^{(j-1)}(y) \right\|_2 \lesssim \varepsilon. \end{aligned}$$

Hence, using Proposition 2.13 we get

$$\sup_{x^* \in \mathcal{N}_\varepsilon, y \in \mathbb{S}^{d-1}, \|x^* - y\|_2 \leq \varepsilon} \text{Tr} \left| D^{(j)}(x^*) - D^{(j)}(y) \right| \leq \delta \cdot N$$

with high probability. We set

$$\mathcal{A} := \{A \in \text{diag}\{0, 1, -1\}^{N \times N} : \text{Tr} |A| \leq \delta \cdot N\}$$

and note that

$$\ln(|\mathcal{A}|) \lesssim N \cdot \delta \ln(1/\delta) \asymp d \cdot \ln^2(N/d),$$

as follows from a standard bound on the binomial coefficient; see Lemma I.6. In the spirit of (4.2), we get

$$\begin{aligned} &\sup_{x^* \in \mathcal{N}_\varepsilon, y \in \mathbb{S}^{d-1}, \|x^* - y\|_2 \leq \varepsilon} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) \left(D^{(j)}(x^*) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right\|_{2 \rightarrow 2} \\ &\leq \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) A \right\|_{2 \rightarrow 2} \right) \cdot \left(\sup_{y \in \mathbb{S}^{d-1}, A \in \mathcal{A}} \left\| A W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right\|_{2 \rightarrow 2} \right). \end{aligned}$$

Using Lemma 3.2 and a union bound, we obtain

$$\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell)}(x^*) A \right\|_{2 \rightarrow 2} \lesssim 1.$$

For the second factor, we condition on the matrices $W^{(0)}, \dots, W^{(j-1)}$ and set

$$\mathcal{V} := \left\{ \overline{D}\Phi^{(j-1)}(y) : y \in \mathbb{S}^{d-1} \right\}.$$

Since the matrices $W^{(0)}, \dots, W^{(j-1)}$ are fixed, we get

$$|\mathcal{V}| \leq \left(\frac{eN}{d+1} \right)^{L(d+1)},$$

as follows, for instance, from [13, Lemma 5.7]. Moreover, using the induction hypothesis, we may assume $\|V\|_{2 \rightarrow 2} \lesssim 1$ for every $V \in \mathcal{V}$. For fixed matrices $A \in \mathcal{A}$ and $V \in \mathcal{V}$ we thus obtain

$$\left\| A W^{(j)} V \right\|_{2 \rightarrow 2} \lesssim \|A\|_{2 \rightarrow 2} \cdot \|V\|_{2 \rightarrow 2} \cdot \frac{\sqrt{\text{rank}(A)} + \sqrt{\text{rank}(V)} + t}{\sqrt{N}} \lesssim \frac{\sqrt{\delta N} + \sqrt{d} + t}{\sqrt{N}}$$

with probability at least $1 - \exp(-c \cdot t^2)$ for every $t \gtrsim 1$; see Lemma I.5. Using a union bound and choosing $t^2 \asymp d \cdot \ln^2(N/d) + Ld \ln(N/d) \asymp \ln(|\mathcal{A}|) + \ln(|\mathcal{V}|)$, we get

$$\sup_{y \in \mathbb{S}^{d-1}, A \in \mathcal{A}} \left\| A W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right\|_{2 \rightarrow 2} \lesssim \frac{\sqrt{d \ln(N/d)} + \sqrt{Ld \ln(N/d)}}{\sqrt{N}} \lesssim 1/L$$

with high probability under the stated conditions.

Overall, this implies

$$\left\| \overline{D}\Phi^{(\ell)}(x^*) - \overline{D}\Phi^{(\ell)}(y) \right\|_{2 \rightarrow 2} \lesssim \sum_{j=0}^{\ell} \frac{1}{L} \lesssim 1,$$

uniformly over $x^* \in \mathcal{N}_\varepsilon$ and $y \in \mathbb{S}^{d-1}$ with $\|x^* - y\|_2 \leq \varepsilon$. Via induction, we can thus bound the norms of the Jacobians of the intermediate maps $\Phi^{(\ell)}$.

In order to uniformly bound the norm of the gradient $\|\bar{\nabla}\Phi(x)\|_2$, we proceed similarly as described above and use Lemma 3.3; see Appendix E for the details. $\square$

5 The lower bound for zero-bias networks

In this section, we state the lower bound of the Lipschitz constant of zero-bias random ReLU networks in the regime $p \in [1, 2]$ and explain the proof idea. The lower bound for the regime $p \in [2, \infty]$ is already contained in Corollary 3.6. For a zero-bias ReLU network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, recall that by Theorem 2.8, we aim to establish a lower bound for

$$\text{Lip}_1(\Phi) = \sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} \|\bar{\nabla}\Phi(x)\|_\infty,$$

where $M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla\Phi(x) = \bar{\nabla}\Phi(x)\}$. In fact, we provide an even stronger result: for any fixed direction $\nu \in \mathbb{S}^{d-1}$, with high probability,

$$\sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} |\langle \bar{\nabla}\Phi(x), \nu \rangle| \gtrsim \frac{\sqrt{d}}{\sqrt{L}}.$$

Taking ν as one of the standard basis vectors immediately yields the desired lower bound on the Lipschitz constant. The following is the main result of this section.

Theorem 5.1. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N as in Assumption 2.1. Assume $N \geq C \cdot d$ and*

$$N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10}, \quad \text{and} \quad d \geq C \cdot L.$$

Then, for every direction $\nu \in \mathbb{S}^{d-1}$, with probability at least $1 - \exp(-c \cdot d/L)$,

$$\sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} \langle \bar{\nabla}\Phi(x), \nu \rangle \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d},$$

where $M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla\Phi(x) = \bar{\nabla}\Phi(x)\}$. In particular, with the same probability,

$$\text{Lip}_1(\Phi) \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}.$$

As a simple corollary, since $\|\cdot\|_p \leq \|\cdot\|_1$ for any $p \geq 1$, we get the same lower bound for $\text{Lip}_p(\Phi)$ for every $p \in [1, 2)$, and in fact, for any $p \in [1, \infty]$. Note, however, that we already established the stronger lower bound of $d^{1-1/p}$ in the latter case, in Section 3.

Below, we provide a proof sketch focusing on the case of shallow networks for which $L = 1$; we refer to Appendix F for the full proof in the case of deep networks.

Proof sketch of Theorem 5.1. The overall strategy is to fix the random vector

$$\vec{W} := (W^{(0)}, \dots, W^{(L-1)})$$

and to then apply Sudakov's minoration inequality (see e.g. [20, Theorem 7.4.1]) when considering the randomness over $W^{(L)}$. To this end, we rewrite

$$\sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} |\langle \bar{\nabla}\Phi(x), \nu \rangle| = \sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} \left| \langle (W^{(L)})^T, \bar{\mathcal{D}}\Phi^{(L-1)}(x)\nu \rangle \right| = \sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle,$$

where

$$\mathcal{L} = \mathcal{L}(\vec{W}) := \left\{ \bar{\mathcal{D}}\Phi^{(L-1)}(x)\nu : x \in M_\Phi \cap \mathbb{S}^{d-1} \right\} \subseteq \mathbb{R}^N.$$

Sudakov's minoration inequality then yields

$$\sup_{x \in M_\Phi \cap \mathbb{S}^{d-1}} |\langle \bar{\nabla} \Phi(x), \nu \rangle| \gtrsim \sup_{\delta > 0} \delta \cdot \sqrt{\ln(\mathcal{P}(\mathcal{L}, \delta))}$$

in expectation, where $\mathcal{P}(\mathcal{L}, \delta)$ denotes the δ -packing number of $\mathcal{L}$. A high probability bound can be obtained using a standard result on Gaussian concentration, see [20, Theorem 5.2.2].

From now on, we focus on the case of shallow networks, since this proof turns out to be much simpler than the one for deep networks. In this case,

$$\mathcal{L} = \left\{ D^{(0)}(x) W^{(0)} \nu : x \in M_\Phi \cap \mathbb{S}^{d-1} \right\}.$$

The idea is that a δ -separated subset P of the unit sphere $\mathbb{S}^{d-1}$ directly yields a $(c\sqrt{\delta})$ -separated subset P' of the set $\mathcal{L}$, where $c > 0$ is an absolute constant, by letting

$$P' := \left\{ D^{(0)}(x^*) W^{(0)} \nu : x^* \in P \right\}.$$

To this end, one has to control expressions of the form

$$\left\| (D^{(0)}(x) - D^{(0)}(y)) W^{(0)} \nu \right\|_2$$

for inputs $x, y \in \mathbb{S}^{d-1}$ that have a Euclidean distance that is larger than δ . Applying Proposition 2.12, we can show that with high probability there exist at least $c \cdot \delta N$ non-zero (diagonal) entries in $D^{(0)}(x) - D^{(0)}(y)$ for an absolute constant $c > 0$. The main difficulty to overcome when considering the expression above is that $D^{(0)}(x) - D^{(0)}(y)$ and $W^{(0)} \nu$ are *not* independent. However, independence is trivial if we consider inputs from $\{0\} \times \mathbb{S}^{d-2}$ and pick $\nu = e_1$. The result then follows from standard bounds on the norm of a Gaussian vector, see, e.g., [20, Theorem 3.1.1]. The generalization to arbitrary $\nu \in \mathbb{S}^{d-1}$ is then obtained by using the rotation invariance of the Gaussian distribution; see Proposition F.11.

Note that this trick to create independence does *not* work anymore in the case of deep networks and more advanced techniques are required. We refer to Appendix F for the details. $\square$

6 Networks with biases

The results presented in Sections 4 and 5 solely dealt with *zero-bias* networks, i.e., networks where all the biases are set to zero. In this section, we discuss the case of networks with *general* biases.

A key observation is that for inputs with a sufficiently large Euclidean norm, the gradient of Φ is similar to the gradient of $\tilde{\Phi}$, the zero-bias network associated to Φ by setting all its biases to zero. To make this intuition precise, we aim to bound

$$\sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \left\| \bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x) \right\|_2$$

from above using Proposition 2.14, where $R > 0$ is chosen appropriately. This is discussed in more detail in Section 6.1. After such a bound is established, we can apply the results from Sections 3 to 5 to obtain bounds on the Lipschitz constant for networks with general biases. More precisely, in order to obtain lower bounds, it suffices to study the gradient of inputs with a sufficiently large Euclidean norm. This gradient is similar to the gradient of the associated homogeneous network, which we can lower bound using Corollary 3.6 and Theorem 5.1. For an upper bound, it essentially suffices to bound

$$\sup_{x \in B_d(0, R)} \left\| \bar{\nabla} \Phi(x) \right\|_2,$$

since the gradient of Φ is similar to the gradient of $\tilde{\Phi}$ on $\mathbb{R}^d \setminus B_d(0, R)$.

6.1 Bounding the difference to the gradient of the associated zero-bias network

In this section, we discuss how the difference between the gradient of a network Φ with possibly non-zero, *deterministic* biases can be related to the gradient of the associated homogeneous network $\tilde{\Phi}$. For detailed proofs of the results presented in the present section, we refer to Appendix G.

Obtaining such a bound crucially relies on the following observation.

Lemma 6.1. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network of width N and depth L as in Assumption 2.1 with constant (deterministic) biases. Let $\lambda' > 0$ such that*

$$\max_{\substack{\ell \in \{0, \dots, L-1\} \\ i \in \{1, \dots, N\}}} |b_i^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}}.$$

Assume $N \geq C \cdot d$ and $N \geq C \cdot L^3 d \ln(N/d)$ and, let $\delta \in (0, e^{-1})$, satisfy

$$N \geq C \cdot d L^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot \max\{d \cdot \ln(1/\delta), dL\}, \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c/L^2.$$

Pick $j \in \{0, \dots, L-1\}$. Then, with probability at least $1 - \exp(-c \cdot \delta N)$,

for all $x \in \mathbb{R}^d$ with $\|x\|_2 \geq C \cdot 3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{1/2}(1/\delta)$:

$$\#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left(\left(W^{(j)}\Phi^{(j-1)}(x)\right)_i + b_i^{(j)}\right) \neq \operatorname{sgn}\left(\left(W^{(j)}\tilde{\Phi}^{(j-1)}(x)\right)_i\right)\right\} \leq \delta \cdot N.$$

Proof sketch. Using [20, Corollary 7.3.3], we have

$$\mathbb{P}\left(\text{for all } \ell \in \{0, \dots, L-1\} : \left\|W^{(\ell)}\right\|_{2 \rightarrow 2} \leq 3\right) \geq 1 - \exp(-c \cdot N).$$

On this event, we get for arbitrary $x \in \mathbb{R}^d$ and $\ell \in \{0, \dots, L-1\}$, using that the ReLU is 1-Lipschitz,

$$\begin{aligned} \left\|\Phi^{(\ell)}(x) - \tilde{\Phi}^{(\ell)}(x)\right\|_2 &= \left\|\operatorname{ReLU}\left(W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)}\right) - \operatorname{ReLU}\left(W^{(\ell)}\tilde{\Phi}^{(\ell-1)}(x)\right)\right\|_2 \\ &\leq \left\|W^{(\ell)}\Phi^{(\ell-1)}(x) + b^{(\ell)} - W^{(\ell)}\tilde{\Phi}^{(\ell-1)}(x)\right\|_2 \\ &\leq 3 \cdot \left\|\Phi^{(\ell-1)}(x) - \tilde{\Phi}^{(\ell-1)}(x)\right\|_2 + \left\|b^{(\ell)}\right\|_2 \\ &\leq 3 \cdot \left\|\Phi^{(\ell-1)}(x) - \tilde{\Phi}^{(\ell-1)}(x)\right\|_2 + \lambda' \sqrt{2} \\ &\leq 3 \cdot \left(3 \cdot \left\|\Phi^{(\ell-2)}(x) - \tilde{\Phi}^{(\ell-2)}(x)\right\|_2 + \lambda' \sqrt{2}\right) + \lambda' \sqrt{2} \\ &\leq \lambda' \sqrt{2} \cdot \sum_{j=0}^{\ell} 3^j \leq 3^L \cdot \lambda', \end{aligned}$$

where the fifth inequality follows from induction.

Based on this observation, we can derive the claim: Note that

$$\begin{aligned} &\#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left(\left(W^{(j)}\Phi^{(j-1)}(x)\right)_i + b_i^{(j)}\right) \neq \operatorname{sgn}\left(\left(W^{(j)}\tilde{\Phi}^{(j-1)}(x)\right)_i\right)\right\} \\ &\leq \#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left(\left(W^{(j)}\Phi^{(j-1)}(x)\right)_i + b_i^{(j)}\right) \neq \operatorname{sgn}\left(\left(W^{(j)}\Phi^{(j-1)}(x)\right)_i\right)\right\} \\ &\quad + \#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left(\left(W^{(j)}\Phi^{(j-1)}(x)\right)_i\right) \neq \operatorname{sgn}\left(\left(W^{(j)}\tilde{\Phi}^{(j-1)}(x)\right)_i\right)\right\}. \end{aligned}$$

The first summand can be bounded by $\delta N/2$ using Proposition 2.14, noting that

$$\begin{aligned} \left\|\Phi^{(j-1)}(x)\right\|_2 &\geq \left\|\tilde{\Phi}^{(j-1)}(x)\right\|_2 - \left\|\Phi^{(j-1)}(x) - \tilde{\Phi}^{(j-1)}(x)\right\|_2 \\ &\stackrel{\text{Cor. E.4}}{\gtrsim} 3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{1/2}(1/\delta) - 3^L \cdot \lambda' \\ &\gtrsim \delta^{-1} \cdot \lambda'. \end{aligned}$$

For the second summand, we apply Proposition 2.13 since we have

$$\begin{aligned}
\left\| \frac{\Phi^{(j-1)}(x)}{\|\Phi^{(j-1)}(x)\|_2} - \frac{\tilde{\Phi}^{(j-1)}(x)}{\|\tilde{\Phi}^{(j-1)}(x)\|_2} \right\|_2 &\stackrel{\text{Lem. I.2}}{\lesssim} \frac{\|\Phi^{(j-1)}(x) - \tilde{\Phi}^{(j-1)}(x)\|_2}{\|\tilde{\Phi}^{(j-1)}(x)\|_2} \\
&\stackrel{\text{Cor. E.4}}{\lesssim} \frac{3^L \cdot \lambda'}{\|x\|_2} \\
&\lesssim \frac{3^L \cdot \lambda'}{3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{1/2}(1/\delta)} = \delta \cdot \ln^{-1/2}(1/\delta).
\end{aligned}$$

Hence, we can also bound the second summand by $\delta N/2$, which implies the claim. $\square$

Based on Lemma 6.1, we can then prove the main theorem of this subsection.

Theorem 6.2. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.1 of width N and depth L with constant biases. Let $\lambda' > 0$ satisfy*

$$\max_{\substack{\ell=\{0,\dots,L-1\} \\ i=\{1,\dots,N\}}} |b_i^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}}.$$

Assume

$$N \geq C \cdot dL \quad \text{and} \quad N \geq C \cdot L^3 d \ln(N/d) \ln(N/(dL)).$$

Then, with probability at least $1 - \exp(-c \cdot dL \ln(N/d))$,

$$\sup_{\|x\|_2 \geq R} \left\| \bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x) \right\|_2 \leq C \cdot \frac{dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}},$$

where $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$.

Proof sketch. We write $\tilde{D}^{(\ell)}(x)$ for the D -matrices arising by passing an input x through the homogeneous network $\tilde{\Phi}$.

In the setting of Lemma 6.1, we let $\delta \asymp \frac{dL}{N} \cdot \ln(N/d)$ and hence get

$$3^L \cdot \lambda \cdot \delta^{-1} \cdot \ln^{1/2}(1/\delta) \lesssim 3^L \cdot \lambda \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d).$$

Similar to the proof of Theorem 4.1, we use a decomposition of the difference of the gradients, namely

$$\begin{aligned}
&\sup_{\|x\|_2 \geq R} \left\| \bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x) \right\|_2 \\
&\leq \sum_{j=0}^{L-1} \left(\sup_{\|x\|_2 \geq R} \left\| W^{(L)} \bar{D} \Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - \tilde{D}^{(j)}(x) \right) \bar{D} \tilde{\Phi}^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\
&\leq \sum_{j=0}^{L-1} \left(\sup_{\|x\|_2 \geq R} \left\| W^{(L)} \bar{D} \Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - \tilde{D}^{(j)}(x) \right) \right\|_{2 \rightarrow 2} \right) \\
&\quad \cdot \left(\sup_{\|x\|_2 \geq R} \left\| D^{(j)}(x) - \tilde{D}^{(j)}(x) \right\| \bar{D} \tilde{\Phi}^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right). \tag{6.1}
\end{aligned}$$

We set $\mathcal{A} := \{A \in \text{diag}\{1, -1, 0\}^{N \times N} : \text{Tr}|A| \leq \delta N\}$. Firstly, note that according to Lemma 6.1, we may assume that with high probability $D^{(j)}(x) - \tilde{D}^{(j)}(x) \in \mathcal{A}$ for every $j \in \{0, \dots, L-1\}$.

Each summand of the latter sum is then considered individually. We fix $j \in \{0, \dots, L-1\}$ and get

$$\sup_{\|x\|_2 \geq R} \left\| W^{(L)} \bar{D} \Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - \tilde{D}^{(j)}(x) \right) \right\|_{2 \rightarrow 2}$$

$$\begin{aligned}
&\leq \sup_{\|x\|_2 \geq R, A \in \mathcal{A}} \left\| W^{(L)} \bar{D}\Phi^{(j+1) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\
&\leq \sup_{\|x\|_2 \geq R, A \in \mathcal{A}} \left\| W^{(L)} \left[\bar{D}\Phi^{(j+1) \rightarrow (L-1)}(x) - \bar{D}\tilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) \right] A \right\|_{2 \rightarrow 2} \\
&\quad + \sup_{\|x\|_2 \geq R, A \in \mathcal{A}} \left\| W^{(L)} \bar{D}\tilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2}.
\end{aligned}$$

The first summand can be bounded using an inductive argument. The second summand that only involves the homogeneous network can be bounded using a generalized version of Theorem 4.1; see Theorem E.2. In total we get a bound on the first factor in (6.1).

Note that the second factor in (6.1) only involves the homogeneous network $\tilde{\Phi}$. We may thus use

$$\sup_{\|x\|_2 \geq R} \left\| \left(D^{(j)}(x) - \tilde{D}^{(j)}(x) \right) \bar{D}\tilde{\Phi}^{(j-1)}(x) \right\|_{2 \rightarrow 2} \leq \sup_{\|x\|_2 \geq R, A \in \mathcal{A}} \left\| A \bar{D}\tilde{\Phi}^{(j-1)}(x) \right\|_{2 \rightarrow 2}$$

and bound the latter expression analogously to the technique presented in the proof sketch of Theorem 4.1. Finally, we perform a union bound over $j \in \{0, \dots, L-1\}$ to get the desired result. $\square$

6.2 Bounds for the Lipschitz constant

In this section, we discuss how Theorem 6.2 may be used to establish upper and lower bounds on the Lipschitz constant of random ReLU networks with non-zero biases. Note that we assumed the biases to be constant in Theorem 6.2. When considering randomly drawn biases, the strategy is hence to *condition* on the biases and to assume that they are upper bounded in absolute value by $\frac{\lambda' \sqrt{2}}{\sqrt{N}}$ for some $\lambda' \geq 0$. We refer to Appendix H for detailed proofs of all the results presented in this section.

Our discussion starts with an upper bound. For this we need to assume that the biases are drawn according to Assumption 2.3, i.e., we assume that the biases are drawn according to symmetric distributions and satisfy a small-ball property. Note again that this is equivalent to all the entries of the bias vectors having absolutely continuous distributions with density bounded almost everywhere.

Theorem 6.3. *There exist constants $C, c > 0$ with the following property: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with symmetric biases that satisfy a small-ball condition with constant $C_\tau > 0$, and with L hidden layers of width N as in Assumption 2.3. Moreover, let*

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \lambda \right) \geq 1 - \exp(-d \ln(N/d))$$

for some $\lambda \geq C_\tau^{-1}$. We assume that

$$N \geq C \cdot d, \quad N \geq C \cdot dL^2 \cdot (\max\{L, \ln(\lambda C_\tau N/d)\})^2 \cdot \ln(N/d), \quad N \geq C \cdot dL^3 \ln^2(N/d) \ln^2(N/(dL)).$$

Then, with probability at least $1 - \exp(-c \cdot d \ln(N/d))$,

$$\text{Lip}_2(\Phi) \leq \sup_{x \in \mathbb{R}^d} \|\bar{\nabla}\Phi(x)\|_2 \leq C \cdot \sqrt{d} \cdot \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right).$$

Proof sketch. We set $\lambda' := \sqrt{N/2} \cdot \lambda$. As in Theorem 6.2, we set $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$ and use the estimate

$$\begin{aligned}
\sup_{x \in \mathbb{R}^d} \|\bar{\nabla}\Phi(x)\|_2 &\leq \sup_{x \in B_d(0, R)} \|\bar{\nabla}\Phi(x)\|_2 + \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \left\| \bar{\nabla}\Phi(x) - \bar{\nabla}\tilde{\Phi}(x) \right\|_2 + \sup_{x \in \mathbb{R}^d} \left\| \bar{\nabla}\tilde{\Phi}(x) \right\|_2 \\
&= \sup_{x \in B_d(0, R)} \|\bar{\nabla}\Phi(x)\|_2 + \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \left\| \bar{\nabla}\Phi(x) - \bar{\nabla}\tilde{\Phi}(x) \right\|_2 + \sup_{x \in \mathbb{S}^{d-1}} \left\| \bar{\nabla}\tilde{\Phi}(x) \right\|_2, \quad (6.2)
\end{aligned}$$

where the last equality uses the fact that the formal gradient of the homogeneous network $\tilde{\Phi}$ is invariant with respect to scaling with positive scalars. The second summand can be controlled by conditioning

on the biases and using Theorem 6.2 and the third summand can be bounded using Theorem 4.1. The first summand $\sup_{x \in B_d(0, R)} \|\nabla \Phi(x)\|_2$ can be estimated similarly to the proof of Theorem 4.1 using Proposition 2.15 instead of Proposition 2.13. $\square$

As in Section 4, this result can be used to get an upper bound on $\text{Lip}_p(\Phi)$ for every $p \in [1, \infty]$.

Corollary 6.4. *On the same event as in Theorem 6.3, for every $p \in [1, \infty]$,*

$$\text{Lip}_p(\Phi) \leq C \cdot \begin{cases} \sqrt{d} \cdot \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right) & \text{if } p \in [1, 2], \\ d^{1-1/p} \cdot \left(\sqrt{L} + \sqrt{\ln(\lambda C_\tau N/d)} \right) & \text{if } p \in [2, \infty]. \end{cases}$$

We emphasize that, unlike in the case of zero-bias networks discussed in Corollary 4.2, the upper bound derived in Corollary 6.4 is in fact not entirely independent of the number of hidden layers L . However, the dependence is not exponential as in [13, Theorem 3.3], but only a factor of $\sqrt{L}$ appears. This is due to the fact that bounding the first summand in (6.2) requires an ε -net of $B_d(0, R)$ where R grows exponentially in L , so that the size of the net grows exponentially in L , too. Taking the square root of the logarithm then results in the factor $\sqrt{L}$.

Remark 6.5. (1) We note that the statement of Theorem 6.3 is *invariant* under scaling of the biases, i.e., under replacing $b^{(\ell)}$ by $\alpha \cdot b^{(\ell)}$ for each $\ell \in \{0, \dots, L-1\}$ for a fixed scaling parameter $\alpha > 0$. This is due to the fact that such a scaling changes the parameter λ appearing in Theorem 6.3 by a factor of α and changes the parameter C_τ by a factor of $1/\alpha$.
(2) Let $\mathcal{D}_\tau$ be a real-valued, symmetric, sub-exponential random distribution with small ball constant $C_\tau > 0$. Note that this in particular includes the case that $\mathcal{D}_\tau$ is sub-gaussian. Assume that

$$b_j^{(\ell)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_\tau \quad \text{for } \ell \in \{0, \dots, L-1\} \text{ and } j \in \{1, \dots, N\}.$$

Note that this implies

$$\sup_{\ell \in \{0, \dots, L-1\}, j \in \{1, \dots, N\}} |b_j^{(\ell)}| \leq t$$

with probability at least $1 - \exp(\ln(NL) - c \cdot t / \|\mathcal{D}_\tau\|_{\psi_1})$ for every $t \geq 0$ for an absolute constant $c > 0$. If $N \gtrsim \ln(NL)$, we thus get

$$\sup_{\ell \in \{0, \dots, L-1\}, j \in \{1, \dots, N\}} |b_j^{(\ell)}| \lesssim N \cdot \|\mathcal{D}_\tau\|_{\psi_1}$$

with probability at least $1 - \exp(-N) \geq 1 - \exp(-d \ln(N/d))$. Using Theorem 6.3, we thus get

$$\text{Lip}_2(\Phi) \lesssim \sqrt{d} \cdot \left(\sqrt{L} + \sqrt{\ln(\|\mathcal{D}_\tau\|_{\psi_1} C_\tau N/d)} \right).$$

with probability at least $1 - \exp(-c \cdot d \ln(N/d))$. Note that this in particular covers the cases that $\mathcal{D}_\tau = \mathcal{N}(0, \sigma^2)$ and $\mathcal{D}_\tau = \text{Unif}([- \sigma, \sigma])$ for some $\sigma > 0$. $\diamond$

Fortunately, in the case of lower bounds we can generalize the results from Sections 3 and 5 to the case of networks with general biases without *any* special assumptions on the distribution of the biases. This is due to the fact that, when lower bounding the gradient, we may restrict ourselves to inputs with a potentially very large Euclidean norm. Note that, no matter what distribution the biases have, we can always for every $\eta > 0$ find $\lambda' = \lambda'(\eta) > 0$ such that

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}} \right) \geq 1 - \eta.$$

We start by stating the main theorem about the Lipschitz constant with respect to the ℓ^1 -norm.

Theorem 6.6. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N as in Assumption 2.1. Assume $N \geq C \cdot d$,*

$$N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10}, \quad N \geq C \cdot d \cdot L^5 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL)), \quad \text{and} \quad d \geq C \cdot L.$$

Then, with probability at least $1 - \exp(-c \cdot d/L)$,

$$\text{Lip}_1(\Phi) \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}.$$

Proof sketch. We first pick $\lambda' > 0$ such that

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}} \right) \geq 1 - \exp(-d/L).$$

We then condition on this high probability event. As in Theorem 6.2, we set $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$. With

$$M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \nabla \Phi(x) = \bar{\nabla} \Phi(x)\},$$

we get, using Theorem 2.8,

$$\begin{aligned} \text{Lip}_1(\Phi) &= \sup_{x \in M_\Phi} \|\bar{\nabla} \Phi(x)\|_\infty \geq \sup_{x \in M_\Phi \cap R\mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x)\|_\infty \\ &\geq \sup_{x \in M_\Phi \cap R\mathbb{S}^{d-1}} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty - \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \\ &= \sup_{x \in \mathbb{S}^{d-1} \cap R^{-1}M_\Phi} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty - \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2, \end{aligned}$$

where the last equality is due to scale invariance of $\bar{\nabla} \tilde{\Phi}$. The expression $\sup_{x \in \mathbb{S}^{d-1} \cap R^{-1}M_\Phi} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty$ can be lower bounded as presented in Section 5 and the second expression is upper bounded using Theorem 6.2. For the full proof, we refer to Appendix H.2. $\square$

The bound presented in Theorem 6.6 immediately yields the same lower bound for $\text{Lip}_p(\Phi)$ for all $p \in [1, \infty]$. In the case $p \in [2, \infty]$, we recall that the stronger lower bound of $d^{1-1/p}$ was established in Section 3 by studying the gradient at a single fixed non-zero input. However, the techniques used in Section 3 are restricted to networks with *symmetric* biases. Using Theorem 6.2, however, we can establish the same lower bound for networks with arbitrary, possibly non-symmetric biases.

Theorem 6.7. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N as in Assumption 2.1. We assume*

$$N \geq C \cdot dL, \quad N \geq C \cdot d \cdot L^4 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL)), \quad \text{and} \quad d \geq C.$$

Then, with probability at least $1 - \exp(-c \cdot d)$,

$$\text{Lip}_p(\Phi) \geq c \cdot d^{1-1/p} \quad \text{for all } p \in [2, \infty].$$

Since the proof of Theorem 6.7 is very similar to the proof of Theorem 6.6, we omit a proof sketch here and refer to Appendix H.2 for a detailed proof.

We emphasize that there is still a qualitative difference between the results from Section 3 and Theorem 6.7: in Theorem 3.4, it is shown that at an *arbitrary* and *fixed* non-zero input $x_0 \in \mathbb{R}^d$ we have

$$\|\bar{\nabla} \Phi(x_0)\|_{p'} \gtrsim d^{1-1/p}$$

with high probability, where $1/p + 1/p' = 1$. The proof of Theorem 6.7 recovers that lower bound for networks with possibly non-symmetrically distributed biases, but only for inputs x_0 with sufficiently large Euclidean norm (exponential in the number of hidden layers L).

6.3 Matching bounds for shallow networks

From [13, Theorems 3.1&3.2], we recall that for a *shallow* random ReLU network following Assumption 2.1,

$$\text{Lip}_2(\Phi) \asymp \sqrt{d}$$

with high probability. Using Theorems 6.6 and 6.7 and standard norm inequalities, we can generalize this estimate to arbitrary ℓ^p -norms.

Corollary 6.8. *There exist absolute constants $C, c > 0$ with the following property. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a shallow random ReLU network with N neurons in the hidden layer as in Assumption 2.1. We assume*

$$N \geq C \cdot d, \quad N \geq C \cdot d \cdot \ln^4(N/d), \quad \text{and} \quad d \geq C.$$

Then, with probability $1 - \exp(-c \cdot d)$,

$$c \cdot \sqrt{d} \leq \text{Lip}_p(\Phi) \leq C \cdot \sqrt{d} \quad \text{for } p \in [1, 2]$$

and

$$c \cdot d^{1-1/p} \leq \text{Lip}_p(\Phi) \leq C \cdot d^{1-1/p} \quad \text{for } p \in [2, \infty].$$

Hence, in the case of shallow networks, we obtain bounds that match up to an absolute multiplicative constant.

Acknowledgments. FV and PG are grateful for the hospitality of Utrecht University, where parts of this project were developed. FV and PG acknowledge support by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1. FV acknowledges support by the Hightech Agenda Bavaria.

References

- [1] Z. Allen-Zhu, Y. Li, and Z. Song. *A Convergence Theory for Deep Learning via Over-Parameterization*. 2019. arXiv: 1811.03962.
- [2] P. Bartlett, S. Bubeck, and Y. Cherapanamjeri. “Adversarial examples in multi-layer random ReLU networks.” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9241–9252.
- [3] J. Berner, D. Elbrächter, P. Grohs, and A. Jentzen. “Towards a regularity theory for ReLU networks—chain rule and global error estimates.” In: *2019 13th International conference on Sampling Theory and Applications (SampTA)*. IEEE. 2019, pp. 1–5.
- [4] S. Buchanan, D. Gilboa, and J. Wright. “Deep Networks and the Multiple Manifold Problem.” In: *International Conference on Learning Representations*. 2021.
- [5] S. Dirksen, M. Genzel, L. Jacques, and A. Stollenwerk. “The separation capacity of random neural networks.” In: *Journal of Machine Learning Research* 23.309 (2022), pp. 1–47.
- [6] S. Dirksen and S. Mendelson. “Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing.” In: *Journal of the European Mathematical Society* 23.9 (2021), pp. 2913–2947.
- [7] S. Dirksen, S. Mendelson, and A. Stollenwerk. “Sharp estimates on random hyperplane tessellations.” In: *SIAM Journal on Mathematics of Data Science* 4.4 (2022), pp. 1396–1419.
- [8] R. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. ISBN: 9780521007542.
- [9] Y. Ebihara, X. Dai, T. Yuno, V. Magron, D. Peaucelle, and S. Tarbouriech. “Local Lipschitz constant computation of ReLU-FNNs: Upper bound computation with exactness verification.” In: *2024 European Control Conference (ECC)*. IEEE. 2024, pp. 2506–2511.
- [10] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. Boca Raton: CRC Press, 1992. 268 pp. ISBN: 978-0-8493-7157-8.
- [11] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. “Efficient and accurate estimation of lipschitz constants for deep neural networks.” In: *Advances in neural information processing systems* 32 (2019).
- [12] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1984. ISBN: 9780471809586.

- [13] P. Geuchen, D. Stöger, T. Telaar, and F. Voigtlaender. “Upper and lower bounds for the Lipschitz constant of random neural networks.” In: *Information and Inference: A Journal of the IMA* 14.2 (Apr. 2025). DOI: 10.1093/imaiai/iaaf009.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.
- [15] S. Oymak and B. Recht. *Near-Optimal Bounds for Binary Embeddings of Arbitrary Sets*. 2015. arXiv: 1512.04433.
- [16] G. Paouris, P. Valettas, and J. Zinn. “Random version of Dvoretzky’s theorem in ℓ_p^n .” In: *Stochastic Processes and their Applications* 127.10 (2017), pp. 3187–3227. DOI: 10.1016/j.spa.2017.02.007.
- [17] Y. Plan and R. Vershynin. “Dimension reduction by random hyperplane tessellations.” In: *Discrete & Computational Geometry* 51.2 (2014), pp. 438–461.
- [18] R. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 2015. ISBN: 9781400873173.
- [19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. “Intriguing properties of neural networks.” In: *2nd International Conference on Learning Representations, ICLR*. 2014.
- [20] R. Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge series in statistical and probabilistic mathematics 47. Cambridge ; New York, NY: Cambridge University Press, 2018. 284 pp. ISBN: 978-1-108-41519-4.

(S. Dirksen) MATHEMATICAL INSTITUTE, UTRECHT UNIVERSITY, BUDAPESTLAAN 6, 3584 CD UTRECHT, NETHERLANDS
Email address: s.dirksen@uu.nl

(P. Finke) MATHEMATICAL INSTITUTE, UTRECHT UNIVERSITY, BUDAPESTLAAN 6, 3584 CD UTRECHT, NETHERLANDS
Email address: p.g.finke@uu.nl

(P. Geuchen) MATHEMATICAL INSTITUTE FOR MACHINE LEARNING AND DATA SCIENCE (MIDS), CATHOLIC UNIVERSITY OF EICHSTÄTT–INGOLSTADT (KU), AUF DER SCHANZ 49, 85049 INGOLSTADT, GERMANY
Email address: paul.geuchen@ku.de

(D. Stöger) MATHEMATICAL INSTITUTE FOR MACHINE LEARNING AND DATA SCIENCE (MIDS), CATHOLIC UNIVERSITY OF EICHSTÄTT–INGOLSTADT (KU), AUF DER SCHANZ 49, 85049 INGOLSTADT, GERMANY
Email address: dominik.stoeger@ku.de

(F. Voigtlaender) MATHEMATICAL INSTITUTE FOR MACHINE LEARNING AND DATA SCIENCE (MIDS), CATHOLIC UNIVERSITY OF EICHSTÄTT–INGOLSTADT (KU), AUF DER SCHANZ 49, 85049 INGOLSTADT, GERMANY
Email address: felix.voigtlaender@ku.de

Appendix

Appendix Contents

A. Proofs of preliminary results regarding the Jacobian of ReLU networks	26
B. Proof of distributional equivalence	30
C. Proofs for the results on random hyperplane tessellations	34
D. Proof of pointwise bounds	42
E. Proof of Theorem 4.1	51
F. Proof of Theorem 5.1	61
G. Proof of Theorem 6.2	82
H. Bounds on the Lipschitz constants for networks with biases	90
I. Other auxiliary results	101

A Proofs of preliminary results regarding the Jacobian of ReLU networks

In this section, we provide proofs for the statements from Section 2. We start with Proposition 2.6, which states that every intermediate layer $\Phi^{(\ell)}$ of a ReLU network is almost everywhere differentiable and the true Jacobian coincides almost everywhere with the formal Jacobian introduced in Definition 2.5. This proof is essentially the same as the one in [3, Theorem III.1].

Proof of Proposition 2.6. The proof is by induction over $\ell \in \{-1, \dots, L-1\}$. For $\ell = -1$, there is nothing to show.

Next, let $\ell \in \{0, \dots, L-1\}$ and assume that there exists a set $\tilde{N} \subseteq \mathbb{R}^d$ with $\mathbb{A}^d(\tilde{N}) = 0$ and such that $\Phi^{(\ell-1)}$ is differentiable on $\mathbb{R}^d \setminus \tilde{N}$ with

$$D\Phi^{(\ell-1)}(x) = \overline{D}\Phi^{(\ell-1)}(x) \quad \text{for all } x \in \mathbb{R}^d \setminus \tilde{N}.$$

We fix an index $j \in \{1, \dots, N\}$ and define

$$Z_j := \left\{ x \in \mathbb{R}^d : \Phi_j^{(\ell)}(x) = 0 \right\}.$$

As the composition of affine maps and the componentwise application of the ReLU, $\Phi_j^{(\ell)}$ is Lipschitz continuous. Hence, we can apply [10, Corollary 1 on p.84] and conclude the existence of a set $N_j \subseteq Z_j$ with $\mathbb{A}^d(N_j) = 0$ such that $\Phi_j^{(\ell)}$ is differentiable on $Z_j \setminus N_j$ with

$$\left(D\Phi_j^{(\ell)} \right) (z) = 0 = \left(D^{(\ell)}(z) \cdot W^{(\ell)} \dots D^{(0)}(z) \cdot W^{(0)} \right)_{j,:} = \left(\overline{D}\Phi^{(\ell)}(x) \right)_{j,:},$$

where the second equality stems from $D^{(\ell)}(z)_{j,j} = 0$ for $z \in Z_j$.

It remains to analyze the derivative on $\mathbb{R}^d \setminus Z_j$. Let

$$\mu := \begin{cases} d & \text{if } \ell = 0 \\ N & \text{if } \ell > 0. \end{cases}$$

We decompose

$$\Phi_j^{(\ell)} = \varphi_j \circ \Phi^{(\ell-1)} \quad \text{with} \quad \varphi_j : \mathbb{R}^\mu \rightarrow \mathbb{R}, \quad \varphi_j(z) = \text{ReLU}(W_{j,:}^{(\ell)} \cdot z + b_j^{(\ell)}).$$

Note that $\widetilde{Z}_j := \{z \in \mathbb{R}^d : \varphi_j(z) > 0\} \subseteq \mathbb{R}^d$ is open and therefore, φ_j is differentiable on $\widetilde{Z}_j$ with

$$D\varphi_j(z) = W_{j,:}^{(\ell)} \quad \text{for } z \in \widetilde{Z}_j.$$

Recall by the induction hypothesis that $\Phi^{(\ell-1)}$ is differentiable on $\mathbb{R}^d \setminus \widetilde{N}$ with

$$D\Phi^{(\ell-1)}(x) = \overline{D}\Phi^{(\ell-1)}(x), \quad x \in \mathbb{R}^d \setminus \widetilde{N}.$$

Note by definition that $x \in \mathbb{R}^d \setminus Z_j$ implies $\Phi^{(\ell-1)}(x) \in \widetilde{Z}_j$. We employ the chain rule and get

$$\begin{aligned} \left(D\Phi_j^{(\ell)}\right)(x) &= (D\varphi_j)\left(\underbrace{\Phi^{(\ell-1)}(x)}_{\in \widetilde{Z}_j}\right) \cdot \left(D\Phi^{(\ell-1)}\right)(x) = W_{j,:}^{(\ell)} \cdot \overline{D}\Phi^{(\ell-1)}(x) \\ &= \left(D^{(\ell)}(x) \cdot W^{(\ell)} \cdot \overline{D}\Phi^{(\ell-1)}(x)\right)_{j,:} = \left(\overline{D}\Phi^{(\ell)}(x)\right)_{j,:}, \end{aligned}$$

for every $x \in \mathbb{R}^d \setminus (\widetilde{N} \cup Z_j)$, where the third equality follows from $(D^{(\ell)}(x))_{j,j} = 1$ for $x \in \mathbb{R}^d \setminus Z_j$.

Overall, we define $\mathcal{N} := \widetilde{N} \cup \bigcup_{j=1}^N N_j$, note $\mathfrak{N}^d(\mathcal{N}) = 0$, and get

$$D\Phi^{(\ell)}(x) = \overline{D}\Phi^{(\ell)}(x), \quad x \in \mathbb{R}^d \setminus \mathcal{N},$$

which concludes the induction step.

The conclusion for the final layer follows by noting that

$$\left(D\Phi^{(L-1)}\right)(x) = D^{(L-1)}(x)W^{(L-1)} \dots D^{(0)}(x)W^{(0)} \quad \text{for almost every } x \in \mathbb{R}^d$$

and an application of the chain rule. $\square$

Next, we aim to prove Proposition 2.4 which is well-known and was for the case of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $p = 2$ already shown in [13, Theorem 4.2]. We include the proof here again to show that a similar reasoning also works in the case of functions $f : K \rightarrow \mathbb{R}^k$, where $K \subseteq \mathbb{R}^d$ is convex and has non-empty interior, and arbitrary $p, q \in [1, \infty]$.

Proof of Proposition 2.4. By [18, Theorem 6.3], we conclude

$$\overline{\text{int}(K)} = \overline{K}.$$

Note here that the interior of K coincides with its relative interior (as defined on [18, p. 44]) since we assume that K has non-empty interior. Hence, since f is continuous by assumption, we get

$$\text{Lip}_{p \rightarrow q}\left(f|_{\text{int}(K)}\right) = \text{Lip}_{p \rightarrow q}\left(f|_{\overline{\text{int}(K)}}\right) = \text{Lip}_{p \rightarrow q}\left(f|_{\overline{K}}\right) = \text{Lip}_{p \rightarrow q}\left(f|_K\right).$$

Hence, from now on we may assume without loss of generality that K is open.

We first fix $i \in \{1, \dots, k\}$. With f_i denoting the i -th component of f , we have $f_i \in W_{\text{loc}}^{1,\infty}(\mathbb{R}^d)$ by [10, Section 4.2.3], where $W_{\text{loc}}^{1,\infty}(\mathbb{R}^d)$ denotes the set of functions that are locally (on every open bounded set $U \subseteq \mathbb{R}^d$) in the Sobolev space $W^{1,\infty}$. Then, [10, Theorem 1 in Section 6.2] implies that there exists a null set $N_i \subseteq \mathbb{R}^d$ such that for every $x \in \mathbb{R}^d \setminus N_i$, f_i is differentiable at x , and the partial derivatives of f_i agree with the weak partial derivatives of f_i on $\mathbb{R}^d \setminus N_i$. For $j \in \{1, \dots, d\}$ let $\widetilde{\partial}_j f_i$ be an explicit weak j -th partial derivative of f_i satisfying

$$\widetilde{\partial}_j f_i = \partial_j f_i \quad \text{on } \mathbb{R}^d \setminus N_i,$$

where ∂_j denotes the usual j -th partial derivative. We further write

$$\widetilde{D}f := \left(\widetilde{\partial}_j f_i\right)_{\substack{i=1,\dots,k \\ j=1,\dots,d}} \in \mathbb{R}^{k \times d}.$$

With $N := \bigcup_{i=1}^k N_i$, we note that $N \subseteq \mathbb{R}^d$ is a null set and

$$Df = \widetilde{D}f \quad \text{on } M \setminus N.$$

Now, let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be a smooth function with compact support satisfying $\varphi \geq 0$ and furthermore $\int_{\mathbb{R}^d} \varphi(x) \, dx = 1$; see, e.g., [10, Section 4.2.1] for an explicit example of such a function. For $\varepsilon > 0$ let

$\varphi_\varepsilon(x) := \varepsilon^{-d} \varphi(x/\varepsilon)$. We again fix $i \in \{1, \dots, k\}$. Since f_i is continuous, the convolution $(f_i)_\varepsilon := f_i * \varphi_\varepsilon$ converges pointwise (even locally uniformly) to f_i as $\varepsilon \rightarrow 0$; see [10, Theorem 1 in Section 4.2]. The same theorem also shows that $(f_i)_\varepsilon \in C^\infty(\mathbb{R}^d)$ with

$$\partial_j [(f_i)_\varepsilon] = (\tilde{\partial}_j f_i) * \varphi_\varepsilon.$$

Letting $f_\varepsilon := ((f_1)_\varepsilon, \dots, (f_k)_\varepsilon)^T$, we hence get

$$D(f_\varepsilon) = (\tilde{D}f) * \varphi_\varepsilon \quad \text{on } \mathbb{R}^d,$$

where the convolution on the right hand side is to be understood in a componentwise sense, i.e., each entry of the matrix $(\tilde{D}f)$ is convolved with φ_ε .

After these preparations, we fix $\delta > 0$ and set

$$K_\delta := \{x \in K : \overline{B}_d(x, \delta) \subseteq K\}.$$

Since φ has compact support and by definition of φ_ε , there exists $\varepsilon' = \varepsilon'(\delta) > 0$ such that for all $\varepsilon \in (0, \varepsilon']$ we have $\text{supp}(\varphi_\varepsilon) \subseteq \overline{B}_d(0, \delta)$. By definition of K_δ , this implies

$$\varphi_\varepsilon(x - y) = 0 \quad \text{for all } x \in K_\delta, y \in \mathbb{R}^d \setminus K \quad (\text{A.1})$$

for $\varepsilon \in (0, \varepsilon']$. Then, since $M \subseteq \mathbb{R}^d$ is a set of full measure, and N is a null set, we obtain for $x \in K_\delta$ and $\varepsilon \in (0, \varepsilon']$ that²

$$\begin{aligned} \|Df_\varepsilon(x)\|_{p \rightarrow q} &= \left\| \left[(\tilde{D}f) * \varphi_\varepsilon \right] (x) \right\|_{p \rightarrow q} = \left\| \int_{\mathbb{R}^d} \varphi_\varepsilon(x - y) \cdot \tilde{D}f(y) \, dy \right\|_{p \rightarrow q} \\ &= \left\| \int_{M \setminus N} \varphi_\varepsilon(x - y) \cdot \tilde{D}f(y) \, dy \right\|_{p \rightarrow q} \stackrel{(\text{A.1})}{=} \left\| \int_{(K \cap M) \setminus N} \varphi_\varepsilon(x - y) \cdot \tilde{D}f(y) \, dy \right\|_{p \rightarrow q} \\ &\leq \int_{(K \cap M) \setminus N} \left\| \varphi_\varepsilon(x - y) \cdot \tilde{D}f(y) \right\|_{p \rightarrow q} \, dy \leq \int_{(K \cap M) \setminus N} |\varphi_\varepsilon(x - y)| \cdot \sup_{\xi \in K \cap M} \left\| \tilde{D}f(\xi) \right\|_{p \rightarrow q} \, dy \\ &\leq \sup_{\xi \in K \cap M} \left\| \tilde{D}f(\xi) \right\|_{p \rightarrow q} \cdot \int_{\mathbb{R}^d} \varphi_\varepsilon(z) \, dz = \sup_{\xi \in K \cap M} \left\| \tilde{D}f(\xi) \right\|_{p \rightarrow q} =: L. \end{aligned}$$

Fix $x, y \in K_\delta$ and $\varepsilon > 0$. Let further $q' \in [1, \infty]$ be conjugate to q , i.e., $1/q + 1/q' = 1$. By duality, we can pick $z \in \mathbb{R}^k$ with

$$\|z\|_{q'} = 1 \quad \text{and} \quad \langle z, f_\varepsilon(y) - f_\varepsilon(x) \rangle = \|f_\varepsilon(y) - f_\varepsilon(x)\|_q.$$

We then consider the function

$$g : \mathbb{R} \rightarrow \mathbb{R}, \quad g(t) = \langle z, f_\varepsilon((1 - t)x + ty) \rangle.$$

The mean value theorem implies the existence of a number $\xi \in (0, 1)$ with

$$g'(\xi) = \frac{g(1) - g(0)}{1 - 0}.$$

By plugging in the definition of g and using the chain rule, we get

$$\|f_\varepsilon(y) - f_\varepsilon(x)\|_q = \langle z, [(Df_\varepsilon)((1 - \xi)x + \xi y)](y - x) \rangle.$$

²For the first inequality, let $A : \mathbb{R}^d \rightarrow \mathbb{R}^{k \times d}$ with $A_{i,j} \in L^1(\mathbb{R}^d)$ for every $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, d\}$ be given. We can then bound the $\|\cdot\|_{p \rightarrow q}$ -norm of the entrywise integral of A via

$$\left\| \int_{\mathbb{R}^d} A(\xi) \, d\xi \right\|_{p \rightarrow q} = \sup_{\substack{x \in \mathbb{R}^d, \|x\|_p = 1 \\ y \in \mathbb{R}^k, \|y\|_{q'} = 1}} \left\langle \left(\int_{\mathbb{R}^d} A(\xi) \, d\xi \right) \cdot x, y \right\rangle = \sup_{\substack{x \in \mathbb{R}^d, \|x\|_p = 1 \\ y \in \mathbb{R}^k, \|y\|_{q'} = 1}} \int_{\mathbb{R}^d} \langle A(\xi) \cdot x, y \rangle \, d\xi \leq \int_{\mathbb{R}^d} \|A(\xi)\|_{p \rightarrow q} \, d\xi,$$

where $q' \in [1, \infty]$ is conjugate to q .

For every $t \in \mathbb{R}^d$ with $\|t\|_2 \leq \delta$, we have

$$(1 - \xi)x + \xi y + t = (1 - \xi)\underbrace{(x + t)}_{\in K} + \xi\underbrace{(y + t)}_{\in K} \in K,$$

where the last step uses the convexity of K . Hence, by definition of K_δ and since t was arbitrary, we have

$$(1 - \xi)x + \xi y \in K_\delta.$$

Using Hölder's inequality, this implies

$$\|f_\varepsilon(x) - f_\varepsilon(y)\|_q \leq \underbrace{\|z\|_{q'}}_{=1} \cdot \|[(Df_\varepsilon)((1 - \xi)x + \xi y)](y - x)\|_q \leq L \cdot \|x - y\|_p$$

for all $x, y \in K_\delta$ and $\varepsilon \in (0, \varepsilon']$. But then it also follows

$$\|f(x) - f(y)\|_q = \lim_{\varepsilon \rightarrow 0} \|f_\varepsilon(x) - f_\varepsilon(y)\|_q \leq L \cdot \|x - y\|_p$$

and thus $\text{Lip}_{p \rightarrow q}(f|_{K_\delta}) \leq L$.

Since δ was arbitrary and $K = \text{int}(K) = \bigcup_{\delta > 0} K_\delta$ (recall that K is assumed to be open), we thus obtain $\text{Lip}_{p \rightarrow q}(f|_K) \leq L$. Hence, the inequality “ $\leq$ ” follows.

To prove the inequality in the other direction, let $x \in K \cap M$. Let $\nu \in \mathbb{R}^d$ be arbitrary with $\|\nu\|_p = 1$. Then it holds

$$\begin{aligned} \|Df(x)\nu\|_q &= \left\| \lim_{t \rightarrow 0} \frac{f(x + t\nu) - f(x)}{t} \right\|_q = \lim_{t \rightarrow 0} \frac{\|f(x + t\nu) - f(x)\|_q}{|t|} \\ &= \lim_{t \rightarrow 0} \frac{\|f(x + t\nu) - f(x)\|_q}{\|(x + t\nu) - x\|_p} \leq \text{Lip}_{p \rightarrow q}(f|_K), \end{aligned}$$

where the last inequality uses that K is open. The claim now follows, since by definition we have

$$\|Df(x)\|_{p \rightarrow q} = \sup_{\nu \in \mathbb{R}^d, \|\nu\|_p = 1} \|Df(x)\nu\|_q. \quad \square$$

Next, we provide the proof for Theorem 2.8, which mainly relies on the fact that the formalized gradient of a zero-bias ReLU network is invariant under scaling by positive constants.

Proof of Theorem 2.8. The first and third part of the theorem follow directly from [13, Theorem E.1] and Propositions 2.4 and 2.6.

For the second part, we note that the lower bound again follows from [13, Theorem E.1]. Further, according to Proposition 2.6, M_Φ is a set of full measure and so is $M_\Phi \setminus \{0\}$. In view of Proposition 2.4, it hence suffices to show that

$$\{\bar{\nabla}\Phi(x) : x \in M_\Phi \setminus \{0\}\} = \{\bar{\nabla}\Phi(x) : x \in M_\Phi \cap \mathbb{S}^{d-1}\}.$$

To this end, let $x \in M_\Phi \setminus \{0\}$ and define $z := \frac{x}{\|x\|_2} \in \mathbb{S}^{d-1}$. Since Φ is a zero-bias network,

$$\bar{\nabla}\Phi(ay) = \bar{\nabla}\Phi(y) \quad \text{for every } a > 0 \text{ and } y \in \mathbb{R}^d.$$

Hence, we get

$$\bar{\nabla}\Phi(x) = \bar{\nabla}\Phi(z).$$

It remains to show that $z \in M_\Phi$. To this end, note that we have

$$\Phi(z) = \frac{1}{\|x\|_2} \cdot \Phi(z \cdot \|x\|_2)$$

since the ReLU is positively homogeneous. Since Φ is differentiable at $z \cdot \|x\|_2 = x$, we may apply the chain rule and get

$$\nabla\Phi(z) = \frac{1}{\|x\|_2} \cdot \|x\|_2 \cdot \nabla\Phi(x) \stackrel{x \in M_\Phi}{=} \bar{\nabla}\Phi(x) = \bar{\nabla}\Phi(z).$$

Therefore, we have $z \in M_\Phi$, which proves the second part. The fourth part follows analogously. $\square$

Next, we prove Proposition 2.10. It states that for a random ReLU network Φ with symmetric biases, at a fixed input $x \neq 0$, with high probability, we may replace each matrix $D^{(\ell)}(x)$ by its randomized counterpart $\overline{D}^{(\ell)}(x)$ as defined in Definition 2.9. In fact, we show that, with high probability, all the preactivations occurring in Φ when x is passed through the network are nonzero, which implies the desired statement.

Proof of Proposition 2.10. We show via induction that for every $\ell \in \{-1, \dots, L-1\}$, with probability $(1 - \frac{1}{2^N})^{\ell+1}$, we have

$$\Phi^{(\ell)}(x) \neq 0 \quad \text{and} \quad \text{for all } \ell' \in \{0, \dots, \ell\}, j \in \{1, \dots, N\} : \quad (W^{(\ell')} \Phi^{(\ell'-1)}(x))_j + b_j^{(\ell')} \neq 0.$$

This implies the claim by the definition of the matrices $D^{(i)}(x)$ and $\overline{D}^{(i)}(x)$.

Note because of $x \neq 0$ that there is nothing to show in the case $\ell = -1$. We now take $\ell \in \{0, \dots, L-1\}$ and assume via induction that

$$\Phi^{(\ell-1)}(x) \neq 0 \quad \text{and} \quad \text{for all } \ell' \in \{0, \dots, \ell-1\}, j \in \{1, \dots, N\} : \quad (W^{(\ell')} \Phi^{(\ell'-1)}(x))_j + b_j^{(\ell')} \neq 0 \quad (\text{A.2})$$

occurs with probability at least $(1 - \frac{1}{2^N})^\ell$. Note that (A.2) only depends on $W^{(0)}, \dots, W^{(\ell-1)}$ and $b^{(0)}, \dots, b^{(\ell-1)}$. We condition on $W^{(0)}, \dots, W^{(\ell-1)}$ and $b^{(0)}, \dots, b^{(\ell-1)}$ and assume that (A.2) is satisfied. Since $\Phi^{(\ell-1)}(x) \neq 0$, the components of $W^{(\ell)} \Phi^{(\ell-1)}(x) + b^{(\ell)}$ have independent, symmetric and absolutely continuous random distributions, as follows from [8, Proposition 9.1.6]. Hence,

$$\text{for all } j \in \{1, \dots, N\} : \quad (W^{(\ell)} \Phi^{(\ell-1)}(x))_j + b_j^{(\ell)} \neq 0$$

with probability 1. Moreover, we have

$$\Phi^{(\ell)}(x) = 0 \quad \Leftrightarrow \quad \text{for all } j \in \{1, \dots, N\} : \quad (W^{(\ell)} \Phi^{(\ell-1)}(x))_j + b_j^{(\ell)} \leq 0.$$

Since the components of $W^{(\ell)} \Phi^{(\ell-1)}(x) + b^{(\ell)}$ follow independent, symmetric, continuous distributions, we conclude $\Phi^{(\ell)}(x) \neq 0$ with probability $1 - \frac{1}{2^N}$. Up to now, we have conditioned on $W^{(0)}, \dots, W^{(\ell-1)}$ and $b^{(0)}, \dots, b^{(\ell-1)}$. Reintroducing the randomness over these random matrices yields the claim. $\square$

B Proof of distributional equivalence

In this appendix, we prove the distributional equivalence stated in Proposition 2.11. Let us recall that this proposition essentially states that in expressions such as

$$\left\| \prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right\|_{2 \rightarrow 2},$$

where the z_i are arbitrary inputs from $\mathbb{R}^d$, we can — in distribution — replace the matrices $\overline{D}^{(i)}(z_i)$ by independent random diagonal matrices $\widehat{D}^{(i)}$, where each entry on the diagonal of $\widehat{D}^{(i)}$ is either 1 or 0, each with probability 1/2. This heavily simplifies the analysis of such expressions.

We start with an elementary lemma, which is probably well-known but for which we could not locate a convenient reference.

Lemma B.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and*

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow (\Omega_1, \mathcal{A}_1), \quad Y : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow (\Omega_2, \mathcal{A}_2)$$

be two independent random variables taking values in the measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$. Let

$$\varphi_1 : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \rightarrow (\Omega_3, \mathcal{A}_3) \quad \text{and} \quad \varphi_2 : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \rightarrow (\Omega_3, \mathcal{A}_3)$$

be measurable functions, taking values in some measurable space $(\Omega_3, \mathcal{A}_3)$. Assume that for every fixed $x \in \Omega_1$,

$$\varphi_1(x, Y) \stackrel{d}{=} \varphi_2(x, Y).$$

Then,

$$\varphi_1(X, Y) \stackrel{d}{=} \varphi_2(X, Y).$$

Proof. Let $A \in \mathcal{A}_3$ be arbitrary. Since X and Y are independent, we get

$$\mathbb{P}(\varphi_1(X, Y) \in A) = \mathbb{E}[\mathbb{1}_{\varphi_1(X, Y) \in A}] = \mathbb{E}_{x \sim X} \mathbb{E}_Y[\mathbb{1}_{\varphi_1(x, Y) \in A}] = \mathbb{E}_{x \sim X} \mathbb{E}_Y[\mathbb{1}_{\varphi_2(x, Y) \in A}] = \mathbb{P}(\varphi_2(X, Y) \in A).$$

Since A was arbitrary, equality in distribution is shown. $\square$

The following lemma provides a first statement in the direction of Proposition 2.11.

Lemma B.2. *Let $V \in \mathbb{R}^{\ell \times N}$, $W \in \mathbb{R}^{N \times k}$ be random matrices and $b \in \mathbb{R}^N$ a random vector, where all their entries are jointly independent and follow (possibly different) symmetric distributions. Moreover, let $\varepsilon_1, \dots, \varepsilon_N \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{0, 1\}$ and assume that the random vector $(\varepsilon_1, \dots, \varepsilon_N)$ is also independent of W , V and b . For $y \in \mathbb{R}^N$, define*

$$\tilde{\Delta}(y) := \Delta(y) + \text{diag}(\mathbb{1}_{y_1=0} \cdot \varepsilon_1, \dots, \mathbb{1}_{y_N=0} \cdot \varepsilon_N)$$

and let further

$$\hat{D} := \text{diag}(\varepsilon_1, \dots, \varepsilon_N).$$

Then, for any $x \in \mathbb{R}^k$,

$$V\tilde{\Delta}(Wx + b)W \stackrel{d}{=} V\hat{D}W.$$

Moreover, for any deterministic matrix $Z \in \mathbb{R}^{k \times m}$,

$$\left\| \tilde{\Delta}(Wx + b)WZ \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \hat{D}WZ \right\|_{2 \rightarrow 2}.$$

Proof. The argument for both statements is similar, but with a subtle difference. Let us start with the first claim. We introduce an additional source of randomness $S \in \mathbb{R}^{N \times N}$, which is a diagonal matrix with independent $\text{Unif}\{1, -1\}$ -entries on the diagonal, independent of all the other occurring random variables. Since all the entries of W , V and b are independent and symmetrically distributed,³

$$(W, V, b, \hat{D}) \stackrel{d}{=} (SW, VS, Sb, \hat{D}).$$

Therefore, we get

$$V\tilde{\Delta}(Wx + b)W \stackrel{d}{=} VS\tilde{\Delta}(SWx + Sb)SW.$$

Since $\tilde{\Delta}(SWx + Sb)$ is a diagonal matrix, it commutes with the diagonal matrix S . Thus, since $S^2 = I_{N \times N}$, we conclude

$$VS\tilde{\Delta}(SWx + Sb)SW = V\tilde{\Delta}(SWx + Sb)W.$$

Let $X := (W, V, b)$ and $Y := (S, \varepsilon)$ with $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)$. Note that X and Y are independent. For a fixed realization of X , consider the matrix

$$T := \tilde{\Delta}(SWx + Sb) = \Delta(SWx + Sb) + \text{diag}(\mathbb{1}_{(SWx+Sb)_1=0} \cdot \varepsilon_1, \dots, \mathbb{1}_{(SWx+Sb)_N=0} \cdot \varepsilon_N).$$

The diagonal entries of this matrix are independent, as follows from the independence of the components of S and ε . Moreover, for fixed $i \in \{1, \dots, N\}$, if $(Wx + b)_i = 0$ we have $T_{i,i} = \varepsilon_i$ and if $(Wx + b)_i \leq 0$ we have $T_{i,i} = \mathbb{1}_{S_{i,i} \leq 0}$. Hence, we get $T_{i,i} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{0, 1\}$. This shows

$$V\tilde{\Delta}(SWx + Sb)W \stackrel{d}{=} V\hat{D}W \quad \text{for fixed } V, W \text{ and } b.$$

³More precisely, for each fixed realization of S the distributional equivalence is satisfied by symmetry and independence of the entries of W , V and b , and Lemma B.1 then yields the distributional equivalence when the randomness is with respect to W , V , b and S .

Applying Lemma B.1, we conclude that this distributional equivalence is even satisfied if the randomness with respect to W , V and b is also incorporated. Thus, the first claim is shown.

For the second claim, we introduce the same diagonal matrix $S \in \mathbb{R}^{N \times N}$. We further use the distributional equivalence $(W, b, \widehat{D}) \stackrel{d}{=} (SW, Sb, \widehat{D})$ and therefore get

$$\left\| \widetilde{\Delta}(Wx + b)WZ \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \widetilde{\Delta}(SWx + Sb)SWZ \right\|_{2 \rightarrow 2} = \left\| S\widetilde{\Delta}(SWx + Sb)SWZ \right\|_{2 \rightarrow 2},$$

where for the last equality we used that the spectral norm of a matrix is not changed if multiplied with any realization of S . Identically to the first part, one can now show

$$S\widetilde{\Delta}(SWx + Sb)SWZ = \widetilde{\Delta}(SWx + Sb)WZ \stackrel{d}{=} \widehat{D}WZ,$$

which gives us the claim. $\square$

The proof of Proposition 2.11 now follows from an inductive argument, where we iteratively apply Lemma B.2.

Proof of Proposition 2.11. Let $\nu \in \mathbb{N}$ and let $A = A(W^{(0)}, \dots, W^{(\ell_1-1)}, b^{(0)}, \dots, b^{(\ell_1-1)}) \in \mathbb{R}^{\mu \times \nu}$ be any matrix that might possibly depend on $W^{(0)}, \dots, W^{(\ell_1-1)}$ and $b^{(0)}, \dots, b^{(\ell_1-1)}$.

Auxiliary claim: We start by proving the auxiliary claim that for every $\ell_1 \in \{0, \dots, L-1\}$ and $\ell_2 \in \{0, \dots, L\}$,

$$W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \stackrel{d}{=} W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A.$$

We fix $\ell_1 \in \{0, \dots, L-1\}$ and prove the claim via induction over $\ell_2 \in \{0, \dots, L\}$. Note that in the case $\ell_2 \leq \ell_1$ there is nothing to show, since then, by definition of the reverse matrix product in (1.2),

$$\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} = \prod_{i=\ell_2-1}^{\ell_1} \widehat{D}^{(i)} W^{(i)} = I_{\mu \times \mu}.$$

For the induction step, let $\ell_2 \in \{\ell_1 + 1, \dots, L\}$ and assume that the claim is already satisfied for $\ell_2 - 1$. We then let $\vec{W} := (W^{(0)}, \dots, W^{(\ell_2-2)})$, $\vec{b} := (b^{(0)}, \dots, b^{(\ell_2-2)})$, $\vec{W}' := (W^{(\ell_2-1)}, W^{(\ell_2)})$, and $\vec{\varepsilon} := (\varepsilon^{(0)}, \dots, \varepsilon^{(\ell_2-2)})$. Consider the two independent random variables $X := (\vec{W}, \vec{b}, \vec{\varepsilon})$ and $Y := (\vec{W}', b^{(\ell_2-1)}, \varepsilon^{(\ell_2-1)})$ and the two measurable maps

$$\varphi_1(X, Y) := W^{(\ell_2)} \overline{D}^{(\ell_2-1)}(z_{\ell_2-1}) W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \quad \text{and}$$

$$\varphi_2(X, Y) := W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A.$$

We apply Lemma B.2 (with $\varepsilon = \varepsilon^{(\ell_2-1)}$) to conclude that for each fixed realization of X (such that $\Phi^{(\ell_2-2)}(z_{\ell_2-1})$ and A are fixed), we get

$$\begin{aligned} W^{(\ell_2)} \overline{D}^{(\ell_2-1)}(z_{\ell_2-1}) W^{(\ell_2-1)} &= W^{(\ell_2)} \widetilde{\Delta}(W^{(\ell_2-1)} \Phi^{(\ell_2-2)}(z_{\ell_2-1})) W^{(\ell_2-1) + b^{(\ell_2-1)}} \\ &\stackrel{d}{=} W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)}, \end{aligned}$$

whence for each fixed realization x_0 of X we have

$$\varphi_1(x_0, Y) \stackrel{d}{=} \varphi_2(x_0, Y).$$

From Lemma B.1 we then conclude

$$W^{(\ell_2)} \overline{D}^{(\ell_2-1)}(z_{\ell_2-1}) W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \stackrel{d}{=} W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A,$$

where the randomness is taken with respect to all weights $W^{(i)}$ and all Bernoulli variables $\varepsilon^{(i)}$. We then define the random variables

$$X' := (\varepsilon^{(\ell_2-1)}, W^{(\ell_2)}) \quad \text{and} \quad Y' := (W^{(0)}, \dots, W^{(\ell_2-1)}, b^{(0)}, \dots, b^{(\ell_2-2)}, \varepsilon^{(0)}, \dots, \varepsilon^{(\ell_2-2)}).$$

Moreover, we let

$$\begin{aligned} \varphi'_1(X', Y') &:= W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A, \\ \varphi'_2(X', Y') &:= W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A. \end{aligned}$$

From the induction hypothesis, we infer that for each fixed realization x' of X' , we have

$$\varphi'_1(x', Y') \stackrel{d}{=} \varphi'_2(x', Y').$$

Since X' and Y' are independent, Lemma B.1 yields

$$W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \stackrel{d}{=} W^{(\ell_2)} \widehat{D}^{(\ell_2-1)} W^{(\ell_2-1)} \left[\prod_{i=\ell_2-2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A,$$

as desired.

Final conclusion: We start with the first claim. It is trivial if $\ell_2 = -1$, so we may assume $\ell_2 \geq 0$. Note that

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} = \left\| \overline{D}^{(\ell_2)}(z_{\ell_2}) W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2}.$$

Let $\vec{W} := (W^{(0)}, \dots, W^{(\ell_2-1)})$, $\vec{b} := (b^{(0)}, \dots, b^{(\ell_2-1)})$ and $\vec{\varepsilon} := (\varepsilon^{(0)}, \dots, \varepsilon^{(\ell_2-1)})$. We define the independent random variables $X := (\vec{W}, \vec{b}, \vec{\varepsilon})$ and $Y := (W^{(\ell_2)}, b^{(\ell_2)}, \varepsilon^{(\ell_2)})$. We now assume that the realization of the random variable X is fixed and, for convenience, let $Z := \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A$.

From the second part of Lemma B.2 (with $\varepsilon = \varepsilon^{(\ell_2)}$), we infer

$$\begin{aligned} \left\| \overline{D}^{(\ell_2)}(z_{\ell_2}) W^{(\ell_2)} Z \right\|_{2 \rightarrow 2} &= \left\| \widetilde{\Delta}(W^{(\ell_2)} \Phi^{(\ell_2-1)+b^{(\ell_2)}}(z_{\ell_2})) W^{(\ell_2)} Z \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \widehat{D}^{(\ell_2)} W^{(\ell_2)} Z \right\|_{2 \rightarrow 2} \\ &= \left\| \widehat{D}^{(\ell_2)} W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \end{aligned}$$

for every fixed realization of X . From Lemma B.1, we then obtain

$$\left\| \overline{D}^{(\ell_2)}(z_{\ell_2}) W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \widehat{D}^{(\ell_2)} W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2},$$

where the randomness is with respect to all the occurring random variables. From the auxiliary claim and the independence of the occurring random variables, we then get

$$\left\| \widehat{D}^{(\ell_2)} W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \widehat{D}^{(\ell_2)} W^{(\ell_2)} \left[\prod_{i=\ell_2-1}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2},$$

as claimed.

The second claim follows from the auxiliary claim, by picking A as the identity matrix and $\ell_2 = L$. $\square$

C Proofs for the results on random hyperplane tessellations

In this appendix, we prove the hyperplane tessellation results from Section 2.4 and prove special cases that will be used in the present paper.

We start by defining the angular distance between two points on the sphere, which turns out to be the correct measure to describe the probability that a Gaussian hyperplane with zero bias separates these two points.

Definition C.1. For two points $x, y \in \mathbb{S}^{n-1}$ we define the angular distance of these points via

$$\text{ang}(x, y) := \frac{1}{\pi} \cdot \arccos(\langle x, y \rangle) \in [0, 1].$$

The following folklore lemma establishes that the angular distance is equivalent to the Euclidean distance.

Lemma C.2. For $x, y \in \mathbb{S}^{n-1}$ we have

$$\frac{1}{\pi} \cdot \|x - y\|_2 \leq \text{ang}(x, y) \leq \|x - y\|_2.$$

Proof. We start by showing that for all $z \in [0, 1]$ we have

$$z \leq \sqrt{2} \cdot \sqrt{1 - \cos(\pi z)} \leq \pi \cdot z.$$

Here, we use the identity $1 - \cos(\pi z) = 2 \cdot \sin^2(\pi z/2)$ and hence get that the above is equivalent to

$$z \leq 2 \cdot \sin(\pi z/2) \leq \pi \cdot z.$$

The upper bound follows directly from the well-known fact that $\sin(t) \leq t$ for $t \geq 0$. For the lower bound, we define $f(z) := 2 \cdot \sin(\pi z/2) - z$ and note that $f(0) = 0$ and $f(1) = 1$. Moreover, since $f''(z) = -\frac{\pi^2}{2} \cdot \sin(\pi z/2) < 0$ for $z \in (0, 1)$, we conclude that f is concave on $(0, 1)$, which implies $f(z) \geq 0$ for every $z \in [0, 1]$.

We now substitute $z = \text{ang}(x, y)$ and first obtain

$$\sqrt{2} \cdot \sqrt{1 - \cos(\pi \cdot \text{ang}(x, y))} = \sqrt{2} \cdot \sqrt{1 - \langle x, y \rangle} = \sqrt{\|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle} = \|x - y\|_2.$$

Hence, we have

$$\text{ang}(x, y) \leq \|x - y\|_2 \leq \pi \cdot \text{ang}(x, y),$$

which gives us the claim. $\square$

The central property of the angular distance related to random hyperplane tessellations is stated in the following proposition.

Proposition C.3 (cf. [17, Equation (5.1)]). Let $x, y \in \mathbb{S}^{d-1}$ and $W \sim \mathcal{N}(0, I_d)$. Then

$$\mathbb{P}(\text{sgn}(\langle W, x \rangle) \neq \text{sgn}(\langle W, y \rangle)) = \text{ang}(x, y).$$

Using Chernoff's inequality [20, Theorem 2.3.1] we can now prove Proposition 2.12.

Proof of Proposition 2.12. We may assume $\sigma^2 = 1$. For $i \in \{1, \dots, m\}$ set $X_i := \mathbb{1}_{\text{sgn}(\langle Ax \rangle_i) \neq \text{sgn}(\langle Ay \rangle_i)}$ and $S_m := \sum_{i=1}^m X_i$. By Proposition C.3 we have

$$\mathbb{E}[S_m] = m \cdot \text{ang}(x, y).$$

Hence, applying [20, Theorem 2.3.1] with $t := em \text{ang}(x, y)$ we get

$$\mathbb{P}(S_m \geq em \text{ang}(x, y)) \leq \exp(-m \cdot \text{ang}(x, y)) \cdot \left(\frac{em \text{ang}(x, y)}{em \text{ang}(x, y)} \right)^t = \exp(-m \cdot \text{ang}(x, y)).$$

Conversely, applying [20, Ex. 2.3.2] with $t := \frac{e}{3} \cdot m \cdot \text{ang}(x, y)$, we get

$$\mathbb{P}\left(S_m \leq \frac{e}{3} m \text{ang}(x, y)\right) \leq \exp(-m \cdot \text{ang}(x, y)) \cdot 3^{\frac{e}{3} m \cdot \text{ang}(x, y)}$$

$$\begin{aligned}
&= \exp\left(\ln(3) \cdot \frac{e}{3} \cdot m \cdot \text{ang}(x, y) - m \cdot \text{ang}(x, y)\right) \\
&\leq \exp(-c \cdot m \cdot \text{ang}(x, y)),
\end{aligned}$$

where $c := 1 - \ln(3) \cdot \frac{e}{3} > 0$. In combination with Lemma C.2 we obtain the claim. $\square$

For the proofs of Propositions 2.13 to 2.15 we need the following auxiliary statement.

Lemma C.4. *There exist absolute constants $C, c > 0$ such that the following holds. Let $A \in \mathbb{R}^{m \times n}$ be a random matrix with entries $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, let $K \subseteq \mathbb{S}^{n-1}$, $\alpha \in (1/m, e^{-1})$ and $\beta > 0$. Suppose that $0 < \varepsilon \leq c \cdot \beta \cdot \ln^{-1/2}(1/\alpha)$ and*

$$m \geq C \cdot w^2((K - K) \cap \overline{B}_n(0, \varepsilon)) \cdot \alpha^{-1} \cdot \beta^{-2}.$$

Then

$$\sup_{w \in (K - K) \cap \overline{B}_n(0, \varepsilon)} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \beta\} \leq \alpha \cdot m$$

with probability at least $1 - 2 \exp(-c \cdot \alpha m)$.

Proof. We set $K_\varepsilon := (K - K) \cap \overline{B}_n(0, \varepsilon)$ and $M := \lfloor \alpha \cdot m \rfloor \geq \frac{\alpha m}{2}$. By [7, Theorem SM1.1], with the setting $u = 1$,

$$\sup_{w \in K_\varepsilon} \max_{I \subseteq \{1, \dots, m\}, |I| \leq M} \left(\sum_{i \in I} (Aw)_i^2 \right)^{1/2} \leq C_1 \cdot \left(w(K_\varepsilon) + \varepsilon \cdot \sqrt{M \ln(em/M)} \right) \quad (\text{C.1})$$

with probability at least $1 - 2 \exp(-c_1 \cdot M \ln(em/M)) \geq 1 - 2 \exp(-c_2 \cdot \alpha m)$ with absolute constants $C_1, c_1, c_2 > 0$. Here, we also used that $-K_\varepsilon = K_\varepsilon$, whence we may replace the Gaussian complexity (which appears in the formulation of [7, Theorem SM1.1]) by the Gaussian width. By the assumption on m and by taking $C > 0$ large enough, we may assume

$$C_1 \cdot w(K_\varepsilon) \leq \sqrt{\alpha m/2} \cdot \frac{\beta}{4} \leq \sqrt{M} \cdot \frac{\beta}{4} \quad (\text{C.2})$$

since $m \geq C \cdot w^2(K_\varepsilon) \cdot \alpha^{-1} \cdot \beta^{-2}$. Further, since $\varepsilon \leq c \cdot \beta \cdot \ln^{-1/2}(1/\alpha)$, we get

$$\begin{aligned}
C_1 \cdot \varepsilon \cdot \sqrt{M \cdot \ln(em/M)} &= C_1 \cdot \varepsilon \cdot \sqrt{M} \cdot \ln^{1/2} \left(\frac{e}{M} \cdot \underbrace{\frac{\alpha m}{2} \cdot \frac{2}{\alpha}}_{\leq M} \right) \leq C_1 \cdot \varepsilon \cdot \sqrt{M \cdot \ln(2e/\alpha)} \\
&\stackrel{\text{Lem. I.3}}{\leq} \sqrt{2e} \cdot C_1 \cdot \varepsilon \cdot \sqrt{M \cdot \ln(1/\alpha)} \leq \sqrt{M} \cdot \frac{\beta}{4}
\end{aligned} \quad (\text{C.3})$$

if c is small enough. Combining (C.1), (C.2) and (C.3) yields

$$\sup_{w \in K_\varepsilon} \max_{I \subseteq \{1, \dots, m\}, |I| \leq M} \left(\frac{1}{M} \cdot \sum_{i \in I} (Aw)_i^2 \right)^{1/2} \leq \frac{\beta}{2}. \quad (\text{C.4})$$

with probability at least $1 - 2 \exp(-c_2 \cdot \alpha m)$. Now note that if we would have

$$\sup_{w \in K_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \beta\} > \alpha \cdot m \geq M$$

then there would exist $w \in K_\varepsilon$ and a set $I \subseteq \{1, \dots, m\}$ with $|I| = M$ and $|(Aw)_i| \geq \beta$ for every $i \in I$. In particular,

$$\left(\frac{1}{M} \cdot \sum_{i \in I} (Aw)_i^2 \right)^{1/2} \geq \left(\frac{1}{M} \cdot \sum_{i \in I} \beta^2 \right)^{1/2} = \left(\frac{1}{M} \cdot M \cdot \beta^2 \right)^{1/2} = \beta > \frac{\beta}{2}$$

in contradiction to (C.4). This completes the proof. $\square$

For the proofs of Propositions 2.13 and 2.14 we additionally need the following statement, which is based on an application of Chernoff's inequality.

Lemma C.5. *There exists an absolute constant $c > 0$ such that the following holds. For arbitrary $\alpha \in (0, 1]$, $x \in \mathbb{S}^{n-1}$ and a random matrix $A \in \mathbb{R}^{m \times n}$ with $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, with probability at least $1 - \exp(-c \cdot \alpha m)$,*

$$\#\{i \in \{1, \dots, m\} : |(Ax)_i| \leq \alpha\} \leq 2e \cdot \alpha m.$$

Proof. Fix $i \in \{1, \dots, m\}$. Then $(Ax)_i \sim \mathcal{N}(0, 1)$. Moreover, for $w \in [-1, 1]$ we have

$$c_1/2 := \frac{1}{\sqrt{2\pi e}} \leq \frac{1}{\sqrt{2\pi}} \exp(-w^2/2) \leq \frac{1}{\sqrt{2\pi}} \leq 1,$$

which implies

$$c_1 \cdot \alpha \leq \mathbb{P}(|(Ax)_i| \leq \alpha) \leq 2 \cdot \alpha.$$

For $i \in \{1, \dots, m\}$, set $X_i := \mathbb{1}_{|(Ax)_i| \leq \alpha}$ and further

$$S_m := \sum_{i=1}^m X_i = \#\{i \in \{1, \dots, m\} : |(Ax)_i| \leq \alpha\}.$$

Note that the X_i are independent Bernoulli variables and

$$c_1 \cdot \alpha m \leq \mathbb{E}[S_m] \leq 2 \cdot \alpha m.$$

By Chernoff's bound (see, e.g., [20, Theorem 2.3.1]) with $t := 2e \cdot \alpha m$,

$$\mathbb{P}(S_m \geq 2e \cdot \alpha m) \leq \exp(-\mathbb{E}[S_m]) \cdot \left(\frac{e \cdot \mathbb{E}[S_m]}{t}\right)^t \leq \exp(-c_1 \cdot \alpha m) \cdot \left(\frac{2e \cdot \alpha m}{2e \cdot \alpha m}\right)^t = \exp(-c_1 \cdot \alpha m). \quad \square$$

In the remainder of this section, we will frequently make use of the following observation: Given $x, y, x^* \in \mathbb{R}$ with $|x^*| > \tau$ and $|x - x^*|, |y - x^*| \leq \tau$ for some $\tau > 0$, we have

$$\text{sgn}(x) = \text{sgn}(x^*) = \text{sgn}(y).$$

We now have all the tools together to prove Proposition 2.13.

Proof of Proposition 2.13. We may assume $\sigma^2 = 1$.

Let $K_\varepsilon := (K - K) \cap \overline{B}_n(0, \varepsilon)$. Let $C_1, c_1 > 0$ be the constants from Lemma C.4. Set $\alpha := \delta/3$ and $\beta := \frac{\delta}{6e}$. Note that $\delta m \geq C$ by assumption. Since we can choose $C > 3$, this yields $1/m < \alpha < \delta < e^{-1}$. Moreover, taking $c > 0$ small enough,

$$\varepsilon \leq c \cdot \delta \cdot \ln^{-1/2}(1/\delta) = 6ec \cdot \beta \cdot \ln^{-1/2}(1/(3\alpha)) \stackrel{\text{Lem. 1.3}}{\leq} 6e\sqrt{3} \cdot c \cdot \beta \cdot \ln^{-1/2}(1/\alpha) \leq c_1 \cdot \beta \cdot \ln^{-1/2}(1/\alpha).$$

Furthermore, by possibly enlarging $C > 0$, we can ensure that

$$C_1 \cdot w^2(K_\varepsilon) \cdot \alpha^{-1} \cdot \beta^{-2} = 3(6e)^2 \cdot C_1 \cdot w^2(K_\varepsilon) \cdot \delta^{-3} \leq m.$$

The conditions of Lemma C.4 are thus satisfied and we get

$$\sup_{w \in K_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \delta/(6e)\} \leq \frac{\delta m}{3}$$

with probability at least $1 - 2\exp(-c_1 \cdot \alpha m) = 1 - 2\exp(-c_2 \cdot \delta m)$ with $c_2 := c_1/3$ and we call the event defined by that property $\mathcal{E}_1$.

Finally, let $\mathcal{N}_\varepsilon$ be an $\varepsilon/2$ -net of K with $|\mathcal{N}_\varepsilon| = \mathcal{N}(K, \varepsilon/2)$. Using Lemma C.5 (with $\alpha = \delta/(6e)$) and a union bound, we obtain the existence of an absolute constant $c_3 > 0$ such that

$$\text{for all } x^* \in \mathcal{N}_\varepsilon : \quad \#\{i \in \{1, \dots, m\} : |(Ax^*)_i| \leq \delta/(6e)\} \leq m \cdot \frac{\delta}{3}$$

with probability at least $1 - \exp(\ln(\mathcal{N}(K, \varepsilon/2)) - c_3 \cdot \delta m) \geq 1 - \exp(-c_4 \cdot \delta m)$ by letting $c_4 := c_3/2$ and by using that

$$\ln(\mathcal{N}(K, \varepsilon/2)) \leq \ln(e \cdot \mathcal{N}(K, \varepsilon/2)) \leq \frac{\delta m}{C} \leq \frac{c_3}{2} \cdot \delta m,$$

which holds for C large enough. We call the event defined by that property $\mathcal{E}_2$. We assume that the defining properties of $\mathcal{E}_1$ and $\mathcal{E}_2$ are satisfied. Let $x, y \in K$ with $\|x - y\|_2 \leq \varepsilon/2$ be arbitrary and pick $x^* \in \mathcal{N}_\varepsilon$ that satisfies $\|x - x^*\|_2 \leq \varepsilon/2$. Clearly,

$$\|x^* - y\|_2 \leq \|x^* - x\|_2 + \|x - y\|_2 \leq \varepsilon$$

and hence $x^* - x, x^* - y \in K_\varepsilon$. Observe for any $i \in \{1, \dots, m\}$ that $\text{sgn}((Ax)_i) = \text{sgn}((Ay)_i)$ if $|(Ax^*)_i| > t$, $|(A(x - x^*))_i| \leq t$ and $|(A(x^* - y))_i| \leq t$ for some $t > 0$. Hence,

$$\begin{aligned} & \#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i) \neq \text{sgn}((Ay)_i)\} \\ & \leq \#\{i \in \{1, \dots, m\} : |(Ax^*)_i| \leq \delta/(6e)\} + \#\{i \in \{1, \dots, m\} : |(A(x^* - x))_i| \geq \delta/(6e)\} \\ & \quad + \#\{i \in \{1, \dots, m\} : |(A(x^* - y))_i| \geq \delta/(6e)\} \leq \delta m. \end{aligned}$$

Moreover, letting $c_5 := \min\{c_2, c_4\}$, we have

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 3 \exp(-c_5 \cdot \delta m) \geq 1 - \exp(-c \cdot \delta m)$$

by picking $c \leq c_5/2$, since, by using $\delta m \geq C$ and taking $C > 0$ large enough, we have $\ln(3) \leq \frac{c_5}{2} \cdot \delta m$. This yields the claim. $\square$

We prove a special case of Proposition 2.13, where the considered inputs are taken from a set $f(\mathbb{S}^{n-1}) \subseteq \mathbb{S}^{\ell-1}$, where $f : \mathbb{S}^{n-1} \rightarrow \mathbb{S}^{\ell-1}$ is Lipschitz. Crucially, the obtained bound does not depend on ℓ , but only on the initial dimension n . Hence, in cases where $\ell \gg n$, we can use this approach to preserve the underlying n -dimensionality of the inputs.

Lemma C.6. *There exist absolute constants $C, c > 0$ with the following property: Given natural numbers $\ell, m, n \in \mathbb{N}$ and a random matrix $A \in \mathbb{R}^{m \times \ell}$ with $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for some fixed variance $\sigma^2 > 0$, a Lipschitz-continuous map $f : \mathbb{S}^{n-1} \rightarrow \mathbb{S}^{\ell-1}$, $M \geq \max\{1, \text{Lip}_{2 \rightarrow 2}(f)\}$ and $\delta \in (0, e^{-1})$ with*

$$m \geq C \cdot n \cdot \delta^{-1} \cdot \ln(M/\delta),$$

the event defined by

$$\begin{aligned} & \text{for all } x, y \in f(\mathbb{S}^{n-1}) \text{ with } \|x - y\|_2 \leq c \cdot \delta \cdot \ln^{-1/2}(1/\delta) : \\ & \#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i) \neq \text{sgn}((Ay)_i)\} \leq \delta \cdot m \end{aligned}$$

occurs with probability at least $1 - \exp(-c \cdot \delta m)$.

Proof. Let $\tilde{C}, \tilde{c} > 0$ be the constants provided by Proposition 2.13. Let $\delta \in (0, e^{-1})$ be arbitrary and set $\varepsilon := \tilde{c} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$. Moreover, let $K := f(\mathbb{S}^{n-1})$. The goal is now to check that the condition from Proposition 2.13 is satisfied, i.e., that

$$m \geq \tilde{C} \cdot \max\{\delta^{-3} w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon), \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon/2)))\}. \quad (\text{C.5})$$

This then proves the claim of the lemma upon choosing $c \leq \tilde{c}/2$.

To this end, we take an arbitrary $\theta \in (0, 1)$. Let $\mathcal{N} \subseteq \mathbb{S}^{n-1}$ be a $\frac{\theta}{M}$ -net of $\mathbb{S}^{n-1}$ with

$$|\mathcal{N}| \leq \left(\frac{3M}{\theta}\right)^n,$$

which exists according to [20, Corollary 4.2.13] and noting $\theta/M \leq 1$. Then one easily sees that $f(\mathcal{N})$ is a θ -net of K , whence

$$\mathcal{N}(K, \theta) \leq \left(\frac{3M}{\theta}\right)^n. \quad (\text{C.6})$$

Moreover, by [20, Exercise 4.2.10],

$$\mathcal{N}((K - K) \cap \overline{B}_\ell(0, \varepsilon), \|\cdot\|_2, \theta) \leq \mathcal{N}((K - K), \|\cdot\|_2, \theta/2) \leq \mathcal{N}(K, \|\cdot\|_2, \theta/4)^2,$$

since taking the Minkowski difference of a $\frac{\theta}{4}$ -net of K with itself canonically yields a $\frac{\theta}{2}$ -net of $K - K$. Therefore, since $\text{diam}((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \leq 2\varepsilon$, Dudley's inequality (see, e.g., [20, Theorem 8.1.10]) yields

$$\begin{aligned} w((K - K) \cap \overline{B}_\ell(0, \varepsilon)) &\leq C_1 \cdot \int_0^{2\varepsilon} \ln^{1/2}(\mathcal{N}((K - K) \cap \overline{B}_\ell(0, \varepsilon), \|\cdot\|_2, \theta)) \, d\theta \\ &\leq \sqrt{2} \cdot C_1 \cdot \int_0^{2\varepsilon} \ln^{1/2}(\mathcal{N}(K, \|\cdot\|_2, \theta/4)) \, d\theta \end{aligned}$$

with an absolute constant $C_1 > 0$. Since $\frac{\varepsilon}{2} = \frac{1}{2} \cdot \tilde{c} \cdot \delta \cdot \ln^{-1/2}(1/\delta) \leq 1$ (since we may assume $\tilde{c} \leq 1$), we can further bound

$$w((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \leq \sqrt{2} \cdot C_1 \cdot \sqrt{n} \cdot \int_0^{2\varepsilon} \ln^{1/2}\left(\frac{12M}{\theta}\right) \, d\theta.$$

Using Lemma I.4, we get

$$\int_0^{2\varepsilon} \ln^{1/2}\left(\frac{12M}{\theta}\right) \, d\theta \leq 4\varepsilon \cdot \ln^{1/2}(6M/\varepsilon).$$

Note that

$$\begin{aligned} \ln^{1/2}(6M/\varepsilon) &= \ln^{1/2}\left(\frac{6M \cdot \ln^{1/2}(1/\delta)}{\tilde{c} \cdot \delta}\right) \stackrel{\ln^{1/2}(1/\delta) \leq 1/\delta}{\leq} \ln^{1/2}\left(\frac{6M}{\tilde{c} \cdot \delta^2}\right) \leq \sqrt{2} \cdot \ln^{1/2}\left(\frac{6M}{\tilde{c}\delta}\right) \\ &\stackrel{\text{Lem. I.3}}{\leq} \frac{2\sqrt{3}}{\tilde{c}} \cdot \ln^{1/2}(M/\delta), \end{aligned} \tag{C.7}$$

whence

$$4\varepsilon \cdot \ln^{1/2}(6M/\varepsilon) \leq 4\tilde{c} \cdot \delta \cdot \ln^{-1/2}(1/\delta) \cdot \frac{2\sqrt{3}}{\tilde{c}} \cdot \ln^{1/2}(M/\delta) \leq 8\sqrt{3} \cdot \delta \cdot \ln^{1/2}(M/\delta).$$

Overall, we get

$$w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \leq C_2 \cdot n \cdot \delta^2 \cdot \ln(M/\delta)$$

with a suitable absolute constant $C_2 > 0$. According to (C.5), we want to ensure

$$w^2((K - K) \cap \overline{B}_n(0, \varepsilon)) \cdot \delta^{-3} \leq C_2 \cdot n \cdot \delta^{-1} \cdot \ln(M/\delta) \stackrel{!}{\leq} \tilde{C}^{-1} \cdot m,$$

which is satisfied taking $C \geq \tilde{C} \cdot C_2$ in our assumptions.

Moreover, we consider

$$\begin{aligned} \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \|\cdot\|_2, \varepsilon/2)) &\stackrel{(C.6)}{\leq} \delta^{-1} \cdot n \cdot \ln\left(\frac{6eM}{\varepsilon}\right) \stackrel{\text{Lem. I.3}}{\leq} e \cdot \delta^{-1} \cdot n \cdot \ln\left(\frac{6M}{\varepsilon}\right) \\ &\stackrel{(C.7)}{\leq} \delta^{-1} \cdot n \cdot \frac{12e}{(\tilde{c})^2} \cdot \ln(M/\delta) \leq C_3 \cdot n \cdot \delta^{-1} \cdot \ln(M/\delta) \leq \tilde{C}^{-1} \cdot m \end{aligned}$$

for an absolute constant $C_3 > 0$, where the last inequality is satisfied if we take $C \geq \tilde{C} \cdot C_3$ in our assumptions. The lemma is proven. $\square$

Let us now turn towards the non-homogeneous setting, i.e., the setting of non-zero biases. We first prove Proposition 2.14, which states that for inputs with large Euclidean norm the bias term may be neglected.

Proof of Proposition 2.14. Set $K_\varepsilon := (K - K) \cap \overline{B}_n(0, \varepsilon)$. Let $C_1, c_1 > 0$ be the constants from Lemma C.4. Set $\alpha = \delta/2$ and $\beta = \frac{\delta}{8e}$. Note that $\delta m \geq C$ by assumption; by choosing $C > 2$, this yields $1/m < \alpha < \delta < e^{-1}$. Moreover, taking $c > 0$ small enough,

$$\varepsilon \leq c \cdot \delta \cdot \ln^{-1/2}(1/\delta) = 8e \cdot c \cdot \beta \cdot \ln^{-1/2}(1/(2\alpha)) \stackrel{\text{Lem. I.3}}{\leq} 8e\sqrt{2} \cdot c \cdot \beta \cdot \ln^{-1/2}(1/\alpha) \leq c_1 \cdot \beta \cdot \ln^{-1/2}(1/\alpha).$$

Furthermore, by possibly enlarging $C > 0$,

$$C_1 \cdot w^2(K_\varepsilon) \cdot \alpha^{-1} \cdot \beta^{-2} = 2(8e)^2 \cdot C_1 \cdot w^2(K_\varepsilon) \cdot \delta^{-3} \leq m.$$

The conditions of Lemma C.4 are thus satisfied and we get

$$\sup_{w \in K_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \delta/(4e)\} \leq \frac{\delta m}{2}$$

with probability at least $1 - 2\exp(-c_1 \cdot \alpha m) = 1 - 2\exp(-c_2 \cdot \delta m)$, where $c_2 := c_1/2$. We call the event defined by that property $\mathcal{E}_1$.

Finally, let $\mathcal{N}_\varepsilon$ be an ε -net of K with $\#\mathcal{N}_\varepsilon = \mathcal{N}(K, \varepsilon)$. Using Lemma C.5 (with $\alpha = \delta/(4e)$) and a union bound, we obtain the existence of an absolute constant $c_3 > 0$ such that

$$\text{for all } x^* \in \mathcal{N}_\varepsilon : \quad \#\{i \in \{1, \dots, m\} : |(Ax^*)_i| \leq \delta/(4e)\} \leq m \cdot \frac{\delta}{2} \quad (\text{C.8})$$

with probability at least $1 - \exp(\ln(\mathcal{N}(K, \varepsilon)) - c_3 \cdot \delta m) \geq 1 - \exp(-c_4 \cdot \delta m)$ by letting $c_4 := c_3/2$ and by choosing $C > 0$ so large that

$$\ln(\mathcal{N}(K, \varepsilon)) \leq \ln(e \cdot \mathcal{N}(K, \varepsilon)) \leq C^{-1} \cdot \delta m \leq \frac{c_3}{2} \cdot \delta m.$$

We write $\mathcal{E}_2$ for the event defined in (C.8). On $\mathcal{E}_1 \cap \mathcal{E}_2$, we observe

$$\begin{aligned} & \sup_{x \in \text{cone}(K), \|x\|_2 \geq 8e\delta^{-1}\lambda} \#\{i \in \{1, \dots, m\} : \text{sgn}((Ax)_i + \tau_i) \neq \text{sgn}((Ax)_i)\} \\ & \leq \sup_{x \in \text{cone}(K), \|x\|_2 \geq 8e\delta^{-1}\lambda} \#\{i \in \{1, \dots, m\} : |(Ax)_i| \leq |\tau_i|\} \\ & \leq \sup_{x \in \text{cone}(K), \|x\|_2 \geq 8e\delta^{-1}\lambda} \#\{i \in \{1, \dots, m\} : |(Ax)_i| \leq \lambda\} \\ & = \sup_{x \in \text{cone}(K), \|x\|_2 \geq 8e\delta^{-1}\lambda} \#\left\{i \in \{1, \dots, m\} : \left| \left(A \frac{x}{\|x\|_2} \right)_i \right| \leq \frac{\lambda}{\|x\|_2} \right\} \\ & \leq \sup_{z \in K} \#\{i \in \{1, \dots, m\} : |(Az)_i| \leq \delta/(8e)\} \\ & \leq \sup_{x^* \in \mathcal{N}_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Ax^*)_i| \leq \delta/(4e)\} + \sup_{w \in K_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \delta/(8e)\} \leq \delta \cdot m \end{aligned}$$

and, letting $c_5 := \min\{c_2, c_4\}$, we have

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 3\exp(-c_5 \cdot \delta m) \geq 1 - \exp(-c \cdot \delta m)$$

by picking $c \leq c_5/2$, since we may assume because of $\delta m \geq C$ that $\ln(3) \leq \frac{c_5}{2} \cdot \delta m$. This yields the claim. $\square$

We continue by proving Proposition 2.15 which may be seen as a version of Proposition 2.13 for the case with non-zero biases.

Proof of Proposition 2.15. Pointwise estimate: Let $z \in K$, let $W \in \mathbb{R}^d$ be a random vector that is independent of b , and with $W \sim \mathcal{N}(0, I_d)$. By conditioning on W and using the small-ball property of the distribution of b_i , we obtain for $i \in \{1, \dots, m\}$,

$$\mathbb{P}\left(|\langle W, z \rangle + b_i| \leq \frac{\delta}{3e \cdot C_\tau}\right) \leq \frac{\delta}{3e}.$$

For $i \in \{1, \dots, m\}$, set $X_i := \mathbb{1}_{|(Az)_i + b_i| \leq \delta/(3eC_\tau)}$. Note by what we have just shown that we have $\mathbb{P}(X_i = 1) \leq \frac{\delta}{3e}$ for $i \in \{1, \dots, m\}$. Hence⁴, we can pick $Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\delta/(3e))$ with $X_i \leq Y_i$ for every $i \in \{1, \dots, m\}$. We set $S_m := \sum_{i=1}^m Y_i$ and note that

$$\mathbb{E}[S_m] = m \cdot \frac{\delta}{3e} =: \mu.$$

We may hence apply [20, Theorem 2.3.1] with $t := m \cdot \frac{\delta}{3}$ and get

$$\mathbb{P}\left(\sum_{i=1}^m X_i \geq m \cdot \frac{\delta}{3}\right) \leq \mathbb{P}\left(S_m \geq m \cdot \frac{\delta}{3}\right) \leq \exp(-\mu) \cdot \left(\frac{e\mu}{t}\right)^t = \exp(-c_1 \cdot \delta m)$$

with an absolute constant $c_1 > 0$. In other words, we have with probability at least $1 - \exp(-c_1 \cdot \delta m)$,

$$\#\{i \in \{1, \dots, m\} : |(Az)_i + b_i| \leq \delta/(3eC_\tau)\} \leq m \cdot \frac{\delta}{3}. \quad (\text{C.9})$$

Uniformization: Let $C_2, c_2 > 0$ be the constants from Lemma C.4. Set $\alpha = \delta/3$ and $\beta = \frac{\delta}{3eC_\tau}$. Note that $\delta m \geq C$ by assumption; by choosing $C > 3$ this yields $1/m < \alpha < \delta < e^{-1}$. Moreover, taking $c > 0$ small enough,

$$\begin{aligned} \varepsilon &\leq c \cdot C_\tau^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta) = c \cdot C_\tau^{-1} \cdot 3e \cdot C_\tau \cdot \beta \cdot \ln^{-1/2}(1/(3\alpha)) \\ &\stackrel{\text{Lem. I.3}}{\leq} 3e\sqrt{3} \cdot c \cdot \beta \cdot \ln^{-1/2}(1/\alpha) \leq c_2 \cdot \beta \cdot \ln^{-1/2}(1/\alpha). \end{aligned}$$

Furthermore, by possibly enlarging $C > 0$, we see for $K_\varepsilon := (K - K) \cap \overline{B}_n(0, \varepsilon)$ that

$$C_2 \cdot w^2(K_\varepsilon) \cdot \alpha^{-1} \cdot \beta^{-2} = 3(3e)^2 \cdot C_\tau^2 \cdot C_2 \cdot w^2(K_\varepsilon) \cdot \delta^{-3} \leq m.$$

The conditions of Lemma C.4 are thus satisfied and we get

$$\sup_{w \in K_\varepsilon} \#\{i \in \{1, \dots, m\} : |(Aw)_i| \geq \delta/(3eC_\tau)\} \leq \frac{\delta m}{3}$$

with probability at least $1 - 2\exp(-c_2 \cdot \alpha m) = 1 - 2\exp(-c_3 \cdot \delta m)$ where $c_3 := c_2/3$. We call the event defined by that property $\mathcal{E}_1$.

Final proof: Let $\mathcal{N}_\varepsilon$ be an $\varepsilon/2$ -net of K with $|\mathcal{N}_\varepsilon| = \mathcal{N}(K, \varepsilon/2)$. Using (C.9) and a union bound, we obtain

$$\text{for all } x^* \in \mathcal{N}_\varepsilon : \quad \#\{i \in \{1, \dots, m\} : |(Ax^*)_i + b_i| \leq \delta/(3eC_\tau)\} \leq m \cdot \frac{\delta}{3}$$

with probability at least $1 - \exp(\ln(\mathcal{N}(K, \varepsilon/2)) - c_1 \cdot \delta m) \geq 1 - \exp(-c_4 \cdot \delta m)$ by letting $c_4 := c_1/2$ and by choosing $C > 0$ so large that

$$\ln(\mathcal{N}(K, \varepsilon/2)) \leq \ln(e \cdot \mathcal{N}(K, \varepsilon/2)) \leq \frac{\delta m}{C} \leq \frac{c_1}{2} \cdot \delta m.$$

We call the event defined by that property $\mathcal{E}_2$. We assume that the defining properties of $\mathcal{E}_1$ and $\mathcal{E}_2$ are satisfied. Let then $x, y \in K$ with $\|x - y\|_2 \leq \varepsilon/2$ be arbitrary and pick $x^* \in \mathcal{N}_\varepsilon$ that satisfies $\|x - x^*\|_2 \leq \varepsilon/2$. This implies

$$\|x^* - y\|_2 \leq \|x^* - x\|_2 + \|x - y\|_2 \leq \varepsilon$$

⁴To see this, for $i \in \{1, \dots, m\}$, let $p_i \in (0, 1)$ be such that $X_i \sim \text{Ber}(p_i)$ and set $q_i := \frac{\frac{\delta}{3e} - p_i}{1 - p_i} \in (0, 1)$. Let $Z_1, \dots, Z_m$ be jointly independent and independent of the X_i with $Z_i \sim \text{Ber}(q_i)$ for $i \in \{1, \dots, m\}$. Finally, set $Y_i := \begin{cases} 1, & X_i = 1, \\ Z_i, & X_i = 0. \end{cases}$ Clearly, the Y_i are independent Bernoulli variables and satisfy $X_i \leq Y_i$. Moreover,

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i = 1) + \mathbb{P}(X_i = 0)\mathbb{P}(Z_i = 1) = p_i + (1 - p_i)q_i = \frac{\delta}{3e}.$$

and hence $x^* - x, x^* - y \in K_\varepsilon$. We hence get

$$\begin{aligned} & \#\{i \in \{1, \dots, m\} : \operatorname{sgn}((Ax)_i + b_i) \neq \operatorname{sgn}((Ay)_i + b_i)\} \\ & \leq \#\{i \in \{1, \dots, m\} : |(Ax^*)_i + b_i| \leq \delta/(3eC_\tau)\} + \#\{i \in \{1, \dots, m\} : |(A(x^* - x))_i| \geq \delta/(3eC_\tau)\} \\ & \quad + \#\{i \in \{1, \dots, m\} : |(A(x^* - y))_i| \geq \delta/(3eC_\tau)\} \leq \delta m, \end{aligned}$$

uniformly over the choice of x and y . Moreover, letting $c_5 := \min\{c_3, c_4\}$, we have

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 3 \exp(-c_5 \cdot \delta m) \geq 1 - \exp(-c \cdot \delta m)$$

by picking $c \leq c_5/2$, since we may assume because of $\delta m \geq C$ that $\ln(3) \leq \frac{c_5}{2} \cdot \delta m$. This yields the claim. $\square$

Analogous to Lemma C.6, we prove a version of Proposition 2.15 for the case where the considered set K is the image of a Euclidean ball under a Lipschitz mapping.

Lemma C.7. *There exist absolute constants $C, c > 0$ with the following property: Let $\ell, m, n \in \mathbb{N}$ and let $A \in \mathbb{R}^{m \times \ell}$ be a random matrix with $A_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Let $b \in \mathbb{R}^m$ be a random vector that is independent of A and with independent entries that satisfy*

$$\sup_{t \in \mathbb{R}, \varepsilon > 0, i \in \{1, \dots, m\}} \varepsilon^{-1} \cdot \mathbb{P}(b_i \in (t - \varepsilon, t + \varepsilon)) \leq C_\tau$$

for a constant $C_\tau > 0$. Let $R > 0$ with $RC_\tau \geq 1$ and let $f : B_n(0, R) \rightarrow \mathbb{R}^\ell$ be Lipschitz continuous, $M \geq \max\{1, \operatorname{Lip}_{2 \rightarrow 2}(f)\}$ and $\delta \in (0, e^{-1})$. Further, we assume

$$m \geq C \cdot \delta^{-1} \cdot n \cdot \ln(MRC_\tau/\delta).$$

Then the event defined by

$$\begin{aligned} & \text{for all } x, y \in f(B_n(0, R)) \quad \text{with} \quad \|x - y\|_2 \leq c \cdot C_\tau^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta) : \\ & \quad \#\{i \in \{1, \dots, m\} : \operatorname{sgn}((Ax)_i + b_i) \neq \operatorname{sgn}((Ay)_i + b_i)\} \leq \delta \cdot m \end{aligned}$$

occurs with probability at least $1 - \exp(-c \cdot \delta m)$.

Proof. Let $C', c' > 0$ be the constants provided by Proposition 2.15. Let $\delta \in (0, e^{-1})$ be arbitrary and set $\varepsilon := c' \cdot C_\tau^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$. Moreover, let $K := f(B_n(0, R))$. The goal is now to check that the condition from Proposition 2.15 is satisfied, i.e., that it holds

$$m \geq C' \cdot \max\{\delta^{-3} w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \cdot C_\tau^2, \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon/2))\}.$$

This proves the claim of the lemma upon choosing $c \leq c'/2$.

To this end, we take an arbitrary $\theta \in (0, MR]$. Let $\mathcal{N} \subseteq B_n(0, R)$ be a $\frac{\theta}{M}$ -net of $B_n(0, R)$ with

$$|\mathcal{N}| \leq \left(\frac{3MR}{\theta}\right)^n,$$

which exists according to [20, Corollary 4.2.13] and noting $\theta/M \leq R$. Then one easily sees that $f(\mathcal{N})$ is a θ -net of K , whence

$$\mathcal{N}(K, \theta) \leq \left(\frac{3MR}{\theta}\right)^n. \quad (\text{C.10})$$

Moreover,

$$\mathcal{N}((K - K) \cap \overline{B}_\ell(0, \varepsilon), \|\cdot\|_2, \theta) \leq \mathcal{N}((K - K), \|\cdot\|_2, \theta/2) \leq \mathcal{N}(K, \|\cdot\|_2, \theta/4)^2,$$

since taking the Minkowski difference of a $\frac{\theta}{4}$ -net of K with itself canonically yields a $\frac{\theta}{2}$ -net of $K - K$ and by [20, Exercise 4.2.10]. Therefore, since $\operatorname{diam}((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \leq 2\varepsilon$, Dudley's inequality (cf. [20, Theorem 8.1.10]) gives us

$$\begin{aligned} & w((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \\ & \leq C_1 \cdot \int_0^{2\varepsilon} \ln^{1/2}(\mathcal{N}((K - K) \cap \overline{B}_\ell(0, \varepsilon), \|\cdot\|_2, \theta)) \, d\theta \end{aligned}$$

$$\leq \sqrt{2} \cdot C_1 \cdot \int_0^{2\varepsilon} \ln^{1/2}(\mathcal{N}(K, \|\cdot\|_2, \theta/4)) \, d\theta$$

with an absolute constant $C_1 > 0$.

By using the bound established in (C.10) and since

$$\frac{\varepsilon}{2} \leq C_\tau^{-1} \stackrel{RC_\tau \geq 1}{\leq} R \leq MR,$$

we obtain

$$\begin{aligned} \int_0^{2\varepsilon} \ln^{1/2}(\mathcal{N}(K, \|\cdot\|_2, \theta/4)) \, d\theta &\leq \sqrt{n} \cdot \int_0^{2\varepsilon} \ln^{1/2}(12MR/\theta) \, d\theta \stackrel{\text{Lem. I.4}}{\leq} 4 \cdot \varepsilon \cdot \sqrt{n} \cdot \ln^{1/2}(6MR/\varepsilon) \\ &\stackrel{\text{Lem. I.3}}{\leq} 4\sqrt{6} \cdot \varepsilon \cdot \sqrt{n} \cdot \ln^{1/2}(MR/\varepsilon). \end{aligned}$$

Hence, in total we obtain the bound

$$w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \leq C_2 \cdot n \cdot \varepsilon^2 \cdot \ln(MR/\varepsilon)$$

with an absolute constant $C_2 > 0$. We compute

$$\begin{aligned} \ln(MR/\varepsilon) &= \ln\left(\frac{MRC_\tau \cdot \ln^{1/2}(1/\delta)}{c' \cdot \delta}\right) \stackrel{\ln^{1/2}(1/\delta) \leq 1/\delta}{\leq} \ln\left(\frac{MRC_\tau}{c' \cdot \delta^2}\right) \stackrel{MRC_\tau/c' \geq 1}{\leq} \ln\left(\left(\frac{MRC_\tau}{c' \cdot \delta}\right)^2\right) \\ &= 2 \cdot \ln\left(\frac{MRC_\tau}{c' \cdot \delta}\right) \stackrel{\text{Lem. I.3}}{\leq} 2 \cdot (c')^{-1} \cdot \ln\left(\frac{MRC_\tau}{\delta}\right). \end{aligned} \tag{C.11}$$

This gives us

$$\begin{aligned} \delta^{-3} \cdot w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \cdot C_\tau^2 &\leq C_2 \cdot n \cdot \varepsilon^2 \cdot \delta^{-3} \cdot \ln(MR/\varepsilon) \cdot C_\tau^2 \\ &\leq C_2 \cdot (c')^2 \cdot n \cdot \delta^2 \cdot \ln^{-1}(1/\delta) \cdot C_\tau^{-2} \cdot \delta^{-3} \cdot \ln(MR/\varepsilon) \cdot C_\tau^2 \\ &\stackrel{(C.11)}{\leq} 2C_2 \cdot c' \cdot n \cdot \delta^{-1} \cdot \ln^{-1}(1/\delta) \cdot \ln(MRC_\tau/\delta) \\ &\leq C_3 \cdot n \cdot \delta^{-1} \cdot \ln(MRC_\tau/\delta) \end{aligned}$$

with $C_3 := 2 \cdot C_2 \cdot c'$.

Moreover, we consider

$$\begin{aligned} \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon/2)) &\leq \delta^{-1} \cdot n \cdot \ln\left(\frac{6eMR}{\varepsilon}\right) \stackrel{\text{Lem. I.3}}{\leq} 6e \cdot \delta^{-1} \cdot n \cdot \ln(MR/\varepsilon) \\ &\stackrel{(C.11)}{\leq} 12 \cdot e \cdot (c')^{-1} \cdot \delta^{-1} \cdot n \cdot \ln(MRC_\tau/\delta) = C_4 \cdot \delta^{-1} \cdot n \cdot \ln(MRC_\tau/\delta) \end{aligned}$$

with $C_4 := 12e \cdot (c')^{-1}$. Hence, we have

$$\begin{aligned} &C' \cdot \max\{\delta^{-3}w^2((K - K) \cap \overline{B}_\ell(0, \varepsilon)) \cdot C_\tau^2, \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(K, \varepsilon/2))\} \\ &\leq C' \cdot \max\{C_3, C_4\} \cdot n \cdot \delta^{-1} \cdot \ln(MRC_\tau/\delta) \leq m \end{aligned}$$

by assumption, by picking $C \geq C' \cdot \max\{C_3, C_4\}$. This proves the lemma. $\square$

D Proof of pointwise bounds

In this appendix, we provide proofs for the statements from Section 3. All the proofs rely on the techniques discussed in Section 2.3, i.e., the fact that the D -matrices in the definition of the formal Jacobian and gradient can effectively be replaced by independent random diagonal matrices $\hat{D}$ with either 1 or 0 on the diagonal, each with probability $1/2$.

We start with the proof of Lemma 3.1.

Proof of Lemma 3.1. Since the claim is trivial in the case $\ell_2 < \ell_1$, we may assume $\ell_1 \leq \ell_2$ from now on. Thus, let $\ell_1, \ell_2 \in \{0, \dots, L-1\}$ with $\ell_1 \leq \ell_2$ be fixed. Note that

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \bar{D}^{(i)}(z_i) W^{(i)} \right] v \right\|_2 \stackrel{d}{=} \left\| \left[\prod_{i=\ell_2}^{\ell_1} \hat{D}^{(i)} W^{(i)} \right] v \right\|_2$$

according to Proposition 2.11. Regarding the matrices $\hat{D}^{(\ell_2)}, \dots, \hat{D}^{(\ell_1)}$, we condition on the event

$$\text{for all } \ell' \in \{\ell_1, \dots, \ell_2\}: \quad \left(1 - \frac{1}{4L}\right) \cdot \frac{N}{2} \leq \text{Tr}(\hat{D}^{(\ell')}) \leq \left(1 + \frac{1}{4L}\right) \cdot \frac{N}{2}, \quad (\text{D.1})$$

which occurs with probability at least $(1 - 2\exp(-c_1 \cdot N/L^2))^{\ell_2 - \ell_1 + 1} \geq 1 - 2L\exp(-c_1 \cdot N/L^2)$, where $c_1 > 0$ is an absolute constant. Here, we used that $\text{Tr}(\hat{D}^{(\ell')}) - \frac{N}{2} = \sum_{j=1}^N (\varepsilon_j^{(\ell')} - \frac{1}{2})$ and applied Bernoulli's inequality and Hoeffding's inequality for bounded random variables; see, e.g., [20, Theorem 2.2.6]. Since we condition on the matrices $\hat{D}^{(\ell')}$ for $\ell' \in \{\ell_1, \dots, \ell_2\}$, we can from now on assume that they are fixed. Note that then

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \hat{D}^{(i)} W^{(i)} \right] v \right\|_2 \stackrel{d}{=} \left\| \left[\prod_{i=\ell_2}^{\ell_1} W_{\dagger}^{(i)} \right] v \right\|_2 \quad (\text{D.2})$$

with $W_{\dagger}^{(\ell_1)} \in \mathbb{R}^{\text{Tr}(\hat{D}^{(\ell_1)}) \times \mu}$ and $W_{\dagger}^{(i)} \in \mathbb{R}^{\text{Tr}(\hat{D}^{(i)}) \times \text{Tr}(\hat{D}^{(i-1)})}$ for every $i \in \{\ell_1 + 1, \dots, \ell_2\}$. Here, the random matrices $W_{\dagger}^{(i)}$ are independent, have independent $\mathcal{N}(0, 2/N)$ entries, and are independent of v . We now show via induction over $\ell' \in \{\ell_1 - 1, \dots, \ell_2\}$ that

$$\left(1 - \frac{1}{2L}\right)^{\ell' - \ell_1 + 1} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell'}^{\ell_1} W_{\dagger}^{(i)} \right] v \right\|_2 \leq \left(1 + \frac{1}{2L}\right)^{\ell' - \ell_1 + 1} \cdot \|v\|_2 \quad (\text{D.3})$$

with probability at least $(1 - 2\exp(-c_2 \cdot N/L^2))^{\ell' - \ell_1 + 1}$ with an absolute constant $c_2 > 0$ that will be specified later. Note that the case $\ell' = \ell_1 - 1$ is trivial, so we can move to the induction step. We hence fix $\ell' \in \{\ell_1, \dots, \ell_2\}$ and assume via induction that

$$\left(1 - \frac{1}{2L}\right)^{\ell' - \ell_1} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell'-1}^{\ell_1} W_{\dagger}^{(i)} \right] v \right\|_2 \leq \left(1 + \frac{1}{2L}\right)^{\ell' - \ell_1} \cdot \|v\|_2 \quad (\text{D.4})$$

with probability at least $(1 - 2\exp(-c_2 \cdot N/L^2))^{\ell' - \ell_1}$. We condition on $W_{\dagger}^{(\ell_1)}, \dots, W_{\dagger}^{(\ell'-1)}$ and v and assume that (D.4) is satisfied. For brevity, write

$$v' := \left[\prod_{i=\ell'-1}^{\ell_1} W_{\dagger}^{(i)} \right] v,$$

which is now a fixed vector. Note that $W_{\dagger}^{(\ell')} v' \in \mathbb{R}^{\text{Tr}(\hat{D}^{(\ell')})}$ is a random vector with independent $\mathcal{N}(0, 2\|v'\|_2^2/N)$ -entries. Hence, [20, Theorem 3.1.1] yields for arbitrary $t > 0$ that

$$\frac{\sqrt{2}}{\sqrt{N}} \|v'\|_2 \cdot \left(\sqrt{\text{Tr}(\hat{D}^{(\ell')})} - t \right) \leq \|W_{\dagger}^{(\ell')} v'\|_2 \leq \frac{\sqrt{2}}{\sqrt{N}} \|v'\|_2 \cdot \left(\sqrt{\text{Tr}(\hat{D}^{(\ell')})} + t \right)$$

with probability at least $1 - 2\exp(-c_3 \cdot t^2)$, with an absolute constant $c_3 > 0$. We set $t := \frac{\sqrt{N}}{4\sqrt{2} \cdot L}$ and obtain, using (D.1),

$$\begin{aligned} \frac{\sqrt{2}}{\sqrt{N}} \cdot \left(\sqrt{\text{Tr}(\hat{D}^{(\ell')})} + t \right) &\leq \frac{\sqrt{2}}{\sqrt{N}} \cdot \left(\frac{\sqrt{N}}{\sqrt{2}} \cdot \sqrt{1 + \frac{1}{4L}} + \frac{\sqrt{N}}{4\sqrt{2} \cdot L} \right) \\ &\stackrel{\sqrt{x} \leq x \text{ for } x \geq 1}{\leq} 1 + \frac{1}{2L} \end{aligned}$$

and analogously

$$\frac{\sqrt{2}}{\sqrt{N}} \cdot \left(\sqrt{\text{Tr}(\widehat{D}^{(\ell')})} - t \right) \geq 1 - \frac{1}{2L}.$$

Hence, using (D.4), we observe

$$\left(1 - \frac{1}{2L}\right)^{\ell' - \ell_1 + 1} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell'}^{\ell_1} W_{\dagger}^{(i)} \right] v \right\|_2 = \|W_{\dagger}^{(\ell')} v'\|_2 \leq \left(1 + \frac{1}{2L}\right)^{\ell' - \ell_1 + 1} \cdot \|v\|_2$$

with probability at least $1 - 2\exp(-c_2 \cdot N/L^2)$, where $c_2 := c_3/32$ and the randomness is solely with respect to $W_{\dagger}^{(\ell')}$. Reintroducing the randomness over $W_{\dagger}^{(\ell_1)}, \dots, W_{\dagger}^{(\ell'-1)}$ and v , we obtain (D.3).

In the end, picking $\ell' = \ell_2$ gives us

$$\frac{1}{2} \cdot \|v\|_2 \leq \left(1 - \frac{1}{2L}\right)^{\ell_2 - \ell_1 + 1} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell_2}^{\ell_1} W_{\dagger}^{(i)} \right] v \right\|_2 \leq \left(1 + \frac{1}{2L}\right)^{\ell_2 - \ell_1 + 1} \cdot \|v\|_2 \leq e \cdot \|v\|_2$$

with probability at least

$$(1 - 2\exp(-c_2 \cdot N/L^2))^{\ell_2 - \ell_1 + 1} \geq (1 - 2\exp(-c_2 \cdot N/L^2))^L \geq 1 - 2\exp(\ln(L) - c_2 \cdot N/L^2).$$

Keeping in mind that we conditioned on the $\widehat{D}$ -matrices (see (D.1) and recalling (D.2)), we get

$$\frac{1}{2} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] v \right\|_2 \leq e \cdot \|v\|_2$$

with probability at least $1 - 4\exp(\ln(L) - c_4 \cdot N/L^2)$, where $c_4 := \min\{c_1, c_2, \ln(2)\}$. Finally, we use Proposition 2.10 combined with a union bound over $z_0, \dots, z_{L-1}$ to conclude

$$\frac{1}{2} \cdot \|v\|_2 \leq \left\| \left[\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i) W^{(i)} \right] v \right\|_2 \leq e \cdot \|v\|_2$$

with probability at least

$$\begin{aligned} 1 - 4\exp(\ln(L) - c_4 \cdot N/L^2) - \frac{L^2}{2^N} &= 1 - 4\exp(\ln(L) - c_4 \cdot N/L^2) - \exp(\ln(L^2) - \ln(2) \cdot N) \\ &\geq 1 - 5\exp(2\ln(L) - c_4 \cdot N/L^2) \\ &= 1 - \exp(\ln(5) + 2\ln(L) - c_4 \cdot N/L^2), \end{aligned}$$

where we used $c_4 \leq \ln(2)$.

Since by assumption $N \geq C \cdot L^2 \cdot \ln(e \cdot L)$, we see by taking $C > 0$ large enough that

$$\ln(5) + 2\ln(L) \leq 3 \cdot \ln(5L) \stackrel{\text{Lem. 1.3}}{\leq} \frac{15}{e} \cdot \ln(e \cdot L) \leq \frac{c_4}{2} \cdot \frac{N}{L^2},$$

so the claim follows letting $c := c_4/2$. □

We continue by proving Lemma 3.2.

Proof of Lemma 3.2. We may assume $\ell_1 \leq \ell_2$, since in the case $\ell_1 > \ell_2$ we have

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} = \|A\|_{2 \rightarrow 2}.$$

Firstly, note that according to Proposition 2.11,

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \left[\prod_{i=\ell_2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2}.$$

We condition on the $\widehat{D}$ -matrices and assume

$$\text{for all } i \in \{\ell_1, \dots, \ell_2\} : \quad \text{Tr}(\widehat{D}^{(i)}) \leq \frac{N}{2} \left(1 + \frac{1}{L}\right),$$

which holds with probability at least $(1 - \exp(-c_1 \cdot N/L^2))^{\ell_2 - \ell_1 + 1} \geq 1 - L \cdot \exp(-c_1 \cdot N/L^2)$ for some absolute constant $c_1 > 0$, which follows from Hoeffding's inequality for bounded random variables [20, Theorem 2.2.6] after noting that $\text{Tr}(\widehat{D}^{(i)}) - \frac{N}{2} = \sum_{j=1}^N \left(\varepsilon_j^{(i)} - \frac{1}{2}\right)$. Fixing the $\widehat{D}$ -matrices, we observe

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2} \stackrel{d}{=} \left\| \left[\prod_{i=\ell_2}^{\ell_1} W_{\dagger}^{(i)} \right] A \right\|_{2 \rightarrow 2}, \quad (\text{D.5})$$

where $W_{\dagger}^{(i)} \in \mathbb{R}^{\text{Tr}(\widehat{D}^{(i)}) \times \text{Tr}(\widehat{D}^{(i-1)})}$ for $i \in \{\ell_1 + 1, \dots, \ell_2\}$ and $W_{\dagger}^{(\ell_1)} \in \mathbb{R}^{\text{Tr}(\widehat{D}^{(\ell_1)}) \times \mu}$. Here, the random matrices $W_{\dagger}^{(i)}$ are independent and have i.i.d. $\mathcal{N}(0, 2/N)$ entries. Hence, the goal is to bound

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} W_{\dagger}^{(i)} \right] A \right\|_{2 \rightarrow 2}.$$

To this end, we show via induction over $\ell' \in \{\ell_1 - 1, \dots, \ell_2\}$ that

$$\left\| \left[\prod_{i=\ell'}^{\ell_1} W_{\dagger}^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq \left(1 + \frac{5}{L}\right)^{\ell' - \ell_1 + 1} \cdot \|A\|_{2 \rightarrow 2} \quad (\text{D.6})$$

with probability at least $(1 - \exp(-c_2 \cdot N/L^2))^{\ell' - \ell_1 + 1}$ for every $\ell' \in \{\ell_1 - 1, \dots, \ell_2\}$, where $c_2 > 0$ is the absolute constant occurring in the formulation of Lemma I.5.

Note that the case $\ell' = \ell_1 - 1$ is trivial. We hence choose $\ell' \in \{\ell_1, \dots, \ell_2\}$ and assume via induction that

$$\left\| \left[\prod_{i=\ell' - 1}^{\ell_1} W_{\dagger}^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq \left(1 + \frac{5}{L}\right)^{\ell' - \ell_1} \cdot \|A\|_{2 \rightarrow 2} \quad (\text{D.7})$$

with probability at least $(1 - \exp(-c_2 \cdot N/L^2))^{\ell' - \ell_1}$. We condition on $W_{\dagger}^{(\ell_1)}, \dots, W_{\dagger}^{(\ell' - 1)}$ and assume that (D.7) is satisfied. We then set $V' := \left[\prod_{i=\ell' - 1}^{\ell_1} W_{\dagger}^{(i)} \right] A$, which is now a fixed matrix. By Lemma I.5, there exists an absolute constant $C_2 > 0$ such that for any $t \geq C_2$,

$$\left\| W_{\dagger}^{(\ell')} V' \right\|_{2 \rightarrow 2} \leq \sqrt{2/N} \cdot \|V'\|_{2 \rightarrow 2} \cdot \left(\sqrt{\text{Tr}(\widehat{D}^{(\ell')})} + \sqrt{\text{rank}(A)} + t \right)$$

with probability at least $1 - \exp(-c_2 \cdot t^2)$. We set $t := \sqrt{N}/L$ and get, using our assumptions,

$$\sqrt{\text{Tr}(\widehat{D}^{(\ell')})} + \sqrt{\text{rank}(A)} + t \leq \sqrt{N/2} \cdot \sqrt{1 + \frac{1}{L}} + \frac{2\sqrt{N}}{L} \leq \sqrt{N/2} \cdot \left(1 + \frac{1 + 2\sqrt{2}}{L}\right) \leq \sqrt{\frac{N}{2}} \cdot \left(1 + \frac{5}{L}\right).$$

Overall, we thus get

$$\left\| \left[\prod_{i=\ell'}^{\ell_1} W_{\dagger}^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq \left(1 + \frac{5}{L}\right)^{\ell' - \ell_1 + 1} \cdot \|A\|_{2 \rightarrow 2}$$

with probability at least $1 - \exp(-c_2 \cdot N/L^2)$. We reintroduce the randomness over $W_{\dagger}^{(\ell_1)}, \dots, W_{\dagger}^{(\ell' - 1)}$ and A and thus obtain (D.6). Via induction, (D.6) is then shown for every $\ell' \in \{\ell_1 - 1, \dots, \ell_2\}$.

In particular, letting $\ell' = \ell_2$ and recalling (D.5), we get

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \widehat{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq \left(1 + \frac{5}{L}\right)^{\ell_2 - \ell_1 + 1} \cdot \|A\|_{2 \rightarrow 2} \leq e^5 \cdot \|A\|_{2 \rightarrow 2}$$

with probability at least

$$(1 - \exp(-c_2 \cdot N/L^2))^{\ell_2 - \ell_1 + 1} \geq 1 - L \exp(-c_2 \cdot N/L^2).$$

Recall that we have conditioned on the matrices $\widehat{D}^{(i)}$ so far. Lifting the conditioning on these matrices, we get

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} \overline{D}^{(i)} W^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq e^5 \cdot \|A\|_{2 \rightarrow 2}$$

with probability at least $1 - 2L \cdot \exp(-c_3 \cdot N/L^2)$ with $c_3 := \min\{c_1, c_2\}$.

In the end, we use Proposition 2.10 and a union bound over $z_0, \dots, z_{L-1}$ to argue that, after possibly shrinking c_3 to guarantee that $c_3 < \ln(2)$, with probability at least

$$1 - \frac{L^2}{2^N} = 1 - L^2 \exp(-\ln(2) \cdot N) \geq 1 - L^2 \exp(-c_3 \cdot N/L^2)$$

we have $D^{(i)}(z_i) = \overline{D}^{(i)}(z_i)$ for every $i \in \{0, \dots, L-1\}$, which gives us

$$\left\| \left[\prod_{i=\ell_2}^{\ell_1} D^{(i)}(z_i) W^{(i)} \right] A \right\|_{2 \rightarrow 2} \leq e^5 \cdot \|A\|_{2 \rightarrow 2}$$

with probability at least $1 - 3L^2 \exp(-c_3 \cdot N/L^2) = 1 - \exp(\ln(3L^2) - c_3 \cdot N/L^2)$. Since $N \geq C \cdot L^2 \cdot \ln(eL)$, we have for $C > 0$ large enough that

$$\ln(3L^2) = 2 \cdot \ln(3L) \stackrel{\text{Lem. 1.3}}{\leq} \frac{6}{e} \cdot \ln(eL) \leq \frac{2}{C} \cdot \frac{N}{L^2} \leq \frac{c_3}{2} \cdot \frac{N}{L^2},$$

so the claim follows by taking $c := c_3/2$. $\square$

Lemma 3.3 is now an immediate consequence of Lemma 3.2.

Proof of Lemma 3.3. Note that we may assume $A \neq 0$ without loss of generality. We condition on the matrices $W^{(0)}, \dots, W^{(L-1)}$ and the biases $b^{(0)}, \dots, b^{(L-1)}$ and write

$$V' := \left[\prod_{i=L-1}^{\ell} D^{(i)}(z_i) W^{(i)} \right] A \in \mathbb{R}^{N \times \nu},$$

noting that $\|V'\|_{2 \rightarrow 2} \leq C_1 \cdot \|A\|_{2 \rightarrow 2}$ with probability at least $1 - \exp(-c_1 \cdot N/L^2)$ with absolute constants $C_1, c_1 > 0$ according to Lemma 3.2. We will assume that this holds in the following. Note that $W^{(L)}$ is independent of $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$.

By Lemma I.5 and recalling that $W^{(L)} \in \mathbb{R}^{1 \times N}$, we obtain absolute constants $C_2, c_2 > 0$ such that for every $t \geq C_2$,

$$\begin{aligned} \|W^{(L)} V'\|_{2 \rightarrow 2} &\leq \|V'\|_{2 \rightarrow 2} \cdot \left(1 + \sqrt{\text{rank}(V')} + t\right) \leq C_1 \cdot \|A\|_{2 \rightarrow 2} \cdot \left(1 + \sqrt{\text{rank}(A)} + t\right) \\ &\leq 2C_1 \cdot \|A\|_{2 \rightarrow 2} \cdot \left(\sqrt{\text{rank}(A)} + t\right) \end{aligned}$$

with probability at least $1 - \exp(-c_2 \cdot t^2)$.

Up to now, we have conditioned on $W^{(0)}, \dots, W^{(L-1)}$ and $b^{(0)}, \dots, b^{(L-1)}$. Reintroducing the randomness over these random variables yields

$$\|W^{(L)} V'\|_{2 \rightarrow 2} \leq 2C_1 \cdot \|A\|_{2 \rightarrow 2} \cdot \left(\sqrt{\text{rank}(A)} + t\right)$$

with probability at least

$$1 - \exp(-c_1 \cdot N/L^2) - \exp(-c_2 \cdot t^2) \geq 1 - 2 \exp(-c_3 \cdot t^2) \geq 1 - \exp(-c_4 \cdot t^2)$$

with $c_3 := \min\{c_1, c_2\}$ and $c_4 := c_3/2$, using that $C \leq t \leq \sqrt{N}/L$ by assumption. $\square$

Next, we prove Theorem 3.4.

Proof of Theorem 3.4. According to Proposition 2.11,

$$\left\| W^{(L)} \prod_{i=L-1}^0 \bar{D}^{(i)}(x_0) W^{(i)} \right\|_p \stackrel{d}{=} \left\| W^{(L)} \prod_{i=L-1}^0 \hat{D}^{(i)} W^{(i)} \right\|_p.$$

We now condition on the $\hat{D}$ -matrices and assume

$$\frac{N}{2} \cdot \left(1 - \frac{1}{4L}\right) \leq \text{Tr}(\hat{D}^{(i)}) \leq \frac{N}{2} \cdot \left(1 + \frac{1}{4L}\right) \quad \text{for all } i \in \{0, \dots, L-1\},$$

which occurs with probability at least $(1 - 2\exp(-c_1 \cdot N/L^2))^L \geq 1 - 2L\exp(-c_1 \cdot N/L^2)$ due to Hoeffding's inequality for bounded random variables [20, Theorem 2.2.6] with an absolute constant $c_1 > 0$. For fixed $\hat{D}$ -matrices,

$$\left\| W^{(L)} \left[\prod_{i=L-1}^0 \hat{D}^{(i)} W^{(i)} \right] \right\|_p \stackrel{d}{=} \left\| \prod_{i=L}^0 W_{\dagger}^{(i)} \right\|_p,$$

with $W_{\dagger}^{(0)} \in \mathbb{R}^{\text{Tr}(\hat{D}^{(0)}) \times d}$, $W_{\dagger}^{(i)} \in \mathbb{R}^{\text{Tr}(\hat{D}^{(i)}) \times \text{Tr}(\hat{D}^{(i-1)})}$ for $i \in \{1, \dots, L-1\}$ and $W_{\dagger}^{(L)} \in \mathbb{R}^{1 \times \text{Tr}(\hat{D}^{(L-1)})}$. Here, the matrices $W_{\dagger}^{(i)}$ are independent with entries satisfying $(W_{\dagger}^{(i)})_{j,k} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N)$ for every $i \in \{0, \dots, L-1\}$ and $(W_{\dagger}^{(L)})_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We first show via induction over $\ell' \in \{0, \dots, L-1\}$ that

$$\sqrt{\frac{N}{2}} \cdot \left(1 - \frac{1}{2L}\right)^{\ell'+1} \leq \left\| \prod_{i=L}^{L-\ell'} W_{\dagger}^{(i)} \right\|_2 \leq \sqrt{\frac{N}{2}} \cdot \left(1 + \frac{1}{2L}\right)^{\ell'+1}$$

with probability at least $(1 - 2\exp(-c_2 \cdot N/L^2))^{\ell'+1}$ with an absolute constant $c_2 > 0$. To this end, we begin with the case $\ell' = 0$. From [20, Theorem 3.1.1] we obtain

$$\sqrt{\text{Tr}(\hat{D}^{(L-1)})} - \frac{\sqrt{N}}{4\sqrt{2}L} \leq \|W_{\dagger}^{(L)}\|_2 \leq \sqrt{\text{Tr}(\hat{D}^{(L-1)})} + \frac{\sqrt{N}}{4\sqrt{2}L}$$

with probability at least $1 - 2\exp(-c_2 \cdot N/L^2)$. From our assumption on $\hat{D}^{(L-1)}$ we can further bound

$$\sqrt{\text{Tr}(\hat{D}^{(L-1)})} + \frac{\sqrt{N}}{4\sqrt{2}L} \leq \sqrt{\frac{N}{2} \left(1 + \frac{1}{4L}\right)} + \frac{\sqrt{N}}{4\sqrt{2}L} \leq \sqrt{\frac{N}{2} \left(1 + \frac{1}{4L} + \frac{1}{4L}\right)} = \sqrt{\frac{N}{2} \left(1 + \frac{1}{2L}\right)}$$

and similarly

$$\sqrt{\text{Tr}(\hat{D}^{(L-1)})} - \frac{\sqrt{N}}{4\sqrt{2}L} \geq \sqrt{\frac{N}{2} \left(1 - \frac{1}{2L}\right)}.$$

Hence, the case $\ell' = 0$ is shown.

We now let $\ell' \in \{1, \dots, L-1\}$ and assume that the claim is satisfied by $\ell' - 1$. We then condition on the matrices $W_{\dagger}^{(L-\ell'+1)}, \dots, W_{\dagger}^{(L)}$ and let $Z := \prod_{i=L}^{L-\ell'+1} W_{\dagger}^{(i)}$. From the rotation invariance of the Gaussian distribution we see

$$(ZW_{\dagger}^{(L-\ell')})^T \sim \mathcal{N}\left(0, \frac{2}{N} \cdot \|Z\|_2^2 \cdot I_{\text{Tr}(\hat{D}^{(L-\ell'-1)})}\right).$$

We apply [20, Theorem 3.1.1] and get

$$\sqrt{\frac{2}{N}} \cdot \|Z\|_2 \cdot \left(\sqrt{\text{Tr}(\hat{D}^{(L-\ell'-1)})} - \frac{\sqrt{N}}{4\sqrt{2}L} \right) \leq \|ZW_{\dagger}^{(L-\ell')}\|_2 \leq \sqrt{\frac{2}{N}} \cdot \|Z\|_2 \cdot \left(\sqrt{\text{Tr}(\hat{D}^{(L-\ell'-1)})} + \frac{\sqrt{N}}{4\sqrt{2}L} \right)$$

with probability at least $1 - 2\exp(-c_2 \cdot N/L^2)$. Identically to above, we bound

$$\sqrt{\text{Tr}(\hat{D}^{(L-\ell'-1)})} + \frac{\sqrt{N}}{4\sqrt{2}L} \leq \sqrt{\frac{N}{2} \left(1 + \frac{1}{2L}\right)}$$

and

$$\sqrt{\text{Tr}(\widehat{D}^{(L-\ell'-1)})} - \frac{\sqrt{N}}{4\sqrt{2}L} \geq \sqrt{\frac{N}{2}} \left(1 - \frac{1}{2L}\right).$$

We now inductively employ the bound on $\|Z\|_2$ and get the desired result using the independence of the random matrices $W_{\dagger}^{(i)}$.

We have shown that

$$\frac{1}{2} \cdot \sqrt{N/2} \leq \sqrt{\frac{N}{2}} \cdot \left(1 - \frac{1}{2L}\right)^L \leq \left\| \prod_{i=L}^1 W_{\dagger}^{(i)} \right\|_2 \leq \sqrt{\frac{N}{2}} \cdot \left(1 + \frac{1}{2L}\right)^L \leq \sqrt{N/2} \cdot e$$

with probability at least $(1 - 2\exp(-c_2 \cdot N/L^2))^L \geq 1 - 2L\exp(-c_2 \cdot N/L^2)$. For the final step, we condition on the matrices $W_{\dagger}^{(1)}, \dots, W_{\dagger}^{(L)}$, let $Z := \prod_{i=L}^1 W_{\dagger}^{(i)}$ and assume

$$\frac{1}{2} \cdot \sqrt{\frac{N}{2}} \leq \|Z\|_2 \leq \sqrt{\frac{N}{2}} \cdot e.$$

Since $(ZW_{\dagger}^{(0)})^T \sim \mathcal{N}\left(0, \frac{2}{N} \cdot \|Z\|_2^2 \cdot I_d\right)$, Lemma 3.5 provides absolute constants $C', c' > 0$ such that:

(1) If $p \in [1, 2]$ and $d \geq C'$, we have

$$\mathbb{P}\left(c' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot d^{1/p} \leq \|ZW_{\dagger}^{(0)}\|_p \leq C' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot d^{1/p}\right) \geq 1 - C' \exp(-c' \cdot d).$$

(2) For $p \in (2, c' \cdot \ln(d))$ and $d \geq C'$ we have

$$\begin{aligned} \mathbb{P}\left(c' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot \sqrt{p} \cdot d^{1/p} \leq \|ZW_{\dagger}^{(0)}\|_p \leq C' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot \sqrt{p} \cdot d^{1/p}\right) \\ \geq 1 - C' \exp(-c' \cdot p \cdot d^{2/p}). \end{aligned}$$

(3) If $p \geq \ln(d)$ and $d \geq C'$, we have

$$\begin{aligned} \mathbb{P}\left(c' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot \sqrt{\ln(d)} \leq \|ZW_{\dagger}^{(0)}\|_p \leq C' \cdot \frac{\sqrt{2} \cdot \|Z\|_2}{\sqrt{N}} \cdot \sqrt{\ln(d)}\right) \\ \geq 1 - C' \exp(-c' \cdot \ln(d)). \end{aligned}$$

Up to now we conditioned on $W_{\dagger}^{(1)}, \dots, W_{\dagger}^{(L)}$. Reintroducing the randomness over those random matrices, we find

(1) If $p \in [1, 2]$ and $d \geq C'$, we have

$$\mathbb{P}\left(\frac{c'}{2} \cdot d^{1/p} \leq \left\| \prod_{i=L}^0 W_{\dagger}^{(i)} \right\|_p \leq C' \cdot e \cdot d^{1/p}\right) \geq 1 - C' \exp(-c' \cdot d) - 2L \exp(-c_2 \cdot N/L^2).$$

(2) For $p \in (2, c' \cdot \ln(d))$ and $d \geq C'$ we have

$$\begin{aligned} \mathbb{P}\left(\frac{c'}{2} \cdot \sqrt{p} \cdot d^{1/p} \leq \left\| \prod_{i=L}^0 W_{\dagger}^{(i)} \right\|_p \leq C' \cdot e \cdot \sqrt{p} \cdot d^{1/p}\right) \\ \geq 1 - C' \exp(-c' \cdot p \cdot d^{2/p}) - 2L \exp(-c_2 \cdot N/L^2). \end{aligned}$$

(3) If $p \geq \ln(d)$ and $d \geq C'$, we have

$$\begin{aligned} \mathbb{P}\left(\frac{c'}{2} \cdot \sqrt{\ln(d)} \leq \left\| \prod_{i=L}^0 W_{\dagger}^{(i)} \right\|_p \leq C' \cdot e \cdot \sqrt{\ln(d)}\right) \\ \geq 1 - C' \exp(-c' \cdot \ln(d)) - 2L \exp(-c_2 \cdot N/L^2). \end{aligned}$$

Finally, we reintroduce the randomness over the $\widehat{D}$ -matrices and apply Proposition 2.10, which yields the following with $c_3 := \min\{c_1, c_2, \ln(2)\}$:

(1) If $p \in [1, 2]$ and $d \geq C'$, we have

$$\mathbb{P}\left(\frac{c'}{2} \cdot d^{1/p} \leq \|\overline{\nabla}\Phi(x_0)\|_p \leq C' \cdot e \cdot d^{1/p}\right) \geq 1 - C' \exp(-c' \cdot d) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N}.$$

(2) For $p \in (2, c' \cdot \ln(d))$ and $d \geq C'$ we have

$$\begin{aligned} & \mathbb{P}\left(\frac{c'}{2} \cdot \sqrt{p} \cdot d^{1/p} \leq \|\overline{\nabla}\Phi(x_0)\|_p \leq C' \cdot e \cdot \sqrt{p} \cdot d^{1/p}\right) \\ & \geq 1 - C' \exp(-c' \cdot p \cdot d^{2/p}) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N}. \end{aligned}$$

(3) If $p \geq \ln(d)$ and $d \geq C'$, we have

$$\begin{aligned} & \mathbb{P}\left(\frac{c'}{2} \cdot \sqrt{\ln(d)} \leq \|\overline{\nabla}\Phi(x_0)\|_p \leq C' \cdot e \cdot \sqrt{\ln(d)}\right) \\ & \geq 1 - C' \exp(-c' \cdot \ln(d)) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N}. \end{aligned}$$

Since $c_3 \leq \ln(2)$, we get

$$\begin{aligned} 4L \exp(-c_3 \cdot N/L^2) + \frac{L}{2^N} &= \exp(\ln(4L) - c_3 \cdot N/L^2) + \exp(\underbrace{\ln(L)}_{\leq \ln(4L)} - \underbrace{\ln(2) \cdot N}_{\geq c_3 \cdot N/L^2}) \\ &\leq \exp(\underbrace{\ln(2)}_{\leq \ln(4L)} + \ln(4L) - c_3 \cdot N/L^2) \\ &\leq \exp(2 \ln(4L) - c_3 \cdot N/L^2) \\ &\leq \exp(-c_4 \cdot N/L^2) \end{aligned}$$

with $c_4 := c_3/2$ and since $N \geq C \cdot L^2 \ln(eL)$, which ensures for C large enough that

$$2 \ln(4L) \stackrel{\text{Lem. 1.3}}{\leq} \frac{8}{e} \cdot \ln(eL) \leq \frac{c_3}{2} \cdot \frac{N}{L^2}.$$

To conclude the proof, we again distinguish the three cases.

(1) Let us first consider the case $p \in [1, 2]$. Note that $d \geq C$ and hence $\ln(C') \leq \frac{c'}{2} \cdot d$, for C large enough. We then get, with $c_5 := c'/2$,

$$\begin{aligned} 1 - C' \exp(-c' \cdot d) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N} &\geq 1 - C' \exp(-c' \cdot d) - \exp(-c_4 \cdot N/L^2) \\ &\geq 1 - \exp(-c_5 \cdot d) - \exp(-c_4 \cdot N/L^2) \\ &\geq 1 - \exp(-c_6 \cdot \min\{d, N/L^2\}) \end{aligned}$$

with $c_6 := \min\{c_4, c_5\}/2$ and using that $\min\{d, N/L^2\} \geq C$ by assumption.

(2) We now consider the case $p \in (2, c' \cdot \ln(d))$. Using the well-known estimate $\exp(2t) \geq 1 + 2t \geq t$ for $t > 0$, we find

$$\exp\left(\frac{2}{p} \cdot \ln(d)\right) \geq \frac{\ln(d)}{p},$$

which is equivalent to

$$p \cdot d^{2/p} \geq \ln(d).$$

We choose C large enough, such that for $d \geq C$,

$$\ln(C') \leq \frac{c'}{2} \cdot \ln(d) \leq \frac{c'}{2} \cdot p \cdot d^{2/p}.$$

We then get

$$\begin{aligned}
& 1 - C' \exp(-c' \cdot p \cdot d^{2/p}) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N} \\
& \geq 1 - C' \exp(-c' \cdot p \cdot d^{2/p}) - \exp(-c_4 \cdot N/L^2) \\
& \geq 1 - \exp(-c_5 \cdot p \cdot d^{2/p}) - \exp(-c_4 \cdot N/L^2) \\
& \geq 1 - \exp(-c_6 \cdot \min\{p \cdot d^{2/p}, N/L^2\}).
\end{aligned}$$

- (3) Lastly, if $p \geq \ln(d)$ we note as in Case (1) that we may assume $\ln(C') \leq \frac{c'}{2} \cdot \ln(d)$. This gives us

$$\begin{aligned}
& 1 - C' \exp(-c' \cdot \ln(d)) - 4L \exp(-c_3 \cdot N/L^2) - \frac{L}{2^N} \\
& \geq 1 - C' \exp(-c' \ln(d)) - \exp(-c_4 \cdot N/L^2) \\
& \geq 1 - \exp(-c_5 \cdot \ln(d)) - \exp(-c_4 \cdot N/L^2) \\
& \geq 1 - \exp(-c_6 \cdot \min\{\ln(d), N/L^2\}),
\end{aligned}$$

where in the last step we again used that $\min\{\ln(d), N/L^2\} \geq \ln(C)$ by assumption.

Overall, the claim is shown. $\square$

We conclude this appendix by proving Lemma 3.5, which follows by combining two results from [16].

Proof of Lemma 3.5. According to [16, Proposition 2.4], there exist absolute constants $C_1, c_1 > 0$, such that for every $d \in \mathbb{N}$ and $p \in [1, \infty)$ with $p < \ln(d)$,

$$c_1 \cdot \sqrt{p} \cdot d^{1/p} \leq \mathbb{E}_{X \sim \mathcal{N}(0, I_d)} [\|X\|_p] \leq C_1 \cdot \sqrt{p} \cdot d^{1/p},$$

whereas for $d \in \mathbb{N}$ and $p \in [1, \infty]$ with $p \geq \ln(d)$, one has

$$c_1 \cdot \sqrt{\ln(d)} \leq \mathbb{E}_{X \sim \mathcal{N}(0, I_d)} [\|X\|_p] \leq C_1 \cdot \sqrt{\ln(d)}.$$

An application of [16, Theorem 4.11] (with $\varepsilon = 1/2$) yields the existence of absolute constants $C_2, c_2, c_3 > 0$ such that for every $d \geq C_2$ and $p \in [1, \infty]$ one has

$$\mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{1}{2} \cdot \mathbb{E} [\|X\|_p] \leq \|X\|_p \leq \frac{3}{2} \cdot \mathbb{E} [\|X\|_p] \right) \geq 1 - C_2 \exp(-c_2 \cdot \beta(d, p, 1/2)),$$

where

$$\beta(d, p, 1/2) := \begin{cases} \frac{d}{4} & \text{if } 1 \leq p \leq 2, \\ \max \left\{ \min \left\{ p^2 \cdot 2^{-p} \cdot \frac{d}{4}, \left(\frac{d}{2} \right)^{2/p} \right\}, \frac{1}{2} \cdot p \cdot d^{2/p} \right\} & \text{if } 2 < p \leq c_3 \cdot \ln(d), \\ \ln(d) \cdot \frac{1}{2} & \text{if } p > c_3 \cdot \ln(d). \end{cases}$$

If $p \in [1, 2]$ and $d \geq \max\{C_2, 8\}$, we get $\ln(d) \geq \ln(8) > 2 \geq p$ and hence

$$c_1 \cdot d^{1/p} \leq c_1 \cdot \sqrt{p} \cdot d^{1/p} \leq \mathbb{E}_{X \sim \mathcal{N}(0, I_d)} [\|X\|_p] \leq C_1 \cdot \sqrt{p} \cdot d^{1/p} \leq C_1 \cdot \sqrt{2} \cdot d^{1/p}.$$

We thus get

$$\begin{aligned}
\mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{c_1}{2} \cdot d^{1/p} \leq \|X\|_p \leq \frac{3}{\sqrt{2}} \cdot C_1 \cdot d^{1/p} \right) & \geq \mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{1}{2} \cdot \mathbb{E} [\|X\|_p] \leq \|X\|_p \leq \frac{3}{2} \cdot \mathbb{E} [\|X\|_p] \right) \\
& \geq 1 - C_2 \exp\left(-\frac{c_2}{4} \cdot d\right)
\end{aligned}$$

in the case $p \in [1, 2]$ and $d \geq \max\{C_2, 8\}$.

In the case $p \in (2, c_3 \cdot \ln(d))$ we have $\ln(d) > p$, and therefore

$$c_1 \cdot \sqrt{p} \cdot d^{1/p} \leq \mathbb{E}_{X \sim \mathcal{N}(0, I_d)} [\|X\|_p] \leq C_1 \cdot \sqrt{p} \cdot d^{1/p}.$$

Moreover, we get

$$\beta(d, p, 1/2) = \max \left\{ \min \left\{ p^2 \cdot 2^{-p} \cdot \frac{d}{4}, \left(\frac{d}{2} \right)^{2/p} \right\}, \frac{1}{2} \cdot p \cdot d^{2/p} \right\} \geq \frac{1}{2} \cdot p \cdot d^{2/p}.$$

Therefore, for $d \geq C_2$,

$$\begin{aligned} & \mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{c_1}{2} \cdot \sqrt{p} \cdot d^{1/p} \leq \|X\|_p \leq \frac{3}{2} \cdot C_1 \cdot \sqrt{p} \cdot d^{1/p} \right) \\ & \geq \mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{1}{2} \cdot \mathbb{E} [\|X\|_p] \leq \|X\|_p \leq \frac{3}{2} \cdot \mathbb{E} [\|X\|_p] \right) \\ & \geq 1 - C_2 \exp \left(-\frac{c_2}{2} \cdot p \cdot d^{2/p} \right). \end{aligned}$$

Lastly, in the case $p \geq \ln(d) \geq \ln(8) > 2$, we have

$$c_1 \cdot \sqrt{\ln(d)} \leq \mathbb{E} [\|X\|_p] \leq C_1 \cdot \sqrt{\ln(d)}$$

and

$$\beta(d, p, 1/2) \geq \min \left\{ \ln(d) \cdot \frac{1}{2}, \frac{1}{2} \cdot p \cdot d^{2/p} \right\} = \ln(d) \cdot \frac{1}{2}.$$

Hence, we get

$$\begin{aligned} & \mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{c_1}{2} \cdot \sqrt{\ln(d)} \leq \|X\|_p \leq \frac{3}{2} \cdot C_1 \cdot \sqrt{\ln(d)} \right) \\ & \geq \mathbb{P}_{X \sim \mathcal{N}(0, I_d)} \left(\frac{1}{2} \cdot \mathbb{E} [\|X\|_p] \leq \|X\|_p \leq \frac{3}{2} \cdot \mathbb{E} [\|X\|_p] \right) \\ & \geq 1 - C_2 \exp \left(-\frac{c_2}{2} \cdot \ln(d) \right) \end{aligned}$$

for every $d \geq \max\{C_2, 8\}$.

In the end, the claim follows letting $C := \max\{\frac{3}{\sqrt{2}} \cdot C_1, C_2, 8\}$ and $c := \min\{\frac{c_1}{2}, \frac{c_2}{4}, c_3\}$. $\square$

E Proof of Theorem 4.1

In this appendix, we prove Theorem 4.1, i.e., the upper bound for the Lipschitz constant of random *zero-bias* ReLU networks $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ following Assumption 2.1. In fact, we even aim to show a more general statement that will be used later in Section 6.1 to tackle the case of random networks with general biases. To be precise, using the notation $\overline{\text{D}}\Phi^{(\ell_1) \rightarrow (\ell_2)}$ introduced in Definition 2.5, we aim to bound expressions of the form

$$\sup_{x \in \mathbb{S}^{d-1}, A} \left\| \overline{\text{D}}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2} \quad \text{and} \quad \sup_{x \in \mathbb{S}^{d-1}, A} \left\| W^{(L)} \overline{\text{D}}\Phi^{(\ell) \rightarrow (L-1)}(x)A \right\|_{2 \rightarrow 2}$$

for suitable indices $\ell, \ell_1, \ell_2 \in \{0, \dots, L-1\}$, where A runs over all diagonal matrices where the diagonal entries have absolute value either zero or one and that satisfy a certain sparsity condition; see Equation (E.1) for a precise definition. We note that the special case of $\ell = 0$ in the second expression yields a bound for

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| \overline{\nabla}\Phi(x) \right\|_2,$$

which implies a bound for $\text{Lip}_2(\Phi)$ according to Theorem 2.8.

First, we provide a decomposition of the difference between the formal gradients of the network at two input points; a similar decomposition has been used in [2].

Lemma E.1. *Let $x, y \in \mathbb{R}^d$ and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a ReLU network with L hidden layers of width N as in (2.1). Then, for every $\ell_1 \in \{0, \dots, L-1\}$, $\ell_2 \in \{-1, \dots, L-1\}$,*

$$\begin{aligned} & \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)} \right) (x) - \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)} \right) (y) \\ &= \sum_{j=\ell_1}^{\ell_2} \left(\overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(y) \right). \end{aligned}$$

Proof. Fix $\ell_1 \in \{0, \dots, L-1\}$. The proof is via induction over $\ell_2 \in \{-1, \dots, L-1\}$, where the case $\ell_2 < \ell_1$ is trivial. Therefore, we take $\ell_2 \in \{\ell_1, \dots, L-1\}$ and assume that $\ell_2 - 1$ satisfies the claim. We then get

$$\begin{aligned} & \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)} \right) (x) - \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)} \right) (y) \\ &= D^{(\ell_2)}(x) W^{(\ell_2)} \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (x) - D^{(\ell_2)}(y) W^{(\ell_2)} \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (y) \\ &= D^{(\ell_2)}(x) W^{(\ell_2)} \left(\left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (x) - \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (y) \right) \\ & \quad + \left(D^{(\ell_2)}(x) - D^{(\ell_2)}(y) \right) W^{(\ell_2)} \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (y) \\ &\stackrel{\text{IH}}{=} D^{(\ell_2)}(x) W^{(\ell_2)} \left(\sum_{j=\ell_1}^{\ell_2-1} \left(\overline{D}\Phi^{(j+1) \rightarrow (\ell_2-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(y) \right) \right) \\ & \quad + \left(D^{(\ell_2)}(x) - D^{(\ell_2)}(y) \right) W^{(\ell_2)} \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (y) \\ &= \sum_{j=\ell_1}^{\ell_2-1} \left(\overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(y) \right) \\ & \quad + \underbrace{\overline{D}\Phi^{(\ell_2+1) \rightarrow (\ell_2)}(x)}_{=I_{N \times N}} \left(D^{(\ell_2)}(x) - D^{(\ell_2)}(y) \right) W^{(\ell_2)} \left(\overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)} \right) (y) \\ &= \sum_{j=\ell_1}^{\ell_2} \left(\overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(y) \right). \end{aligned}$$

By induction, the claim is shown. $\square$

For given $\delta > 0$, $\ell \in \mathbb{N}$, we let

$$\mathcal{A}_\delta^{(0)} := \{I_{d \times d}\} \quad \text{and} \quad \mathcal{A}_\delta^{(\ell)} := \{A \in \text{diag}\{0, 1, -1\}^{N \times N} : \text{Tr}|A| \leq \delta N\}. \quad (\text{E.1})$$

The following result is the main theorem of this section. Theorem 4.1 follows immediately from it by taking $\delta \asymp \frac{d}{N} \ln(N/d)$.

Theorem E.2. *There exist absolute constants $C > 0$ and $c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network satisfying Assumption 2.1 with L hidden layers of width N satisfying $N \geq C \cdot L^3 d \ln(N/d)$. Moreover, let $\delta \in (0, e^{-1})$ with*

$$N \geq C \cdot d L^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot d \cdot \ln(1/\delta) \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c/L^2.$$

Then with probability at least $1 - \exp(-c \cdot \delta N)$,

$$\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x) A \right\|_{2 \rightarrow 2} \leq C \quad \text{for all } \ell_1 \in \{0, \dots, L-1\} \text{ and } \ell_2 \in \{-1, \dots, L-1\}.$$

Moreover, with probability at least $1 - \exp(-c \cdot \delta N)$,

$$\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell)}}} \left\| W^{(L)} \overline{D} \Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \leq C \cdot \sqrt{N \cdot \delta \ln(1/\delta)} \quad \text{for all } \ell \in \{0, \dots, L-1\}.$$

Finally, under the additional assumptions

$$\delta \ln^{1/2}(1/\delta) \leq c \cdot \frac{\sqrt{d}}{L\sqrt{N}}, \quad \delta \leq \frac{c}{L^3 \ln(N/d)} \quad \text{and} \quad d \ln(1/\delta) \geq C,$$

we obtain

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| \overline{\nabla} \Phi(x) \right\|_2 \leq C \cdot \sqrt{d \ln(1/\delta)}$$

with probability at least $1 - \exp(-c \cdot d \ln(1/\delta))$.

We first prove the following auxiliary statement.

Lemma E.3. *There exist absolute constants $C > 0$ and $c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network satisfying Assumption 2.1 with L hidden layers of width N satisfying $N \geq C \cdot L^3 d \ln(N/d)$. Moreover, let $\delta \in (0, e^{-1})$ with*

$$N \geq C \cdot d L^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot d \cdot \ln(1/\delta) \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c \cdot 1/L^2.$$

We let $C_1 > 0$ be the absolute constant from Lemma 3.2 and $c' \in (0, 1)$ be the absolute constant from Lemma C.6 and set $\varepsilon := \frac{c'}{16C_1} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$. Then, for every $\ell_2 \in \{-1, \dots, L-1\}$ with probability at least $1 - 4(\ell_2 + 1) \exp(-c \cdot \delta N)$, all of the following statements are satisfied simultaneously:

- (1) $\sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(j)}(x) - D^{(j)}(y) \right] \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2\},$
- (2) $\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D} \Phi^{(\ell_1) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \leq C \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{L d \ln(N/d)}{N}} \right)$
for all $\ell_1 \in \{0, \dots, L-1\}$ and $j \in \{0, \dots, \ell_2\},$
- (3) $\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D} \Phi^{(\ell_1) \rightarrow (j)}(x) A \right\|_{2 \rightarrow 2} \leq C \quad \text{for all } \ell_1 \in \{0, \dots, L-1\} \text{ and } j \in \{-1, \dots, \ell_2\},$
- (4) $\left\| \Phi^{(j)}(x) \right\|_2 \geq \frac{\|x\|_2}{4} \quad \text{for every } x \in \mathbb{R}^d \text{ and } j \in \{0, \dots, \ell_2\}.$

Proof. The proof is via induction over ℓ_2 . In the case $\ell_2 = -1$, only (3) needs to be considered. However, here we have, since necessarily $j = -1$,

$$\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D} \Phi^{(\ell_1) \rightarrow (j)}(x) A \right\|_{2 \rightarrow 2} = \sup_{A \in \mathcal{A}_\delta^{(\ell_1)}} \|A\|_{2 \rightarrow 2} = 1$$

for every $\ell_1 \in \{0, \dots, L-1\}$.

Let us move to the induction step. Pick $\ell_2 \in \{0, \dots, L-1\}$ and assume that with probability at least $1 - 4\ell_2 \exp(-c \cdot \delta N)$, all of the following are satisfied:

- (a) $\sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(j)}(x) - D^{(j)}(y) \right] \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\},$
- (b) $\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D} \Phi^{(\ell_1) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{L d \ln(N/d)}{N}} \right)$
for all $\ell_1 \in \{0, \dots, L-1\}$ and $j \in \{0, \dots, \ell_2 - 1\},$

- (c) $\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (j)}(x)A \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } \ell_1 \in \{0, \dots, L-1\} \text{ and } j \in \{0, \dots, \ell_2-1\},$
- (d) $\left\| \Phi^{(j)}(x) \right\|_2 \geq \frac{\|x\|_2}{4} \quad \text{for every } x \in \mathbb{R}^d \text{ and } j \in \{0, \dots, \ell_2-1\},$

where $C_2, c > 0$ are appropriate constants to be exactly determined later and we recall that $C_1 > 0$ is the absolute constant from Lemma 3.2.

We now condition on the weights $W^{(0)}, \dots, W^{(\ell_2-1)}$ and assume that properties (a)-(d) are satisfied. Note that, by picking $\ell_1 = 0$ in (c) and by Theorem 2.8, this in particular implies

$$\text{Lip}_{2 \rightarrow 2}(\Phi^{(j)}) \leq \sup_{x \in \mathbb{S}^{d-1}} \left\| \overline{D}\Phi^{(j)}(x) \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } j \in \{0, \dots, \ell_2-1\}. \quad (\text{E.2})$$

We set

$$\mu := \begin{cases} d & \text{if } \ell_2 = 0, \\ N & \text{if } \ell_2 > 0. \end{cases}$$

Let

$$f : \mathbb{S}^{d-1} \rightarrow \mathbb{S}^{\mu-1}, \quad f(x) = \frac{\Phi^{(\ell_2-1)}(x)}{\left\| \Phi^{(\ell_2-1)}(x) \right\|_2},$$

which is well-defined according to (d). For $x, y \in \mathbb{S}^{d-1}$ we have

$$\|f(x) - f(y)\|_2 \stackrel{\text{Lem. I.2}}{\leq} 2 \cdot \frac{\left\| \Phi^{(\ell_2-1)}(x) - \Phi^{(\ell_2-1)}(y) \right\|_2}{\left\| \Phi^{(\ell_2-1)}(x) \right\|_2} \stackrel{(\text{E.2}), (d)}{\leq} 16C_1 \cdot \|x - y\|_2.$$

Thus, f is $(16C_1)$ -Lipschitz. Note that for $x, y \in \mathbb{S}^{d-1}$ with $\|x - y\|_2 \leq \varepsilon$ we have

$$\|f(x) - f(y)\|_2 \leq 16C_1 \cdot \varepsilon = c' \cdot \delta \cdot \ln^{-1/2}(1/\delta).$$

Hence, we may apply Lemma C.6 by using that

$$N \geq C \cdot d \cdot \delta^{-1} \cdot \ln(1/\delta)$$

and get

$$\begin{aligned} & \sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \text{Tr} \left| D^{(\ell_2)}(y) - D^{(\ell_2)}(x) \right| \\ & \leq \sup_{\substack{z_1, z_2 \in f(\mathbb{S}^{d-1}), \\ \|z_1 - z_2\|_2 \leq c' \cdot \delta \cdot \ln^{-1/2}(1/\delta)}} \# \left\{ i \in \{1, \dots, N\} : \text{sgn}((W^{(\ell_2)} z_1)_i) \neq \text{sgn}((W^{(\ell_2)} z_2)_i) \right\} \leq \delta N \end{aligned} \quad (\text{E.3})$$

with probability at least $1 - \exp(-c' \cdot \delta N)$ over the randomness in $W^{(\ell_2)}$. We denote the event on which (E.3) holds $\mathcal{E}_1$.

Further, fix $\ell_1 \in \{0, \dots, L-1\}$ and set

$$\mathcal{V} := \left\{ \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)}(x)A_1 : x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \right\}.$$

Note that we have $\|M\|_{2 \rightarrow 2} \leq 2C_1$ and $\text{rank}(M) \leq \delta N$ for every $M \in \mathcal{V}$ by (c) and by definition of $\mathcal{A}_\delta^{(\ell_1)}$. Here, we used that $d \leq C \cdot d \cdot \ln(1/\delta) \leq \delta N$. Hence, for fixed $x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)}$, and $A_2 \in \mathcal{A}_\delta^{(1)}$, Lemma I.5 shows for absolute constants $C_3, c_1 > 0$ that for every $t \geq C_3$, the condition

$$\left\| A_2 W^{(\ell_2)} \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2-1)}(x)A_1 \right\|_{2 \rightarrow 2} \leq \sqrt{2} \cdot 2C_1 \cdot \frac{\sqrt{\delta N} + \sqrt{\delta N} + t}{\sqrt{N}}$$

is true with probability at least $1 - \exp(-c_1 \cdot t^2)$. Moreover, combining Lemma I.6 and [13, Lemma 5.7], we get

$$\ln(|\mathcal{V}|) + \ln \left(\left| \mathcal{A}_\delta^{(1)} \right| \right) \leq L(d+1) \cdot \ln \left(\frac{eN}{d+1} \right) + 2\delta N \cdot \ln(4e/\delta)$$

$$\begin{aligned}
&\stackrel{\text{Lem. I.3}}{\leq} 2eLd \cdot \ln(N/d) + 8e \cdot \delta N \cdot \ln(1/\delta) \\
&\leq 8e(Ld \ln(N/d) + \delta N \cdot \ln(1/\delta)).
\end{aligned} \tag{E.4}$$

Hence, via a union bound, we observe for every $t \geq C_3$ that the condition

$$\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(\ell_2)} \overline{D} \Phi^{(\ell_1) \rightarrow (\ell_2-1)}(x) A_1 \right\|_{2 \rightarrow 2} \leq \sqrt{2} \cdot 2C_1 \cdot \frac{\sqrt{\delta N} + \sqrt{\delta N} + t}{\sqrt{N}}$$

is satisfied with probability at least

$$1 - \exp\left(8e(Ld \ln(N/d) + \delta N \cdot \ln(1/\delta)) - c_1 t^2\right).$$

We then explicitly pick

$$t := \sqrt{c_1^{-1} \cdot 9e \cdot (Ld \ln(N/d) + \delta N \cdot \ln(1/\delta))},$$

which yields the existence of an absolute constant $C_2 > 0$, such that

$$\begin{aligned}
&\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(\ell_2)} \overline{D} \Phi^{(\ell_1) \rightarrow (\ell_2-1)}(x) A_1 \right\|_{2 \rightarrow 2} \\
&\leq C_2 \cdot \frac{\sqrt{\delta N \cdot \ln(1/\delta)} + \sqrt{Ld \ln(N/d)}}{\sqrt{N}}
\end{aligned}$$

with probability at least

$$1 - \exp(-e \cdot (Ld \ln(N/d) + \delta N \cdot \ln(1/\delta)))$$

over the randomness in $W^{(\ell_2)}$.

Until now, $\ell_1 \in \{0, \dots, L-1\}$ was fixed. By performing a union bound over ℓ_1 , we note that

$$\begin{aligned}
&\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(\ell_2)} \overline{D} \Phi^{(\ell_1) \rightarrow (\ell_2-1)}(x) A_1 \right\|_{2 \rightarrow 2} \\
&\leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \quad \text{for all } \ell_1 \in \{0, \dots, L-1\}
\end{aligned} \tag{E.5}$$

with probability at least

$$1 - \exp(\ln(L) - e \cdot (Ld \ln(N/d) + \delta N \cdot \ln(1/\delta))) \stackrel{\ln(L) \leq L}{\geq} 1 - \exp(-\delta N)$$

over the randomness in $W^{(\ell_2)}$. We denote the event defined by property (E.5) $\mathcal{E}_2$.

Moreover, let $\mathcal{N}_\varepsilon \subseteq \mathbb{S}^{d-1}$ be an ε -net of $\mathbb{S}^{d-1}$ with

$$\begin{aligned}
\ln(|\mathcal{N}_\varepsilon|) &\leq d \cdot \ln(3/\varepsilon) = d \cdot \ln\left(\frac{48C_1 \cdot \ln^{1/2}(1/\delta)}{c' \cdot \delta}\right) \stackrel{\ln^{1/2}(1/\delta) \leq 1/\delta}{\leq} d \cdot \ln\left(\frac{48C_1}{c' \cdot \delta^2}\right) \\
&\leq 2 \cdot d \cdot \ln(48C_1/(c' \delta)) \stackrel{\text{Lem. I.3}}{\leq} C_4 \cdot d \cdot \ln(1/\delta)
\end{aligned} \tag{E.6}$$

with an absolute constant $C_4 > 0$. For $x \in \mathbb{S}^{d-1}$ we denote by $\pi(x) \in \mathcal{N}_\varepsilon$ a net point with $\|x - \pi(x)\|_2 \leq \varepsilon$. On $\mathcal{E}_1 \cap \mathcal{E}_2$, by Lemma E.1, for arbitrary $\ell_1 \in \{0, \dots, L-1\}$,

$$\begin{aligned}
&\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \overline{D} \Phi^{(\ell_1) \rightarrow (\ell_2)}(\pi(x)) A - \overline{D} \Phi^{(\ell_1) \rightarrow (\ell_2)}(x) A \right\|_{2 \rightarrow 2} \\
&\leq \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \sum_{j=\ell_1}^{\ell_2} \left\| \overline{D} \Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) W^{(j)} \overline{D} \Phi^{(\ell_1) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2}
\end{aligned}$$

$$\begin{aligned}
&\leq \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left(\sum_{j=\ell_1}^{\ell_2} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) \right\|_{2 \rightarrow 2} \right. \\
&\quad \cdot \left. \left\| \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2} \right) \\
&\leq \sum_{j=\ell_1}^{\ell_2} \left(\sup_{x \in \mathbb{S}^{d-1}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) \right\|_{2 \rightarrow 2} \right) \\
&\quad \cdot \left(\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2} \right) \\
&\stackrel{\mathcal{E}_{1,(a)}}{\leq} \sum_{j=\ell_1}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right) \\
&\quad \cdot \left(\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \right) \\
&\stackrel{\mathcal{E}_{2,(b)}}{\leq} C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \cdot \sum_{j=\ell_1}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right).
\end{aligned}$$

Since $\delta \ln(1/\delta) \leq c/L^2$ and $N \geq C \cdot L^3 d \ln(N/d)$,

$$C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \leq \frac{1}{L}.$$

Hence, on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned}
&\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(\pi(x)) A - \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x) A \right\|_{2 \rightarrow 2} \\
&\leq \frac{1}{L} \cdot \sum_{j=\ell_1}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right).
\end{aligned}$$

Recall that up to now we have conditioned on $W^{(0)}, \dots, W^{(\ell_2-1)}$. Reintroducing the randomness over these random matrices, we note that with probability at least

$$1 - 4\ell_2 \exp(-c \cdot \delta N) - \exp(-c' \cdot \delta N) - \exp(-\delta N),$$

all of the following properties are simultaneously satisfied:

- (i) $\sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left(D^{(j)}(x) - D^{(j)}(y) \right) \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2\},$
- (ii) $\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)} A_1 \right\|_{2 \rightarrow 2} \leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right)$
for all $\ell_1 \in \{0, \dots, L-1\}$ and $j \in \{0, \dots, \ell_2\},$

- (iii) $\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(\pi(x))A - \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2}$
 $\leq \frac{1}{L} \cdot \sum_{j=\ell_1}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*)A_2 \right\|_{2 \rightarrow 2} \right) \quad \text{for all } \ell_1 \in \{0, \dots, L-1\},$
(iv) $\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (j)}(x)A \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } \ell_1 \in \{0, \dots, L-1\} \text{ and } j \in \{-1, \dots, \ell_2-1\},$
(v) $\left\| \Phi^{(j)}(x) \right\|_2 \geq \frac{\|x\|_2}{4} \quad \text{for every } x \in \mathbb{R}^d \text{ and } j \in \{0, \dots, \ell_2-1\}.$

We call the event defined by properties (i)-(v) $\mathcal{E}_3$, noting that the randomness is now with respect to $W^{(0)}, \dots, W^{(\ell_2)}$.

For fixed $x^* \in \mathcal{N}_\varepsilon$, $j \in \{0, \dots, L\}$ and $A_2 \in \mathcal{A}_\delta^{(j)}$, using Lemma 3.2 we get

$$\left\| \overline{D}\Phi^{(j) \rightarrow (\ell_2)}(x^*)A_2 \right\|_{2 \rightarrow 2} \leq C_1$$

with probability at least $1 - \exp(-c_1 \cdot N/L^2)$ with an absolute constant $c_1 > 0$. Note here that $d \leq C \cdot d \cdot \ln(1/\delta) \leq \delta N \leq N/L^2$, since $\delta \leq \delta \ln(1/\delta) \leq c/L^2$. Since

$$\ln \left(|\mathcal{A}_\delta^{(j)}| \right) \stackrel{\text{Lem. I.6}}{\leq} \delta N \cdot \ln(4e/\delta) \stackrel{\text{Lem. I.3}}{\leq} 4e \cdot \delta N \cdot \ln(1/\delta) \quad \text{and} \quad \ln(|\mathcal{N}_\varepsilon|) \stackrel{\text{(E.6)}}{\leq} C_4 \cdot d \cdot \ln(1/\delta),$$

a union bound yields that

$$\left\| \overline{D}\Phi^{(j) \rightarrow (\ell_2)}(x^*)A_2 \right\|_{2 \rightarrow 2} \leq C_1 \quad \text{for all } x^* \in \mathcal{N}_\varepsilon, j \in \{0, \dots, L\} \text{ and } A_2 \in \mathcal{A}_\delta^{(j)}, \quad (\text{E.7})$$

with probability at least

$$1 - \exp(\ln(L+1) + 4e \cdot \delta N \cdot \ln(1/\delta) + C_4 \cdot d \cdot \ln(1/\delta) - c_1 \cdot N/L^2) \geq 1 - \exp(-c_2 \cdot N/L^2).$$

For the last step, we used $N \geq C \cdot L^3 d \ln(N/d) \geq C \cdot L^2 \ln(eL)$, as well as $\delta \ln(1/\delta) \leq c/L^2$ and $N \geq C \cdot L^2 d \ln(1/\delta)$ and set $c_2 := c_1/2$. We call the event defined by (E.7) $\mathcal{E}_4$. On $\mathcal{E}_5 := \mathcal{E}_3 \cap \mathcal{E}_4$, by (iii),

$$\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(\pi(x))A - \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2} \leq C_1$$

for every $\ell_1 \in \{0, \dots, L-1\}$ and therefore

$$\begin{aligned} & \sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2} \\ & \leq \sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x^*)A \right\|_{2 \rightarrow 2} + \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell_1)}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(\pi(x))A - \overline{D}\Phi^{(\ell_1) \rightarrow (\ell_2)}(x)A \right\|_{2 \rightarrow 2} \leq 2C_1 \end{aligned}$$

for every $\ell_1 \in \{0, \dots, L-1\}$.

All together, we get that on $\mathcal{E}_5$, all of the following properties are simultaneously satisfied:

- (A) $\sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left(D^{(j)}(x) - D^{(j)}(y) \right) \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2\},$
(B) $\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell_1)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D}\Phi^{(\ell_1) \rightarrow (j-1)} A_1 \right\|_{2 \rightarrow 2} \leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right)$
for all $\ell_1 \in \{0, \dots, L-1\}$ and $j \in \{0, \dots, \ell_2\},$
(C) $\sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell_1)}}} \left\| \overline{D}\Phi^{(\ell_1) \rightarrow (j)}(x)A \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } \ell_1 \in \{0, \dots, L-1\} \text{ and } j \in \{-1, \dots, \ell_2\},$

$$(D) \quad \left\| \Phi^{(j)}(x) \right\|_2 \geq \frac{\|x\|_2}{4} \quad \text{for every } x \in \mathbb{R}^d \text{ and } j \in \{0, \dots, \ell_2 - 1\}.$$

It thus remains to prove (D) for the case $j = \ell_2$. By Lemma 3.1 and a union bound (using (E.6)),

$$\left\| \Phi^{(\ell_2)}(x^*) \right\|_2 \geq 1/2 \quad \text{for all } x^* \in \mathcal{N}_\varepsilon$$

with probability at least

$$1 - \exp(C_4 \cdot d \cdot \ln(1/\delta) - c_3 \cdot N/L^2) \geq 1 - \exp(-c_4 \cdot N/L^2).$$

Here, $c_3 > 0$ is an absolute constant, $c_4 := c_3/2$, and we used $N \geq C \cdot L^2 d \ln(1/\delta)$. We call the event defined by that property $\mathcal{E}_6$. Note that on $\mathcal{E}_5$ we have $\text{Lip}_{2 \rightarrow 2}(\Phi^{(\ell_2)}) \leq 2C_1$, as follows from (C) and Theorem 2.8. On $\mathcal{E}_7 := \mathcal{E}_5 \cap \mathcal{E}_6$, we thus get for arbitrary $x \in \mathbb{S}^{d-1}$ that

$$\left\| \Phi^{(\ell_2)}(x) \right\|_2 \geq \left\| \Phi^{(\ell_2)}(\pi(x)) \right\|_2 - \left\| \Phi^{(\ell_2)}(\pi(x)) - \Phi^{(\ell_2)}(x) \right\|_2 \geq \frac{1}{2} - 2C_1\varepsilon \geq \frac{1}{4}.$$

Since a zero-bias ReLU network is positively homogeneous, we get

$$\left\| \Phi^{(\ell_2)}(x) \right\|_2 \geq \|x\|_2/4 \quad \text{for every } x \in \mathbb{R}^d,$$

as desired.

Hence, on $\mathcal{E}_7$ all the desired properties for the induction step are satisfied. We get

$$\begin{aligned} \mathbb{P}(\mathcal{E}_7) &= \mathbb{P}(\mathcal{E}_5 \cap \mathcal{E}_6) = \mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_6) \\ &\geq 1 - 4\ell_2 \exp(-c \cdot \delta N) - \exp(-c' \cdot \delta N) - \exp(-\delta N) - \exp(-c_2 \cdot N/L^2) - \exp(-c_4 \cdot N/L^2) \\ &\geq 1 - 4(\ell_2 + 1) \exp(-c \cdot \delta N) \end{aligned}$$

by taking $c := \min\{c', c_2, c_4\}$ and using $\delta \leq \delta \cdot \ln(1/\delta) \leq c/L^2 \leq 1/L^2$. $\square$

Having shown the auxiliary statement, we can now move to the proof of Theorem E.2.

Proof of Theorem E.2. We first note that the first claim is an immediate consequence of Lemma E.3 using that $C \cdot \ln(eL) \leq \delta N$. Therefore, let us move to the second and third claim.

Step 1 (Pointwise estimate): We start with a pointwise estimate: Let $x \in \mathbb{S}^{d-1}$ be fixed, $j \in \{0, \dots, L-1\}$ and $A \in \mathcal{A}_\delta^{(j)}$. Using Lemma 3.3, we get the existence of absolute constants $C_1, c_1 > 0$ such that for every $C_1 \leq t \leq \sqrt{N}/L$, the estimate

$$\left\| W^{(L)} \bar{D} \Phi^{(j) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{\text{Tr}|A|} + t)$$

is true with probability at least $1 - \exp(-c_1 \cdot t^2)$. Note that we can ensure

$$\delta N \leq \delta \cdot \ln(1/\delta) \cdot N \leq c \cdot N/L^2 \leq N/L^2,$$

by taking C sufficiently large and c sufficiently small, which implies $\text{Tr}|A| \leq N/L^2$ for every $A \in \mathcal{A}_\delta^{(j)}$.

Step 2 (Union bound): We continue by taking a union bound: Let ε be as in Lemma E.3 and let $\mathcal{N}_\varepsilon \subseteq \mathbb{S}^{d-1}$ be an ε -net for $\mathbb{S}^{d-1}$ with

$$\ln(|\mathcal{N}_\varepsilon|) \leq C_2 \cdot d \cdot \ln(1/\delta) \tag{E.8}$$

with an absolute constant $C_2 > 0$ similar to (E.6). Let $j \in \{0, \dots, L-1\}$. Then we have

$$\ln \left(\left| \mathcal{A}_\delta^{(j)} \right| \right) \leq 4e \cdot \delta N \cdot \ln(1/\delta)$$

according to Lemmas I.3 and I.6. Since by assumption $\delta N \geq C \cdot d \cdot \ln(1/\delta) \geq C \cdot d$, we get (after possibly enlarging C_2)

$$\ln(|\mathcal{N}_\varepsilon|) + \ln \left(\left| \mathcal{A}_\delta^{(j)} \right| \right) \leq C_2 \cdot \delta N \cdot \ln(1/\delta).$$

Hence, taking a union bound, we get for every $C_1 \leq t \leq \sqrt{N}/L$ that the estimate

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \overline{D} \Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{\delta N} + t),$$

is satisfied with probability at least $1 - \exp(C_2 \cdot \delta N \cdot \ln(1/\delta) - c_1 t^2)$. We explicitly pick

$$t := \sqrt{2C_2 c_1^{-1} \cdot \delta N \cdot \ln(1/\delta)}$$

and hence get

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \overline{D} \Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_3 \cdot \sqrt{\delta N \cdot \ln(1/\delta)} \quad (\text{E.9})$$

with probability at least

$$1 - \exp(-\delta N \cdot \ln(1/\delta)) \geq 1 - \exp(-\delta N)$$

with an absolute constant $C_3 > 0$. Note that $t \leq \frac{\sqrt{N}}{L}$ can be ensured by taking C large and c small enough, since we have

$$\delta \ln(1/\delta) \leq c/L^2$$

and $t \geq C_1$ is trivially satisfied, because of $\delta N \geq C \cdot d \cdot \ln(1/\delta) \geq C \cdot d \geq C$.

In the case $j = 0$, we may omit the union bound over $\mathcal{A}_\delta^{(j)}$ (since $|\mathcal{A}_\delta^{(0)}| = 1$) and get for arbitrary $C_1 \leq t \leq \sqrt{N}/L$ that

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \left\| \overline{D} \Phi(x^*) \right\|_2 = \sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(0)}}} \left\| W^{(L)} \overline{D} \Phi^{(L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{d} + t)$$

with probability at least $1 - \exp(C_2 \cdot d \cdot \ln(1/\delta) - c_1 t^2)$. Under the additional assumption $d \cdot \ln(1/\delta) \geq C$ (see the last part of Theorem E.2), we can thus take

$$t := \sqrt{2C_2 c_1^{-1} \cdot d \ln(1/\delta)},$$

to obtain

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \left\| \overline{D} \Phi(x) \right\|_2 \leq C_3 \cdot \sqrt{d \cdot \ln(1/\delta)} \quad (\text{E.10})$$

with probability at least $1 - \exp(-d \ln(1/\delta))$, after possibly enlarging the constant C_3 .

Step 3 (Controlling the deviation):

It remains to control the deviation when moving from an arbitrary point $x \in \mathbb{S}^{d-1}$ to a net point $\pi(x) \in \mathcal{N}_\varepsilon$ with $\|x - \pi(x)\|_2 \leq \varepsilon$.

Let $C_4, c' > 0$ be the absolute constants provided by Lemma E.3. We let $\mathcal{E}_1$ be the event defined by the properties

- (1) $\sup_{x, y \in \mathbb{S}^{d-1}, \|x - y\|_2 \leq \varepsilon} \left| \text{Tr} \left(D^{(j)}(x) - D^{(j)}(y) \right) \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, L-1\},$
- (2) $\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell)} \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D} \Phi^{(\ell) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \leq C_4 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right)$
for all $\ell, j \in \{0, \dots, L-1\}$.

By Lemma E.3 we have $\mathbb{P}(\mathcal{E}_1) \geq 1 - 4L \exp(-c' \cdot \delta N)$. Using Lemma E.1, on $\mathcal{E}_1$ we get

$$\begin{aligned} & \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \overline{D} \Phi^{(\ell) \rightarrow (L-1)}(\pi(x)) A - W^{(L)} \overline{D} \Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\ & \leq \sum_{j=\ell}^{L-1} \left(\sup_{x \in \mathbb{S}^{d-1}} \left\| W^{(L)} \overline{D} \Phi^{(j+1) \rightarrow (L-1)}(\pi(x)) \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) \right\|_{2 \rightarrow 2} \right) \end{aligned}$$

$$\begin{aligned}
& \cdot \left(\sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell)}} \left\| \left(D^{(j)}(\pi(x)) - D^{(j)}(x) \right) W^{(j)} \overline{D}\Phi^{(\ell) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2} \right) \\
& \leq \sum_{j=\ell}^{L-1} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{D}\Phi^{(j+1) \rightarrow (L-1)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right) \\
& \quad \cdot \left(\sup_{\substack{x \in \mathbb{S}^{d-1}, A_1 \in \mathcal{A}_\delta^{(\ell)}, \\ A_2 \in \mathcal{A}_\delta^{(1)}}} \left\| A_2 W^{(j)} \overline{D}\Phi^{(\ell) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \right) \\
& \stackrel{(2)}{\leq} C_4 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \cdot \sum_{j=\ell}^{L-1} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{D}\Phi^{(j+1) \rightarrow (L-1)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right)
\end{aligned}$$

for arbitrary $\ell \in \{0, \dots, L-1\}$.

We let $\mathcal{E}_2$ be the event defined by

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \overline{D}\Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_3 \cdot \sqrt{N \cdot \delta \ln(1/\delta)} \quad \text{for all } j \in \{0, \dots, L\}.$$

Using (E.9) and a union bound,

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - \exp(\ln(L+1) - \delta N) \geq 1 - \exp(-c_2 \cdot \delta N)$$

with an absolute constant $c_2 > 0$, since $\delta N \geq C \cdot \ln(eL)$. On $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned}
& \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(\pi(x)) A - W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\
& \leq C_3 C_4 \cdot L \cdot \left(\sqrt{N \cdot \delta \ln(1/\delta)} + \sqrt{Ld \ln(N/d) \cdot \delta \ln(1/\delta)} \right)
\end{aligned} \tag{E.11}$$

for every $\ell \in \{0, \dots, L-1\}$. Since by assumption $\delta \ln(1/\delta) \leq c/L^2$ and $N \geq C \cdot L^3 d \ln(N/d)$, we get

$$\begin{aligned}
& \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(\pi(x)) A - W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\
& \leq \sqrt{N \cdot \delta \ln(1/\delta)}
\end{aligned}$$

for every $\ell \in \{0, \dots, L-1\}$ on $\mathcal{E}_1 \cap \mathcal{E}_2$.

Step 4 (Concluding the proof): For $\ell \in \{0, \dots, L-1\}$, we see that

$$\begin{aligned}
& \sup_{\substack{x \in \mathbb{S}^{d-1} \\ A \in \mathcal{A}_\delta^{(\ell)}}} \left\| W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\
& \leq \sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(\ell)}}} \left\| W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \\
& \quad + \sup_{x \in \mathbb{S}^{d-1}, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(\pi(x)) A - W^{(L)} \overline{D}\Phi^{(\ell) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \\
& \leq (C_3 + 1) \cdot \sqrt{N \cdot \delta \ln(1/\delta)}
\end{aligned}$$

on $\mathcal{E}_1 \cap \mathcal{E}_2$ and for $c_3 := c/2$ we see

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 4L \exp(-c' \cdot \delta N) - \exp(-c_2 \cdot \delta N) \geq 1 - \exp(-c_3 \cdot \delta N),$$

since $\delta N \geq C \cdot \ln(eL)$. This proves the second part of the theorem.

For the third part, let $\mathcal{E}_3$ be the event defined by

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{\nabla} \Phi(x)\|_2 \leq C_3 \cdot \sqrt{d \cdot \ln(1/\delta)}.$$

By (E.10),

$$\mathbb{P}(\mathcal{E}_3) \geq 1 - \exp(-d \ln(1/\delta)).$$

Moreover, under the additional assumptions $\delta \ln^{1/2}(1/\delta) \leq c \cdot \frac{\sqrt{d}}{L\sqrt{N}}$ and $\delta \leq \frac{c}{L^3 \ln(N/d)}$, on $\mathcal{E}_1 \cap \mathcal{E}_2$, using (E.11),

$$\sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \Phi(\pi(x))\|_2 \leq \sqrt{d \cdot \ln(1/\delta)}.$$

On $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we thus have

$$\sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x)\|_2 \leq \sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{\nabla} \Phi(x)\|_2 + \sup_{x \in \mathbb{S}^{d-1}} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \Phi(\pi(x))\|_2 \leq (C_3 + 1) \cdot \sqrt{d \cdot \ln(1/\delta)}.$$

To obtain the claim, note because of $\delta N \geq C \cdot d \cdot \ln(1/\delta)$ that

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \exp(-c_3 \cdot \delta N) - \exp(-d \ln(1/\delta)) \geq 1 - 2 \exp(-c_3 \cdot d \ln(1/\delta)) \geq 1 - \exp(-c_4 \cdot d \ln(1/\delta))$$

with an absolute constant $c_4 > 0$, since $d \ln(1/\delta) \geq C$ by assumption. $\square$

As a direct consequence of Lemma E.3 we explicitly state the following corollary which we make use of in Appendix G.

Corollary E.4. *There exist absolute constants $C > 0$ and $c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network satisfying Assumption 2.1 with L hidden layers of width N satisfying $N \geq C \cdot d$ and $N \geq C \cdot L^3 d \ln(N/d)$. Moreover, let $\delta \in (0, e^{-1})$ with*

$$N \geq C \cdot d L^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot d \cdot \ln(1/\delta), \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c/L^2.$$

Moreover, let $\ell \in \{-1, \dots, L-1\}$. Then with probability at least $1 - \exp(-c \cdot \delta N)$,

$$\text{for all } x \in \mathbb{R}^d : \quad \left\| \Phi^{(\ell)}(x) \right\|_2 \geq \frac{\|x\|_2}{4}.$$

F Proof of Theorem 5.1

In this section, we prove a lower bound of $\frac{\sqrt{d}}{\sqrt{L}}$ for the ℓ^1 -Lipschitz constant of a random zero-bias ReLU neural network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$. As in the proof sketch for shallow networks (see Section 5), the ultimate goal is to apply Sudakov's minoration inequality, which yields to lower bounding the packing number of the set

$$\mathcal{L} := \left\{ \bar{\mathbf{D}}\Phi^{(L-1)}(x)\nu : x \in \mathbb{S}^{d-1} \right\}.$$

We hence show that if two points $x, y \in \mathbb{S}^{d-1}$ satisfy $\|x - y\|_2 \asymp \delta$, we get

$$\left\| (\bar{\mathbf{D}}\Phi^{(L-1)}(x) - \bar{\mathbf{D}}\Phi^{(L-1)}(y))\nu \right\|_2^2 \gtrsim L \cdot \delta$$

with high probability if δ is chosen appropriately and $\nu \in \mathbb{S}^{d-1}$ is a fixed direction vector.

In order to obtain that bound, we note that, using Lemma E.1,

$$\begin{aligned} & \left\| (\bar{\mathbf{D}}\Phi^{(L-1)}(x) - \bar{\mathbf{D}}\Phi^{(L-1)}(y))\nu \right\|_2^2 \\ &= \left\| \sum_{j=0}^{L-1} \left(\bar{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \bar{\mathbf{D}}\Phi^{(j-1)}(y) \right) \nu \right\|_2^2 \end{aligned}$$

$$= \sum_{j=0}^{L-1} \|\theta_j\|_2^2 + 2 \sum_{\substack{j,k=0 \\ j < k}}^{L-1} \langle \theta_j, \theta_k \rangle,$$

where

$$\theta_j := \left(\overline{D}\Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \right) \nu \quad (\text{F.1})$$

for every $j \in \{0, \dots, L-1\}$. The idea is now to show that if $\|x - y\|_2 \asymp \delta$, we get

$$\|\theta_j\|_2^2 \gtrsim \delta \quad \text{and} \quad \langle \theta_j, \theta_k \rangle \gtrsim -\delta^{3/2} \quad \text{for every } j, k \in \{0, \dots, L-1\} \text{ with } j \neq k$$

with high probability. This yields

$$\sum_{j=0}^{L-1} \|\theta_j\|_2^2 + 2 \sum_{\substack{j,k=0 \\ j < k}}^{L-1} \langle \theta_j, \theta_k \rangle \gtrsim L \cdot \delta - L^2 \cdot \delta^{3/2} \gtrsim L \cdot \delta$$

with high probability, as soon as we pick $\delta \lesssim \frac{1}{L^2}$.

We start with two auxiliary statements taken from [1]. The formulation of the first statement in [1] is slightly different to the version presented here, which is why we include a short proof for clarification.

Lemma F.1 (cf. [1, Lemma 7.1]). *There exist constants $C, c > 0$ such that the following holds. If $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random zero-bias ReLU network following Assumption 2.1 with L hidden layers of width N with $N \geq C \cdot \varepsilon^{-2} \cdot L \cdot \ln(eL)$, $x \in \mathbb{S}^{d-1}$ with $x_d = \frac{1}{\sqrt{2}}$ and $\varepsilon \in (0, 1]$, then*

$$\left\| \Phi^{(\ell)}(x) \right\|_2 \in [1 - \varepsilon, 1 + \varepsilon] \quad \text{for all } \ell \in \{0, \dots, L-1\}$$

with probability at least $1 - \exp(-c \cdot \varepsilon^2 N / L)$.

Proof. [1, Lemma 7.1] shows that there exist absolute constants $C_1, c_1 > 0$ such that the event

$$\left\| \Phi^{(\ell)}(x) \right\|_2 \in [1 - \varepsilon, 1 + \varepsilon] \quad \text{for all } \ell \in \{0, \dots, L-1\}$$

occurs with probability at least $1 - C_1 \cdot L \cdot \exp(-c_1 \cdot N \varepsilon^2 / L)$. We may assume

$$\ln(C_1 L) \stackrel{\text{Lem. 1.3}}{\leq} \frac{C_1}{e} \cdot \ln(eL) \leq \frac{c_1}{2} \cdot N \varepsilon^2 / L$$

if we take $C \geq \frac{2C_1}{e c_1}$. Hence, the claim follows letting $c := c_1 / 2$. $\square$

The next lemma is taking verbatim from [1].

Lemma F.2 (cf. [1, Lemma 7.5]). *There exist constants $C, c > 0$ such that the following holds. If $\delta > 0$ and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a random zero-bias ReLU network with L hidden layers of width N according to Assumption 2.1 with $N \geq C \cdot L \cdot \ln(2L) \cdot \delta^{-6}$ and $\delta \leq \frac{1}{CL}$, then for any choice of $x, y \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$ and $\|x - y\|_2 \geq \delta$,*

$$\left\| \left(I_{\mu \times \mu} - \frac{\Phi^{(\ell)}(x) \Phi^{(\ell)}(x)^T}{\|\Phi^{(\ell)}(x)\|_2^2} \right) \Phi^{(\ell)}(y) \right\|_2 \geq \frac{\delta}{2} \quad \text{for all } \ell \in \{-1, \dots, L-1\}$$

with probability at least $1 - \exp(-c \cdot \delta^6 N / L)$, where we set

$$\mu := \begin{cases} d, & \text{if } \ell = -1, \\ N, & \text{otherwise.} \end{cases}$$

We remark that the expression

$$\frac{\Phi^{(\ell)}(x) \Phi^{(\ell)}(x)^T}{\|\Phi^{(\ell)}(x)\|_2^2} \Phi^{(\ell)}(y)$$

is the *orthogonal projection* of $\Phi^{(\ell)}(y)$ onto the space spanned by $\Phi^{(\ell)}(x)$. Therefore, Lemma F.2 in particular implies that

$$\text{for all } \ell \in \{-1, \dots, L-1\} : \min_{\alpha \in \mathbb{R}} \left\| \Phi^{(\ell)}(y) - \alpha \cdot \Phi^{(\ell)}(x) \right\|_2 \geq \frac{\delta}{2}$$

with high probability.

Combining these two results with Proposition 2.12 we get the following lemma.

Lemma F.3. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1. We pick $\delta \leq \frac{1}{CL}$ and $x, y \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$ and $24 \cdot \delta \geq \|x - y\|_2 \geq \delta$. Furthermore, we assume*

$$N \geq C \cdot d, \quad N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6}.$$

Then, for every $\ell \in \{0, \dots, L-1\}$, with probability at least $1 - \exp(-c \cdot d \ln(N/d))$, we have

$$c \cdot \delta N \leq \text{Tr} \left| D^{(\ell)}(x) - D^{(\ell)}(y) \right| \leq C \cdot \delta N.$$

Proof. If well defined, we set

$$\delta' := \left\| \frac{\Phi^{(\ell-1)}(x)}{\|\Phi^{(\ell-1)}(x)\|_2} - \frac{\Phi^{(\ell-1)}(y)}{\|\Phi^{(\ell-1)}(y)\|_2} \right\|_2.$$

We now derive an upper and a lower bound for δ' .

Upper bound: We let $\mathcal{E}_1$ be the event defined by

$$\left\| \Phi^{(\ell-1)}(x) \right\|_2, \left\| \Phi^{(\ell-1)}(y) \right\|_2 \geq \frac{1}{2}, \quad (\text{F.2})$$

which occurs with probability at least $1 - 2 \exp(-c_1 N/L)$, according to Lemma F.1 for an absolute constant $c_1 > 0$. Note that Lemma F.1 is applicable by taking C large enough and using the assumption

$$d \ln^2(N/d) \geq d \ln(N/d) \geq L \cdot \ln(eL) \geq \ln(eL)$$

which implies

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6} \geq C \cdot L \cdot \ln(eL).$$

Moreover, we let $C_2, c_2 > 0$ be the constants appearing in the formulation of Theorem 4.1. Taking $C \geq C_2$, the assumptions of Theorem 4.1 are satisfied. Note here in particular that

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6} \stackrel{\delta \leq L^{-1}}{\geq} C \cdot d \cdot L \cdot \ln^2(N/d) \cdot L^6 \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot L^3.$$

We thus get

$$\left\| \Phi^{(\ell-1)}(x) - \Phi^{(\ell-1)}(y) \right\|_2 \leq 24C_2 \cdot \delta,$$

with probability at least $1 - \exp(-c_2 \cdot d \cdot \ln(N/d))$. We call this event $\mathcal{E}_2$. We use Lemma I.2 and (F.2) to obtain that on $\mathcal{E}_{\text{up}} := \mathcal{E}_1 \cap \mathcal{E}_2$ the expression δ' is well-defined and we have

$$\delta' = \left\| \frac{\Phi^{(\ell-1)}(x)}{\|\Phi^{(\ell-1)}(x)\|_2} - \frac{\Phi^{(\ell-1)}(y)}{\|\Phi^{(\ell-1)}(y)\|_2} \right\|_2 \leq 96C_2 \cdot \delta,$$

with probability at least $1 - 2 \exp(-c_1 N/L) - \exp(-c_2 \cdot d \cdot \ln(N/d))$.

Lower bound: Since $d \ln^2(N/d) \geq \ln(eL)$ (see above), we get

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6} \geq C \cdot L \cdot \delta^{-2} \cdot \ln(eL).$$

We may hence apply Lemma F.1, which yields that

$$\left\| \Phi^{(\ell-1)}(x) \right\|_2, \left\| \Phi^{(\ell-1)}(y) \right\|_2 \in [1 - \delta/16, 1 + \delta/16] \quad (\text{F.3})$$

happens with probability at least $1 - 2 \exp(-c_1 \cdot \delta^2 \cdot N/L)$, after possibly decreasing c_1 . We call this event $\mathcal{E}_3$. On this event, writing $z := \frac{\Phi^{(\ell-1)}(y)}{\|\Phi^{(\ell-1)}(y)\|_2}$, we obtain

$$\begin{aligned} \left\| \frac{\Phi^{(\ell-1)}(x)}{\|\Phi^{(\ell-1)}(x)\|_2} - z \right\|_2 &\geq \underbrace{\frac{1}{\|\Phi^{(\ell-1)}(x)\|_2}}_{\leq 2} \cdot \left\| \Phi^{(\ell-1)}(x) - z \right\|_2 - \left| 1 - \frac{1}{\|\Phi^{(\ell-1)}(x)\|_2} \right| \cdot \underbrace{\|z\|_2}_{=1} \\ &\geq \frac{1}{2} \cdot \left\| \Phi^{(\ell-1)}(x) - z \right\|_2 - \left| 1 - \frac{1}{\|\Phi^{(\ell-1)}(x)\|_2} \right|. \end{aligned}$$

Note that $\|\Phi^{(\ell-1)}(x)\|_2 \in [1 - \delta/16, 1 + \delta/16]$ implies

$$\frac{1}{\|\Phi^{(\ell-1)}(x)\|_2} \in \left[\frac{1}{1 + \delta/16}, \frac{1}{1 - \delta/16} \right] = \left[1 - \frac{\delta/16}{1 + \delta/16}, 1 + \frac{\delta/16}{1 - \delta/16} \right] \subseteq [1 - \delta/16, 1 + \delta/8]$$

using $1 - \delta/16 \geq 1/2$ at the last step. Therefore, we observe

$$\left\| \frac{\Phi^{(\ell-1)}(x)}{\|\Phi^{(\ell-1)}(x)\|_2} - z \right\|_2 \geq \frac{1}{2} \cdot \left\| \Phi^{(\ell-1)}(x) - z \right\|_2 - \frac{\delta}{8}$$

on $\mathcal{E}_3$. Moreover, Lemma F.2 (note that $N \geq C \cdot \delta^{-6} \cdot L \cdot \ln(CL)$) tells us that

$$\min_{\alpha \in \mathbb{R}} \left\| \Phi^{(\ell-1)}(x) - \alpha \cdot \Phi^{(\ell-1)}(y) \right\|_2 \geq \frac{\delta}{2}$$

holds with probability at least $1 - \exp(-c_3 \cdot \delta^6 \cdot N/L)$ with an absolute constant $c_3 > 0$. We call this event $\mathcal{E}_4$. In summary,

$$\delta' = \left\| \frac{\Phi^{(\ell-1)}(x)}{\|\Phi^{(\ell-1)}(x)\|_2} - \frac{\Phi^{(\ell-1)}(y)}{\|\Phi^{(\ell-1)}(y)\|_2} \right\|_2 \geq \frac{\delta}{8}$$

on $\mathcal{E}_{\text{low}} := \mathcal{E}_3 \cap \mathcal{E}_4$ and the probability of this event can be lower bounded by

$$1 - 2 \cdot \exp(-c_1 \cdot \delta^2 \cdot N/L) - \exp(c_3 \cdot \delta^6 \cdot N/L).$$

Conclusion: Overall, letting $\mathcal{E} := \mathcal{E}_{\text{up}} \cap \mathcal{E}_{\text{low}}$, we get that on $\mathcal{E}$ the expression δ' is well-defined and satisfies

$$\delta/8 \leq \delta' \leq 96C_2 \cdot \delta.$$

The probability of this event can be lower bounded by

$$1 - 2 \exp(-c_1 N/L) - \exp(-c_2 \cdot d \cdot \ln(N/d)) - 2 \cdot \exp(-c_1 \cdot \delta^2 \cdot N/L) - \exp(c_3 \cdot \delta^6 \cdot N/L).$$

We now condition on $W^{(0)}, \dots, W^{(\ell-1)}$ and assume that $\mathcal{E}$ is satisfied. Let $C_4, c_4 > 0$ be the constants from Proposition 2.12. When considering the randomness with respect to $W^{(\ell)}$, we can thus apply Proposition 2.12 and get

$$c_4 \cdot \delta' \leq \text{Tr} \left| D^{(\ell)}(x) - D^{(\ell)}(y) \right| \leq C_4 \cdot \delta'$$

with probability at least $1 - 2 \exp(-c_4 \cdot \delta' N)$. Incorporating the bounds defined by $\mathcal{E}$, we overall get

$$\frac{c_4}{8} \cdot \delta \leq \text{Tr} \left| D^{(\ell)}(x) - D^{(\ell)}(y) \right| \leq 96C_2C_4 \cdot \delta$$

with probability at least

$$1 - 2 \exp(-c_1 N/L) - \exp(-c_2 \cdot d \cdot \ln(N/d)) - 2 \cdot \exp(-c_1 \cdot \delta^2 \cdot N/L) - \exp(-c_3 \cdot \delta^6 \cdot N/L) - 2 \exp(c_4 \cdot \delta \cdot N).$$

We firstly let $c' := \min\{c_1, c_2, c_3, c_4\}$. Further, we note that

$$\min\{N/L, \delta N\} \geq \delta^6 \cdot N/L \geq d \cdot \ln(N/d),$$

as follows by assumption, since

$$N \geq C \cdot L \cdot \delta^{-6} \cdot d \cdot \ln(N/d).$$

Therefore, we can bound the overall probability by

$$1 - 8 \exp(-c' \cdot d \cdot \ln(N/d)) = 1 - \exp(\ln(8) - c' \cdot d \cdot \ln(N/d)) \geq 1 - \exp(-c'' \cdot d \ln(N/d)),$$

where the last step uses $N \geq C \cdot d$. $\square$

In order to establish bounds on the norms and inner products of the θ_j , the main difficulty to overcome is the fact that the D -matrices are not independent of the other occurring random variables in the definition of the θ_j . To circumvent this issue, we use the fundamental fact that for two *orthogonal* vectors v_1, v_2 and a Gaussian matrix A , the two random variables Av_1 and Av_2 are independent. Assume for the moment that we consider the two random variables $D^{(\ell)}(x)$ and $W^{(\ell)}v$ for a random neural network according to Assumption 2.1 and some arbitrary vector v of suitable size. These two random variables are a priori not independent. However, conditioning on $W^{(0)}, \dots, W^{(\ell-1)}$, the randomness in $D^{(\ell)}(x)$ solely depends on $W^{(\ell)}\Phi^{(\ell-1)}(x)$. Therefore, we would obtain independence if $\Phi^{(\ell-1)}(x) \perp v$. While this might in general not be the case, we can consider the orthogonal projection π onto the space spanned by $\Phi^{(\ell-1)}(x)$ and decompose using the identity $W^{(\ell)}v = W^{(\ell)}\pi(v) + W^{(\ell)}\pi^\perp(v)$. Now, $W^{(\ell)}\pi^\perp(v)$ is independent of $D^{(\ell)}(x)$. It remains to show that the contribution of $W^{(\ell)}\pi(v)$ is not too significant, i.e., that the norm $\|\pi(v)\|_2$ is not too large. This is, for specific vectors v , governed by the following lemma.

Lemma F.4. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1 with $N \geq C \cdot L \cdot \ln(eL) \cdot \delta^{-6}$. We fix $\delta \leq \frac{1}{CL}$ and $x, y \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$ and $\|x - y\|_2 \geq \delta$ and pick $\nu \in \mathbb{S}^{d-1}$ with $\nu \perp \text{span}\{x, y\}$. For $\ell \in \{-1, \dots, L-1\}$, we let*

$$V^{(\ell)} := \text{span}\{\Phi^{(\ell)}(x), \Phi^{(\ell)}(y)\}$$

and let $\pi^{(\ell)} : \mathbb{R}^\mu \rightarrow \mathbb{R}^\mu$ denote the orthogonal projection onto $V^{(\ell)}$, where

$$\mu := \begin{cases} d, & \text{if } \ell = -1, \\ N, & \text{otherwise.} \end{cases}$$

Then, for any $\ell \in \{-1, \dots, L-1\}$ and $z_0, \dots, z_\ell \in \{x, y\}$ we get

$$\left\| \pi^{(\ell)} \left(\left[\prod_{j=\ell}^0 D^{(j)}(z_j) W^{(j)} \right] \nu \right) \right\|_2^2 \leq C^L \cdot \delta^{-2L} \cdot \frac{t}{N}$$

with probability at least $1 - \exp(-c \cdot t)$ for every $C \cdot \ln(eL) \leq t \leq \delta^6 N/L$.

Proof. The proof is via induction over ℓ . Note that in the case $\ell = -1$ we have

$$\prod_{j=\ell}^0 D^{(j)}(z_j) W^{(j)} = I_{d \times d}.$$

Therefore, since $\nu \perp \text{span}\{x, y\}$, we get $\pi^{(-1)}(\nu) = 0$, which implies the claim.

We now take $\ell \in \{0, \dots, L-1\}$ and assume via induction that

$$\left\| \pi^{(\ell-1)} \left(\left[\prod_{j=\ell-1}^0 D^{(j)}(z_j) W^{(j)} \right] \nu \right) \right\|_2^2 \leq C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N}$$

holds with probability at least $1 - 22 \cdot \ell \cdot \exp(-c' \cdot t)$ for every $0 \leq t \leq \delta^6 N/L^2$ for specific constants $C, c' > 0$ that we explicitly define at the end of the proof. For simplicity, we write

$$v := \left[\prod_{j=\ell}^0 D^{(j)}(z_j) W^{(j)} \right] \nu.$$

By Lemma F.2, we may assume

$$2 \geq \left\| \Phi^{(\ell)}(x) \right\|_2 \geq \frac{1}{2} \quad \text{and} \quad 2 \geq \left\| \Phi^{(\ell)}(y) \right\|_2 \geq \frac{1}{2}, \quad (\text{F.4})$$

which occurs with probability at least $1 - 2 \exp(-c_1 N/L)$ as follows from Lemma F.1 for an absolute constant $c_1 > 0$. Lemma F.1 may be applied by taking C large enough since

$$N \geq C \cdot L \cdot \ln(eL) \cdot \delta^{-6} \geq C \cdot L \cdot \ln(eL).$$

We call this event $\mathcal{E}_1$. Moreover, we let $\mathcal{E}_2$ be the event defined by

$$\min_{\alpha \in \mathbb{R}} \left\| \Phi^{(\ell)}(y) - \alpha \cdot \Phi^{(\ell)}(x) \right\|_2 \geq \delta/2,$$

which occurs with probability at least $1 - \exp(-c_2 \cdot \delta^6 \cdot N/L)$, using Lemma F.2. Here, we used $N \geq C \cdot L \cdot \ln(eL) \cdot \delta^{-6}$. By definition, on $\mathcal{E}_3 := \mathcal{E}_1 \cap \mathcal{E}_2$, letting

$$w := \Phi^{(\ell)}(y) - \langle \Phi^{(\ell)}(x), \Phi^{(\ell)}(y) \rangle \cdot \frac{\Phi^{(\ell)}(x)}{\left\| \Phi^{(\ell)}(x) \right\|_2^2},$$

we get

$$\|w\|_2 \geq \frac{\delta}{2}.$$

Note that $\{\Phi^{(\ell)}(x), w\}$ forms an orthogonal basis of $V^{(\ell)}$. On $\mathcal{E}_3$, we thus have

$$\left\| \pi^{(\ell)}(v) \right\|_2^2 = \left\langle v, \frac{\Phi^{(\ell)}(x)}{\left\| \Phi^{(\ell)}(x) \right\|_2} \right\rangle^2 + \left\langle v, \frac{w}{\|w\|_2} \right\rangle^2,$$

which implies

$$\left\| \pi^{(\ell)}(v) \right\|_2^2 \leq 4\delta^{-2} \cdot \left(\left\langle v, \Phi^{(\ell)}(x) \right\rangle^2 + \langle v, w \rangle^2 \right)$$

since $\delta < 1$. Moreover, by Cauchy-Schwarz,

$$|\langle v, w \rangle| \leq \left| \langle v, \Phi^{(\ell)}(y) \rangle \right| + \left| \langle v, \Phi^{(\ell)}(x) \rangle \right| \cdot \frac{\left\| \Phi^{(\ell)}(y) \right\|_2}{\left\| \Phi^{(\ell)}(x) \right\|_2} \leq 4 \cdot \left(\left| \langle v, \Phi^{(\ell)}(y) \rangle \right| + \left| \langle v, \Phi^{(\ell)}(x) \rangle \right| \right).$$

All in all, we hence get

$$\langle v, w \rangle^2 \leq 32 \cdot \left(\langle v, \Phi^{(\ell)}(x) \rangle^2 + \langle v, \Phi^{(\ell)}(y) \rangle^2 \right),$$

which then gives us

$$\begin{aligned} \left\| \pi^{(\ell)}(v) \right\|_2^2 &\leq 4 \cdot \delta^{-2} \cdot \left(\langle v, \Phi^{(\ell)}(x) \rangle^2 + 32 \langle v, \Phi^{(\ell)}(x) \rangle^2 + 32 \langle v, \Phi^{(\ell)}(y) \rangle^2 \right) \\ &\leq 132 \cdot \delta^{-2} \cdot \left(\langle v, \Phi^{(\ell)}(x) \rangle^2 + \langle v, \Phi^{(\ell)}(y) \rangle^2 \right) \end{aligned} \quad (\text{F.5})$$

on $\mathcal{E}_3$. The probability of $\mathcal{E}_3$ can be lower bounded by

$$1 - 2 \exp(-c_1 \cdot N/L) - \exp(-c_2 \cdot \delta^{-6} \cdot N/L).$$

In the following, it therefore suffices to control $\langle v, \Phi^{(\ell)}(x) \rangle^2$ and $\langle v, \Phi^{(\ell)}(y) \rangle^2$.

To this end, we fix a number $0 \leq t \leq \delta^6 \cdot N/L$ and set

$$v_1 := \left[\prod_{j=\ell-1}^0 D^{(j)}(z_j) W^{(j)} \right] \nu$$

and $\pi := \pi^{(\ell-1)}$. We now define $\mathcal{E}_4$ to be the event where

$$\left\| \pi(v_1) \right\|_2^2 \leq C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N} \quad \text{and} \quad \max \left\{ \left\| v_1 \right\|_2^2, \left\| \Phi^{(\ell-1)}(x) \right\|_2^2, \left\| \Phi^{(\ell-1)}(y) \right\|_2^2 \right\} \leq e^2$$

which occurs with probability at least $1 - 22 \cdot \ell \cdot \exp(-c' \cdot t) - 3 \exp(-c_3 \cdot N/L^2)$ according to the induction hypothesis and Lemma 3.1, with an absolute constant $c_3 > 0$. We may apply Lemma 3.1, since

$$N \geq C \cdot L \cdot \ln(eL) \cdot \underbrace{\delta^{-6}}_{\geq L^6} \geq C \cdot L^2 \cdot \ln(eL).$$

We condition on the matrices $W^{(0)}, \dots, W^{(\ell-1)}$ and assume that they satisfy the event $\mathcal{E}_4$. Due to symmetry, we only consider $\langle v, \Phi^{(\ell)}(x) \rangle^2$. We get

$$\langle v, \Phi^{(\ell)}(x) \rangle^2 = \left(v_1^T (W^{(\ell)})^T D^{(\ell)}(z_\ell) D^{(\ell)}(x) W^{(\ell)} \Phi^{(\ell-1)}(x) \right)^2,$$

which implies

$$\langle v, \Phi^{(\ell)}(x) \rangle^2 = \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} v_1 \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2.$$

Decomposing $v_1 = \pi(v_1) + \pi^\perp(v_1)$, we obtain

$$\begin{aligned} & \langle v, \Phi^{(\ell)}(x) \rangle^2 \\ &= \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \pi(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k + \left(W^{(\ell)} \pi^\perp(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2 \\ &\leq 2 \cdot \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \pi(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2 \\ &\quad + 2 \cdot \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \pi^\perp(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2. \end{aligned}$$

Considering the first summand, we obtain by Cauchy-Schwarz

$$\begin{aligned} & \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \pi(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2 \\ &\leq \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \pi(v_1) \right)_k^2 \right) \cdot \left(\sum_{\substack{k=1 \\ D^{(\ell)}(z_\ell)_k, k=D^{(\ell)}(x)_k, k=1}}^N \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k^2 \right) \\ &\leq \left\| W^{(\ell)} \pi(v_1) \right\|_2^2 \cdot \left\| W^{(\ell)} \Phi^{(\ell-1)}(x) \right\|_2^2. \end{aligned}$$

Since we take the randomness only with respect to $W^{(\ell)}$, we note that $W^{(\ell)} \pi(v_1) \sim \mathcal{N}\left(0, \frac{2}{N} \cdot \|\pi(v_1)\|_2^2\right)$ and $W^{(\ell)} \Phi^{(\ell-1)}(x) \sim \mathcal{N}\left(0, \frac{2}{N} \cdot \|\Phi^{(\ell-1)}(x)\|_2^2\right)$. We apply [20, Theorem 3.1.1] and obtain

$$\left\| W^{(\ell)} \pi(v_1) \right\|_2 \leq \sqrt{\frac{2}{N}} \cdot \|\pi(v_1)\|_2 \cdot (\sqrt{N} + u)$$

with probability at least $1 - 2\exp(-c_4 u^2)$ for every $u \geq 0$ with an absolute constant $c_4 > 0$. Letting $u := \sqrt{N}$, we get

$$\left\| W^{(\ell)} \pi(v_1) \right\|_2^2 \leq 8 \cdot \left\| \pi(v_1) \right\|_2^2$$

with probability at least $1 - 2\exp(-c_4 N)$. Similarly, we get

$$\left\| W^{(\ell)} \Phi^{(\ell-1)}(x) \right\|_2^2 \leq 8 \cdot \left\| \Phi^{(\ell-1)}(x) \right\|_2^2$$

with probability at least $1 - 2\exp(-c_4 N)$. Summarizing, we find

$$\left(\sum_{k=1}^N \left(W^{(\ell)} \pi(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2 \leq 64 \cdot \left\| \pi(v_1) \right\|_2^2 \cdot \left\| \Phi^{(\ell-1)}(x) \right\|_2^2$$

$D^{(\ell)}(z_\ell)_{k,k} = D^{(\ell)}(x)_{k,k} = 1$

with probability at least $1 - 4\exp(-c_4 N)$. Since we assume that $\mathcal{E}_4$ is satisfied, this may be further bounded by

$$\left(\sum_{k=1}^N \left(W^{(\ell)} \pi(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k \right)^2 \leq 64e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N}.$$

$D^{(\ell)}(z_\ell)_{k,k} = D^{(\ell)}(x)_{k,k} = 1$

Let us now consider the other summand. Note that according to [20, Theorem 3.1.1], we have (see above)

$$\left\| W^{(\ell)} \Phi^{(\ell-1)}(x) \right\|_2^2 \leq 8 \cdot \left\| \Phi^{(\ell-1)}(x) \right\|_2^2 \leq 8e^2 \quad (\text{F.6})$$

with probability at least $1 - 2\exp(-c_4 N)$. Since $W^{(\ell)} \pi^\perp(v_1)$ is independent of the random variable $(W^{(\ell)} \Phi^{(\ell-1)}(x), W^{(\ell)} \Phi^{(\ell-1)}(y))$, we may condition on the latter random variable and assume that (F.6) is satisfied.

For brevity, we write $X_k := (W^{(\ell)} \pi^\perp(v_1))_k$, $a_k := (W^{(\ell)} \Phi^{(\ell-1)}(x))_k$ and

$$K := \{k \in \{1, \dots, N\} : D^{(\ell)}(z_\ell)_{k,k} = D^{(\ell)}(x)_{k,k} = 1\}.$$

Note that all the a_k as well as K are fixed since we conditioned on $(W^{(\ell)} \Phi^{(\ell-1)}(x), W^{(\ell)} \Phi^{(\ell-1)}(y))$ and $W^{(0)}, \dots, W^{(\ell-1)}$. Moreover, we note that $X_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{2}{N} \cdot \left\| \pi^\perp(v_1) \right\|_2^2\right)$ for every $k \in \{1, \dots, N\}$. Hence,

$$\sum_{k=1}^N \left(W^{(\ell)} \pi^\perp(v_1) \right)_k \cdot \left(W^{(\ell)} \Phi^{(\ell-1)}(x) \right)_k = \sum_{k \in K} X_k \cdot a_k$$

$D^{(\ell)}(z_\ell)_{k,k} = D^{(\ell)}(x)_{k,k} = 1$

is Gaussian with mean zero and

$$\mathbb{V} \left[\sum_{k \in K} X_k \cdot a_k \right] = \frac{2}{N} \cdot \left\| \pi^\perp(v_1) \right\|_2^2 \cdot \sum_{k \in K} a_k^2 \leq \frac{2}{N} \cdot \|v_1\|_2^2 \cdot \left\| W^{(\ell)} \Phi^{(\ell-1)}(x) \right\|_2^2 \leq \frac{16e^4}{N}.$$

Hence, we get, for $u \geq 0$,

$$\mathbb{P} \left(\left| \sum_{k \in K} X_k \cdot a_k \right| \geq u \right) \leq 2 \exp(-c_6 N u^2),$$

with an absolute constant $c_6 > 0$. This directly implies

$$\mathbb{P} \left(\left(\sum_{k \in K} X_k \cdot a_k \right)^2 \geq t/N \right) \leq 2 \exp(-c_6 t).$$

Up to now, we have conditioned on the random variable $(W^{(\ell)}\Phi^{(\ell-1)}(x), W^{(\ell)}\Phi^{(\ell-1)}(y))$. Reintroducing the randomness over $(W^{(\ell)}\Phi^{(\ell-1)}(x), W^{(\ell)}\Phi^{(\ell-1)}(y))$, we obtain

$$\left(\sum_{k=1}^N \left(W^{(\ell)}\pi^\perp(v_1) \right)_k \cdot \left(W^{(\ell)}\Phi^{(\ell-1)}(x) \right)_k \right)^2 \leq t/N$$

$D^{(\ell)}(z_\ell)_{k,k} = D^{(\ell)}(x)_{k,k} = 1$

with probability at least

$$1 - 2\exp(-c_4N) - 2\exp(-c_6t).$$

Therefore, we overall get when conditioning on $W^{(0)}, \dots, W^{(\ell-1)}$, assuming that $\mathcal{E}_4$ is satisfied and only considering the randomness with respect to $W^{(\ell)}$ that

$$\langle v, \Phi^{(\ell)}(x) \rangle^2 \leq 64e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N} + \frac{t}{N} \leq 65e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N}$$

with probability at least

$$1 - 4\exp(-c_4N) - 2\exp(-c_4N) - 2\exp(-c_6t).$$

Since $t \leq N$ and by taking $c_7 := \min\{c_4, c_6\}$, we get

$$1 - 4\exp(-c_4N) - 2\exp(-c_4N) - 2\exp(-c_6t) \geq 1 - 8\exp(-c_7t).$$

Completely analogously, we obtain

$$\langle v, \Phi^{(\ell)}(y) \rangle^2 \leq 65e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N}$$

with probability at least

$$1 - 8\exp(-c_7t).$$

Reintroducing the randomness over $W^{(0)}, \dots, W^{(\ell-1)}$ and noting that we assumed $\mathcal{E}_4$ to be satisfied, we observe that

$$\max\{\langle v, \Phi^{(\ell)}(x) \rangle^2, \langle v, \Phi^{(\ell)}(y) \rangle^2\} \leq 65e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N}$$

occurs with probability at least

$$1 - 22 \cdot \ell \cdot \exp(-c' \cdot t) - 3\exp(-c_3 \cdot N/L^2) - 16\exp(-c_7t).$$

We call this event $\mathcal{E}_5$. By taking $c' \leq \min\{c_3, c_7\}$ and since $t \leq \delta^6 N/L \leq N/L^2$, we can bound

$$\mathbb{P}(\mathcal{E}_5) \geq 1 - 22 \cdot \ell \cdot \exp(-c \cdot t) - 19\exp(-ct).$$

Using (F.5), we now obtain that on $\mathcal{E}_3 \cap \mathcal{E}_5$ it holds

$$\begin{aligned} \left\| \pi^{(\ell)}(v) \right\|_2^2 &\leq 132 \cdot \delta^{-2} \cdot \left(\langle v, \Phi^{(\ell)}(x) \rangle^2 + \langle v, \Phi^{(\ell)}(y) \rangle^2 \right) \\ &\leq 264 \cdot \delta^{-2} \cdot 65e^2 \cdot C^\ell \cdot \delta^{-2\ell} \cdot \frac{t}{N} \\ &= 17160e^2 \cdot C^\ell \cdot \delta^{-2(\ell+1)} \cdot \frac{t}{N} \end{aligned}$$

and we can bound the probability of $\mathcal{E}_3 \cap \mathcal{E}_5$ from below by

$$1 - 22 \cdot \ell \cdot \exp(-c' \cdot t) - 19\exp(-ct) - 2\exp(-c_1 \cdot N/L) - \exp(-c_2 \cdot \delta^6 \cdot N/L).$$

Taking $c' \leq \min\{c_1, c_2\}$ and since $t \leq \delta^6 \cdot N/L$, we bound the probability further by

$$1 - 22 \cdot \ell \cdot \exp(-c' \cdot t) - 22\exp(-ct) = 1 - 22 \cdot (\ell + 1) \cdot \exp(-c' \cdot t)$$

Taking $C \geq 17160e^2$, we therefore see via induction that for every $\ell \in \{-1, \dots, L-1\}$, we get

$$\left\| \pi^{(\ell)}(v) \right\|_2^2 \leq C^{\ell+1} \cdot \delta^{-2(\ell+1)} \cdot \frac{t}{N}$$

with probability at least

$$1 - 22 \cdot (\ell + 1) \cdot \exp(-c' \cdot t)$$

for every $\ell \in \{-1, \dots, L-1\}$.

In the end, take $\ell \in \{-1, \dots, L-1\}$ arbitrary. By using $\ell+1 \leq L$, this implies

$$\left\| \pi^{(\ell)}(v) \right\|_2^2 \leq C^L \cdot \delta^{-2L} \cdot \frac{t}{N}$$

with probability at least

$$1 - 22L \cdot \exp(-c' \cdot t) = 1 - \exp(\ln(22L) - c' \cdot t).$$

If we now additionally assume $\ln(22L) \stackrel{\text{Lem. 1.3}}{\leq} \frac{22}{e} \cdot \ln(eL) \leq \frac{c'}{2}t$ (which we may do by taking $C \geq \frac{44}{ec'}$), we obtain the final claim by letting $c := c'/2$. $\square$

We can now prove an upper and a lower bound for the Euclidean norm of θ_j .

Lemma F.5. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1. Moreover, we take $\delta > 0$ with $\delta \leq \frac{1}{CL}$ and $x, y, \nu \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$, $24\delta \geq \|x - y\|_2 \geq \delta$ and $\nu \perp \text{span}\{x, y\}$. We assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6}, \quad \text{and} \quad N \geq C^L \cdot \delta^{-2L-5} \cdot d \cdot \ln(N/d).$$

Then for every $j \in \{0, \dots, L-1\}$ we have

$$c \cdot \delta \leq \|\theta_j\|_2^2 \leq C \cdot \delta$$

with probability at least $1 - \exp(-c \cdot d \ln(N/d))$. Here, θ_j is as defined in (F.1).

Proof. Recall that by definition we have

$$\theta_j = \overline{D}\Phi^{(j+1) \rightarrow (L-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu.$$

We set

$$v := \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu.$$

Applying Lemma 3.1, we conclude

$$\frac{1}{2} \cdot \|v\|_2 \leq \|\theta_j\|_2 \leq e \cdot \|v\|_2$$

with probability at least $1 - \exp(-c_1 \cdot N/L^2)$ with an absolute constant $c_1 > 0$. Here, we used

$$d \ln^2(N/d) \geq d \ln(N/d) \geq L \cdot \ln(eL)$$

and hence

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6} \geq C \cdot L^2 \cdot \max \ln(eL)$$

by taking C large enough. We call this event $\mathcal{E}_1$.

We let $\pi := \pi^{(j-1)}$ be the orthogonal projection onto $\text{span}\{\Phi^{(j-1)}(x), \Phi^{(j-1)}(y)\}$ as defined in Lemma F.4 and set

$$v' := \overline{D}\Phi^{(j-1)}(y) \nu.$$

We then get, by decomposing $v' = \pi(v') + \pi^\perp(v')$,

$$\begin{aligned} \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 - \left\| W^{(j)} \pi(v') \right\|_2 &\leq \|v\|_2 \quad \text{and} \\ \|v\|_2 &\leq \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 + \left\| W^{(j)} \pi(v') \right\|_2 \end{aligned}$$

using the (reverse) triangle inequality.

We note that according to Lemma F.4 we have

$$\|\pi(v')\|_2 \leq C_2^L \cdot \delta^{-L} \cdot \frac{\sqrt{t}}{\sqrt{N}}$$

with probability at least $1 - \exp(-c_2 \cdot t)$ for every $C_2 \cdot \ln(eL) \leq t \leq \delta^6 N/L$, where $C_2, c_2 > 0$ are absolute constants. As above, we here used that $d \ln^2(N/d) \geq L \cdot \ln(eL)$ which implies

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6} \geq C \cdot L \cdot \ln(eL) \cdot \delta^{-6}.$$

Let $c_3 > 0$ be another constant which is determined later. By plugging in explicitly

$$t := c_3^2 \cdot C_2^{-2L} \cdot \delta^{2L+4} \cdot N \cdot \frac{\delta}{16},$$

we observe

$$\|\pi(v')\|_2 \leq c_3 \cdot \frac{\sqrt{\delta}}{4} \cdot \delta^2 \leq c_3 \cdot \frac{\sqrt{\delta}}{4} \quad (\text{F.7})$$

with probability at least $1 - \exp(-c_2 \cdot c_3^2 \cdot C_2^{-2L} \cdot \delta^{2L+4} \cdot N \cdot \frac{\delta}{16})$. Note here that we use

$$c_3^2 \cdot C_2^{-2L} \cdot \delta^{2L+4} \cdot N \cdot \frac{\delta}{16} \leq \delta^7 \cdot N \leq \frac{\delta^6 N}{L},$$

where the last inequality follows from $\delta \leq 1/L$. Moreover, we used

$$\begin{aligned} C_2 \cdot \ln(eL) \leq c_3 \cdot C_2^{-2L} \cdot \delta^{2L+4} \cdot N \cdot \frac{\delta}{16} &\Leftrightarrow N \geq 16C_2^{2L+1} \cdot c_3^{-1} \cdot \delta^{-2L-5} \ln(eL) \\ &\Leftarrow N \geq (16C_2^3 \cdot c_3^{-1})^L \cdot \delta^{-2L-5} \ln(eL), \end{aligned}$$

where the latter is also satisfied by assumption, taking $C \geq 16C_2^3 \cdot c_3^{-1}$, since we even have

$$N \geq C^L \cdot \delta^{-2L-5} \cdot d \ln(N/d).$$

We denote by $\mathcal{E}_2$ the event which is defined by (F.7). Moreover, it holds

$$\frac{1}{2} \leq \|v'\|_2 \leq e$$

with probability at least $1 - \exp(-c_1 \cdot N/L^2)$, again using Lemma 3.1. This event is called $\mathcal{E}_3$. Note that on $\mathcal{E}_2 \cap \mathcal{E}_3$ we have

$$\frac{1}{4} \leq \frac{1}{2} - \delta/16 \leq \|v'\|_2^2 - \|\pi(v')\|_2^2 = \|\pi^\perp(v')\|_2^2 \leq \|v'\|_2^2 \leq e^2,$$

which gives us

$$1/2 \leq \|\pi^\perp(v')\|_2 \leq e.$$

We let $\mathcal{E}_4$ be the event defined by the property

$$\text{Tr} \left| D^{(j)}(x) - D^{(j)}(y) \right| \geq c_4 \cdot \delta \cdot N,$$

which occurs with probability at least $1 - \exp(-c_4 \cdot d \ln(N/d))$ for an absolute constant $c_4 > 0$, following Lemma F.3. Note here that all the conditions of Lemma F.3 are satisfied if we pick C large enough, in particular making δ small enough.

We condition on $W^{(0)}, \dots, W^{(j-1)}$. Firstly, we consider

$$\left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2.$$

Note that the random vector $W^{(j)} \pi^\perp(v')$ is independent of the difference $D^{(j)}(x) - D^{(j)}(y)$. Therefore, we may condition on $D^{(j)}(x) - D^{(j)}(y)$. Since $W^{(j)} \pi^\perp(v') \sim \mathcal{N}(0, \frac{2}{N} \cdot \|\pi^\perp(v')\|_2^2 \cdot I_N)$, we get, using [20, Theorem 3.1.1],

$$\begin{aligned} \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 &\leq \sqrt{\frac{2}{N}} \cdot \|\pi^\perp(v')\|_2 \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} + u) \quad \text{and} \\ \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 &\geq \sqrt{\frac{2}{N}} \cdot \|\pi^\perp(v')\|_2 \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} - u) \end{aligned}$$

with probability at least $1 - 2\exp(-c_5 u^2)$ for every $u \geq 0$, for an absolute constant $c_5 > 0$. We call this event $\mathcal{E}_5 = \mathcal{E}_5(u)$. Moreover, still conditioning on $W^{(0)}, \dots, W^{(j-1)}$, we get

$$\left\| W^{(j)} \pi(v') \right\|_2 \leq 2 \cdot \|\pi(v')\|_2$$

with probability at least $1 - 2\exp(-c_6 N)$ with an absolute constant $c_6 > 0$, as follows from [20, Theorem 3.1.1]. We call this event $\mathcal{E}_6$. Note that the lower bounds on the probabilities of $\mathcal{E}_5$ and $\mathcal{E}_6$ remain the same if we also consider the randomness over $W^{(0)}, \dots, W^{(j-1)}$, since we did not impose any restrictions on $W^{(0)}, \dots, W^{(j-1)}$.

Overall, on the event $\mathcal{E}_7 := \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6$, we get

$$\begin{aligned} \|v\|_2 &\leq \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 + \left\| W^{(j)} \pi(v') \right\|_2 \\ &\stackrel{\mathcal{E}_5 \cap \mathcal{E}_6}{\leq} \sqrt{\frac{2}{N}} \cdot \underbrace{\|\pi^\perp(v')\|_2}_{\leq \|v'\|_2} \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} + u) + 2 \cdot \|\pi(v')\|_2 \\ &\stackrel{\mathcal{E}_2 \cap \mathcal{E}_3}{\leq} \frac{e\sqrt{2}}{\sqrt{N}} \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} + u) + c_3 \cdot \frac{\sqrt{\delta}}{2} \\ &\stackrel{\mathcal{E}_4}{\leq} \frac{e\sqrt{2}}{\sqrt{N}} \cdot (\sqrt{C_4 \cdot \delta \cdot N} + u) + c_3 \cdot \frac{\sqrt{\delta}}{2}. \end{aligned}$$

and

$$\begin{aligned} \|v\|_2 &\geq \left\| \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \pi^\perp(v') \right\|_2 - \left\| W^{(j)} \pi(v') \right\|_2 \\ &\stackrel{\mathcal{E}_5 \cap \mathcal{E}_6}{\geq} \sqrt{\frac{2}{N}} \cdot \|\pi^\perp(v')\|_2 \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} - u) - 2 \cdot \|\pi(v')\|_2 \\ &\stackrel{\mathcal{E}_2 \cap \mathcal{E}_3}{\geq} \frac{\sqrt{2}}{2\sqrt{N}} \cdot (\sqrt{\text{Tr} |D^{(j)}(x) - D^{(j)}(y)|} - u) - \frac{c_3}{2} \cdot \sqrt{\delta} \\ &\stackrel{\mathcal{E}_4}{\geq} \frac{\sqrt{2}}{2\sqrt{N}} \cdot (\sqrt{c_4 \cdot \delta \cdot N} - u) - \frac{c_3}{2} \cdot \sqrt{\delta}. \end{aligned}$$

We explicitly pick $u := \frac{\sqrt{c_4 \cdot \delta \cdot N}}{2}$ which gives us

$$\frac{c_3}{2} \cdot \sqrt{\delta} \leq \frac{\sqrt{c_4 \cdot \delta \cdot N}}{4} \cdot \sqrt{\delta} - \frac{c_3}{2} \cdot \sqrt{\delta} \leq \|v\|_2 \leq C_5 \cdot \sqrt{\delta}$$

by defining $c_3 := \frac{\sqrt{c_4 \cdot \delta \cdot N}}{4}$ and by picking $C_5 > 0$ appropriately. On $\mathcal{E}_1 \cap \mathcal{E}_7$ we therefore get

$$\left(\frac{c_3}{4} \right)^2 \cdot \delta \leq \|\theta_j\|_2^2 \leq C_5^2 \cdot \delta.$$

It therefore remains to prove a lower bound for the probability of $\mathcal{E}_1 \cap \mathcal{E}_7$. By using the bounds we have shown before, we get the lower bound of

$$\begin{aligned} &1 - 2\exp(-c_1 \cdot N/L^2) - \exp\left(-c_2 c_3 \cdot C_2^{-2L} \cdot \delta^{2L+4} \cdot N \cdot \frac{\delta}{16}\right) \\ &- \exp(-c_4 \cdot d \ln(N/d)) - 2\exp(-c_7 \cdot \delta N) - 2\exp(-c_6 \cdot N) \end{aligned}$$

for an absolute constant $c_7 > 0$. By picking $c' := \min\{c_1, c_2 c_3/16, c_4, c_6, c_7\}$ and using

$$d \cdot \ln(N/d) \leq C_2^{-2L} \cdot \delta^{2L+5} \cdot N \leq \delta N \quad \text{and} \quad d \cdot \ln(N/d) \leq N/L^2 \leq N$$

we may bound this probability by

$$1 - 8\exp(-c' \cdot d \cdot \ln(N/d)).$$

The final claim follows since we may assume

$$\ln(8) \leq \frac{c'}{2} \cdot d \cdot \ln(N/d)$$

and by taking $c := \min\{\frac{c'}{2}, \frac{c_3^2}{16}\}$. $\square$

According to the outline described in the beginning of this section, we now aim to study the inner product $\langle \theta_j, \theta_k \rangle$. The following lemma is an auxiliary statement about the concentration of the inner product $\langle Vx, Vy \rangle$ (where V is a specific random matrix) around the inner product $\langle x, y \rangle$. It follows from an application of Bernstein's inequality.

Lemma F.6. *There exists a constant $c > 0$ such that the following holds. Let $W \in \mathbb{R}^{N \times k}$ be a random matrix with $W_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/N)$ for all indices $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, k\}$. Let $x_0 \in \mathbb{R}^k \setminus \{0\}$ be arbitrary and let $V := \Delta(Wx_0)W$ with Δ introduced as in Section 2.2. Then for every $x, y \in \mathbb{R}^k \setminus \{0\}$ and $t \geq 0$,*

$$\mathbb{P}\left(|\langle Vx, Vy \rangle - \langle x, y \rangle| \geq t\right) \leq 2 \exp\left(-c \cdot N \min\left(\frac{t^2}{\|x\|_2^2 \|y\|_2^2}, \frac{t}{\|x\|_2 \|y\|_2}\right)\right).$$

Proof. We first note that according to [13, Lemma 6.4], the matrix $\sqrt{N}V$ has isotropic rows. By definition, this implies $\mathbb{E}[V_{i,j_1} \cdot V_{i,j_2}] = \frac{1}{N} \cdot \mathbb{1}_{j_1=j_2}$ for every $i \in \{1, \dots, N\}$ and $j_1, j_2 \in \{1, \dots, k\}$. Therefore, for arbitrary $i \in \{1, \dots, N\}$ we get

$$\begin{aligned} \mathbb{E}[(Vx)_i \cdot (Vy)_i] &= \mathbb{E}\left[\left(\sum_{j_1=1}^k V_{i,j_1} x_{j_1}\right) \left(\sum_{j_2=1}^k V_{i,j_2} y_{j_2}\right)\right] = \sum_{j_1, j_2=1}^k \mathbb{E}[V_{i,j_1} \cdot V_{i,j_2}] \cdot x_{j_1} \cdot y_{j_2} \\ &= \frac{1}{N} \cdot \sum_{j=1}^k x_j y_j = \frac{1}{N} \langle x, y \rangle. \end{aligned}$$

We can thus rewrite

$$\mathbb{P}\left(|\langle Vx, Vy \rangle - \langle x, y \rangle| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^N ((Vx)_i (Vy)_i - \mathbb{E}[(Vx)_i (Vy)_i])\right| \geq t\right).$$

Note that the summands are independent. We therefore estimate the right-hand side using Bernstein's inequality [20, Theorem 2.8.1]. To this end, note that for any $i \in \{1, \dots, N\}$ we have $|(Vx)_i| \leq |(Wx)_i|$ by definition of $\Delta(Wx_0)$ and therefore

$$\|(Vx)_i\|_{\psi_2} \leq \|(Wx)_i\|_{\psi_2} \leq C_1 \cdot \frac{\|x\|_2}{\sqrt{N}}$$

and similarly $\|(Vy)_i\|_{\psi_2} \leq C_1 \cdot \frac{\|y\|_2}{\sqrt{N}}$ with an absolute constant $C_1 > 0$. Here we used that $(Wx)_i \sim \mathcal{N}(0, \frac{2}{N} \|x\|_2^2)$ due to rotational invariance of the Gaussian distribution. We can thus apply [20, Lemma 2.7.7] to conclude that the product $(Wx)_i (Wy)_i$ is sub-exponential with

$$\|(Vx)_i (Vy)_i\|_{\psi_1} \leq \|(Vx)_i\|_{\psi_2} \cdot \|(Vy)_i\|_{\psi_2} \leq \frac{C_1}{N} \cdot \|x\|_2 \cdot \|y\|_2,$$

after possibly increasing the constant C_1 . Moreover, applying the centering technique [20, Exercise 2.7.10] we obtain

$$\|(Vx)_i (Vy)_i - \mathbb{E}[(Vx)_i (Vy)_i]\|_{\psi_1} \leq \frac{C_1}{N} \cdot \|x\|_2 \cdot \|y\|_2$$

after possibly increasing the constant C_1 again. Hence, by Bernstein's inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^N ((Vx)_i (Vy)_i - \mathbb{E}[(Vx)_i (Vy)_i])\right| \geq t\right) \leq 2 \exp\left(-c \cdot \min\left(\frac{N \cdot t^2}{\|x\|_2^2 \cdot \|y\|_2^2}, \frac{Nt}{\|x\|_2 \cdot \|y\|_2}\right)\right),$$

where $c > 0$ is an absolute constant, as was to be shown. $\square$

This leads to the desired upper bound on the absolute value of the inner product $|\langle \theta_j, \theta_k \rangle|$, which is contained in the following lemma.

Lemma F.7. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1. Moreover, let $\delta > 0$ with $\delta \leq \frac{1}{CL}$ be chosen and let $x, y, \nu \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$, $24 \cdot \delta \geq \|x - y\|_2 \geq \delta$ and $\nu \perp \text{span}\{x, y\}$. We assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6}, \quad \text{and} \quad N \geq C^L \cdot \delta^{-2L-5} \cdot d \cdot \ln(N/d).$$

Then for every $j, k \in \{0, \dots, L-1\}$ with $j \neq k$ we have

$$|\langle \theta_j, \theta_k \rangle| \leq C \cdot \delta^{3/2}$$

with probability at least $1 - \exp(-c \cdot d \cdot \ln(N/d))$. Here, θ_j and θ_k are as defined in (F.1).

Proof. By symmetry, we may assume $j < k$.

We first show via induction over $\ell \in \{0, \dots, L-1-k\}$ that with probability at least $1 - 6\ell \cdot \exp(-c_1 \cdot \delta^3 N/L^2)$,

$$\begin{aligned} & |\langle \theta_j, \theta_k \rangle| \\ & \leq \left| \left\langle \overline{D}\Phi^{(j+1) \rightarrow (L-1-\ell)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu, \right. \right. \\ & \quad \left. \left. \overline{D}\Phi^{(k+1) \rightarrow (L-1-\ell)}(x) \left(D^{(k)}(x) - D^{(k)}(y) \right) W^{(k)} \overline{D}\Phi^{(k-1)}(y) \nu \right\rangle \right| + \delta^{3/2} \cdot \ell/L. \end{aligned}$$

Note that case $\ell = 0$ is trivial and follows directly from the definition of θ_j and θ_k . Hence, we move to the induction step. Take $\ell \in \{1, \dots, L-1-k\}$ and assume that the claim is true for $\ell-1$. Moreover, let

$$\begin{aligned} V &:= D^{(L-\ell)}(x) W^{(L-\ell)}, \\ v_1 &:= \overline{D}\Phi^{(j+1) \rightarrow (L-\ell-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu, \\ v_2 &:= \overline{D}\Phi^{(k+1) \rightarrow (L-\ell-1)}(x) \left(D^{(k)}(x) - D^{(k)}(y) \right) W^{(k)} \overline{D}\Phi^{(k-1)}(y) \nu. \end{aligned}$$

By the induction hypothesis,

$$|\langle \theta_j, \theta_k \rangle| \leq \langle V v_1, V v_2 \rangle + \frac{\ell-1}{L} \cdot \delta^{3/2} \quad (\text{F.8})$$

with probability at least $1 - 6(\ell-1) \exp(-c_1 \cdot \delta^3 N/L^2)$. Firstly, we note that

$$\|v_1\|_2 \leq \left\| \overline{D}\Phi^{(j+1) \rightarrow (L-\ell-1)}(x) \overline{D}\Phi^{(j-1)}(y) \nu \right\|_2 + \left\| \overline{D}\Phi^{(j+1) \rightarrow (L-\ell-1)}(x) \overline{D}\Phi^{(j)}(y) \nu \right\|_2 \leq 4$$

with probability at least $1 - 2 \exp(-c_1 \cdot N/L^2)$ for some absolute constant $c_1 > 0$, using Lemma 3.1. The same bound can be obtained for v_2 . Conditioning on $W^{(0)}, \dots, W^{(L-\ell-2)}$ and assuming $\|v_1\|_2, \|v_2\|_2 \leq 4$, we can thus infer

$$|\langle V v_1, V v_2 \rangle| \leq |\langle v_1, v_2 \rangle| + t$$

with probability at least $1 - 2 \exp(-c_1 \cdot N \cdot \min(t, t^2))$ for any $t \geq 0$, after possibly decreasing c_1 . Here, we applied Lemma F.6. Note here that in order to apply Lemma F.6, we need to require $v_1 \neq 0 \neq v_2$, but the high probability bound stated above remains true if $v_1 = 0$ or $v_2 = 0$. By reintroducing the randomness over $W^{(0)}, \dots, W^{(L-\ell-1)}$, we get that the event above holds with probability at least $1 - 2 \exp(-c_1 \cdot N \cdot \min(t, t^2)) - 4 \exp(-c_1 \cdot N/L^2)$. Letting $t := \delta^{3/2}/L$ (which implies $t^2 \leq t$), we obtain

$$|\langle V v_1, V v_2 \rangle| \leq |\langle v_1, v_2 \rangle| + \delta^{3/2}/L$$

with probability at least $1 - 6 \exp(-c_1 \cdot N \delta^3/L^2)$, where we also used $N \delta^3/L^2 \leq N/L^2$. Combining this with (F.8) proves the induction step.

We then let

$$\begin{aligned} w_1 &:= \overline{D}\Phi^{(j+1) \rightarrow (k)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu, \\ w_2 &:= \left(D^{(k)}(x) - D^{(k)}(y) \right) W^{(k)} \overline{D}\Phi^{(k-1)}(y) \nu \end{aligned}$$

and define $\mathcal{E}_1$ to be the event on which

$$|\langle \theta_j, \theta_k \rangle| \leq |\langle w_1, w_2 \rangle| + \delta^{3/2}.$$

By what we have just shown,

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - 6L \cdot \exp(-c_1 \cdot \delta^3 N / L^2).$$

Let $\tilde{D} := D^{(k)}(x) (D^{(k)}(x) - D^{(k)}(y))$ and moreover, set

$$\begin{aligned} z_1 &:= \overline{D}\Phi^{(j+1) \rightarrow (k-1)}(x) \left(D^{(j)}(x) - D^{(j)}(y) \right) W^{(j)} \overline{D}\Phi^{(j-1)}(y) \nu \\ z_2 &:= \overline{D}\Phi^{(k-1)}(y) \nu. \end{aligned}$$

Let $\pi = \pi^{(k-1)}$ denote the orthogonal projection onto $\text{span}\{\Phi^{(k-1)}(x), \Phi^{(k-1)}(y)\}$ as it was defined in Lemma F.4. We then get

$$\begin{aligned} |\langle w_1, w_2 \rangle| &= \left| z_1^T (W^{(k)})^T \tilde{D} W^{(k)} z_2 \right| \\ &= \left| \pi(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi(z_2) \right| + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi(z_2) \right| \\ &\quad + \left| \pi(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| \\ &\leq \left\| W^{(k)} \pi(z_1) \right\|_2 \left\| W^{(k)} \pi(z_2) \right\|_2 + \left\| W^{(k)} \pi^\perp(z_1) \right\|_2 \left\| W^{(k)} \pi(z_2) \right\|_2 \\ &\quad + \left\| W^{(k)} \pi(z_1) \right\|_2 \left\| W^{(k)} \pi^\perp(z_2) \right\|_2 + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| \\ &\leq 4 \left\| \pi(z_1) \right\|_2 \cdot \left\| \pi(z_2) \right\|_2 + 4 \left\| \pi^\perp(z_1) \right\|_2 \cdot \left\| \pi(z_2) \right\|_2 + 4 \left\| \pi(z_1) \right\|_2 \cdot \left\| \pi^\perp(z_2) \right\|_2 \\ &\quad + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| \\ &\leq 4 \left\| \pi(z_1) \right\|_2 \cdot \left\| \pi(z_2) \right\|_2 + 4 \left\| z_1 \right\|_2 \cdot \left\| \pi(z_2) \right\|_2 + 4 \left\| \pi(z_1) \right\|_2 \cdot \left\| z_2 \right\|_2 \\ &\quad + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right|, \end{aligned} \tag{F.9}$$

where the second last inequality holds with probability $1 - 12 \exp(-c_2 N)$ by conditioning on the weight matrices $W^{(0)}, \dots, W^{(k-1)}$ and only taking the randomness with respect to $W^{(k)}$ [20, Theorem 3.1.1]. Here, $c_2 > 0$ is an absolute constant. We call $\mathcal{E}_2$ the event defined by that property. The intuition is now that in the above decomposition only the last summand actually matters and the other summands can be upper bounded using Lemma F.4.

Let us therefore start by studying the last summand. Set

$$I := \left\{ i \in \{1, \dots, N\} : \tilde{D}_{i,i} \neq 0 \right\}.$$

We firstly then note that

$$\begin{aligned} \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| &= \left| \sum_{i \in I} \left(W^{(k)} \pi^\perp(z_1) \right)_i \left(W^{(k)} \pi^\perp(z_2) \right)_i \right| \\ &\leq \left(\sum_{i \in I} \left(W^{(k)} \pi^\perp(z_1) \right)_i^2 \right)^{1/2} \left(\sum_{i \in I} \left(W^{(k)} \pi^\perp(z_2) \right)_i^2 \right)^{1/2}, \end{aligned}$$

using the Cauchy-Schwarz inequality.

We now condition on $W^{(0)}, \dots, W^{(k-1)}$. Note that, when considering randomness only with respect to $W^{(k)}$, the matrix $\tilde{D}$ is then independent of $(W^{(k)} \pi^\perp(z_1), W^{(k)} \pi^\perp(z_2))$. Since $W^{(k)} \pi^\perp(z_1) \sim \mathcal{N}(0, \frac{2}{N} \left\| \pi^\perp(z_1) \right\|_2^2)$ and $W^{(k)} \pi^\perp(z_2) \sim \mathcal{N}(0, \frac{2}{N} \left\| \pi^\perp(z_2) \right\|_2^2)$, we conclude, by conditioning on $\tilde{D}$, that

$$\left(\sum_{i \in I} \left(W^{(k)} \pi^\perp(z_1) \right)_i^2 \right)^{1/2} \left(\sum_{i \in I} \left(W^{(k)} \pi^\perp(z_2) \right)_i^2 \right)^{1/2} \leq \frac{2}{N} \cdot \left\| \pi^\perp(z_1) \right\|_2 \cdot \left\| \pi^\perp(z_2) \right\|_2 \cdot (\sqrt{I} + u)^2$$

$$\leq \frac{4}{N} \cdot \|\pi^\perp(z_1)\|_2 \cdot \|\pi^\perp(z_2)\|_2 \cdot (I + u^2) \quad (\text{F.10})$$

with probability at least $1 - 4\exp(-c_3 \cdot u^2)$ for every $u \geq 0$ for an absolute constant $c_3 > 0$ [20, Theorem 3.1.1]. Note that this bound remains true if we reintroduce the randomness over $\tilde{D}$ and $W^{(0)}, \dots, W^{(k-1)}$, since we did not impose any restrictions on these random variables. We call the event defined by property (F.10) $\mathcal{E}_3(u)$. Moreover, note that

$$\begin{aligned} \#I &= \#\left\{i \in \{1, \dots, N\} : D^{(k)}(x)_{i,i} = 1 \text{ and } D^{(k)}(y)_{i,i} = 0\right\} \\ &\leq \#\left\{i \in \{1, \dots, N\} : D^{(k)}(x)_{i,i} \neq D^{(k)}(y)_{i,i}\right\} \leq C_4 \cdot \delta N \end{aligned}$$

holds with probability at least $1 - \exp(-c_4 \cdot d \cdot \ln(N/d))$, using Lemma F.3 with absolute constants $C_4, c_4 > 0$. We call this event $\mathcal{E}_4$. Therefore, with $u := \sqrt{N} \cdot \delta$ and $\mathcal{E}_3 = \mathcal{E}_3(u)$, on $\mathcal{E}_5 := \mathcal{E}_3 \cap \mathcal{E}_4$ we obtain

$$\left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| \leq C_5 \cdot \delta \cdot \|z_1\|_2 \|z_2\|_2$$

for an absolute constant $C_5 > 0$. The probability of that event can be lower bounded by

$$1 - 4\exp(-c_5 \cdot \delta N) - \exp(-c_5 \cdot d \ln(N/d)) \geq 1 - 5\exp(-c_5 \cdot d \ln(N/d))$$

by letting $c_5 := \min\{c_3, c_4\}$, since we assume $d \cdot \ln(N/d) \leq \delta N$.

We now deal with $\|z_1\|_2$ and $\|z_2\|_2$. z_1 may be viewed as an expression θ_j for a network with $k - 1$ hidden layers. Since $k \leq L - 1$, we may therefore use Lemma F.5 and infer

$$\|z_1\|_2 \leq C_6 \cdot \sqrt{\delta}$$

with probability at least $1 - \exp(-c_6 \cdot d \ln(N/d))$ for absolute constants $C_6, c_6 > 0$. We call the event defined by that property $\mathcal{E}_6$. Moreover,

$$\|z_2\|_2 \leq e$$

with probability at least $1 - \exp(-c_7 \cdot N/L^2)$ according to Lemma 3.1, and we call that event $\mathcal{E}_7$. Therefore, on $\mathcal{E}_8 := \mathcal{E}_5 \cap \mathcal{E}_6 \cap \mathcal{E}_7$, we obtain

$$\left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| \leq C_8 \cdot \delta^{3/2}$$

with an absolute constant $C_8 > 0$ and we can lower bound the probability of that event by

$$1 - 5\exp(-c_5 \cdot d \ln(N/d)) - \exp(-c_6 \cdot d \ln(N/d)) - \exp(-c_7 \cdot N/L^2) \geq 1 - 7\exp(-c_8 \cdot d \ln(N/d)).$$

We have thus established a convenient bound on the last summand in (F.9).

The other three summands can be handled simultaneously; by linearity of π ,

$$\|\pi(z_1)\|_2 \leq \left\| \pi \left(\bar{D}\Phi^{(j) \rightarrow (k-1)}(x) \bar{D}\Phi^{(j-1)}(y) \nu \right) \right\|_2 + \left\| \pi \left(\bar{D}\Phi^{(j+1) \rightarrow (k-1)}(x) \bar{D}\Phi^{(j)}(y) \nu \right) \right\|_2,$$

and hence by Lemma F.4,

$$\max\{\|\pi(z_1)\|_2, \|\pi(z_2)\|_2\} \leq C_9^L \cdot \delta^{-L} \cdot \frac{\sqrt{t}}{\sqrt{N}}$$

with probability at least $1 - 3\exp(-c_9 \cdot t)$ for absolute constants $C_9, c_9 > 0$ for every choice of t satisfying $C_9 \cdot \ln(eL) \leq t \leq \delta^6 \cdot N/L$. We plug in $t := N \cdot C_9^{-2L} \cdot \delta^{2L+5}$, which gives us

$$\max\{\|\pi(z_1)\|_2, \|\pi(z_2)\|_2\} \leq \delta^{5/2} \leq \delta^{3/2}$$

with probability at least $1 - 3\exp(-c_9 \cdot C_9^{-2L} \cdot \delta^{2L+5} \cdot N)$. Here, we used once more that

$$t = N \cdot C_9^{-2L} \cdot \delta^{2L+5} \leq \delta^7 N \leq \delta^6 N/L$$

and

$$t = N \cdot C_9^{-2L} \cdot \delta^{2L+5} \geq C \cdot \ln(eL),$$

where the last inequality is satisfied since we even have

$$N \geq C^L \cdot \delta^{-2L-5} \cdot d \cdot \ln(N/d) \geq C \cdot \ln(eL).$$

We call the event on which this property is satisfied $\mathcal{E}_9$. Therefore, on $\mathcal{E}_{10} := \mathcal{E}_8 \cap \mathcal{E}_9$ we get, since $\mathcal{E}_8 \subseteq \mathcal{E}_6 \cap \mathcal{E}_7$, that

$$\begin{aligned} 4 \|\pi(z_1)\|_2 \cdot \|\pi(z_2)\|_2 + 4 \|z_1\|_2 \cdot \|\pi(z_2)\|_2 + 4 \|\pi(z_1)\|_2 \cdot \|z_2\|_2 &\leq 4\delta^3 + 4C_6 \cdot \delta^2 + 4e \cdot \delta^{3/2} \\ &\leq C_{10} \cdot \delta^{3/2} \end{aligned}$$

for some absolute constant $C_{10} > 0$. Overall, on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_{10}$, we have

$$\begin{aligned} &|\langle \theta_j, \theta_k \rangle| \\ &\leq |\langle w_1, w_2 \rangle| + \delta^{3/2} \\ &\leq 4 \|\pi(z_1)\|_2 \cdot \|\pi(z_2)\|_2 + 4 \|z_1\|_2 \cdot \|\pi(z_2)\|_2 + 4 \|\pi(z_1)\|_2 \cdot \|z_2\|_2 \\ &\quad + \left| \pi^\perp(z_1)^T (W^{(k)})^T \tilde{D} W^{(k)} \pi^\perp(z_2) \right| + \delta^{3/2} \\ &\leq C_{10} \cdot \delta^{3/2} + C_8 \cdot \delta^{3/2} + \delta^{3/2} \\ &\leq C_{11} \cdot \delta^{3/2} \end{aligned}$$

for an absolute constant $C_{11} > 0$ and we can lower bound the probability of that event by

$$\begin{aligned} &1 - 4 \cdot L \cdot \exp(-c_1 \cdot \delta^3 N / L^2) - 12 \exp(-c_2 N) - 7 \exp(-c_8 \cdot d \cdot \ln(N/d)) - 3 \exp(-c_9 \cdot C_9^{-2L} \cdot \delta^{2L+5} \cdot N) \\ &\geq 1 - 26L \cdot \exp(-c_{10} \cdot d \cdot \ln(N/d)), \end{aligned}$$

setting $c_{10} := \min\{c_1, c_2, c_8, c_9\}$. Here, we used the condition

$$d \cdot \ln(N/d) \leq C^{-L} \cdot \delta^{2L+5} \cdot N.$$

In the end, the claim follows by letting $c = c_{10}/2$ and since

$$\ln(eL) \leq C \cdot L \cdot \ln(eL) \leq d \cdot \ln(N/d),$$

taking C large enough. $\square$

We can now lower bound the distance $\|\overline{D}\Phi^{(L-1)}(x)\nu - \overline{D}\Phi^{(L-1)}(y)\nu\|_2$ for two fixed inputs x and y satisfying certain conditions.

Theorem F.8. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1. Moreover, let $x, y, \nu \in \mathbb{S}^{d-1}$ with $x_d = y_d = \frac{1}{\sqrt{2}}$, $24 \cdot \delta \geq \|x - y\|_2 \geq \delta$, where $\delta \leq c \cdot \frac{1}{L^2}$, and $\nu \perp \text{span}\{x, y\}$. We assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6}, \quad \text{and} \quad N \geq C^L \cdot \delta^{-2L-5} \cdot d \cdot \ln(N/d).$$

Then we have

$$\left\| \overline{D}\Phi^{(L-1)}(x)\nu - \overline{D}\Phi^{(L-1)}(y)\nu \right\|_2 \geq c \cdot \sqrt{L \cdot \delta}$$

with probability at least $1 - \exp(-c \cdot d \ln(N/d))$.

Proof. We note that, using Lemma E.1,

$$\begin{aligned} \left\| \overline{D}\Phi^{(L-1)}(x)\nu - \overline{D}\Phi^{(L-1)}(y)\nu \right\|_2^2 &= \sum_{j=0}^{L-1} \|\theta_j\|_2^2 + 2 \sum_{j < k} \langle \theta_j, \theta_k \rangle \\ &\geq \sum_{j=0}^{L-1} \|\theta_j\|_2^2 - 2 \sum_{j < k} |\langle \theta_j, \theta_k \rangle|. \end{aligned}$$

Using the bounds established in Lemmas F.5 and F.7 and a union bound, this results in

$$\left\| \overline{D}\Phi^{(L-1)}(x)\nu - \overline{D}\Phi^{(L-1)}(y)\nu \right\|_2^2 \geq c_1 \cdot L \cdot \delta - C_1 \cdot L^2 \cdot \delta^{3/2}$$

with probability at least

$$1 - (L + L^2) \exp(-c_1 \cdot d \ln(N/d)) \geq 1 - \exp(-c \cdot d \ln(N/d))$$

with absolute constants $C_1, c_1, c > 0$, since $d \ln(N/d) \geq C \cdot \ln(eL)$. If $\delta \leq \frac{c_1^2}{(2C_1)^2} \cdot \frac{1}{L^2}$, then

$$C_1 \cdot L^2 \cdot \delta^{3/2} \leq \frac{c_1 \cdot L \cdot \delta}{2}.$$

This gives us

$$c_1 \cdot L \cdot \delta - C_1 \cdot L^2 \cdot \delta^{3/2} \geq \frac{c_1}{2} \cdot L \cdot \delta,$$

which implies the claim. $\square$

We can now establish a lower bound on the packing number of the set $\mathcal{L}$ that was already briefly discussed in the beginning of this section. It is essential to note that the covering number increases *exponentially* in the input dimension d .

Corollary F.9. *There exist absolute constants $C, c > 0$ satisfying the following: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N satisfying Assumption 2.1. Moreover, we let $\delta \leq c \cdot \frac{1}{L^2}$. We assume $N \geq C \cdot d$ and*

$$N \geq C \cdot d \cdot L \cdot \ln^2(N/d) \cdot \delta^{-6}, \quad \text{and} \quad N \geq C^L \cdot \delta^{-2L-5} \cdot d \cdot \ln(N/d).$$

Let $\mathcal{B}_\Phi \subseteq \mathbb{S}^{d-1}$ be a random set such that $\mathbb{P}(x_0 \in \mathcal{B}_\Phi) = 1$ for every $x_0 \in \mathbb{S}^{d-1}$. Consider the set

$$\mathcal{L} := \left\{ \overline{\mathbf{D}}\Phi^{(L-1)}(x) \cdot e_1 : x \in \mathcal{B}_\Phi \right\}$$

with e_1 denoting the first standard basis vector in $\mathbb{R}^d$. Then we get

$$\mathcal{P}(\mathcal{L}, c \cdot \sqrt{L\delta}) \geq c \cdot \delta \cdot 2^d.$$

with probability at least $1 - \exp(-c \cdot d \ln(N/d))$.

Proof. Let $C_1, c_1 > 0$ be the constants from Theorem F.8. We take $\delta \leq c_1 \cdot \frac{1}{L^2}$. Moreover, we define

$$\mathcal{A} := \{0\} \times \left(\frac{1}{\sqrt{2}} \mathbb{S}^{d-3} \cap B_{d-2} \left(\frac{\tilde{e}_1}{\sqrt{2}}, 12\delta \right) \right) \times \left\{ \frac{1}{\sqrt{2}} \right\} \subseteq \mathbb{S}^{d-1}$$

where $\tilde{e}_1$ is the first standard basis vector in $\mathbb{R}^{d-2}$. According to Lemma I.8, we can pick a δ -packing $P \subseteq \mathcal{A}$ with

$$|P| \geq \frac{\sqrt{2} \cdot \delta}{18} \cdot 2^{d-3} = \frac{\sqrt{2} \cdot \delta}{144} \cdot 2^d.$$

Removing some elements from P , we may assume

$$2^d \geq \frac{\sqrt{2} \cdot \delta}{144} \cdot 2^d \geq |P| \geq \frac{\sqrt{2} \cdot \delta}{288} \cdot 2^d.$$

Since P is finite, we may conclude that the event

$$P \subseteq \mathcal{B}_\Phi$$

occurs with probability 1. This yields that

$$\left\{ \overline{\mathbf{D}}\Phi^{(L-1)}(x) \cdot e_1 : x \in P \right\} \subseteq \mathcal{L}$$

with probability 1. Note that for each pair of numbers $x, y \in P$ with $x \neq y$ we have $x_d = y_d = \frac{1}{\sqrt{2}}$, $e_1 \perp \text{span}\{x, y\}$ and $24\delta \geq \|x - y\|_2 \geq \delta$ (since P is a δ -packing). This yields

$$\left\| \overline{\mathbf{D}}\Phi^{(L-1)}(x)e_1 - \overline{\mathbf{D}}\Phi^{(L-1)}(y)e_1 \right\|_2 \geq c_1 \cdot \sqrt{L \cdot \delta}$$

with probability at least $1 - \exp(-c_1 \cdot d \cdot \ln(N/d))$ where we used Theorem F.8. By performing a union bound over all pairs $(x, y) \in P \times P$ with $x \neq y$ and using that $\#P \leq 2^d$, we get that the event

$$\text{for all } x, y \in P \text{ with } x \neq y : \left\| \overline{\mathbf{D}}\Phi^{(L-1)}(x)e_1 - \overline{\mathbf{D}}\Phi^{(L-1)}(y)e_1 \right\|_2 \geq c_1 \cdot \sqrt{L \cdot \delta}$$

occurs with probability at least

$$1 - \exp(2 \ln(|P|) - c_1 \cdot d \cdot \ln(N/d)) \geq 1 - \exp(2d \cdot \ln(2) - c_1 \cdot d \cdot \ln(N/d)) \geq 1 - \exp(-c \cdot d \cdot \ln(N/d))$$

for an absolute constant $c > 0$, where $c \leq c_1/2$. Therefore, with probability at least $1 - \exp(-c \cdot d \cdot \ln(N/d))$, the set

$$\tilde{P} := \left\{ \overline{D}\Phi^{(L-1)}(x)e_1 : x \in P \right\}$$

is a $(c_1 \sqrt{L\delta})$ -packing of $\mathcal{L}$. On the same event, we furthermore get

$$|\tilde{P}| = |P| \geq \frac{\sqrt{2} \cdot \delta}{288} \cdot 2^d,$$

which implies the claim. $\square$

Using Sudakov's minoration inequality and Gaussian concentration, we now get the main result of this section.

Theorem F.10. *There exist absolute constants $C, c > 0$ satisfying the following: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N following Assumption 2.1. Moreover, we assume $N \geq C \cdot d$ and*

$$N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10} \quad \text{and} \quad d \geq C \cdot L.$$

Let $\mathcal{B}_\Phi \subseteq \mathbb{S}^{d-1}$ be a random set such that $\mathbb{P}(x_0 \in \mathcal{B}_\Phi) = 1$ for every $x_0 \in \mathbb{S}^{d-1}$, where the randomness in $\mathcal{B}_\Phi$ only depends on $W^{(0)}, \dots, W^{(L-1)}$. Then, with probability at least $1 - \exp(-c \cdot d/L)$,

$$\sup_{x \in \mathcal{B}_\Phi} \langle \overline{\nabla} \Phi(x), e_1 \rangle \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d},$$

where e_1 denotes the first standard basis vector in $\mathbb{R}^d$.

Proof. Let $C_1, c_1 > 0$ be the constants provided by Corollary F.9. Let $\delta := c_1/L^2$ and $\mathcal{L}$ be as defined in Corollary F.9. We condition on $W^{(0)}, \dots, W^{(L-1)}$ and assume that the property

$$\mathcal{P}(\mathcal{L}, c_1 \cdot \sqrt{L\delta}) \geq c_1 \cdot \delta \cdot 2^d$$

is satisfied. By assuming $(\frac{4}{3})^d \geq (c_1 \cdot \delta)^{-1}$ (which we may do since $d \geq C \cdot L \geq C \cdot \ln(eL)$), we can lower bound the above by $(\frac{3}{2})^d$. Since the weight matrices $W^{(0)}, \dots, W^{(L-1)}$ are fixed, we may use Sudakov's minoration inequality [20, Theorem 7.4.1] and get

$$\begin{aligned} \mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathcal{B}_\Phi} \langle \overline{\nabla} \Phi(x), e_1 \rangle \right] &= \mathbb{E}_{W^{(L)}} \left[\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle \right] \geq c_2 \cdot c_1 \cdot \sqrt{L\delta} \cdot \sqrt{\ln(\mathcal{P}(\mathcal{L}, c_1 \cdot \sqrt{L\delta}))} \\ &\geq c_1^{3/2} c_2 \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{\ln(3/2)} \cdot \sqrt{d} \end{aligned}$$

with an absolute constant $c_2 > 0$ ⁵. Letting $c_3 := \sqrt{\ln(3/2)} \cdot c_1^{3/2} c_2$, we obtain

$$\mathbb{E}_{W^{(L)}} \left[\sup_{x \in \mathcal{B}_\Phi} \langle \overline{\nabla} \Phi(x), e_1 \rangle \right] \geq c_3 \cdot \sqrt{\frac{d}{L}}.$$

We now turn this expectation bound into a high probability bound. To this end, in the spirit of Theorem 4.1, we additionally assume that $\sup_{x \in \mathbb{S}^{d-1}} \|\Phi^{(L-1)}\|_{2 \rightarrow 2} \leq C_1$ is satisfied by the weight matrices $W^{(0)}, \dots, W^{(L-1)}$. Recall that we conditioned on these matrices. We consider the function

$$f : \mathbb{R}^{1 \times N} \rightarrow \mathbb{R}, \quad W^{(L)} \mapsto \sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle.$$

⁵We used the well-known fact that the L^2 -increments of the random process $(\langle (W^{(L)})^T, v \rangle)_{v \in \mathcal{L}}$ are measured by the Euclidean distance on $\mathcal{L}$, see also [20, Chapter 7.1.2].

Let $W^{(L)}, \widetilde{W}^{(L)} \in \mathbb{R}^{1 \times N}$ be arbitrary and assume without loss of generality that $f(W^{(L)}) \geq f(\widetilde{W}^{(L)})$. We choose $v^* \in \mathcal{L}$ with

$$\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle = \langle (W^{(L)})^T, v^* \rangle,$$

which in particular implies

$$f(\widetilde{W}^{(L)}) = \sup_{v \in \mathcal{L}} \langle (\widetilde{W}^{(L)})^T, v \rangle \geq \langle (\widetilde{W}^{(L)})^T, v^* \rangle.$$

Such a vector v^* exists since there are only finitely many activation patterns defined by the weight matrices $W^{(0)}, \dots, W^{(L-1)}$. We hence get

$$\begin{aligned} \left| f(W^{(L)}) - f(\widetilde{W}^{(L)}) \right| &= f(W^{(L)}) - f(\widetilde{W}^{(L)}) \leq \langle (W^{(L)})^T, v^* \rangle - \langle (\widetilde{W}^{(L)})^T, v^* \rangle \\ &\leq \|W^{(L)} - \widetilde{W}^{(L)}\|_2 \cdot \|v^*\|_2 = \|W^{(L)} - \widetilde{W}^{(L)}\|_2 \cdot \sup_{x \in \mathcal{B}_\Phi} \|\overline{D}\Phi^{(L-1)}(x)e_1\|_2 \\ &\leq \|W^{(L)} - \widetilde{W}^{(L)}\|_2 \cdot \sup_{x \in \mathcal{B}_\Phi} \|\overline{D}\Phi^{(L-1)}(x)\|_2 \\ &\leq \|W^{(L)} - \widetilde{W}^{(L)}\|_2 \cdot \sup_{x \in \mathbb{S}^{d-1}} \|\overline{D}\Phi^{(L-1)}(x)\|_2 \\ &\leq C_1 \cdot \|W^{(L)} - \widetilde{W}^{(L)}\|_2. \end{aligned}$$

Therefore, f is Lipschitz continuous with Lipschitz constant C_1 . We can thus apply standard Gaussian concentration for Lipschitz functions, see, e.g., [20, Theorem 5.2.2], to see that for any $t \geq 0$, the event

$$\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle \geq \mathbb{E}_{W^{(L)}} \left[\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle \right] - t \geq c_3 \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d} - t$$

occurs with probability at least $1 - 2\exp(-c_4 \cdot t^2)$, where $c_5 > 0$ is an absolute constant. Letting $t := \frac{c_3}{2} \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}$, we get

$$\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle \geq \frac{c_3}{2} \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d} = c_5 \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}$$

with probability at least $1 - 2\exp(-c_6 \cdot d/L)$ with absolute constants $c_5, c_6 > 0$.

Up to now, we have conditioned on $W^{(0)}, \dots, W^{(L-1)}$. Reintroducing the randomness over these random variables, we get that the event

$$\sup_{v \in \mathcal{L}} \langle (W^{(L)})^T, v \rangle \geq c_6 \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}$$

occurs with probability at least

$$1 - 2\exp(-c_6 \cdot d/L) - \exp(-c_1 \cdot d \ln(N/d)) - \exp(-c_7 \cdot d \ln(N/d))$$

with an absolute constant $c_7 > 0$, where we used Corollary F.9 and Theorem 4.1 and the randomness is now taken with respect to all the occurring random variables. It now suffices to note that $d/L \leq d \cdot \ln(N/d)$. By picking $c_8 = \min\{c_1, c_6, c_7\}$, we therefore bound the above probability from below by

$$1 - 4\exp(-c_8 \cdot d/L).$$

Since by assumption we have

$$d \geq C \cdot L,$$

we may assume

$$\ln(4) \leq \frac{c_8}{2} \cdot L^{-1} \cdot d$$

and we can bound the probability further by

$$1 - \exp(-c \cdot d/L),$$

where we take $c := c_8/2$ and $C \geq c_4^{-1}$. Therefore, the claim is shown. $\square$

We continue by showing that the special case $\nu = e_1$ indeed implies the statement for arbitrary $\nu \in \mathbb{S}^{d-1}$.

Proposition F.11. *Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network satisfying Assumption 2.1 of arbitrary width and depth. Let e_1 denote the first standard basis vector in $\mathbb{R}^d$ and let $\nu \in \mathbb{S}^{d-1}$ be arbitrary. Let*

$$M_\Phi := \{x \in \mathbb{S}^{d-1} : \Phi \text{ differentiable at } x \text{ with } \nabla \Phi(x) = \bar{\nabla} \Phi(x)\}.$$

Then we have

$$\sup_{x \in M_\Phi} \langle \bar{\nabla} \Phi(x), \nu \rangle \stackrel{d}{=} \sup_{x \in M_\Phi} \langle \bar{\nabla} \Phi(x), e_1 \rangle.$$

Proof. Pick an orthogonal matrix $U \in \mathbb{R}^{d \times d}$ that satisfies $U^T e_1 = \nu$. We then set

$$\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}, \quad \tilde{\Phi}(x) := \Phi(Ux).$$

Firstly, we note that $\tilde{\Phi}$ is itself a ReLU network that arises from Φ by replacing the first weight matrix $W^{(0)}$ by $W^{(0)}U$. By rotation invariance of the Gaussian distribution, we conclude $\Phi \stackrel{d}{=} \tilde{\Phi}$. Moreover, by definition it is clear that

$$\bar{\nabla} \tilde{\Phi}(x)^T = \bar{\nabla} \Phi(Ux)^T U^T \quad \text{for all } x \in \mathbb{R}^d. \quad (\text{F.11})$$

Lastly, by using the chain rule, we get

$$U \cdot M_\Phi^\sim = M_\Phi.$$

Indeed, let $x \in M_\Phi^\sim$ and $y := Ux$, which implies

$$\Phi(y) = \Phi(Ux) = \tilde{\Phi}(x) = \tilde{\Phi}(U^T y).$$

Since $\tilde{\Phi}$ is by definition differentiable at $U^T y = x$, we may apply the chain rule and get

$$\nabla \Phi(y) = U \cdot \nabla \tilde{\Phi}(U^T y) = U \cdot \nabla \tilde{\Phi}(x) \stackrel{x \in M_\Phi^\sim}{=} U \cdot \bar{\nabla} \tilde{\Phi}(x) \stackrel{(\text{F.11})}{=} \bar{\nabla} \Phi(Ux) = \bar{\nabla} \Phi(y).$$

This proves $Ux \in M_\Phi$ and hence $U \cdot M_\Phi^\sim \subseteq M_\Phi$. The other direction follows analogously.

We then observe

$$\begin{aligned} \sup_{x \in M_\Phi} \langle \bar{\nabla} \Phi(x), \nu \rangle &= \sup_{x \in M_\Phi} (\bar{\nabla} \Phi(x))^T U^T e_1 \stackrel{d}{=} \sup_{x \in M_\Phi^\sim} (\bar{\nabla} \tilde{\Phi}(x))^T U^T e_1 = \sup_{x \in M_\Phi^\sim} \bar{\nabla} \Phi(Ux)^T e_1 \\ &= \sup_{U^T y \in M_\Phi^\sim} \bar{\nabla} \Phi(y)^T e_1 = \sup_{y \in U M_\Phi^\sim} \bar{\nabla} \Phi(y)^T e_1 = \sup_{y \in M_\Phi} \langle \bar{\nabla} \Phi(y), e_1 \rangle. \end{aligned} \quad \square$$

The statement of Theorem 5.1 is now obtained by studying the set

$$M_\Phi := \{x \in \mathbb{S}^{d-1} : \Phi \text{ differentiable at } x \text{ with } \bar{\nabla} \Phi(x) = \nabla \Phi(x)\}$$

in combination with [13, Theorem E.1].

Proof of Theorem 5.1. In the beginning, let $\nu = e_1$. We set

$$M_\Phi := \{x \in \mathbb{S}^{d-1} : \Phi \text{ differentiable at } x \text{ with } \bar{\nabla} \Phi(x) = \nabla \Phi(x)\}.$$

Then, for every fixed $x_0 \in \mathbb{S}^{d-1}$ we have $x_0 \in M_\Phi$ with probability 1 according to [13, Theorem E.1]. Hence, we may apply Theorem F.10 and get

$$\sup_{x \in M_\Phi} \langle \bar{\nabla} \Phi(x), e_1 \rangle \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d}$$

with probability at least $1 - \exp(-c \cdot d/L)$. The generalization to arbitrary $\nu \in \mathbb{S}^{d-1}$ follows from Proposition F.11. Finally, the bound on the Lipschitz constant is again obtained by choosing $\nu = e_1$ (or any other standard basis vector) and using Theorem 2.8. $\square$

As a corollary, we get the same lower bound for the entire regime $p \in [1, 2]$. Note that we essentially get the same bound for arbitrary $p \in [1, \infty]$, but we already established the stronger lower bound of $d^{1-1/p}$ for $p \in [2, \infty]$ in Section 3.

Corollary F.12. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random zero-bias ReLU network with L hidden layers of width N following Assumption 2.1. Moreover, we assume $N \geq C \cdot d$ and*

$$N \geq C^L \cdot d \cdot \ln^2(N/d) \cdot L^{4L+10} \quad \text{and} \quad d \geq C \cdot L.$$

Then, we have

$$\text{Lip}_p(\Phi) \geq c \cdot \frac{1}{\sqrt{L}} \cdot \sqrt{d} \quad \text{for every } p \in [1, 2)$$

with probability at least $1 - \exp(-c \cdot d/L)$.

G Proof of Theorem 6.2

In this section, we show that for inputs with sufficiently large Euclidean norm, the gradient of a ReLU network (with potentially non-zero biases) at these inputs is, with high probability, very similar to the gradient of the associated *homogeneous* network, i.e., the same network with zero biases. The main tool to derive the result is Proposition 2.14, which states that for inputs with large Euclidean norm the bias term does *not* have a significant influence on whether the ReLU is active or not.

In this section, we study random ReLU networks following Assumption 2.1 with *fixed* bias vectors $b^{(\ell)} \in \mathbb{R}^N$ for $\ell \in \{0, \dots, L-1\}$. We pick $\lambda' > 0$ such that

$$\text{for all } \ell \in \{0, \dots, L-1\}, i \in \{1, \dots, N\} : \left| b_i^{(\ell)} \right| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}}. \quad (\text{G.1})$$

The scaling by $\sqrt{2/N}$ is required to make Proposition 2.14 applicable, since the entries of the weight matrices are not standard gaussian, but scaled by $\sqrt{2/N}$ too.

We denote by $\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}$ the homogeneous network associated to Φ , i.e., the weight matrices of $\tilde{\Phi}$ are the same as for Φ and all the bias vectors are set to zero. For $j \in \{1, \dots, L-1\}$ and $x \in \mathbb{R}^d$, we let

$$\tilde{D}^{(j)}(x) := \Delta(W^{(j)} \tilde{\Phi}^{(j-1)}(x)) \in \mathbb{R}^{N \times N}.$$

We then set

$$A^{(j)}(x) := D^{(j)}(x) - \tilde{D}^{(j)}(x).$$

The first goal is to show that $A^{(j)}(x) \in \mathcal{A}_\delta^{(j)}$ for $x \in \mathbb{R}^d$ with sufficiently large Euclidean norm, where $\mathcal{A}_\delta^{(j)}$ is as defined in (E.1), i.e., $A^{(j)}(x)$ is a diagonal matrix with rather sparse diagonal. As already explained above, the idea is to apply Proposition 2.14. To this end, we aim to bound the covering numbers and the Gaussian width of certain sets.

Lemma G.1. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N satisfying Assumption 2.1 with fixed biases. Assume $N \geq Cd$ and $N \geq C \cdot L^3 d \ln(N/d)$. Let $\varepsilon, \delta \in (0, e^{-1})$ with*

$$N \geq C \cdot d L^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot d \cdot \ln(1/\delta), \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c \cdot 1/L^2.$$

Set $\theta := 48 \cdot 3^L \cdot \lambda' \cdot \varepsilon^{-1}$, where λ' is as in (G.1). Let $j \in \{-1, \dots, L-1\}$ and define

$$T^{(j)} := \left\{ \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \|x\|_2 \geq \theta, \Phi^{(j)}(x) \neq 0 \right\} \cup \left\{ \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \|x\|_2 \geq \theta, \tilde{\Phi}^{(j)}(x) \neq 0 \right\},$$

where $\tilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}$ is the associated homogeneous network. Then, with probability at least $1 - \exp(-c \cdot \delta N)$,

$$\ln(e \cdot \mathcal{N}(T^{(j)}, \varepsilon/2)) \leq C \cdot d \cdot \ln(1/\varepsilon) \quad \text{and} \quad w((T^{(j)} - T^{(j)}) \cap B_\mu(0, \varepsilon)) \leq C \cdot \sqrt{dL} \cdot \varepsilon \cdot \ln^{1/2}(1/\varepsilon).$$

Here,

$$\mu := \begin{cases} d & \text{if } j = -1 \\ N & \text{if } j \geq 0. \end{cases}$$

Proof. We may assume $j \geq 0$, since

$$T^{(-1)} = \mathbb{S}^{d-1} \quad \text{and} \quad (T^{(-1)} - T^{(-1)}) \cap B_d(0, \varepsilon) \subseteq B_d(0, \varepsilon),$$

so that the claim is trivially satisfied by [20, Corollary 4.2.13 & Proposition 7.5.2(f)].

For simplicity, we write $T = T^{(j)}$ for the remainder of the proof. Note that

$$\mathbb{P} \left(\text{for all } \ell \in \{0, \dots, j\} : \|W^{(\ell)}\|_{2 \rightarrow 2} \leq 3 \right) \geq 1 - \exp(-c_1 \cdot N), \quad (\text{G.2})$$

with an absolute constant $c_1 > 0$, as follows by [20, Corollary 7.3.3] and a union bound, noting that $N \geq C \cdot \ln(L)$. Further,

$$\text{for all } x \in \mathbb{R}^d : \|\tilde{\Phi}^{(j)}(x)\|_2 \geq \frac{\|x\|_2}{4} \quad (\text{G.3})$$

with probability at least $1 - \exp(-c_2 \cdot \delta N)$ with an absolute constant $c_2 > 0$ by Corollary E.4. Lastly, note that

$$\text{Lip}_{2 \rightarrow 2}(\tilde{\Phi}^{(j)}) \leq C_1 \quad (\text{G.4})$$

with probability at least $1 - \exp(-c_3 \cdot \delta N)$ with absolute constants $C_1, c_3 > 0$, as follows by Theorem E.2. From now on, we assume that the properties defined in Equations (G.2) to (G.4) are satisfied. This happens with probability at least

$$\begin{aligned} & 1 - \exp(-c_1 \cdot N) - \exp(-c_2 \cdot \delta N) - \exp(-c_3 \cdot \delta N) \\ & \geq 1 - 3 \exp(-c' \cdot \delta N) \\ & = 1 - \exp(\ln(3) - c' \cdot \delta N) \end{aligned}$$

with $c' := \min\{c_1, c_2, c_3\}$. By taking C large enough, we may assume

$$\delta N \geq C \cdot d \cdot \ln(1/\delta) \geq \frac{2}{c'} \cdot \ln(3)$$

and can thus, by letting $c := c'/2$, lower bound the probability by

$$1 - \exp(-c \cdot \delta N).$$

Note that the conditions $\Phi^{(j)}(x) \neq 0$ and $\tilde{\Phi}^{(j)}(x) \neq 0$ in the definition of $T^{(j)}$ are redundant due to Equation (G.3).

By our assumption (G.1),

$$\|b^{(\ell)}\|_2^2 = \sum_{i=1}^N |b_i^{(\ell)}|^2 \leq 2(\lambda')^2$$

for every $\ell \in \{0, \dots, j-1\}$. Using (G.2) we then observe, for any $x \in \mathbb{R}^d$, that

$$\begin{aligned} \|\Phi^{(j-1)}(x) - \tilde{\Phi}^{(j-1)}(x)\|_2 & \leq \|W^{(j-1)}\Phi^{(j-2)}(x) + b^{(j-1)} - W^{(j-1)}\tilde{\Phi}^{(j-2)}(x)\|_2 \\ & \leq \|b^{(j-1)}\|_2 + \|W^{(j-1)}\|_{2 \rightarrow 2} \cdot \|\Phi^{(j-2)}(x) - \tilde{\Phi}^{(j-2)}(x)\|_2 \\ & \leq \|b^{(j-1)}\|_2 + 3 \cdot \|\Phi^{(j-2)}(x) - \tilde{\Phi}^{(j-2)}(x)\|_2 \\ & \leq \sum_{\ell=0}^{j-1} 3^{j-1-\ell} \cdot \|b^{(\ell)}\|_2 \\ & \leq \sqrt{2} \cdot \lambda' \cdot \frac{3^j - 1}{3 - 1} \\ & \leq \lambda' \cdot 3^L. \end{aligned} \quad (\text{G.5})$$

After these preliminary observations, let us move to the actual proof. As a first step, we aim to bound the γ -covering numbers of T for every $\gamma \in (0, \varepsilon/2]$. To this end, note that $T = T_1 \cup T_2$ with

$$T_1 := \left\{ \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \|x\|_2 \geq \theta \right\}$$

and

$$T_2 := \left\{ \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \|x\|_2 \geq \theta \right\} = \left\{ \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} : x \in \mathbb{S}^{d-1} \right\},$$

whence it suffices to bound the γ -covering numbers of T_1 and T_2 separately.

We start with T_1 . Let $\gamma \in (0, \varepsilon/2]$. We then set $\theta_\gamma := 24 \cdot 3^L \cdot \lambda' \cdot \gamma^{-1} \geq \theta_{\varepsilon/2} = \theta$. Further, we set

$$T' := \left\{ \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \|x\|_2 \geq \theta_\gamma \right\} \subseteq T.$$

We now separately bound the γ -covering numbers of T' and $T_1 \setminus T'$, which in total gives us a bound on the γ -covering number of T_1 .

Let us start with T' . Let $\mathcal{N}_\gamma$ be a $\left(\frac{\gamma}{24C_1}\right)$ -net of $\mathbb{S}^{d-1}$ with $|\mathcal{N}_\gamma| \leq (72C_1/\gamma)^d$, which is possible using [20, Corollary 4.2.13], and define

$$\mathcal{N}' := \left\{ \frac{\Phi^{(j)}(\theta_\gamma x^*)}{\|\Phi^{(j)}(\theta_\gamma x^*)\|_2} : x^* \in \mathcal{N}_\gamma \right\} \subseteq T'.$$

We claim that $\mathcal{N}'$ is a γ -net of T' . To this end, let $x \in \mathbb{R}^d$ with $\|x\|_2 \geq \theta_\gamma$ be arbitrary and pick $x^* \in \mathcal{N}_\gamma$ satisfying

$$\left\| \frac{x}{\|x\|_2} - x^* \right\|_2 \leq \frac{\gamma}{24C_1}.$$

We get

$$\begin{aligned} & \left\| \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} - \frac{\Phi^{(j)}(\theta_\gamma x^*)}{\|\Phi^{(j)}(\theta_\gamma x^*)\|_2} \right\|_2 \\ & \leq \left\| \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} - \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} \right\|_2 + \left\| \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} - \frac{\tilde{\Phi}^{(j)}(\theta_\gamma x^*)}{\|\tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2} \right\|_2 \\ & \quad + \left\| \frac{\tilde{\Phi}^{(j)}(\theta_\gamma x^*)}{\|\tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2} - \frac{\Phi^{(j)}(\theta_\gamma x^*)}{\|\Phi^{(j)}(\theta_\gamma x^*)\|_2} \right\|_2 \\ & \stackrel{\text{Lem. I.2}}{\leq} 2 \cdot \frac{\|\Phi^{(j)}(x) - \tilde{\Phi}^{(j)}(x)\|_2}{\|\tilde{\Phi}^{(j)}(x)\|_2} + \left\| \frac{\tilde{\Phi}^{(j)}(x/\|x\|_2)}{\|\tilde{\Phi}^{(j)}(x/\|x\|_2)\|_2} - \frac{\tilde{\Phi}^{(j)}(x^*)}{\|\tilde{\Phi}^{(j)}(x^*)\|_2} \right\|_2 \\ & \quad + 2 \cdot \frac{\|\Phi^{(j)}(\theta_\gamma x^*) - \tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2}{\|\tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2} \\ & \stackrel{\text{Lem. I.2}}{\leq} 2 \cdot \frac{\|\Phi^{(j)}(x) - \tilde{\Phi}^{(j)}(x)\|_2}{\|\tilde{\Phi}^{(j)}(x)\|_2} + 2 \cdot \frac{\|\tilde{\Phi}^{(j)}(x/\|x\|_2) - \tilde{\Phi}^{(j)}(x^*)\|_2}{\|\tilde{\Phi}^{(j)}(x^*)\|_2} \\ & \quad + 2 \cdot \frac{\|\Phi^{(j)}(\theta_\gamma x^*) - \tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2}{\|\tilde{\Phi}^{(j)}(\theta_\gamma x^*)\|_2} \end{aligned}$$

$$\leq \frac{8 \cdot 3^L \cdot \lambda'}{\theta_\gamma} + \frac{8C_1\gamma}{24C_1} + \frac{8 \cdot 3^L \cdot \lambda'}{\theta_\gamma} = \frac{\gamma}{3} + \frac{\gamma}{3} + \frac{\gamma}{3} = \gamma$$

by choice of γ and θ_γ , using Equations (G.3) to (G.5) in the last step. Hence, we have $\mathcal{N}(T', \gamma) \leq \left(\frac{72C_1}{\gamma}\right)^d$.

Let us now turn to $T_1 \setminus T'$. We first observe

$$T_1 \setminus T' \subseteq \left\{ \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} : x \in \mathbb{R}^d, \theta_\gamma \geq \|x\|_2 \geq \theta \right\},$$

whence it suffices to upper bound the covering number of the latter set. To this end, let $\mathcal{N}$ be a $(2\lambda'\gamma)$ -net of the annulus $A := \{x \in \mathbb{R}^d : \theta_\gamma \geq \|x\|_2 \geq \theta\}$ with $|\mathcal{N}| \leq \left(\frac{6\theta_\gamma}{2\lambda'\gamma}\right)^d = \left(\frac{72 \cdot 3^L}{\gamma^2}\right)^d$, which exists according to [20, Ex. 4.2.10 & Cor. 4.2.13]. As above, we set

$$\mathcal{N}' := \left\{ \frac{\Phi^{(j)}(x^*)}{\|\Phi^{(j)}(x^*)\|_2} : x^* \in \mathcal{N} \right\}.$$

For $x \in A$ and $x^* \in \mathcal{N}$ with $\|x - x^*\|_2 \leq 2\lambda'\gamma$, we then get

$$\left\| \frac{\Phi^{(j)}(x)}{\|\Phi^{(j)}(x)\|_2} - \frac{\Phi^{(j)}(x^*)}{\|\Phi^{(j)}(x^*)\|_2} \right\|_2 \stackrel{\text{Lemma I.2}}{\leq} 2 \cdot \frac{\|\Phi^{(j)}(x) - \Phi^{(j)}(x^*)\|_2}{\|\Phi^{(j)}(x^*)\|_2} \leq \frac{2 \cdot 3^L}{\|\Phi^{(j)}(x^*)\|_2} \cdot \|x - x^*\|_2, \quad (\text{G.6})$$

using that $\Phi^{(j)}$ is 3^L -Lipschitz from (G.2). Note further that, using (G.3) and (G.5),

$$\begin{aligned} \|\Phi^{(j)}(x^*)\|_2 &\geq \|\tilde{\Phi}^{(j)}(x^*)\|_2 - \|\Phi^{(j)}(x^*) - \tilde{\Phi}^{(j)}(x^*)\|_2 \\ &\geq \frac{\theta}{4} - 3^L \cdot \lambda' = 12 \cdot 3^L \cdot \lambda' \cdot \underbrace{\varepsilon^{-1}}_{\geq 1} - 3^L \cdot \lambda' \geq 4 \cdot 3^L \cdot \lambda'. \end{aligned}$$

This bound together with (G.6) shows that $\mathcal{N}$ is a γ -net of $T_1 \setminus T'$. Overall, we hence get

$$\mathcal{N}(T_1, \gamma) \leq \mathcal{N}(T', \gamma) + \mathcal{N}(T_1 \setminus T', \gamma) \leq \left(\frac{72C_1}{\gamma}\right)^d + \left(\frac{72 \cdot 3^L}{\gamma^2}\right)^d \leq 2 \cdot \left(\frac{72C_1 \cdot 3^L}{\gamma^2}\right)^d.$$

for $0 < \gamma \leq \varepsilon/2$.

We continue by bounding the γ -covering number of T_2 . Note that we may assume that $\tilde{\Phi}^{(j)}$ is C_1 -Lipschitz by (G.4). If we then choose a $\left(\frac{\gamma}{8C_1}\right)$ -net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \leq \left(\frac{24C_1}{\gamma}\right)^d$, see [20, Cor. 4.2.13], we can for arbitrary $x \in \mathbb{S}^{d-1}$ choose $x^* \in \mathcal{N}$ with $\|x - x^*\| \leq \frac{\gamma}{8C_1}$ and get

$$\left\| \frac{\tilde{\Phi}^{(j)}(x)}{\|\tilde{\Phi}^{(j)}(x)\|_2} - \frac{\tilde{\Phi}^{(j)}(x^*)}{\|\tilde{\Phi}^{(j)}(x^*)\|_2} \right\|_2 \stackrel{\text{Lem. I.2}}{\leq} \frac{2 \cdot \|\tilde{\Phi}^{(j)}(x) - \tilde{\Phi}^{(j)}(x^*)\|_2}{\|\tilde{\Phi}^{(j)}(x^*)\|_2} \stackrel{(\text{G.3}), (\text{G.4})}{\leq} 8C_1 \|x - x^*\|_2 \leq \gamma.$$

This proves that $\mathcal{N}(T_2, \gamma) \leq \left(\frac{24C_1}{\gamma}\right)^d$.

In total, we hence get

$$\mathcal{N}(T, \gamma) \leq \mathcal{N}(T_1, \gamma) + \mathcal{N}(T_2, \gamma) \leq 3 \cdot \left(\frac{72C_1 \cdot 3^L}{\gamma^2}\right)^d \leq \left(\frac{C_2}{\gamma}\right)^{2dL} \quad \text{for } \gamma \in (0, \varepsilon/2] \quad (\text{G.7})$$

with an appropriate constant $C_2 > 0$. Clearly,

$$\text{diam}((T - T) \cap B_N(0, \varepsilon)) \leq \text{diam}(B_N(0, \varepsilon)) \leq 2\varepsilon.$$

Hence, using Dudley's inequality [20, Theorem 8.1.3], we get with an absolute constant $C_3 > 0$ that

$$w((T - T) \cap B_N(0, \varepsilon)) \leq C_3 \cdot \int_0^{2\varepsilon} \sqrt{\ln(\mathcal{N}((T - T) \cap B_N(0, \varepsilon), \gamma))} \, d\gamma$$

$$\begin{aligned}
& \stackrel{[20, \text{Ex. 4.2.10}]}{\leq} C_3 \cdot \int_0^{2\varepsilon} \sqrt{\ln(\mathcal{N}(T - T, \gamma/2))} \, d\gamma \\
& \leq C_3 \cdot \int_0^{2\varepsilon} \sqrt{\ln(\mathcal{N}(T, \gamma/4)^2)} \, d\gamma \\
& \stackrel{(\text{G.7})}{\leq} 2C_3 \cdot \sqrt{dL} \cdot \int_0^{2\varepsilon} \sqrt{\ln(4C_2/\gamma)} \, d\gamma \\
& \stackrel{\text{Lem. I.4}}{\leq} 8C_3 \cdot \sqrt{dL} \cdot \varepsilon \cdot \ln^{1/2}(2C_2/\varepsilon) \\
& \leq C \cdot \sqrt{dL} \cdot \varepsilon \cdot \ln^{1/2}(1/\varepsilon)
\end{aligned}$$

with $C > 0$ chosen large enough, applying Lemma I.3 in the last step. This proves the bound on the Gaussian width.

Moreover, if $\gamma = \varepsilon/2$ we have $T = T'$ and thus

$$\ln(e \cdot \mathcal{N}(T, \varepsilon/2)) \leq \ln \left(e \cdot \left(\frac{144C_1}{\varepsilon} \right)^d \right) \stackrel{\text{Lem. I.3}}{\leq} C_4 \cdot d \cdot \ln(1/\varepsilon)$$

with an absolute constant $C_4 > 0$. $\square$

We can now prove that the matrices $A^{(j)}(x)$ are indeed sparse for inputs with large Euclidean norm. Note that Proposition G.2 below can also be found as Lemma 6.1 in the main part.

Proposition G.2. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N following Assumption 2.1 with fixed biases. Assume $N \geq Cd$ and $N \geq C \cdot L^3 d \ln(N/d)$. Let $\delta \in (0, e^{-1})$ with*

$$N \geq C \cdot dL^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot \max\{d \cdot \ln(1/\delta), dL\} \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c \cdot 1/L^2.$$

Pick $j \in \{0, \dots, L-1\}$. Then with probability at least $1 - \exp(-c \cdot \delta N)$,

$$\text{for all } x \in \mathbb{R}^d, \|x\|_2 \geq C \cdot 3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{1/2}(1/\delta) : \quad A^{(j)}(x) \in \mathcal{A}_\delta^{(1)}.$$

Here, $\mathcal{A}_\delta^{(1)}$ is as in (E.1) and λ' is as in (G.1).

Proof. Set

$$\mu := \begin{cases} d, & j = 0 \\ N, & j > 0. \end{cases}$$

Let $\tilde{c}_1$ be taken according to Proposition 2.13 and $\tilde{C}_2, \tilde{c}_2 > 0$ be taken according to Proposition 2.14. Set $c' := \min\{\tilde{c}_1, \tilde{c}_2\}$ and $\varepsilon' := \frac{c'}{2\sqrt{2}} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$. As in the beginning of the proof of Lemma G.1, we pick constants $c_1, c_2 > 0$ that satisfy Equations (G.2) and (G.3). Further, in view of Lemma G.1, we know

$$\ln(e \cdot \mathcal{N}(T^{(j-1)}, \varepsilon'/2)) \leq C_1 \cdot d \cdot \ln(1/\varepsilon') \quad \text{and} \quad (\text{G.8})$$

$$w((T^{(j-1)} - T^{(j-1)}) \cap B_\mu(0, \varepsilon')) \leq C_1 \cdot \sqrt{dL} \cdot \varepsilon \cdot \ln^{1/2}(1/\varepsilon') \quad (\text{G.9})$$

with probability at least $1 - \exp(-c_3 \cdot \delta N)$ with absolute constants $C_1, c_3 > 0$. Here, $T^{(j-1)}$ is defined as in Lemma G.1. Since Equations (G.2), (G.3), and (G.9) solely depend on the randomness in $W^{(0)}, \dots, W^{(j-1)}$, we may condition on these random variables and assume that the mentioned properties are satisfied. Using (G.5), we see for $x \in \mathbb{R}^d$ with $\|x\|_2 \geq 16\tilde{C}_2 \cdot 3^L \cdot \lambda' \cdot (\varepsilon')^{-1} = \theta$,

$$\begin{aligned}
& \left\| \Phi^{(j-1)}(x) \right\|_2 \geq \left\| \tilde{\Phi}^{(j-1)}(x) \right\|_2 - \left\| \Phi^{(j-1)}(x) - \tilde{\Phi}^{(j-1)}(x) \right\|_2 \\
& \stackrel{(\text{G.3}), (\text{G.5})}{\geq} \frac{\theta}{4} - 3^L \cdot \lambda' \\
& = 4\tilde{C}_2 \cdot 3^L \cdot \lambda' \cdot \varepsilon^{-1} - 3^L \cdot \lambda'
\end{aligned}$$

$$\begin{aligned}
&= 3^L \cdot \lambda' \cdot (4\widetilde{C}_2 \cdot \varepsilon^{-1} - \underbrace{1}_{\leq 2\widetilde{C}_2 \cdot \varepsilon^{-1}}) \\
&\geq 2\widetilde{C}_2 \cdot 3^L \cdot \lambda' \cdot \varepsilon^{-1} \\
&\stackrel{\varepsilon \leq \delta}{\geq} \widetilde{C}_2 \cdot (\delta/2)^{-1} \lambda'.
\end{aligned} \tag{G.10}$$

The idea is to apply Proposition 2.14 with δ replaced by $\delta/2$ to the set $T^{(j-1)}$. First of all, note that

$$\varepsilon' = c' \cdot \frac{\delta}{2} \cdot (2 \cdot \ln(1/\delta))^{-1/2} \stackrel{\text{Lem. I.3}}{\leq} \widetilde{c}_2 \cdot \frac{\delta}{2} \cdot \ln^{-1/2}(2/\delta).$$

Moreover, using (G.8),

$$\begin{aligned}
(\delta/2)^{-1} \cdot \ln(e \cdot \mathcal{N}(T^{(j-1)}, \varepsilon')) &\leq 2 \cdot \delta^{-1} \cdot \ln(e \cdot \mathcal{N}(T^{(j-1)}, \varepsilon'/2)) \\
&\leq 2C_1 \cdot \delta^{-1} \cdot d \cdot \ln(1/\varepsilon') \\
&\leq 2C_1 \cdot \delta^{-1} \cdot d \cdot \ln \left(\frac{2\sqrt{2}}{c'} \cdot \delta^{-1} \cdot \underbrace{\ln^{1/2}(1/\delta)}_{\leq 1/\delta} \right) \\
&\leq 2C_1 \cdot \delta^{-1} \cdot d \cdot \ln \left(\frac{2\sqrt{2}}{c'} \cdot \delta^{-2} \right) \\
&\stackrel{\text{Lem. I.3}}{\leq} C_2 \cdot \delta^{-1} \cdot d \cdot \ln(1/\delta)
\end{aligned} \tag{G.11}$$

with a suitable constant $C_2 > 0$. Further, using (G.9),

$$\begin{aligned}
&(\delta/2)^{-3} \cdot w^2((T^{(j-1)} - T^{(j-1)}) \cap B_\mu(0, \varepsilon')) \\
&\leq 8 \cdot \delta^{-3} \cdot C_1^2 \cdot dL \cdot (\varepsilon')^2 \cdot \ln(1/\varepsilon') \\
&\leq \left(\frac{c'}{2\sqrt{2}} \right)^2 \cdot C_1^2 \cdot dL \cdot \delta^{-3} \cdot \delta^2 \cdot \ln^{-1}(1/\delta) \cdot \ln \left(\frac{c'}{2\sqrt{2}} \cdot \delta^{-1} \cdot \underbrace{\ln^{1/2}(1/\delta)}_{\leq 1/\delta} \right) \\
&\stackrel{\text{Lem. I.3}}{\leq} C_2 \cdot d \cdot L \cdot \delta^{-1}
\end{aligned}$$

after possibly increasing the size of C_2 . Since by assumption

$$\delta N \geq C \cdot d \cdot \ln(1/\delta) \quad \text{and} \quad \delta N \geq C \cdot dL,$$

we note that the assumptions of Proposition 2.14 are satisfied by taking C large enough. Note that

$$\{\alpha \cdot y : y \in T^{(j-1)}, \alpha \geq \widetilde{C}_2 \cdot (\delta/2)^{-1} \lambda'\} \supseteq \{\Phi^{(j-1)}(x) : x \in \mathbb{R}^d, \|x\|_2 \geq \theta\}$$

as follows from (G.10) and the definition of $T^{(j-1)}$, which thus implies

for all $x \in \mathbb{R}^d$ with $\|x\|_2 \geq \theta$:

$$\# \left\{ i \in \{1, \dots, N\} : \text{sgn} \left((W^{(j)} \Phi^{(j-1)}(x) + b^{(j)})_i \right) \neq \text{sgn} \left((W^{(j)} \Phi^{(j-1)}(x))_i \right) \right\} \leq \frac{\delta}{2} \cdot N$$

with probability at least $1 - \exp(-c_4 \cdot \delta N)$, according to Proposition 2.14 with an absolute constant $c_4 > 0$.

Moreover, we compute

$$\begin{aligned}
&\left\| \frac{\Phi^{(j-1)}(x)}{\|\Phi^{(j-1)}(x)\|_2} - \frac{\widetilde{\Phi}^{(j-1)}(x)}{\|\widetilde{\Phi}^{(j-1)}(x)\|_2} \right\|_2 \stackrel{\text{Lem. I.2}}{\leq} \frac{2 \|\Phi^{(j-1)}(x) - \widetilde{\Phi}^{(j-1)}(x)\|_2}{\underbrace{\|\widetilde{\Phi}^{(j-1)}(x)\|_2}_{\geq \theta/4}} \\
&\stackrel{(\text{G.3}), (\text{G.5})}{\leq} \frac{8 \cdot 3^L \cdot \lambda'}{\theta} \leq \varepsilon
\end{aligned}$$

for every $x \in \mathbb{R}^d$ with $\|x\|_2 \geq \theta$. Since the assumptions of Proposition 2.13 are satisfied (since these are the same conditions as for Proposition 2.14 and we may possibly increase C), we get

for all $x \in \mathbb{R}^d$, $\|x\|_2 \geq \theta$:

$$\#\left\{i : \operatorname{sgn}\left((W^{(j)}\Phi^{(j-1)}(x))_i\right) \neq \operatorname{sgn}\left((W^{(j)}\tilde{\Phi}^{(j-1)}(x))_i\right)\right\} \leq \frac{\delta}{2} \cdot N$$

with probability at least $1 - \exp(-c_5 \cdot \delta N)$ over the randomness in $W^{(j)}$ with an absolute constant $c_5 > 0$. Overall, we have

$$\begin{aligned} \operatorname{Tr} \left| A^{(j)}(x) \right| &= \#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left((W^{(j)}\Phi^{(j-1)}(x) + b^{(j)})_i\right) \neq \operatorname{sgn}\left((W^{(j)}\tilde{\Phi}^{(j-1)}(x))_i\right)\right\} \\ &\leq \#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left((W^{(j)}\Phi^{(j-1)}(x) + b^{(j)})_i\right) \neq \operatorname{sgn}\left((W^{(j)}\Phi^{(j-1)}(x))_i\right)\right\} \\ &\quad + \#\left\{i \in \{1, \dots, N\} : \operatorname{sgn}\left((W^{(j)}\Phi^{(j-1)}(x))_i\right) \neq \operatorname{sgn}\left((W^{(j)}\tilde{\Phi}^{(j-1)}(x))_i\right)\right\} \leq \delta N \end{aligned} \quad (\text{G.12})$$

uniformly over $x \in \mathbb{R}^d$ with $\|x\|_2 \geq \theta$ with probability at least $1 - 2\exp(-c_6 \cdot \delta N)$ over the randomness in $W^{(j)}$ with $c_6 := \min\{c_4, c_5\}$.

In the end, the claim follows by reintroducing the randomness in $W^{(0)}, \dots, W^{(j-1)}$: Note that Equations (G.2), (G.3), and (G.9) occur with probability at least

$$1 - \exp(-c_1 \cdot N) - \exp(-c_2 \cdot \delta N) - \exp(-c_3 \cdot \delta N).$$

Adding (G.12), we get

$$\begin{aligned} &1 - \exp(-c_1 \cdot N) - \exp(-c_2 \cdot \delta N) - \exp(-c_3 \cdot \delta N) - 2\exp(-c_6 \cdot \delta N) \\ &\geq 1 - 5\exp(-c_7 \cdot \delta N), \end{aligned}$$

letting $c_7 := \min\{c_1, c_2, c_3, c_6\}$. The claim follows letting $c := c_7/2$, since we may assume

$$\delta N \geq C \geq \frac{2\ln(5)}{c_7}. \quad \square$$

We can now prove the main result of this section. The special case $\delta \asymp \frac{dL}{N} \cdot \ln(N/d)$ yields Theorem 6.2.

Theorem G.3. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network with L hidden layers of width N following Assumption 2.1 with fixed biases. Assume $N \geq Cd$ and $N \geq C \cdot L^3 d \ln(N/d)$. Let $\delta \in (0, e^{-1})$ with*

$$N \geq C \cdot dL^2 \cdot \ln(1/\delta), \quad \delta N \geq C \cdot \max\{d \cdot \ln(1/\delta), dL\} \quad \text{and} \quad \delta \cdot \ln(1/\delta) \leq c \cdot 1/L^2.$$

Let λ' be chosen as in (G.1). Then the event

$$\sup_{\|x\|_2 \geq C \cdot 3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{-1/2}(1/\delta)} \left\| \overline{\nabla} \Phi(x) - \overline{\nabla} \tilde{\Phi}(x) \right\|_2 \leq C \cdot L \cdot (\sqrt{N} \cdot \delta \ln(1/\delta) + \sqrt{\delta \ln(1/\delta) L d \ln(N/d)})$$

occurs with probability at least $1 - \exp(-c \cdot \delta N)$.

Proof. Let $C_1, c_1 > 0$ be chosen according to Proposition G.2 and choose $C_2, c_2 > 0$ according to Theorem E.2. Set $\theta := C_1 \cdot 3^L \cdot \lambda' \cdot \delta^{-1} \cdot \ln^{-1/2}(1/\delta)$. We aim to show the following statement: For every $\ell \in \{0, \dots, L\}$, the event

$$\begin{aligned} &\sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \left[\overline{\mathbf{D}} \Phi^{(j) \rightarrow (L-1)}(x) - \overline{\mathbf{D}} \tilde{\Phi}^{(j) \rightarrow (L-1)}(x) \right] A \right\|_{2 \rightarrow 2} \\ &\leq C' \cdot L \cdot (\sqrt{N} \cdot \delta \ln(1/\delta) + \sqrt{\delta \ln(1/\delta) L d \ln(N/d)}) \quad \text{for every } j \in \{\ell, \dots, L\} \end{aligned} \quad (\text{G.13})$$

occurs with probability at least $1 - 3(L - \ell) \exp(-c' \cdot \delta N)$, where $C', c' > 0$ are absolute constants that are exactly determined later. This will then imply the statement of the theorem by studying the special case $\ell = 0$.

The proof is via induction over ℓ , starting with the case $\ell = L$, for which there is nothing to show, since by definition we have

$$\overline{\mathbf{D}}\Phi^{(L) \rightarrow (L-1)}(x) = \overline{\mathbf{D}}\widetilde{\Phi}^{(L) \rightarrow (L-1)}(x) = I_{N \times N}$$

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

For the induction step, we take $\ell \in \{0, \dots, L-1\}$ and assume that $\ell+1$ satisfies the claim. Let $\mathcal{E}_1$ be the event defined by

$$\text{for all } x \in \mathbb{R}^d \text{ with } \|x\|_2 \geq \theta \text{ and } j \in \{0, \dots, L-1\} : A^{(j)}(x) \in \mathcal{A}_\delta^{(1)},$$

which occurs with probability at least $1 - \exp(-c_1 \cdot \delta N)$ according to Proposition G.2 under the application of a union bound (noting that $\delta N \geq C \cdot \ln(eL)$). Let the event $\mathcal{E}_2$ be defined by the property

$$\sup_{x \in \mathbb{R}^d, A \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\widetilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \leq C_2 \cdot \sqrt{N \cdot \delta \ln(1/\delta)} \quad \text{for all } j \in \{0, \dots, L-1\},$$

which occurs with probability at least $1 - \exp(-c_2 \cdot \delta N)$ according to Theorem E.2. Let $\mathcal{E}_3$ be the event defined by

$$\begin{aligned} & \sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \left[\overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) - \overline{\mathbf{D}}\widetilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) \right] A \right\|_{2 \rightarrow 2} \\ & \leq C' \cdot L \cdot \left(\sqrt{N \cdot \delta \ln(1/\delta)} + \sqrt{\delta \ln(1/\delta) L d \ln(N/d)} \right) \quad \text{for every } j \in \{\ell+1, \dots, L\}. \end{aligned}$$

According to the induction hypothesis, $\mathcal{E}_3$ occurs with probability at least $1 - 3(L - \ell - 1) \exp(-c' \cdot \delta N)$. Lastly, let $\mathcal{E}_4$ be the event on which

$$\sup_{x \in \mathbb{R}^d, A_1 \in \mathcal{A}_\delta^{(1)}, A_2 \in \mathcal{A}_\delta^{(\ell)}} \left\| A_1 W^{(j)} \overline{\mathbf{D}}\widetilde{\Phi}^{(\ell) \rightarrow (j-1)}(x) A_2 \right\|_{2 \rightarrow 2} \leq C_3 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{L d \ln(N/d)}{N}} \right)$$

for every $j \in \{0, \dots, L-1\}$, which occurs with probability at least $1 - \exp(-c_3 \cdot \delta N)$ according to Lemma E.3. On the intersection $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ we use a decomposition similar to Lemma E.1 to get

$$\begin{aligned} & \sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| \left[W^{(L)} \overline{\mathbf{D}}\Phi^{(\ell) \rightarrow (L-1)}(x) - \overline{\mathbf{D}}\widetilde{\Phi}^{(\ell) \rightarrow (L-1)}(x) \right] A \right\|_{2 \rightarrow 2} \\ & \leq \sum_{j=\ell}^{L-1} \sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A^{(j)}(x) W^{(j)} \overline{\mathbf{D}}\widetilde{\Phi}^{(\ell) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2} \\ & \leq \sum_{j=\ell}^{L-1} \left(\sup_{\|x\|_2 \geq \theta} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A^{(j)}(x) \right\|_{2 \rightarrow 2} \right) \\ & \quad \cdot \left(\sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| A^{(j)}(x) W^{(j)} \overline{\mathbf{D}}\widetilde{\Phi}^{(\ell) \rightarrow (j-1)}(x) A \right\|_{2 \rightarrow 2} \right) \\ & \stackrel{\mathcal{E}_1}{\leq} \sum_{j=\ell}^{L-1} \left(\sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A_2 \right\|_{2 \rightarrow 2} \right) \\ & \quad \cdot \left(\sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}, A_1 \in \mathcal{A}_\delta^{(j)}} \left\| A_2 W^{(j)} \overline{\mathbf{D}}\widetilde{\Phi}^{(\ell) \rightarrow (j-1)}(x) A_1 \right\|_{2 \rightarrow 2} \right) \\ & \stackrel{\mathcal{E}_4}{\leq} C_3 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{L d \ln(N/d)}{N}} \right) \cdot \sum_{j=\ell}^{L-1} \left(\sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A_2 \right\|_{2 \rightarrow 2} \right). \end{aligned}$$

For arbitrary $j \in \{\ell, \dots, L-1\}$,

$$\begin{aligned}
& \sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A_2 \right\|_{2 \rightarrow 2} \\
& \leq \sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \left[\overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) - \overline{\mathbf{D}}\tilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) \right] A_2 \right\|_{2 \rightarrow 2} \\
& \quad + \sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\tilde{\Phi}^{(j+1) \rightarrow (L-1)}(x) A_2 \right\|_{2 \rightarrow 2} \\
& \stackrel{\mathcal{E}_2 \cap \mathcal{E}_3}{\leq} C' \cdot L \cdot \left(\sqrt{N} \cdot \delta \ln(1/\delta) + \sqrt{\delta \ln(1/\delta) L d \ln(N/d)} \right) + C_2 \cdot \sqrt{N \cdot \delta \ln(1/\delta)}.
\end{aligned}$$

Since $\delta \cdot \ln(1/\delta) \leq c \cdot 1/L^2$ and $N \geq C \cdot L^3 d \ln(N/d)$, we can, by possibly increasing C and decreasing c , ensure

$$\sqrt{N} \cdot \delta \ln(1/\delta) \leq \frac{C_2}{2C'} \cdot \frac{\sqrt{N \cdot \delta \ln(1/\delta)}}{L} \quad \text{and} \quad \sqrt{\delta \ln(1/\delta) L d \ln(N/d)} \leq \frac{C_2}{2C'} \cdot \frac{\sqrt{N \cdot \delta \ln(1/\delta)}}{L},$$

which implies

$$C' \cdot L \cdot \left(\sqrt{N} \cdot \delta \ln(1/\delta) + \sqrt{\delta \ln(1/\delta) L d \ln(N/d)} \right) \leq C_2 \sqrt{N \cdot \delta \ln(1/\delta)}.$$

We thus get

$$\sup_{\|x\|_2 \geq \theta, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \overline{\mathbf{D}}\Phi^{(j+1) \rightarrow (L-1)}(x) A_2 \right\|_{2 \rightarrow 2} \leq 2C_2 \cdot \sqrt{N \cdot \delta \ln(1/\delta)}$$

for every $j \in \{\ell, \dots, L-1\}$ on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$.

In total, on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$,

$$\begin{aligned}
& \sup_{\|x\|_2 \geq \theta, A \in \mathcal{A}_\delta^{(\ell)}} \left\| \left[W^{(L)} \overline{\mathbf{D}}\Phi^{(\ell) \rightarrow (L-1)}(x) - \overline{\mathbf{D}}\tilde{\Phi}^{(\ell) \rightarrow (L-1)}(x) \right] A \right\|_{2 \rightarrow 2} \\
& \leq 2C_2 C_3 \cdot \left(\sqrt{N} \cdot \delta \ln(1/\delta) + \sqrt{L d \ln(N/d) \delta \ln(1/\delta)} \right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4) & \geq 1 - \exp(-c_1 \cdot \delta N) - \exp(-c_2 \cdot \delta N) - 3(L - \ell - 1) \exp(-c' \cdot \delta N) - \exp(-c_3 \cdot \delta N) \\
& \geq 1 - 3(L - \ell) \exp(-c' \cdot \delta N),
\end{aligned}$$

letting $c' := \min\{c_1, c_2, c_3\}$, which proves the induction step by choosing $C' = 2C_2 C_3$.

The final claim then follows from the case $\ell = 0$, noting that $\delta N \geq C \cdot dL \geq C \cdot \ln(eL)$. $\square$

Note that there is a trade-off between decreasing the size of δ (in order to make the bound sharper) and increasing the size of δ in order to make θ smaller.

H Bounds on the Lipschitz constants for networks with biases

In this section, we use the results from Section 6.1 to prove bounds on the Lipschitz constant of random ReLU nets with general biases. Precisely, in order to establish our upper bound (see Appendix H.1), we assume that the entries of the bias vectors are drawn independently from (possibly different) symmetric distributions and satisfy a small ball property, i.e., we assume that there exists a constant $C_\tau > 0$ for which

$$\sup_{\ell \in \{0, \dots, L-1\}, i \in \{1, \dots, N\}, t \in \mathbb{R}, \varepsilon > 0} \left(\varepsilon^{-1} \cdot \mathbb{P} \left(b_i^{(\ell)} \in (t - \varepsilon, t + \varepsilon) \right) \right) \leq C_\tau.$$

For our lower bounds (see Appendix H.2), no special assumptions on the distribution of the biases are required.

H.1 Upper bound

Let $R > 0$ be arbitrary. The first goal is to establish a bound on

$$\sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x)\|_2.$$

This works very similar to the proof of Theorem E.2, with the difference that Proposition 2.15 is used instead of Proposition 2.13. We start by bounding the Jacobians of the intermediate maps $\Phi^{(\ell)}$.

Lemma H.1. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.3 with L hidden layers of width N and with small ball constant $C_\tau > 0$. Let $C'_\tau := \sqrt{2/N} \cdot C_\tau$. Let further $R \geq (C'_\tau)^{-1}$, $\delta \in (0, e^{-1})$ and assume $N \geq C \cdot d$ and*

$$N \geq C \cdot L^3 d \ln(N/d), \quad N \geq C \cdot L^2 d \ln(RC'_\tau/\delta), \quad \delta N \geq C \cdot d \cdot \ln(RC'_\tau/\delta), \quad \delta \ln(1/\delta) \leq c \cdot 1/L^2.$$

We let $C_1 > 0$ be the absolute constant from Lemma 3.2 and $c' > 0$ be the absolute constant from Lemma C.7 and set $\varepsilon := \frac{c'}{2C_1} \cdot (C'_\tau)^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta)$. Then, for every $\ell_2 \in \{-1, \dots, L-1\}$ with probability at least $1 - 3(\ell_2 + 1) \exp(-c \cdot \delta N)$,

$$\begin{aligned} (1) \quad & \sup_{x, y \in B_d(0, R), \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(j)}(x) - D^{(j)}(y) \right] \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2\}, \\ (2) \quad & \sup_{\substack{x \in B_d(0, R) \\ A \in A_\delta^{(1)}}} \left\| AW^{(j)} \bar{D}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \leq C \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \quad \text{for all } j \in \{0, \dots, \ell_2\}, \\ (3) \quad & \sup_{x \in B_d(0, R)} \left\| \bar{D}\Phi^{(j)}(x) \right\|_{2 \rightarrow 2} \leq C \quad \text{for all } j \in \{-1, \dots, \ell_2\}. \end{aligned}$$

Proof. The proof is via induction over ℓ_2 . In the case $\ell_2 = -1$, only (3) needs to be considered. However, here we have, since necessarily $j = -1$,

$$\sup_{x \in B_d(0, R)} \left\| \bar{D}\Phi^{(j)}(x) \right\|_{2 \rightarrow 2} = 1.$$

Let us move to the induction step. Pick $\ell_2 \in \{0, \dots, L-1\}$ and assume that with probability at least $1 - 3\ell_2 \exp(-c \cdot \delta N)$,

$$\begin{aligned} (1) \quad & \sup_{x, y \in B_d(0, R), \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(j)}(x) - D^{(j)}(y) \right] \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\}, \\ (2) \quad & \sup_{\substack{x \in B_d(0, R) \\ A \in A_\delta^{(1)}}} \left\| AW^{(j)} \bar{D}\Phi^{(j-1)} \right\|_{2 \rightarrow 2} \leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \\ & \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\}, \\ (3) \quad & \sup_{x \in B_d(0, R)} \left\| \bar{D}\Phi^{(j)}(x) \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\} \end{aligned}$$

where $C_2, c > 0$ are appropriate constants to be exactly determined later and we recall that $C_1 > 0$ is the absolute constant from Lemma 3.2.

We now condition on the weights $W^{(0)}, \dots, W^{(\ell_2-1)}$ and biases $b^{(0)}, \dots, b^{(\ell_2-1)}$ and assume that properties (1)-(3) are satisfied. By (3) and Theorem 2.8, this in particular implies

$$\text{Lip}_{2 \rightarrow 2} \left(\Phi^{(j)} \Big|_{B_d(0, R)} \right) \leq \sup_{x \in B_d(0, R)} \left\| \bar{D}\Phi^{(j)} \right\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\}. \quad (\text{H.1})$$

We set $T := \Phi^{(\ell_2-1)}(B_d(0, R))$. Note that for $x, y \in B_d(0, R)$ with $\|x - y\|_2 \leq \varepsilon$ we have

$$\left\| \Phi^{(\ell_2-1)}(x) - \Phi^{(\ell_2-1)}(y) \right\|_2 \leq 2C_1 \cdot \varepsilon = c' \cdot (C'_\tau)^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta).$$

Hence, we may apply Lemma C.7 by using that

$$N \geq C \cdot d \cdot \delta^{-1} \cdot \ln(RC'_\tau/\delta)$$

and get

$$\begin{aligned} & \sup_{x, y \in B_d(0, R), \|x-y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(\ell_2)}(y) - D^{(\ell_2)}(x) \right] \right| \\ & \leq \sup_{\substack{x, y \in T, \\ \|x-y\|_2 \leq c' \cdot (C'_\tau)^{-1} \cdot \delta \cdot \ln^{-1/2}(1/\delta)}} \# \left\{ i \in \{1, \dots, N\} : \text{sgn}((W^{(\ell_2)}x)_i + b_i^{(\ell_2)}) \neq \text{sgn}((W^{(\ell_2)}y)_i + b_i^{(\ell_2)}) \right\} \leq \delta N \end{aligned} \quad (\text{H.2})$$

with probability at least $1 - \exp(-c' \cdot \delta N)$ over the randomness in $W^{(\ell_2)}$ and $b^{(\ell_2)}$. We call the event on which (H.2) holds $\mathcal{E}_1$.

Further, set

$$\mathcal{V} := \left\{ \overline{D}\Phi^{(\ell_2-1)}(x) : x \in B_d(0, R) \right\}.$$

Note that we have $\|M\|_{2 \rightarrow 2} \leq 2C_1$ and $\text{rank}(M) \leq d \leq \delta N$ for every $M \in \mathcal{V}$ by ((c)). Here, we used that $d \leq C \cdot d \cdot \ln(RC'_\tau/\delta) \leq \delta N$. Hence, for fixed $x \in B_d(0, R)$ and $A \in \mathcal{A}_\delta^{(1)}$, we find, using Lemma I.5,

$$\left\| AW^{(\ell_2)} \overline{D}\Phi^{(\ell_2-1)}(x) \right\|_{2 \rightarrow 2} \leq \sqrt{2} \cdot 2C_1 \cdot \frac{\sqrt{\delta N} + \sqrt{\delta N} + t}{\sqrt{N}}$$

with probability at least $1 - \exp(-c_1 \cdot t^2)$ for every $t \geq C_3$ for absolute constants $C_3, c_1 > 0$. Moreover, combining Lemma I.6 and [13, Lemma 5.7], we get

$$\begin{aligned} \ln(|\mathcal{V}|) + \ln \left(\left| \mathcal{A}_\delta^{(1)} \right| \right) & \leq L(d+1) \cdot \ln \left(\frac{eN}{d+1} \right) + \delta N \cdot \ln(4e/\delta) \\ & \stackrel{\text{Lem. I.3}}{\leq} 2eLd \cdot \ln(N/d) + 4e \cdot N \cdot \delta \ln(1/\delta) \\ & \leq 4e(Ld \ln(N/d) + N\delta \ln(1/\delta)). \end{aligned}$$

Hence, via a union bound, we observe

$$\sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \left\| AW^{(\ell_2)} \overline{D}\Phi^{(\ell_2-1)}(x) \right\|_{2 \rightarrow 2} \leq \sqrt{2} \cdot 2C_1 \cdot \frac{\sqrt{\delta N} + \sqrt{\delta N} + t}{\sqrt{N}}$$

with probability at least

$$1 - \exp(4e(Ld \ln(N/d) + N\delta \ln(1/\delta)) - c_1 t^2)$$

for every $t \geq C_3$. We then explicitly pick

$$t := \sqrt{c_1^{-1} \cdot 5e \cdot (Ld \ln(N/d) + N\delta \cdot \ln(1/\delta))},$$

which yields the existence of an absolute constant $C_2 > 0$, such that

$$\begin{aligned} & \sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \left\| AW^{(\ell_2)} \overline{D}\Phi^{(\ell_2-1)}(x) \right\|_{2 \rightarrow 2} \\ & \leq C_2 \cdot \frac{\sqrt{N \cdot \delta \cdot \ln(1/\delta)} + \sqrt{Ld \ln(N/d)}}{\sqrt{N}} \end{aligned}$$

with probability at least

$$1 - \exp(-e \cdot (Ld \ln(N/d) + N\delta \ln(1/\delta))) \geq 1 - \exp(-\delta N)$$

over the randomness in $W^{(\ell_2)}$. We call the event defined by property (E.5) $\mathcal{E}_2$.

Moreover, let $\mathcal{N}_\varepsilon \subseteq B_d(0, R)$ be an ε -net of $B_d(0, R)$ with

$$\begin{aligned} \ln(|\mathcal{N}_\varepsilon|) &\leq d \cdot \ln(3R/\varepsilon) = d \cdot \ln \left(\frac{6C_1 \cdot RC'_\tau \cdot \ln^{1/2}(1/\delta)}{c' \cdot \delta} \right)^{\ln^{1/2}(1/\delta) \leq 1/\delta} \leq d \cdot \ln \left(\frac{6C_1 \cdot RC'_\tau \cdot \ln^{1/2}(1/\delta)}{c' \cdot \delta^2} \right) \\ &\leq 2 \cdot d \cdot \ln(6C_1 \cdot RC'_\tau / (c' \delta)) \leq C_4 \cdot d \cdot \ln(RC'_\tau / \delta) \end{aligned} \quad (\text{H.3})$$

with an absolute constant $C_4 > 0$ using Lemma I.3. For $x \in B_d(0, R)$ we denote by $\pi(x) \in \mathcal{N}_\varepsilon$ a net point with $\|x - \pi(x)\|_2 \leq \varepsilon$. On $\mathcal{E}_1 \cap \mathcal{E}_2$, by Lemma E.1,

$$\begin{aligned} &\sup_{x \in B_d(0, R)} \left\| \overline{D}\Phi^{(\ell_2)}(\pi(x)) - \overline{D}\Phi^{(\ell_2)}(x) \right\|_{2 \rightarrow 2} \\ &\leq \sup_{x \in B_d(0, R)} \sum_{j=0}^{\ell_2} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) (D^{(j)}(\pi(x)) - D^{(j)}(x)) W^{(j)} \overline{D}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \\ &\leq \sup_{x \in B_d(0, R)} \left(\sum_{j=0}^{\ell_2} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) (D^{(j)}(\pi(x)) - D^{(j)}(x)) \right\|_{2 \rightarrow 2} \right. \\ &\quad \cdot \left. \left\| (D^{(j)}(\pi(x)) - D^{(j)}(x)) W^{(j)} \overline{D}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\ &\leq \sum_{j=0}^{\ell_2} \left(\sup_{x \in B_d(0, R)} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(\pi(x)) (D^{(j)}(\pi(x)) - D^{(j)}(x)) \right\|_{2 \rightarrow 2} \right) \\ &\quad \cdot \left(\sup_{x \in B_d(0, R)} \left\| (D^{(j)}(\pi(x)) - D^{(j)}(x)) W^{(j)} \overline{D}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\ &\stackrel{\mathcal{E}_{1,(1)}}{\leq} \sum_{j=0}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A \right\|_{2 \rightarrow 2} \right) \cdot \left(\sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \left\| A W^{(j)} \overline{D}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\ &\stackrel{\mathcal{E}_{2,(2)}}{\leq} C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \cdot \sum_{j=0}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A_2 \right\|_{2 \rightarrow 2} \right). \end{aligned}$$

Since $\delta \ln(1/\delta) \leq c/L^2$ and $N \geq C \cdot L^3 d \ln(N/d)$,

$$C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \leq \frac{1}{L}.$$

Hence, on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned} &\sup_{x \in B_d(0, R)} \left\| \overline{D}\Phi^{(\ell_2)}(\pi(x)) - \overline{D}\Phi^{(\ell_2)}(x) \right\|_{2 \rightarrow 2} \\ &\leq \frac{1}{L} \cdot \sum_{j=0}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}_\delta^{(1)}} \left\| \overline{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*) A \right\|_{2 \rightarrow 2} \right). \end{aligned}$$

Recall that up to now we have conditioned on $W^{(0)}, \dots, W^{(\ell_2-1)}$ and $b^{(0)}, \dots, b^{(\ell_2-1)}$. Reintroducing the randomness over these random matrices, we note that with probability at least $1 - 3\ell_2 \exp(-c \cdot \delta N) - \exp(-c' \cdot \delta N) - \exp(-\delta N)$,

$$(1) \quad \sup_{x, y \in \mathbb{S}^{d-1}, \|x-y\|_2 \leq \varepsilon} \text{Tr} \left| D^{(j)}(x) - D^{(j)}(y) \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, \ell_2\},$$

$$\begin{aligned}
(2) \quad & \sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \|AW^{(j)}\bar{D}\Phi^{(j-1)}\|_{2 \rightarrow 2} \leq C_2 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \\
& \text{for all } j \in \{0, \dots, \ell_2\}, \\
(3) \quad & \sup_{x \in B_d(0, R)} \|\bar{D}\Phi^{(\ell_2)}(\pi(x)) - \bar{D}\Phi^{(\ell_2)}(x)\|_{2 \rightarrow 2} \\
& \leq \frac{1}{L} \cdot \sum_{j=0}^{\ell_2} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A_2 \in \mathcal{A}_\delta^{(1)}} \|\bar{D}\Phi^{(j+1) \rightarrow (\ell_2)}(x^*)A_2\|_{2 \rightarrow 2} \right), \\
(4) \quad & \sup_{x \in B_d(0, R)} \|\bar{D}\Phi^{(j)}(x)\|_{2 \rightarrow 2} \leq 2C_1 \quad \text{for all } j \in \{0, \dots, \ell_2 - 1\}.
\end{aligned}$$

We call the event defined by properties (1)-(4) $\mathcal{E}_3$, noting that the randomness is now with respect to $W^{(0)}, \dots, W^{(\ell_2)}$ and $b^{(0)}, \dots, b^{(\ell_2)}$.

For fixed $x^* \in \mathcal{N}_\varepsilon$, $j \in \{0, \dots, L\}$ and $A_2 \in \mathcal{A}_\delta^{(j)}$, we get

$$\|\bar{D}\Phi^{(j) \rightarrow (\ell_2)}(x^*)A_2\|_{2 \rightarrow 2} \leq C_1$$

with probability at least $1 - \exp(-c_1 \cdot N/L^2)$ using Lemma 3.2 with an absolute constant $c_1 > 0$. Since

$$\ln \left(|\mathcal{A}_\delta^{(j)}| \right) \leq \delta N \cdot \ln(4e/\delta) \stackrel{\text{Lem. 1.3}}{\leq} 4e \cdot \delta N \cdot \ln(1/\delta),$$

a union bound yields that

$$\|\bar{D}\Phi^{(j) \rightarrow (\ell_2)}(x^*)A_2\|_{2 \rightarrow 2} \leq C_1 \quad \text{for all } x^* \in \mathcal{N}_\varepsilon, j \in \{0, \dots, L\} \text{ and } A_2 \in \mathcal{A}_\delta^{(j)}. \quad (\text{H.4})$$

with probability at least

$$1 - \exp(\ln(L+1) + 4e \cdot \delta N \cdot \ln(1/\delta) + C_4 \cdot d \cdot \ln(RC'_\tau/\delta) - c_1 \cdot N/L^2) \geq 1 - \exp(-c_2 \cdot N/L^2).$$

For the last step, we used $N \geq C \cdot L^3 d \ln(N/d) \geq C \cdot L^2 \ln(eL)$, $\delta \ln(1/\delta) \leq c/L^2$ and $N \geq C \cdot L^2 d \ln(RC'_\tau/\delta)$ and set $c_2 := c_1/2$. We call the event defined by (H.4) $\mathcal{E}_4$. On $\mathcal{E}_5 := \mathcal{E}_3 \cap \mathcal{E}_4$, by (3),

$$\sup_{x \in B_d(0, R)} \|\bar{D}\Phi^{(\ell_2)}(\pi(x)) - \bar{D}\Phi^{(\ell_2)}(x)\|_{2 \rightarrow 2} \leq C_1$$

and therefore

$$\begin{aligned}
& \sup_{x \in B_d(0, R)} \|\bar{D}\Phi^{(\ell_2)}(x)\|_{2 \rightarrow 2} \\
& \leq \sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{D}\Phi^{(\ell_2)}(x^*)\|_{2 \rightarrow 2} + \sup_{x \in B_d(0, R)} \|\bar{D}\Phi^{(\ell_2)}(\pi(x)) - \bar{D}\Phi^{(\ell_2)}(x)\|_{2 \rightarrow 2} \leq 2C_1.
\end{aligned}$$

Altogether, all the desired properties are satisfied on $\mathcal{E}_5$. Note that we get

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_5) & \geq 1 - 3\ell_2 \exp(-c \cdot \delta N) - \exp(-c' \cdot \delta N) - \exp(-\delta N) - \exp(-c_2 \cdot N/L^2) \\
& \geq 1 - 3(\ell_2 + 1) \exp(-c \cdot \delta N)
\end{aligned}$$

by taking $c := \min\{c', c_2\}$ and using $\delta \leq 1/L^2$. $\square$

We can now incorporate the final layer to get a bound on the gradient of Φ on the ball $B_d(0, R)$.

Theorem H.2. *There exist absolute constants $C, c > 0$ such that the following holds. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.3 with L hidden layers of width N and with small ball constant $C_\tau > 0$. Set $C'_\tau := \sqrt{2/N} \cdot C_\tau$. Let further $R \geq (C'_\tau)^{-1}$, $\delta \in (0, e^{-1})$ and assume $N \geq Cd$, $N \geq L^3 d \ln(N/d)$ and*

$$\begin{aligned}
\delta N & \geq C \cdot d \cdot \ln(RC'_\tau/\delta), \quad N \geq C \cdot L^2 d \ln(RC'_\tau/\delta), \quad \delta \ln(1/\delta) \leq c \cdot 1/L^2, \\
\delta \ln^{1/2}(1/\delta) & \leq c \cdot \frac{\sqrt{d}}{L\sqrt{N}}, \quad \delta \leq c \cdot \frac{1}{L^3 \ln(N/d)} \quad \text{and} \quad d \ln(RC'_\tau/\delta) \geq C.
\end{aligned}$$

Then the event

$$\sup_{x \in B_d(0, R)} \|\overline{\nabla} \Phi(x)\|_2 \leq C \cdot \sqrt{d \cdot \ln(RC'_\tau/\delta)}$$

occurs with probability at least $1 - \exp(-c \cdot d \ln(RC'_\tau/\delta))$.

Proof. Step 1 (Point estimate): We start with a pointwise estimate: Let $x \in B_d(0, R)$ be fixed, $j \in \{0, \dots, L-1\}$ and $A \in \mathcal{A}_\delta^{(j)}$. Using Lemma 3.3, we get the existence of absolute constants $C_1, c_1 > 0$ such that

$$\left\| W^{(L)} \overline{\mathbf{D}} \Phi^{(j) \rightarrow (L-1)}(x) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{\text{Tr } |A|} + t)$$

with probability at least $1 - \exp(-c_1 t^2)$ for every $C_1 \leq t \leq \frac{\sqrt{N}}{L}$. Note that we may assume $d \leq N/L^2$ and

$$\delta N \leq \delta \cdot \ln(1/\delta) \cdot N \leq c \cdot N/L^2 \leq N/L^2,$$

by taking C sufficiently large and c sufficiently small, which implies $\text{Tr } |A| \leq N/L^2$ for every $A \in \mathcal{A}_\delta^{(j)}$.

Step 2 (Union bound): We continue by taking a union bound: We take $\varepsilon > 0$ as in Lemma H.1 and let $\mathcal{N}_\varepsilon$ be an ε -net of $B_d(0, R)$ with

$$\ln(|\mathcal{N}_\varepsilon|) \leq C_2 \cdot d \cdot \ln(C'_\tau R/\delta)$$

with an absolute constant $C_2 > 0$ as in (H.3). Let $j \in \{0, \dots, L-1\}$. Then we have

$$\ln \left(\left| \mathcal{A}_\delta^{(j)} \right| \right) \leq 4e \cdot \delta N \cdot \ln(1/\delta)$$

according to Lemma I.6. Since by assumption $\delta N \geq C \cdot d \cdot \ln(C'_\tau R/\delta)$, we get

$$\ln(|\mathcal{N}_\varepsilon|) + \ln \left(\left| \mathcal{A}_\delta^{(j)} \right| \right) \leq C_3 \cdot N \cdot \delta \ln(1/\delta)$$

with an absolute constant $C_3 > 0$. Hence, taking a union bound, we get

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \overline{\mathbf{D}} \Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{\delta N} + t),$$

with probability at least $1 - \exp(C_3 \cdot \delta N \cdot \ln(1/\delta) - c_1 t^2)$ for every $C_1 \leq t \leq \frac{\sqrt{N}}{L}$. We explicitly pick

$$t := \sqrt{2C_3 c_1^{-1} \cdot \delta N \cdot \ln(1/\delta)}$$

and hence get

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \overline{\mathbf{D}} \Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_4 \cdot \sqrt{N \cdot \delta \ln(1/\delta)} \quad (\text{H.5})$$

with probability at least

$$1 - \exp(-N \cdot \delta \ln(1/\delta)) \geq 1 - \exp(-\delta N)$$

with a constant $C_4 > 0$. Further, note that $t \leq \frac{\sqrt{N}}{L}$ holds by taking C large and c small enough, since we have

$$\delta \ln(1/\delta) \leq c \cdot 1/L^2 \quad \text{and} \quad N \geq C \cdot L^2 d \ln(1/\delta)$$

and $t \geq C_1$ is trivially satisfied.

In the case $j = 0$, we may leave out the union bound over $\mathcal{A}_\delta^{(j)}$ (since $|\mathcal{A}_\delta^{(0)}| = 1$) and get

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \|\overline{\nabla} \Phi(x)\|_2 = \sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(0)}}} \left\| W^{(L)} \overline{\mathbf{D}} \Phi^{(L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_1 \cdot (\sqrt{d} + t)$$

with probability at least $1 - \exp(C_3 \cdot d \cdot \ln(RC'_\tau/\delta) - c_1 t^2)$. Taking

$$t := \sqrt{2C_3 c_1^{-1} \cdot d \ln(RC'_\tau/\delta)},$$

we obtain

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{\nabla} \Phi(x)\|_2 \leq C_4 \cdot \sqrt{d \cdot \ln(RC'_\tau/\delta)} \quad (\text{H.6})$$

with probability at least $1 - \exp(-d \ln(RC'_\tau/\delta))$, after possibly increasing the constant C_4 .

Step 3 (Controlling the deviation):

It remains to control the deviation when moving from an arbitrary point $x \in B_d(0, R)$ to a net point $\pi(x) \in \mathcal{N}_\varepsilon$ with $\|x - \pi(x)\|_2 \leq \varepsilon$.

We let $\mathcal{E}_1$ be the event defined by the properties

- (1) $\sup_{x, y \in B_d(0, R), \|x - y\|_2 \leq \varepsilon} \left| \text{Tr} \left[D^{(j)}(x) - D^{(j)}(y) \right] \right| \leq \delta N \quad \text{for all } j \in \{0, \dots, L-1\},$
- (2) $\sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \|AW^{(j)}\bar{\text{D}}\Phi^{(j-1)}(x)\|_{2 \rightarrow 2} \leq C \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right)$
for all $j \in \{0, \dots, L-1\}$.

By Lemma H.1, $\mathbb{P}(\mathcal{E}_1) \geq 1 - 3L \exp(-c' \cdot \delta N) \geq 1 - \exp(-c'' \cdot \delta N)$ with absolute constants $c', c'' > 0$, since $\delta N \geq C \cdot \ln(eL)$. Using Lemma E.1, on $\mathcal{E}_1$, we get

$$\begin{aligned} & \sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(\pi(x)) - \bar{\nabla} \Phi(x)\|_2 \\ & \leq \sum_{j=0}^{L-1} \left(\sup_{x \in B_d(0, R)} \left\| W^{(L)} \bar{\text{D}}\Phi^{(j+1) \rightarrow (L-1)}(\pi(x)) (D^{(j)}(\pi(x)) - D^{(j)}(x)) \right\|_{2 \rightarrow 2} \right) \\ & \quad \cdot \left(\sup_{x \in B_d(0, R)} \left\| (D^{(j)}(\pi(x)) - D^{(j)}(x)) W^{(j)} \bar{\text{D}}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\ & \leq \sum_{j=0}^{L-1} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \bar{\text{D}}\Phi^{(j+1) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \right) \\ & \quad \cdot \left(\sup_{\substack{x \in B_d(0, R) \\ A \in \mathcal{A}_\delta^{(1)}}} \left\| AW^{(j)} \bar{\text{D}}\Phi^{(j-1)}(x) \right\|_{2 \rightarrow 2} \right) \\ & \leq C_4 \cdot \left(\sqrt{\delta \ln(1/\delta)} + \sqrt{\frac{Ld \ln(N/d)}{N}} \right) \cdot \sum_{j=0}^{L-1} \left(\sup_{x^* \in \mathcal{N}_\varepsilon, A \in \mathcal{A}_\delta^{(1)}} \left\| W^{(L)} \bar{\text{D}}\Phi^{(j+1) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \right). \end{aligned}$$

We let $\mathcal{E}_2$ be the event defined by

$$\sup_{\substack{x^* \in \mathcal{N}_\varepsilon \\ A \in \mathcal{A}_\delta^{(j)}}} \left\| W^{(L)} \bar{\text{D}}\Phi^{(j) \rightarrow (L-1)}(x^*) A \right\|_{2 \rightarrow 2} \leq C_3 \cdot \sqrt{N \cdot \delta \ln(1/\delta)} \quad \text{for all } j \in \{1, \dots, L\}.$$

Using (H.5) and a union bound,

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - \exp(\ln(L) - \delta N) \geq 1 - \exp(-c_2 \cdot \delta N)$$

with an absolute constant $c_2 > 0$, since $\delta N \geq C \cdot \ln(eL)$. On $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned} & \sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(\pi(x)) - \bar{\nabla} \Phi(x)\|_2 \\ & \leq C_5 \cdot L \cdot \left(\sqrt{N \cdot \delta \ln(1/\delta)} + \sqrt{Ld \ln(N/d) \cdot \delta \ln(1/\delta)} \right). \end{aligned} \quad (\text{H.7})$$

Since by assumption $\delta \ln^{1/2}(1/\delta) \leq c \cdot \frac{\sqrt{d}}{L \cdot \sqrt{N}}$ and $\delta \leq c \cdot \frac{1}{L^3 \cdot \ln(N/d)}$, we get

$$\sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(\pi(x)) - \bar{\nabla} \Phi(x)\|_2 \leq \sqrt{d \cdot \ln(1/\delta)} \leq \sqrt{d \cdot \ln(RC'_\tau/\delta)}$$

on $\mathcal{E}_1 \cap \mathcal{E}_2$.

Step 4 (Concluding the proof): Let $\mathcal{E}_3$ be the event defined by

$$\sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{\nabla} \Phi(x)\|_2 \leq C_3 \cdot \sqrt{d \cdot \ln(RC'_\tau/\delta)}.$$

By (H.6),

$$\mathbb{P}(\mathcal{E}_3) \geq 1 - \exp(-d \ln(RC'_\tau/\delta)).$$

On $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we thus have

$$\sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x)\|_2 \leq \sup_{x^* \in \mathcal{N}_\varepsilon} \|\bar{\nabla} \Phi(x)\|_2 + \sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \Phi(\pi(x))\|_2 \leq (C_3 + 1) \cdot \sqrt{d \cdot \ln(RC'_\tau/\delta)}.$$

To obtain the claim, note

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) &\geq 1 - \exp(-c'' \cdot \delta N) - \exp(-c_2 \cdot \delta N) - \exp(-d \ln(RC'_\tau/\delta)) \\ &\geq 1 - 3 \exp(-d \ln(RC'_\tau/\delta)) \\ &\geq 1 - \exp(-c_4 \cdot d \ln(RC'_\tau/\delta)) \end{aligned}$$

with an absolute constant $c_4 > 0$. Here we used $\delta N \geq C \cdot d \ln(RC'_\tau/\delta)$ and $d \ln(RC'_\tau/\delta) \geq C$. This proves the claim. $\square$

The special case of $\delta \asymp \frac{d}{N} \cdot \ln(RC'_\tau N/d)$ yields the following corollary.

Corollary H.3. *There exist absolute constants $C, c > 0$ satisfying the following: Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a random ReLU network following Assumption 2.3 with L hidden layers of width N satisfying $N \geq Cd$ and with small ball constant $C_\tau > 0$. Set $C'_\tau := \sqrt{2/N} \cdot C_\tau$. Let further $R \geq (C'_\tau)^{-1}$ and assume*

$$\begin{aligned} d \cdot \ln(RC'_\tau N/d) &\geq C, \quad N \geq C \cdot L^2 d \ln^2(RC'_\tau N/d) \ln(N/d), \\ N &\geq C \cdot d L^3 \ln(RC'_\tau N/d) \ln(N/d). \end{aligned}$$

Then the event

$$\sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x)\|_2 \leq C \cdot \sqrt{d \cdot \ln(RC'_\tau N/d)}$$

occurs with probability at least $1 - \exp(-c \cdot d \ln(RC'_\tau N/d))$.

Picking $R = 3^L \cdot \lambda \cdot \sqrt{N/2} \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$ and using Theorem 6.2 and Corollary H.3, we can now bound the global ℓ^2 -Lipschitz constant of a network $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ following Assumption 2.3.

Proof of Theorem 6.3. In the beginning of the proof, we set $\lambda' := \sqrt{N/2} \cdot \lambda$ and $C'_\tau := \sqrt{2/N} \cdot C_\tau$. Let the constant $C_1 > 0$ be taken from Theorem 6.2 and set $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$ as in Theorem 6.2. Moreover, note that $\lambda' C'_\tau = \lambda C_\tau$ by construction.

Step 1: Control the gradient on the ball We check that the conditions of Corollary H.3 are satisfied. To this end, let $C_2, c_2 > 0$ be the constants from Corollary H.3. Note that $R \geq \lambda' \geq (C'_\tau)^{-1}$ by assumption. Moreover, since $N \geq C \cdot d$, we have

$$d \cdot \ln(RC'_\tau N/d) \geq C.$$

Further, note that

$$\begin{aligned} \ln(RC'_\tau N/d) &\leq \ln\left(\frac{3^L \cdot \lambda' \cdot N^2 \cdot C'_\tau}{d^2 L}\right) = \ln(3^L) + \ln\left(\frac{\lambda' \cdot N^2 \cdot C'_\tau}{d^2 L}\right) \\ &\leq 2L + \ln\left(\left(\frac{\lambda' \cdot C'_\tau \cdot N}{d}\right)^2\right) \leq 2 \cdot \left(L + \ln\left(\frac{\lambda \cdot C_\tau \cdot N}{d}\right)\right). \end{aligned} \quad (\text{H.8})$$

This gives us

$$\begin{aligned} C_2 \cdot L^2 d \ln^2(RC'_\tau N/d) \ln(N/d) &\leq 4C_2 \cdot L^2 d (L + \ln(\lambda C_\tau N/d))^2 \ln(N/d) \\ &\stackrel{(a+b)^2 \leq 2a^2 + 2b^2}{\leq} 8C_2 \cdot L^4 d \ln(N/d) + 8C_2 \cdot L^2 d \ln^2(\lambda C_\tau N/d) \ln(N/d) \leq N \end{aligned}$$

by assumption, since $N \geq C \cdot L^4 d \ln(N/d)$ and $N \geq C \cdot L^2 d \ln^2(\lambda C_\tau N/d) \ln(N/d)$. Moreover, we get

$$C_2 \cdot L^3 d \ln(N/d) \ln(RC'_\tau N/d) \leq 2C_2 \cdot L^4 d \ln(N/d) + 2C_2 \cdot L^3 d \ln(N/d) \ln(\lambda C_\tau N/d) \leq N$$

by assumption, since $N \geq C \cdot L^4 d \ln(N/d)$ and $N \geq C \cdot L^3 d \ln(N/d) \ln(\lambda C_\tau N/d)$. We can thus apply Corollary H.3 and get

$$\begin{aligned} \sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x)\|_2 &\leq C_2 \cdot \sqrt{d \cdot \ln(RC'_\tau N/d)} \\ &\stackrel{(H.8)}{\leq} \sqrt{2} \cdot C_2 \cdot \left(\sqrt{dL} + \sqrt{d \cdot \ln(\lambda C_\tau N/d)} \right) \end{aligned}$$

with probability at least $1 - \exp(-c_2 \cdot d \ln(RC'_\tau N/d)) \geq 1 - \exp(-c_2 \cdot d \ln(N/d))$.

Step 2: Control the difference to zero-bias network We use Theorem 6.2 to control the difference of the gradient of Φ to the gradient of its zero-bias counterpart $\tilde{\Phi}$. To this end, we first condition on the bias vectors $b^{(0)}, \dots, b^{(L-1)}$ and assume that

$$\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \lambda = \frac{\lambda' \sqrt{2}}{\sqrt{N}}.$$

Since the biases are now fixed, R was chosen appropriately and the conditions of Theorem 6.2 are per assumption satisfied, we get

$$\sup_{\|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \leq C_3 \cdot \frac{dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}}$$

with probability at least $1 - \exp(-c_3 \cdot dL \ln(N/d))$ with absolute constants $C_3, c_3 > 0$. Note that by assumption we have

$$N \geq C \cdot dL^3 \cdot \ln^2(N/d) \ln^2(N/(dL)),$$

which implies

$$dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL)) = \sqrt{d^2 L^4 \cdot \ln^2(N/d) \ln^2(N/(dL))} \leq \sqrt{dLN}$$

and hence,

$$\sup_{\|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \leq C_3 \cdot \sqrt{dL}$$

with probability at least $1 - \exp(-c_3 \cdot dL \ln(N/d))$. Recall that we conditioned on $b^{(0)}, \dots, b^{(L-1)}$ so far. Reintroducing the randomness over the bias vectors, we obtain

$$\sup_{\|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \leq C_3 \cdot \sqrt{dL}$$

with probability at least $1 - \exp(-c_3 \cdot dL \ln(N/d)) - \exp(-d \ln(N/d))$.

Step 3: Control the gradient of the zero-bias network We use Theorem 4.1 and get

$$\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \tilde{\Phi}(x)\|_2 \leq C_4 \cdot \sqrt{d \ln(N/d)} \leq C_4 \cdot \sqrt{d \cdot \ln(\lambda C_\tau N/d)}$$

with probability at least $1 - \exp(-c_4 \cdot d \ln(N/d))$ for absolute constants $C_4, c_4 > 0$.

Step 4: Conclusion We combine the events defined in Steps 1, 2 and 3 and obtain the existence of an absolute constant $C_6 > 0$ such that

$$\sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \Phi(x)\|_2 \leq \sup_{x \in B_d(0, R)} \|\bar{\nabla} \Phi(x)\|_2 + \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 + \sup_{x \in \mathbb{R}^d} \|\bar{\nabla} \tilde{\Phi}(x)\|_2$$

$$\leq C_6 \cdot \left(\sqrt{dL} + \sqrt{d \cdot \ln(\lambda C_\tau N/d)} \right)$$

with probability at least

$$\begin{aligned} & 1 - \exp(-c_2 \cdot d \ln(N/d)) - \exp(-c_3 \cdot dL \ln(N/d)) - \exp(-d \ln(N/d)) - \exp(-c_4 \cdot d \ln(N/d)) \\ & \geq 1 - 4 \exp(-\tilde{c} \cdot d \ln(N/d)) \end{aligned}$$

with $\tilde{c} := \min\{c_2, c_3, c_4\}$. Since we may assume

$$d \ln(N/d) \geq \frac{2 \ln(4)}{\tilde{c}},$$

we get the claim by picking $c := \tilde{c}/2$. $\square$

H.2 Lower bound

In this section, we prove lower bounds for the Lipschitz constant of random ReLU networks following Assumption 2.1 with general biases. Note that we do not need to impose any assumptions on the distribution of the biases. We start with a lower bound for the Lipschitz constant with respect to the ℓ^p -norm for $p \in [1, 2)$.

Proof of Theorem 6.6. Let $\lambda' > 0$ be such that

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}} \right) \geq 1 - \exp(-d/L).$$

By independence, we may condition on the bias vectors $b^{(0)}, \dots, b^{(L-1)}$ and assume that

$$\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}}$$

is satisfied. Let $C_1, c_1 > 0$ be the constants that are given by Theorem 6.2 and set $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$. Let then

$$M_\Phi := \{x \in \mathbb{R}^d : \Phi \text{ differentiable at } x \text{ with } \bar{\nabla} \Phi(x) = \nabla \Phi(x)\}.$$

Let $\tilde{\Phi}$ be the zero-bias network associated to Φ . Using Theorem 2.8, we get

$$\begin{aligned} \text{Lip}_1(\Phi) &= \sup_{x \in M_\Phi} \|\bar{\nabla} \Phi(x)\|_\infty \geq \sup_{x \in M_\Phi \cap \mathbb{R} S^{d-1}} \|\bar{\nabla} \Phi(x)\|_\infty \\ &\geq \sup_{x \in M_\Phi \cap \mathbb{R} S^{d-1}} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty - \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \\ &= \sup_{x \in \mathbb{S}^{d-1}, x \in R^{-1} M_\Phi} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty - \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2, \end{aligned}$$

where the last step follows due to scale invariance of the formal gradient of a zero-bias ReLU network. Note that for fixed $x_0 \in \mathbb{S}^{d-1}$ we have $Rx_0 \in M_\Phi$ with probability 1 according to [13, Theorem E.1]. Hence, using Theorem F.10, we obtain

$$\sup_{x \in \mathbb{S}^{d-1}, x \in R^{-1} M_\Phi} \|\bar{\nabla} \tilde{\Phi}(x)\|_\infty \geq c_1 \cdot \sqrt{d/L}$$

with probability at least $1 - \exp(-c_1 \cdot d/L)$ with an absolute constant $c_1 > 0$. We call the event defined by that property $\mathcal{E}_1$. Further, we may use Theorem 6.2 and get

$$\sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla} \Phi(x) - \bar{\nabla} \tilde{\Phi}(x)\|_2 \leq C_2 \cdot \frac{dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}}$$

with probability at least $1 - \exp(-c_2 \cdot dL \ln(N/d))$ with absolute constants $C_2, c_2 > 0$. Since by assumption

$$N \geq C \cdot d \cdot L^5 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL)),$$

we may, by taking $C > 0$ large enough, assume that

$$C_2 \cdot \frac{dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}} \leq \frac{c_1}{2} \cdot \frac{\sqrt{d}}{\sqrt{L}}.$$

We call the event defined by that property $\mathcal{E}_2$ and note that on $\mathcal{E}_1 \cap \mathcal{E}_2$ we have

$$\text{Lip}_1(\Phi) \geq \frac{c_1}{2} \cdot \sqrt{d/L}.$$

Until now, we have conditioned on the bias vectors $b^{(0)}, \dots, b^{(L-1)}$. Reintroducing the randomness over these vectors, we obtain

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \exp(-c_1 \cdot d/L) - \exp(-c_2 \cdot dL \ln(N/d)).$$

Letting $\tilde{c} := \min\{c_1, c_2\}$, we get

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 3 \exp(-\tilde{c} \cdot d/L).$$

In the end, the claim follows by letting $c := \tilde{c}/2$ and assuming

$$d \geq \frac{2}{\tilde{c}} \cdot \ln(3) \cdot L. \quad \square$$

Recall that a lower bound for $\text{Lip}_p(\Phi)$ in the case $p \in [2, \infty]$ was established in Corollary 3.6 by studying the gradient at a fixed input $x_0 \in \mathbb{R}^d \setminus \{0\}$. However, Corollary 3.6 requires the biases to be drawn from a symmetric distribution. Using Theorem 6.2 we can now establish such a lower bound for networks with biases drawn from an arbitrary distribution.

Proof of Theorem 6.7. Let $\lambda' > 0$ such that

$$\mathbb{P} \left(\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}} \right) \geq 1 - \exp(-d).$$

By independence, we may condition on the bias vectors $b^{(0)}, \dots, b^{(L-1)}$ and assume that

$$\sup_{\substack{\ell \in \{0, \dots, L-1\} \\ j \in \{1, \dots, N\}}} |b_j^{(\ell)}| \leq \frac{\lambda' \sqrt{2}}{\sqrt{N}}$$

is satisfied. Let $C_1, c_1 > 0$ be the constants that are given by Theorem 6.2 and define the radius $R := 3^L \cdot \lambda' \cdot \frac{N}{dL} \cdot \ln^{-1/2}(N/d)$. Let $x_0 \in \mathbb{R}^d$ with $\|x_0\|_2 \geq R$ be arbitrary. According to Theorem 3.4,

$$\|\nabla \tilde{\Phi}(x_0)\|_1 \geq c_1 \cdot d$$

with probability at least $1 - \exp(-c_1 \cdot d)$ for an absolute constant $c_1 > 0$. Here, $\tilde{\Phi}$ is the zero-bias network associated to Φ . We call the event defined by this property $\mathcal{E}_1$. Since $\|\cdot\|_1 \leq d^{1-1/p'} \cdot \|\cdot\|_{p'}$ for every $p' \in [1, \infty]$, we deduce that on $\mathcal{E}_1$,

$$\|\nabla \tilde{\Phi}(x_0)\|_{p'} \geq c_1 \cdot d^{1/p'} \quad \text{for all } p' \in [1, \infty].$$

We further let $\mathcal{E}_2$ be the event defined by the property

$$\sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\nabla \Phi(x) - \nabla \tilde{\Phi}(x)\|_2 \leq C_2 \cdot \frac{dL^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}}$$

with an absolute constant $C_2 > 0$, which occurs with probability at least $1 - \exp(-c_2 \cdot dL \ln(N/d))$ with an absolute constant $c_2 > 0$ according Theorem 6.2.

For $p \in [2, \infty]$, let $p' \in [1, 2]$ be chosen such that $1/p + 1/p' = 1$. Using Theorem 2.8, on $\mathcal{E}_1 \cap \mathcal{E}_2$ we get for every $p \in [2, \infty]$,

$$\begin{aligned} \text{Lip}_p(\Phi) &\geq \|\bar{\nabla}\Phi(x_0)\|_{p'} \\ &\geq \|\bar{\nabla}\tilde{\Phi}(x_0)\|_{p'} - \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla}\Phi(x) - \bar{\nabla}\tilde{\Phi}(x)\|_{p'} \\ &\geq c_1 \cdot d^{1-1/p} - d^{(1/p')-(1/2)} \cdot \sup_{x \in \mathbb{R}^d, \|x\|_2 \geq R} \|\bar{\nabla}\Phi(x) - \bar{\nabla}\tilde{\Phi}(x)\|_2 \\ &\geq c_1 \cdot d^{1-1/p} - C_2 \cdot d^{3/2-1/p} \cdot \frac{L^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}}. \end{aligned}$$

Since by assumption

$$N \geq C \cdot d \cdot L^4 \cdot \ln^2(N/d) \cdot \ln^2(N/(dL)),$$

we find that

$$C_2 \cdot d^{3/2-1/p} \cdot \frac{L^2 \cdot \ln(N/d) \cdot \ln(N/(dL))}{\sqrt{N}} \leq \frac{c_1}{2} \cdot d^{1-1/p}.$$

Lastly, note that

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \exp(-c_1 \cdot d) - \exp(-c_2 \cdot dL \ln(N/d)) \geq 1 - 2\exp(-c_3 \cdot d)$$

with $c_3 := \min\{c_1, c_2\}$. The claim follows by reintroducing the randomness over the bias vectors $b^{(0)}, \dots, b^{(L-1)}$, since we may assume $\ln(3) \leq \frac{c_3}{2} \cdot d$ and by letting $c := c_3/2$. $\square$

I Other auxiliary results

In this appendix, we provide some additional auxiliary results that are used over the course of the present paper. First, we show that the small-ball property in Assumption 2.3 is indeed equivalent to each bias having an absolutely continuous distribution with probability density function bounded by half the small-ball constant.

Theorem I.1. *Let X be a real-valued random variable and $T > 0$. Then the following are equivalent.*

- (1) *X has an absolutely continuous distribution with probability density function bounded by $T/2$ almost everywhere.*
- (2) *For every $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathbb{P}(X \in (t - \varepsilon, t + \varepsilon)) \leq T \cdot \varepsilon.$$

Proof. The direction “(1) $\Rightarrow$ (2)” is trivial. Hence, it remains to show the opposite direction. As a first step, let $A \subseteq \mathbb{R}$ with $\lambda(A) = 0$ and $\varepsilon > 0$. According to [12, Lemma 1.17], there exist $((a_j, b_j))_{j=1}^\infty$ with

$$A \subseteq \bigcup_{j=1}^\infty (a_j, b_j) \quad \text{and} \quad \sum_{j=1}^\infty (b_j - a_j) \leq \frac{2\varepsilon}{T}.$$

This implies

$$\mathbb{P}(X \in A) \leq \sum_{j=1}^\infty \mathbb{P}(X \in (a_j, b_j)) \leq \frac{T}{2} \cdot \sum_{j=1}^\infty (b_j - a_j) \leq \varepsilon.$$

Hence, since $\varepsilon > 0$ was arbitrary,

$$\mathbb{P}(X \in A) = 0.$$

According to the Radon-Nikodym theorem [12, Theorem 3.8] there exists a measurable and non-negative probability density function $f : \mathbb{R} \rightarrow \mathbb{R}$ of X . According to the Lebesgue differentiation

theorem [12, Theorem 3.21], for almost every $x \in \mathbb{R}$,

$$f(x) = \lim_{\varepsilon \rightarrow 0} \frac{1}{2\varepsilon} \int_{(x-\varepsilon, x+\varepsilon)} f(y) \, dy.$$

This implies the claim, since for every $x \in \mathbb{R}$ and $\varepsilon > 0$,

$$\frac{1}{2\varepsilon} \int_{(x-\varepsilon, x+\varepsilon)} f(y) \, dy = \frac{1}{2\varepsilon} \cdot \mathbb{P}(X \in (t - \varepsilon, t + \varepsilon)) \leq \frac{1}{2\varepsilon} \cdot T \cdot \varepsilon = T/2. \quad \square$$

The next lemma provides a bound on the Euclidean distance of two normalized vectors in terms of the original vectors.

Lemma I.2. *Let $x, y \in \mathbb{R}^d \setminus \{0\}$. Then*

$$\left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 \leq \frac{2}{\max\{\|x\|_2, \|y\|_2\}} \|x - y\|_2.$$

Proof. Without loss of generality, we assume $\|x\|_2 \geq \|y\|_2$. We then get

$$\begin{aligned} \left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 &\leq \left\| \frac{x}{\|x\|_2} - \frac{y}{\|x\|_2} \right\|_2 + \left\| \frac{y}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 \\ &= \frac{1}{\|x\|_2} \cdot \|x - y\|_2 + \|y\|_2 \cdot \left| \frac{1}{\|x\|_2} - \frac{1}{\|y\|_2} \right| \\ &= \frac{1}{\|x\|_2} \cdot \|x - y\|_2 + \|y\|_2 \cdot \frac{|\|y\|_2 - \|x\|_2|}{\|x\|_2 \|y\|_2} \\ &\leq \frac{2}{\|x\|_2} \cdot \|x - y\|_2, \end{aligned}$$

as was to be shown. $\square$

For the sake of completeness, we state and prove the following elementary lemma, which allows to pull constants outside of a logarithm.

Lemma I.3. *Let $x \geq e$ and $C \geq 1$. Then,*

$$\ln(C \cdot x) \leq C \cdot \ln(x).$$

Proof. Using the well-known estimate $\ln(C) \leq C - 1$ for $C > 0$, we compute

$$\ln(C \cdot x) = \ln(C) + \ln(x) \leq C - 1 + \ln(x) \stackrel{x \geq e}{\leq} (C - 1) \ln(x) + \ln(x) = C \cdot \ln(x). \quad \square$$

Next, we provide a useful integral bound which occurs often when bounding the Gaussian width of a set using Dudley's inequality.

Lemma I.4. *Let $C, \alpha > 0$ with $C/\alpha \geq e$. Then,*

$$\int_0^\alpha \sqrt{\ln(C/\gamma)} \, d\gamma \leq 2 \cdot \alpha \cdot \sqrt{\ln(C/\alpha)}.$$

Proof. We first substitute $\sigma = \ln(C/\gamma)$ and get

$$\int_0^\alpha \sqrt{\ln(C/\gamma)} \, d\gamma = -C \cdot \int_\infty^{\ln(C/\alpha)} \sqrt{\sigma} \cdot \exp(-\sigma) \, d\sigma = C \cdot \int_{\ln(C/\alpha)}^\infty \sqrt{\sigma} \cdot \exp(-\sigma) \, d\sigma.$$

Using integration by parts, we further get

$$\begin{aligned} \int_{\ln(C/\alpha)}^\infty \sqrt{\sigma} \cdot \exp(-\sigma) \, d\sigma &= [-\sqrt{\sigma} \cdot \exp(-\sigma)]_{\ln(C/\alpha)}^\infty + \frac{1}{2} \cdot \int_{\ln(C/\alpha)}^\infty \frac{\exp(-\sigma)}{\sqrt{\sigma}} \, d\sigma \\ &= \frac{\alpha}{C} \cdot \ln^{1/2}(C/\alpha) + \frac{1}{2} \cdot \int_{\ln(C/\alpha)}^\infty \frac{\exp(-\sigma)}{\sqrt{\sigma}} \, d\sigma. \end{aligned}$$

We study the last integrand further. Using the substitution $\sigma = u^2/2$, we obtain

$$\begin{aligned} \int_{\ln(C/\alpha)}^{\infty} \frac{\exp(-\sigma)}{\sqrt{\sigma}} d\sigma &= \sqrt{2} \cdot \int_{\sqrt{2\ln(C/\alpha)}}^{\infty} \exp(-u^2/2) du \\ &= \sqrt{2} \cdot \sqrt{2\pi} \cdot \mathbb{P}_{g \sim \mathcal{N}(0,1)}(g \geq \sqrt{2\ln(C/\alpha)}) \\ &\stackrel{[20, \text{Prop. 2.1.2}]}{\leq} \frac{\sqrt{2}}{\sqrt{2\ln(C/\alpha)}} \cdot \frac{\alpha}{C} \stackrel{C/\alpha \geq e}{\leq} \frac{\alpha}{C}. \end{aligned}$$

This proves the claim. $\square$

Next, we provide an auxiliary result about the spectral norm of a Gaussian matrix which is multiplied from the left and the right with (deterministic) matrices.

Lemma I.5. *There exist absolute constants $C, c > 0$ such that the following holds. For $k, \ell, m, n \in \mathbb{N}$, let $A_1 \in \mathbb{R}^{k \times \ell}$, $A_2 \in \mathbb{R}^{m \times n}$ be fixed matrices and $W \in \mathbb{R}^{\ell \times m}$ be a random matrix with $W_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Then, for arbitrary $t \geq C$,*

$$\|A_1 W A_2\|_{2 \rightarrow 2} \leq \|A_1\|_{2 \rightarrow 2} \cdot \|A_2\|_{2 \rightarrow 2} \cdot \left(\sqrt{\text{rank}(A_1)} + \sqrt{\text{rank}(A_2)} + t \right)$$

with probability at least $1 - \exp(-c \cdot t^2)$.

Proof. Let $A_1 = U_1 \Sigma_1 V_1^T$ be the singular value decomposition of A_1 , where $U_1 \in \mathbb{R}^{k \times k}$ and $V_1 \in \mathbb{R}^{\ell \times \ell}$ are orthogonal matrices and $\Sigma_1 \in \mathbb{R}^{k \times \ell}$ contains the singular values of A_1 . Moreover, let $A_2 = U_2 \Sigma_2 V_2^T$ be the singular decomposition of A_2 with orthogonal matrices $U_2 \in \mathbb{R}^{m \times m}$, $V_2 \in \mathbb{R}^{n \times n}$ and $\Sigma_2 \in \mathbb{R}^{m \times n}$ containing the singular values of A_2 . We then get by the rotational invariance of the normal distribution and of the spectral norm that

$$\|A_1 W A_2\|_{2 \rightarrow 2} = \|U_1 \Sigma_1 V_1^T W U_2 \Sigma_2 V_2^T\|_{2 \rightarrow 2} \stackrel{d}{=} \|\Sigma_1 W \Sigma_2\|_{2 \rightarrow 2}.$$

Let $\alpha := \text{rank}(A_1)$, $\beta := \text{rank}(A_2)$ and let $\sigma_1 \geq \dots \geq \sigma_\alpha > 0$ and $\sigma'_1 \geq \dots \geq \sigma'_\beta > 0$ be the singular values of A_1 and A_2 , respectively. We set $\widetilde{\Sigma}_1 := \text{diag}(\sigma_1, \dots, \sigma_\alpha) \in \mathbb{R}^{\alpha \times \alpha}$ and, moreover, $\widetilde{\Sigma}_2 := \text{diag}(\sigma'_1, \dots, \sigma'_\beta) \in \mathbb{R}^{\beta \times \beta}$. We then get $\Sigma_1 = I_{k \times \alpha} \widetilde{\Sigma}_1 I_{\alpha \times \ell}$ and $\Sigma_2 = I_{m \times \beta} \widetilde{\Sigma}_2 I_{\beta \times n}$. Hence, we see

$$\begin{aligned} \|\Sigma_1 W \Sigma_2\|_{2 \rightarrow 2} &= \left\| I_{k \times \alpha} \widetilde{\Sigma}_1 I_{\alpha \times \ell} W I_{m \times \beta} \widetilde{\Sigma}_2 I_{\beta \times n} \right\|_{2 \rightarrow 2} \leq \|I_{\alpha \times \ell} W I_{m \times \beta}\|_{2 \rightarrow 2} \cdot \sigma_1 \cdot \sigma'_1 \\ &= \|I_{\alpha \times \ell} W I_{m \times \beta}\|_{2 \rightarrow 2} \cdot \|A_1\|_{2 \rightarrow 2} \cdot \|A_2\|_{2 \rightarrow 2} \stackrel{d}{=} \|W'\|_{2 \rightarrow 2} \cdot \|A_1\|_{2 \rightarrow 2} \cdot \|A_2\|_{2 \rightarrow 2} \end{aligned}$$

where $W' \in \mathbb{R}^{\alpha \times \beta}$ has independent $\mathcal{N}(0, 1)$ -entries. Hence, from [20, Corollary 7.3.3] we see

$$\|W'\|_{2 \rightarrow 2} \leq \sqrt{\text{rank}(A_1)} + \sqrt{\text{rank}(A_2)} + t$$

with probability at least $1 - 2 \exp(-\tilde{c} \cdot t^2) = 1 - \exp(\ln(2) - \tilde{c} \cdot t^2)$ with an absolute constant $\tilde{c} > 0$. By taking C sufficiently large, we have

$$\ln(2) \leq \frac{\tilde{c}}{2} \cdot t^2 \quad \text{for } t \geq C,$$

so that the claim follows letting $c := \tilde{c}/2$. $\square$

In the following lemma, we bound the cardinality of the set of $(N \times N)$ -diagonal matrices with entries on the diagonal either 0, 1, or -1 and where the number of non-zero entries on the diagonal is bounded by $\delta \cdot N$.

Lemma I.6. *For given $\delta \in (0, 1)$ and $N \in \mathbb{N}$, let*

$$\mathcal{A} := \{A \in \text{diag}\{0, 1, -1\}^{N \times N} : \text{Tr } |A| \leq \delta \cdot N\}.$$

Then,

$$\#\mathcal{A} \leq \left(\frac{4e}{\delta} \right)^{\delta \cdot N}.$$

Proof. In the case $\delta N < 1$, the set $\mathcal{A}$ only consists of the zero matrix, so that the claim is satisfied since $4e/\delta \geq 1$. Hence, we may assume $\delta N \geq 1$.

Note that the cardinality of $\mathcal{A}$ can be upper bounded by the number of subsets of $\{1, \dots, N\}$ with cardinality at most $\lfloor \delta \cdot N \rfloor$, multiplied by $2^{\lfloor \delta \cdot N \rfloor}$, since we can choose between 1 and -1 for each diagonal entry that is nonzero. Therefore, using [20, Ex. 0.0.5], we observe

$$\begin{aligned} \#\mathcal{A} &\leq 2^{\lfloor \delta \cdot N \rfloor} \cdot \sum_{k=0}^{\lfloor \delta N \rfloor} \binom{N}{k} \leq 2^{\lfloor \delta \cdot N \rfloor} \cdot \left(\frac{eN}{\lfloor \delta \cdot N \rfloor} \right)^{\lfloor \delta \cdot N \rfloor} \\ &\stackrel{|x| \geq x/2 \text{ for } x \geq 1}{\leq} 2^{\delta \cdot N} \cdot \left(\frac{2eN}{\delta \cdot N} \right)^{\delta \cdot N} = \left(\frac{4e}{\delta} \right)^{\delta \cdot N}. \end{aligned} \quad \square$$

Next, we show a lower bound on the covering number of the Euclidean sphere $\mathbb{S}^{n-1}$. This statement is probably well-known, but since we could not locate a convenient reference, we include the short proof.

Lemma I.7. *For any $\varepsilon \in (0, 1)$,*

$$\mathcal{N}(\mathbb{S}^{n-1}, \varepsilon) \geq \frac{1}{6} \cdot \left(\frac{1}{2\varepsilon} \right)^{n-1}.$$

Proof. Let Z_1 be an ε -net of $\mathbb{S}^{n-1}$ and Z_2 be an ε -net of $[-1, 1]$ with $|Z_2| \leq \frac{3}{\varepsilon}$. Z_2 exists due to [20, Corollary 4.2.13]. We define the surjective mapping

$$\varphi : [-1, 1] \times \mathbb{S}^{n-1} \rightarrow \overline{B}_n(0, 1), \quad (t, x) \mapsto tx$$

and set

$$Z_3 := \{\varphi(t, x) : x \in Z_1, t \in Z_2\}.$$

Let $y_0 \in \overline{B}_n(0, 1)$ be arbitrary and pick $t_0 \in [-1, 1], x_0 \in \mathbb{S}^{n-1}$ with $\varphi(t_0, x_0) = y_0$. Moreover, pick $\tilde{t} \in Z_2, \tilde{x} \in Z_1$ with $|\tilde{t} - t_0|, \|\tilde{x} - x_0\|_2 \leq \varepsilon$. Then,

$$\|\varphi(\tilde{t}, \tilde{x}) - \varphi(t_0, x_0)\|_2 = \|\tilde{t}\tilde{x} - t_0x_0\|_2 \leq \|\tilde{t}\tilde{x} - \tilde{t}x_0\|_2 + \|\tilde{t}x_0 - t_0x_0\|_2 \leq \|\tilde{x} - x_0\|_2 + |\tilde{t} - t_0| \leq 2\varepsilon.$$

Hence, Z_3 is a (2ε) -net of $\overline{B}_n(0, 1)$. By [20, Corollary 4.2.13], we get $|Z_3| \geq \left(\frac{1}{2\varepsilon}\right)^n$. On the other hand, by the definition of Z_3 ,

$$|Z_3| \leq |Z_1| \cdot |Z_2| \leq |Z_1| \cdot \frac{3}{\varepsilon}.$$

Altogether, we get

$$|Z_1| \geq \frac{\varepsilon}{3} \cdot \left(\frac{1}{2\varepsilon} \right)^n = \frac{1}{6} \cdot \left(\frac{1}{2\varepsilon} \right)^{n-1}.$$

Since Z_1 was an arbitrary ε -net of $\mathbb{S}^{n-1}$, the claim is shown. $\square$

Lemma I.7 immediately implies that

$$\mathcal{N}\left(\frac{1}{\sqrt{2}}\mathbb{S}^{n-1}, \varepsilon\right) \geq \frac{1}{6} \cdot \left(\frac{1}{2\sqrt{2} \cdot \varepsilon} \right)^{n-1} \quad (\text{I.1})$$

for every $\varepsilon \in (0, 1/\sqrt{2})$, which follows since if $\mathcal{N}$ is an ε -net of $\frac{1}{\sqrt{2}}\mathbb{S}^{n-1}$ then $\sqrt{2}\mathcal{N}$ is an $\sqrt{2}\varepsilon$ -net of $\mathbb{S}^{n-1}$. For the proof of Theorem 5.1, we in fact need to lower bound the covering numbers of intersections of the unit sphere with Euclidean balls. Such a lower bound is provided by the following lemma.

Lemma I.8. *Let e_1 be the first standard basis vector in $\mathbb{R}^n$ and let $\varepsilon, \delta \in (0, 1/\sqrt{2})$. Then there exists an ε -packing P of $\overline{B}_n\left(\frac{1}{\sqrt{2}} \cdot e_1, \delta\right) \cap \frac{1}{\sqrt{2}}\mathbb{S}^{n-1}$ with*

$$|P| \geq \frac{\sqrt{2} \cdot \delta}{18} \cdot \left(\frac{\delta}{6\varepsilon} \right)^{n-1}.$$

Proof. According to [20, Corollary 4.2.13] we can pick points $\tilde{x}_1, \dots, \tilde{x}_M \in \mathbb{S}^{n-1}$ with

$$\mathbb{S}^{n-1} \subseteq \bigcup_{i=1}^M (\bar{B}_n(\tilde{x}_i, \sqrt{2}\delta) \cap \mathbb{S}^{n-1})$$

and $M \leq \left(\frac{3}{\sqrt{2}\delta}\right)^n$. Letting $x_i := \frac{\tilde{x}_i}{\sqrt{2}}$ we get

$$\frac{1}{\sqrt{2}}\mathbb{S}^{n-1} \subseteq \bigcup_{i=1}^M \left(\bar{B}_n(x_i, \delta) \cap \frac{1}{\sqrt{2}}\mathbb{S}^{n-1} \right).$$

Let Z_i be an ε -net of $\bar{B}_n(x_i, \delta) \cap \frac{1}{\sqrt{2}}\mathbb{S}^{n-1}$. We let $K_i := |Z_i|$ for every i . It follows immediately that $\bigcup_{i=1}^M Z_i$ is an ε -net of $\frac{1}{\sqrt{2}}\mathbb{S}^{n-1}$. According to (I.1) this implies

$$\max_i K_i \cdot \left(\frac{3}{\sqrt{2}\delta}\right)^n \geq \max_i K_i \cdot M \geq \left| \bigcup_{i=1}^M Z_i \right| \geq \frac{1}{6} \cdot \left(\frac{1}{2\sqrt{2}\varepsilon}\right)^{n-1}.$$

Rearranging this yields

$$\max_i K_i \geq \frac{1}{6} \cdot \left(\frac{\sqrt{2}\delta}{3}\right)^n \cdot \left(\frac{1}{2\sqrt{2}\varepsilon}\right)^{n-1} = \frac{\sqrt{2} \cdot \delta}{18} \cdot \left(\frac{\delta}{6\varepsilon}\right)^{n-1}.$$

The claim then follows from the equivalence of covering and packing numbers [20, Lemma 4.2.8] and due to rotation invariance. $\square$

B.3. Optimal approximation using complex-valued neural networks

Optimal approximation using complex-valued neural networks

Paul Geuchen
MIDS,
KU Eichstätt-Ingolstadt,
Auf der Schanz 49,
85049 Ingolstadt, Germany
paul.geuchen@ku.de

Felix Voigtlaender
MIDS,
KU Eichstätt-Ingolstadt,
Auf der Schanz 49,
85049 Ingolstadt, Germany
felix.voigtlaender@ku.de

Abstract

Complex-valued neural networks (CVNNs) have recently shown promising empirical success, for instance for increasing the stability of recurrent neural networks and for improving the performance in tasks with complex-valued inputs, such as in MRI fingerprinting. While the overwhelming success of Deep Learning in the real-valued case is supported by a growing mathematical foundation, such a foundation is still largely lacking in the complex-valued case. We thus analyze the expressivity of CVNNs by studying their approximation properties. Our results yield the first quantitative approximation bounds for CVNNs that apply to a wide class of activation functions including the popular modReLU and complex cardioid activation functions. Precisely, our results apply to any activation function that is smooth but not polyharmonic on some non-empty open set; this is the natural generalization of the class of smooth and non-polynomial activation functions to the complex setting. Our main result shows that the error for the approximation of C^k -functions scales as $m^{-k/(2n)}$ for $m \rightarrow \infty$ where m is the number of neurons, k the smoothness of the target function and n is the (complex) input dimension. Under a natural continuity assumption, we show that this rate is optimal; we further discuss the optimality when dropping this assumption. Moreover, we prove that the problem of approximating C^k -functions using continuous approximation methods unavoidably suffers from the curse of dimensionality.

1 Introduction

Deep Learning currently predominantly relies on real-valued neural networks, which have led to breakthroughs in fields like image classification or speech recognition [23, 27, 43]. However, recent work has uncovered several application areas in which the use of complex-valued neural networks (CVNNs) leads to better results than the use of their real-valued counterparts. These application areas mainly include tasks where complex numbers inherently occur as inputs of a machine learning model such as Magnetic Resonance Imaging (MRI) [16, 38, 44] and Polarimetric Synthetic Aperture Radar (PolSAR) Imaging [8, 9, 48]. Moreover, CVNNs have been used to improve the stability of recurrent neural networks [5] and have been successfully applied in various other fields [32, 37]. The mathematical theory of these complex-valued neural networks, however, is still in its infancy. There is therefore a great interest in studying CVNNs and in particular in uncovering the differences and commonalities between CVNNs and their real-valued counterparts. A prominent example highlighting the unexpected differences between both network classes is the *universal approximation theorem* for neural networks, whose most general real-valued version was proven in 1993 [28] (with a more restricted version appearing earlier [17]) and which was recently generalized to the case of CVNNs [45]. The two theorems describe necessary and sufficient conditions on

	Condition on activation function	Continuity of weight selection	Approximation Error
Theorem 3.2	smooth & non-polyharmonic	possible	$\mathcal{O}(m^{-k/(2n)})$
Consequence of Theorem 4.1	continuous	assumed	$\Omega(m^{-k/(2n)})$
Theorem 4.2	very special activation function	not assumed	$\mathcal{O}(m^{-k/(2n-1)})$
Theorem 4.3	$\frac{1}{1 + e^{-\operatorname{Re}(z)}}$	not assumed	$\tilde{\Omega}(m^{-k/(2n)})$

Table 1: Overview of the proven approximation bounds. k is the regularity of the approximated functions (which are assumed to be C^k), n the (complex) input dimension and m the number of neurons in the hidden layer of the network. The notation $\tilde{\Omega}$ is similar to Ω , but ignoring log factors.

an activation function which guarantee that arbitrarily wide neural networks of a fixed depth can approximate any continuous function on any compact set arbitrarily well (with respect to the uniform norm). Already here it was shown that complex-valued networks behave significantly different from real-valued networks: While real-valued networks are universal if and only if the activation function is non-polynomial, complex-valued networks with a single hidden layer are universal if and only if the activation function is non-polyharmonic (see below for a definition). Furthermore, there exist continuous activation functions for which deep CVNNs are universal but shallow CVNNs are not, whereas the same cannot happen for real-valued neural networks. This example shows that it is highly relevant to study the properties of CVNNs and to examine which of the fundamental properties of real-valued networks extend to complex-valued networks.

Essentially the only existing *quantitative* result regarding the approximation-theoretical properties of CVNNs is [14], which provides results for approximating C^k -functions by *deep* CVNNs using the modReLU activation function. However, for real-valued NNs it is known that already *shallow* NNs can approximate C^k -functions at the optimal rate. Precisely, Mhaskar showed in [33] that one can approximate C^k -functions on $[-1, 1]^n$ with an error of order $m^{-k/n}$ as $m \rightarrow \infty$, where m is the number of neurons in the hidden layer. Here he assumed that the activation function is smooth on an open interval and that at some point of this interval no derivative vanishes. This is equivalent to the activation function being smooth and non-polynomial on that interval, cf. [20, p. 53].

The present paper shows that a comparable result holds in the setting of complex-valued networks, by proving that one can approximate every function in $C^k(\Omega_n; \mathbb{C})$ (where differentiability is understood in the sense of real variables) with an error of the order $m^{-k/(2n)}$ (as $m \rightarrow \infty$) using shallow complex-valued neural networks with m neurons in the hidden layer. Here we define the cube $\Omega_n := [-1, 1]^n + i[-1, 1]^n$. The result holds whenever the activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ is smooth and non-polyharmonic on some non-empty open set. This is a very natural condition, since for polyharmonic activation functions there exist C^k -functions that cannot be approximated at all below some error threshold using shallow neural networks with this activation function [45].

Furthermore, the present paper studies in how far the approximation order of $m^{-k/(2n)}$ is *optimal*, meaning that an order of $m^{-(k/(2n))-\alpha}$ (where $\alpha > 0$) cannot be achieved. Here it turns out that the derived order of approximation is indeed optimal (even in the class of CVNNs with possibly more than one hidden layer) in the setting that the weight selection is *continuous*, meaning that the map that assigns to a function $f \in C^k(\Omega_n; \mathbb{C})$ the weights of the approximating network is continuous with respect to some norm on $C^k(\Omega_n; \mathbb{C})$. This continuity assumption is natural since typical learning algorithms such as (stochastic) gradient descent use samples $f(x_j)$ of the target function and then apply continuous operations to them to update the network weights.

We investigate this optimality result further by dropping the continuity assumption and constructing two special smooth and non-polyharmonic activation functions with the first one having the property that the order of approximation can indeed be strictly improved via a *discontinuous* selection of

the related weights. For the second activation function we show that the order of $m^{-k/(2n)}$ *cannot* be improved, even if one allows a discontinuous weight selection. This in particular shows that in the given generality of arbitrary smooth, non-polyharmonic activation functions, the upper bound $\mathcal{O}(m^{-k/(2n)})$ cannot be improved, even for a possibly discontinuous choice of the weights. An overview of the approximation bounds proven in this paper can be found in Table 1.

Moreover, we analyze the *tractability* (in terms of the input dimension n) of the considered problem of approximating C^k -functions using neural networks. Theorem 5.1 shows that one necessarily needs a number of parameters *exponential* in n to obtain a non-trivial approximation error. To the best of our knowledge, Theorem 5.1 is the first result showing that the problem of approximating C^k -functions using continuous approximation methods is intractable (in terms of the input dimension n).

1.1 Related Work

Real-valued neural networks. By now, the approximation properties of real-valued neural networks are quite well-studied (cf. [10, 28, 33, 40, 41, 46, 47] and the references therein). We here only discuss a few papers that are most closely related to the present work.

In [33], Mhaskar analyzes the rate of approximation of shallow real-valued neural networks for target functions of regularity C^k . Our results can be seen as the generalization of [33] to the complex setting. Our proofs rely on several techniques from [33]; however, significant modifications are required to make the proofs work for general smooth non-polyharmonic functions.

One of the first papers to observe that neural networks with general (smooth) activation function can be surprisingly expressive is [31] where it was shown that a neural network of *constant size* can be universal. One of the activation functions in Section 4 is based on a similar idea.

The importance of distinguishing between continuous and discontinuous weight selection (which in our setting is discussed in Section 4) was observed for ReLU-networks in [47].

The paper [25] shows that neural network approximation is *not* continuous in the following sense: The best approximating neural network $\Phi(f)$ of a given size does not depend continuously on $f \in C^k$. This result, however, is not in conflict with our results: We want to assign to any C^k -function f a network $\tilde{\Phi}(f)$ that approximates f below the error threshold $m^{-k/(2n)}$. The network $\tilde{\Phi}(f)$, however, does *not* have to coincide with the best approximating network $\Phi(f)$.

Complex-valued neural networks. When it comes to general literature about mathematical properties of complex-valued neural networks, surprisingly little work can be found. The *Universal Approximation Theorem for Complex-Valued Neural Networks* [45] has already been mentioned above. In particular, it has been shown that shallow CVNNs are universal if and only if the activation function ϕ is not polyharmonic. Thus, the condition assumed in the present paper (that ϕ should be smooth, but not polyharmonic) is quite natural.

Regarding *quantitative* approximation results for CVNNs, the only existing work of which we are aware is [14], analyzing the approximation capabilities of *deep* CVNNs where the modReLU is used as activation function. Since the modReLU satisfies our condition regarding the activation function, the present work can be seen as an improvement to [14]. Precisely, (i) we consider general activation functions, including but not limited to the modReLU, (ii) we improve the order of approximation by a log factor, and (iii) we show that this order of approximation can be achieved using shallow networks instead of the deep networks used in [14]. We stress that our proof techniques differ significantly from the ones applied in [14]: The arguments in [14] take their main ideas from [46] making heavy use of the specific definition of the modReLU. In contrast, since we consider quite general activation functions, we necessarily follow a much more general approach following the ideas from [33].

2 Preliminaries

Shallow complex-valued neural networks. In this paper we mainly consider so-called shallow complex-valued neural networks, meaning complex-valued neural networks with a single hidden layer. Precisely, we consider functions of the form

$$\mathbb{C}^n \ni z \mapsto \sum_{j=1}^m \sigma_j \phi(\rho_j^T \cdot z + \eta_j) \in \mathbb{C},$$

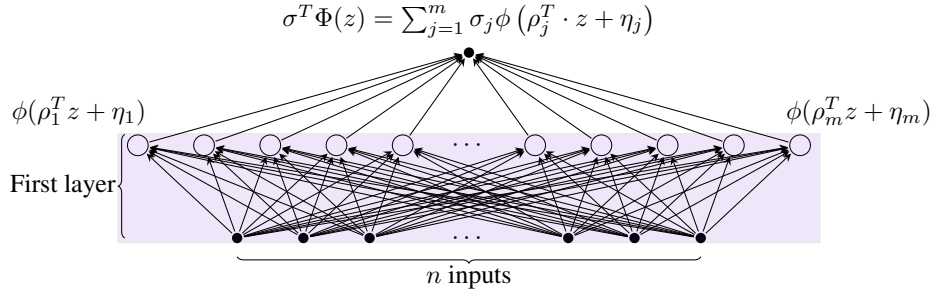


Figure 1: Graphical representation of a shallow neural network. Input and output neurons are depicted as dots, hidden neurons are depicted as circles. The term *first layer* describes the transformation from the input to the hidden neurons, including the application of the activation function.

with $\rho_1, \dots, \rho_m \in \mathbb{C}^n$, $\sigma_1, \dots, \sigma_m, \eta_1, \dots, \eta_m \in \mathbb{C}$ and an *activation function* $\phi : \mathbb{C} \rightarrow \mathbb{C}$. Here, $m \in \mathbb{N}$ denotes the number of neurons of the network and we write v^T for the transpose of a vector v .

To simplify the formulation of the results, we introduce the following notation: We write $\mathcal{F}_{n,m}^\phi$ for the set of *first layers* of shallow complex-valued neural networks with activation function ϕ , with n input neurons and m hidden neurons, meaning

$$\mathcal{F}_{n,m}^\phi := \left\{ z \mapsto \left(\phi \left(\rho_j^T \cdot z + \eta_j \right) \right)_{j=1}^m : \rho_j \in \mathbb{C}^n, \eta_j \in \mathbb{C} \right\} \subseteq \{F : \mathbb{C}^n \rightarrow \mathbb{C}^m\}.$$

Hence, each shallow CVNN can be written as $\sigma^T \Phi$ with $\sigma \in \mathbb{C}^m$ and $\Phi \in \mathcal{F}_{n,m}^\phi$; see Figure 1 for a graphical representation of a shallow CVNN.

Approximation. The paper aims to analyze the approximation of C^k -functions on the complex cube

$$\Omega_n := [-1, 1]^n + i[-1, 1]^n$$

using shallow CVNNs. Here, we say that a function $f : \Omega_n \rightarrow \mathbb{C}$ is in $C^k(\Omega_n; \mathbb{C})$ if and only if f is k times continuously differentiable on Ω_n , where the derivative is to be understood in the sense of real variables, i.e., in the sense of interpreting f as a function $[-1, 1]^{2n} \rightarrow \mathbb{R}^2$ and taking usual real derivatives. We further define the C^k -norm of a function $f \in C^k(\Omega_n; \mathbb{C})$ as

$$\|f\|_{C^k(\Omega_n; \mathbb{C})} := \sup_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ |\mathbf{k}| \leq k}} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega_n; \mathbb{C})}, \quad \text{where} \quad \|g\|_{L^\infty(\Omega_n; \mathbb{C})} := \sup_{z \in \Omega_n} |g(z)| \quad (2.1)$$

for any function $g : \Omega_n \rightarrow \mathbb{C}$. Note that we write $\mathbb{N} = \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$. Using the previously introduced notation, we thus seek to bound the worst-case approximation error, i.e.,

$$\sup_{\substack{f \in C^k(\Omega_n; \mathbb{C}) \\ \|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1}} \inf_{\substack{\Phi \in \mathcal{F}_{n,m}^\phi \\ \sigma \in \mathbb{C}^m}} \|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})}.$$

Wirtinger calculus and polyharmonic functions. For a function $f : \mathbb{C} \rightarrow \mathbb{C}$ which is differentiable in the sense of real variables at a point $z_0 \in \mathbb{C}$ we define its *Wirtinger derivatives* at z_0 as

$$\partial_{\text{wirt}} f(z_0) := \frac{1}{2} \left(\frac{\partial f}{\partial x}(z_0) - i \cdot \frac{\partial f}{\partial y}(z_0) \right) \quad \text{and} \quad \bar{\partial}_{\text{wirt}} f(z_0) := \frac{1}{2} \left(\frac{\partial f}{\partial x}(z_0) + i \cdot \frac{\partial f}{\partial y}(z_0) \right).$$

Here, $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ denote the usual partial derivatives in the sense of real variables. We extend this definition to multivariate functions defined on open subsets of $\mathbb{C}^n$ by considering *coordinatewise* Wirtinger derivatives.

A function $f : U \rightarrow \mathbb{C}$, where $U \subseteq \mathbb{C}$ is an open set, is called *smooth* if it is differentiable arbitrarily many times (in the sense of real variables). We write $f \in C^\infty(U; \mathbb{C})$ in that case. Moreover, f is called *polyharmonic* (on U) if it is smooth and if there exists $m \in \mathbb{N}_0$ satisfying

$$\Delta^m f \equiv 0 \quad \text{on } U.$$

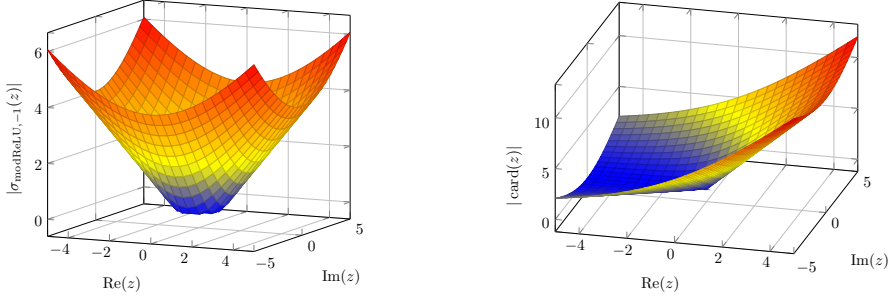


Figure 2: Absolute value of the activation functions $\sigma_{\text{modReLU}, -1}$ (left) and card (right).

Here, $\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = 4\partial_{\text{wirt}}\bar{\partial}_{\text{wirt}}$ denotes the usual Laplace-Operator.

The following Proposition 2.1 describes a property of non-polyharmonic functions which is crucial for proving the approximation results of this paper.

Proposition 2.1. *Let $\emptyset \neq U \subseteq \mathbb{C}$ be an open set and let $\phi \in C^\infty(U; \mathbb{C})$ be non-polyharmonic. Then for every $M \in \mathbb{N}_0$ there exists a point $z_M \in U$ satisfying*

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \phi(z_M) \neq 0 \quad \text{for all } m, \ell \in \mathbb{N}_0 \text{ with } m, \ell \leq M.$$

The proof of Proposition 2.1 is an application of the Baire category theorem; see Appendix B.2.

Important complex activation functions. We briefly discuss in how far two commonly used complex activation functions satisfy our assumptions: The *modReLU* proposed in [5] and the *complex cardioid* used in [44] for MRI fingerprinting where the performance could be significantly improved using complex-valued neural networks. The modReLU is defined as

$$\sigma_{\text{modReLU}, b} : \mathbb{C} \rightarrow \mathbb{C}, \quad \sigma_{\text{modReLU}, b}(z) := \begin{cases} (|z| + b) \frac{z}{|z|}, & \text{if } |z| + b \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $b < 0$ is a fixed parameter. The complex cardioid is defined as

$$\text{card} : \mathbb{C} \rightarrow \mathbb{C}, \quad \text{card}(z) := \frac{1}{2}(1 + \cos(\angle z))z.$$

Here, $\angle z = \theta \in \mathbb{R}$ denotes the polar angle of a complex number $z = re^{i\theta}$, where we define $\angle 0 := 0$; see Figure 2 for plots of the absolute value of the two functions.

Both functions are smooth and non-polyharmonic on a non-empty open subset of $\mathbb{C}$, which is proven in Appendix A.2. Furthermore, they are both continuous on $\mathbb{C}$. Therefore, our approximation bounds established in Theorems 3.1 and 3.2 in particular apply to those two functions.

3 Main results

In this section we state the main results of this paper and provide proof sketches for them. Detailed proofs of the two statements can be found in Appendices B.3 and C.3.

First we show in Theorem 3.1 that it is possible to approximate any complex polynomial in z and $\bar{z}$ arbitrarily well using shallow CVNNs with the size of the networks only depending on the degree of the polynomial (not on the desired approximation accuracy). Using this result we can prove the main approximation bound, Theorem 3.2, by first approximating a given C^k -function using a polynomial in z and $\bar{z}$ and then approximating this polynomial using Theorem 3.1.

For $m, n \in \mathbb{N}$ let

$$\mathcal{P}_m^n := \left\{ \mathbb{C}^n \rightarrow \mathbb{C}, z \mapsto \sum_{\mathbf{m} \leq m} \sum_{\ell \leq n} a_{\mathbf{m}, \ell} z^{\mathbf{m}} \bar{z}^\ell : a_{\mathbf{m}, \ell} \in \mathbb{C} \right\}$$

denote the space of all complex polynomials on $\mathbb{C}^n$ in z and $\bar{z}$ of componentwise degree at most m . Here, we are summing over all multi-indices $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $\mathbf{m}_j, \ell_j \leq m$ for every $j \in \{1, \dots, n\}$ and use the notation

$$z^{\mathbf{m}} := \prod_{j=1}^n z_j^{\mathbf{m}_j} \quad \text{and} \quad \bar{z}^{\ell} := \prod_{j=1}^n \bar{z}_j^{\ell_j}.$$

The space $\mathcal{P}_m^n$ is finite-dimensional; hence, it makes sense to talk about bounded subsets of $\mathcal{P}_m^n$ without specifying a norm.

Theorem 3.1. *Let $m, n \in \mathbb{N}$, $\varepsilon > 0$ and $\phi : \mathbb{C} \rightarrow \mathbb{C}$ be smooth and non-polyharmonic on an open set $\emptyset \neq U \subseteq \mathbb{C}$. Let $\mathcal{P}' \subseteq \mathcal{P}_m^n$ be bounded and set $N := (4m + 1)^{2n}$. Then there exists a first layer $\Phi \in \mathcal{F}_{n,N}^\phi$ with the following property: For each polynomial $p \in \mathcal{P}'$ there exists $\sigma \in \mathbb{C}^N$, such that*

$$\|p - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \varepsilon.$$

Sketch of proof. For any multi-indices $\mathbf{m}, \ell \in \mathbb{N}_0^n$ an inductive argument shows for every fixed $z \in \Omega_n$ and $b \in \mathbb{C}$ that

$$\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} [w \mapsto \phi(w^T z + b)] = z^{\mathbf{m}} \bar{z}^{\ell} \cdot \left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b).$$

Here, $\partial_{\text{wirt}}^{\mathbf{m}}$ and $\bar{\partial}_{\text{wirt}}^{\ell}$ denote the multivariate Wirtinger derivatives with respect to w according to the multi-indices $\mathbf{m}$ and ℓ , respectively. Evaluating this at $w = 0$ and taking $b \in \mathbb{C}$ such that none of the mixed Wirtinger derivatives of ϕ at b up to a sufficiently high order vanish (where such a b exists by Proposition 2.1) shows that we can rewrite

$$z^{\mathbf{m}} \bar{z}^{\ell} = \left(\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} [w \mapsto \phi(w^T z + b)] \right) \Big|_{w=0} \cdot \left(\left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b) \right)^{-1}. \quad (3.1)$$

The mixed Wirtinger derivatives can by definition be expressed as linear combinations of usual partial derivatives. Those partial derivatives can be approximated using a generalized version of *divided differences*: If $g \in C^k((-r, r)^s; \mathbb{R})$ and $\mathbf{p} \in \mathbb{N}_0^s$ with $|\mathbf{p}| \leq k$ we have

$$\partial^{\mathbf{p}} g(0) \approx (2h)^{-|\mathbf{p}|} \sum_{0 \leq \mathbf{r} \leq \mathbf{p}} (-1)^{|\mathbf{p}| - |\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} g(h(2\mathbf{r} - \mathbf{p})) \quad \text{for } h \searrow 0. \quad (3.2)$$

See Appendix B.1 for a proof of this approximation. Note that when one takes $g(w) = \phi(w^T z + b)$, the right-hand side of (3.2) is a shallow neural network, as a function of z .

Combining (3.1) and (3.2) yields the desired result; see Appendix B.3 for the details. $\square$

It is crucial that the size of the networks considered in Theorem 3.1 is independent of the approximation accuracy ε . Moreover, the first layer Φ can be chosen independently of the particular polynomial p . Only the weights σ connecting hidden layer and output neuron have to be adapted to p .

The final approximation result is as follows. Its full proof can be found in Appendix C.3.

Theorem 3.2. *Let $n, k \in \mathbb{N}$. Then there exists a constant $c = c(n, k) > 0$ with the following property: For any activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ that is smooth and non-polyharmonic on an open set $\emptyset \neq U \subseteq \mathbb{C}$ and for any $m \in \mathbb{N}$ there exists a first layer $\Phi \in \mathcal{F}_{n,m}^\phi$ with the following property: For any $f \in C^k(\Omega_n; \mathbb{C})$ there exist coefficients $\sigma = \sigma(f) \in \mathbb{C}^m$, such that*

$$\|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c \cdot m^{-k/(2n)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}.$$

Furthermore, the map $f \mapsto \sigma(f)$ is a continuous linear operator with respect to the L^∞ -norm.

Sketch of proof. By splitting f into real and imaginary part we may assume that f is real-valued. Fourier-analytical results (recalled in Appendix C.1) imply that each C^k -function f can be well approximated by a linear combination of (multivariate) Chebyshev polynomials. Precisely, we have

$$\|f - P\|_{L^\infty(\Omega_n; \mathbb{R})} \leq \frac{c}{m^k} \cdot \|f\|_{C^k(\Omega_n; \mathbb{R})},$$

where P is given via the formula

$$P(z) = \sum_{0 \leq \mathbf{k} \leq 2m-1} \mathcal{V}_{\mathbf{k}}^m(f) \cdot T_{\mathbf{k}}(z), \quad z \in \Omega_n.$$

Here, the functions $T_{\mathbf{k}}$ are multivariate versions of Chebyshev polynomials and $\mathcal{V}_{\mathbf{k}}^m(f)$ are continuous linear functionals in f . The constant $c > 0$ only depends on n and k . See Appendix C.1 for a rigorous proof of this approximation property. Approximating the polynomials $T_{\mathbf{k}}$ by neural networks according to Theorem 3.1 yields the desired claim; see Appendix C.3 for the details. $\square$

Remark 3.3. Theorem 3.1 can not only be used to derive approximation rates for the approximation of C^k -functions but can be applied to any function class that is well approximable by algebraic polynomials. For example, it can be used to prove an order of approximation of $\nu^{-m^{1/(2n)}/17}$ for the approximation of functions $f : \Omega_n \rightarrow \mathbb{C}$ that can be *holomorphically extended* onto some polyellipse in $\mathbb{C}^{2n}$. The parameter $\nu > 1$ specifies the size of this polyellipse. We refer the interested reader to Appendix D for detailed definitions, statements and proofs for this fact.

4 Optimality of the derived approximation rate

In this section we discuss the optimality of the approximation rate proven in Theorem 3.2. We first deduce from general results by DeVore et al. [19] that the rate is optimal in the setting that the map which assigns to a function $f \in C^k(\Omega_n; \mathbb{C})$ the weights of the approximating network is continuous, as is the case in Theorem 3.2. However, it might be possible to achieve a better degree of approximation if this map is *not* required to be continuous, which is discussed in the second part of this section. Proofs for all the statements in this section are given in Appendices E.1, F.2 and F.3.

Continuous weight selection. We begin by considering the case of continuous weight selection. As mentioned in the introduction, this is a natural assumption, since in classical training algorithms such as (S)GD, continuous operations based on samples $f(x_j)$ are used to adjust the weights.

In [19, Theorem 4.2] a lower bound of $m^{-k/s}$ is established for the rate of approximating functions f of Sobolev regularity $W^{k,\infty}$ in the following very general setting: The set of functions that is used for approximation can be parametrized using m (real) parameters and the map that assigns to f the parameters of the approximating function is continuous with respect to *some* norm on $W^{k,\infty}([-1, 1]^s; \mathbb{R})$. A detailed version of the proof of that statement (for C^k instead of $W^{k,\infty}$) is contained in Appendix E.1. A careful transfer of this result to the complex-valued setting yields the following theorem (see also Appendix E.1).

Theorem 4.1. *Let $n, k \in \mathbb{N}$. Then there exists a constant $c = c(n, k) > 0$ with the following property: For any $m \in \mathbb{N}$, any map $\bar{a} : C^k(\Omega_n; \mathbb{C}) \rightarrow \mathbb{C}^m$ that is continuous with respect to some norm on $C^k(\Omega_n; \mathbb{C})$ and any map $M : \mathbb{C}^m \rightarrow C(\Omega_n; \mathbb{C})$ we have*

$$\sup_{f \in C^k(\Omega_n; \mathbb{C}), \|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1} \|f - M(\bar{a}(f))\|_{L^\infty(\Omega_n; \mathbb{C})} \geq c \cdot m^{-k/(2n)}.$$

The interpretation of Theorem 4.1 is as follows: If one approximates C^k -functions on Ω_n using a set of functions that can be parametrized using m (complex) parameters and one assumes that the weight assignment is continuous, one cannot achieve a better rate of approximation than $m^{-k/(2n)}$. As a special case it can be deduced that the approximation rate is at most $m^{-k/(2n)}$ when approximating C^k -functions using shallow CVNNs with m parameters under continuous weight assignment (see Corollary E.3). One can show that this even holds for *deep* CVNNs. Hence, for continuous weight selection the rate proven in Theorem 3.2 is optimal, even in the class of (possibly) deep networks.

Discontinuous weight selection. When we drop the continuity assumption on the selection of the weights, the behavior of the optimal approximation rate is more subtle. Precisely, we show that there are activation functions for which the rate of approximation can be improved to $m^{-k/(2n-1)}$. On the other hand, we also show that there are activation functions for which an improvement of the approximation rate (up to logarithmic factors) is not possible.

Theorem 4.2. *There exists a function $\phi \in C^\infty(\mathbb{C}; \mathbb{C})$ which is non-polyharmonic with the following additional property: For every $k \in \mathbb{N}$ and $n \in \mathbb{N}$ there exists a constant $c = c(n, k) > 0$ such that*

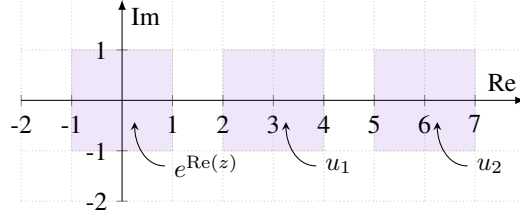


Figure 3: Illustration of the construction of the activation function in Theorem 4.2.

for any $m \in \mathbb{N}$ and $f \in C^k(\Omega_n; \mathbb{C})$ there is a first layer $\Phi \in \mathcal{F}_{n,m}^\phi$ and a vector $\sigma \in \mathbb{C}^m$ such that

$$\|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}.$$

Sketch of proof. The function ϕ is constructed in the following way: Take a countable dense subset $\{u_\ell\}_{\ell \in \mathbb{N}}$ of $C(\Omega_1; \mathbb{C})$, for instance the set of all polynomials in z and $\bar{z}$ with coefficients in $\mathbb{Q} + i\mathbb{Q}$. Define ϕ in a way such that

$$\phi(z + 3\ell) = u_\ell(z)$$

for every $z \in \Omega_1$ and $\ell \in \mathbb{N}$. Furthermore, to ensure that ϕ is non-polyharmonic, let $\phi(z) = e^{\text{Re}(z)}$ for every $z \in \Omega_1$. The smoothness of ϕ can be accomplished by multiplying with a smooth *bump function*; see Lemma F.4 for details. The construction of ϕ is illustrated in Figure 3.

Let then $f \in C^k(\Omega_n; \mathbb{C})$ be arbitrary. General results from the theory of *ridge functions* [22] show that there exist $b_1, \dots, b_m \in \mathbb{C}^n$ and $g_1, \dots, g_m \in C(\Omega_1; \mathbb{C})$ such that

$$\left\| f(z) - \sum_{j=1}^m g_j(b_j^T \cdot z) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})};$$

see Proposition F.3 and Appendix F.1 for a detailed proof of this fact. Approximating the functions g_j by suitable functions u_{ℓ_j} and expressing those functions via $\phi(\bullet + 3\ell_j)$ yields the claim. $\square$

The preceding theorem showed that there exists an activation function for which the rate in Theorem 3.2 can be strictly improved, if one allows a discontinuous weight selection. In contrast, the following theorem shows for a certain (quite natural) activation function that the rate $m^{-k/(2n)}$ *cannot* be improved (up to logarithmic factors), even if one allows a discontinuous weight assignment.

Theorem 4.3. *Let $n, k \in \mathbb{N}$ and*

$$\phi : \mathbb{C} \rightarrow \mathbb{C}, \quad \phi(z) := \frac{1}{1 + e^{-\text{Re}(z)}}.$$

Then ϕ is smooth but non-polyharmonic. Furthermore, there exists a constant $c = c(n, k) > 0$ with the following property: For any $m \in \mathbb{N}_{\geq 2}$ there exists a function $f \in C^k(\Omega_n; \mathbb{C})$ with $\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1$, such that for every $\Phi \in \mathcal{F}_{n,m}^\phi$ and $\sigma \in \mathbb{C}^m$ we have

$$\|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \geq c \cdot (m \cdot \ln(m))^{-k/(2n)}.$$

Sketch of proof. The idea of the proof is based on that of [46, Theorem 4] but instead of the bound for the VC dimension of ReLU networks used in [46], we will employ a bound for the VC dimension stated in [4, Theorem 8.11] using the real sigmoid function. For a detailed introduction to the concept of the VC dimension and related topics, see for instance [39, Chapter 6].

A technical reduction from the complex to the real case (see Appendix F.3) shows that it suffices to show the following: If $\varepsilon \in (0, \frac{1}{2})$ and $m \in \mathbb{N}$ are such that for every $f \in C^k([-1, 1]^n; \mathbb{R})$ with $\|f\|_{C^k([-1, 1]^n; \mathbb{R})} \leq 1$ there exists a real-valued shallow network $\mathcal{N}$ with $\gamma(x) = \frac{1}{1+e^{-x}}$ as activation function satisfying $\|f - \mathcal{N}\|_{L^\infty([-1, 1]^n; \mathbb{R})} \leq \varepsilon$, then necessarily

$$m \geq c \cdot \frac{\varepsilon^{-n/k}}{\ln(1/\varepsilon)},$$

where the constant c only depends on n and k .

To show that the latter claim holds, we assume that ε and m have the property stated above. Take $N \in \mathbb{N}$ such that $N^{-k} \asymp \varepsilon$ and consider the grid

$$\mathcal{G} := \frac{1}{N} \{-N, \dots, N\}^n \subseteq [-1, 1]^n.$$

For every $\alpha \in \mathcal{G}$ we pick a number $z_\alpha \in \{0, 1\}$ arbitrarily and construct a map $f \in C^\infty([-1, 1]^n; \mathbb{R})$ satisfying $f(\alpha) = z_\alpha$ for every $\alpha \in \mathcal{G}$. Scaling of f to $\tilde{f}$ ensures $\|\tilde{f}\|_{C^k([-1, 1]^n; \mathbb{R})} \leq 1$, but then $\tilde{f}(\alpha) = c_0 \cdot z_\alpha \cdot N^{-k}$ where $c_0 = c_0(n, k) > 0$. By assumption, we can infer the existence of a shallow real-valued neural network $\mathcal{N}$ with γ as activation function and m hidden neurons satisfying $\|\tilde{f} - \mathcal{N}\|_{L^\infty([-1, 1]^n; \mathbb{R})} \leq \varepsilon$. But this shows

$$\mathcal{N}(\alpha) \begin{cases} > \tilde{c}N^{-k}, & \text{if } z_\alpha = 1, \\ < \tilde{c}N^{-k}, & \text{if } z_\alpha = 0 \end{cases} \quad \text{for all } \alpha \in \mathcal{G}$$

with a constant $\tilde{c} = \tilde{c}(n, k) > 0$. Since the z_α are arbitrary, it follows that the set

$$H := \left\{ \mathbb{1}(\mathcal{N} > \tilde{c}N^{-k}) \Big|_{\mathcal{G}} : \mathcal{N} \text{ shallow NN with activation } \gamma \text{ and } m \text{ hidden neurons} \right\}$$

shatters the whole grid $\mathcal{G}$. This yields $\text{VC}(H) \geq |\mathcal{G}| = (2N + 1)^n$. On the other hand, the bound provided by [4, Theorem 8.11] for linear threshold networks yields $\text{VC}(H) \lesssim m \cdot \ln(N)$. Combining the two bounds and using $N^{-k} \asymp \varepsilon$ yields the claim. $\square$

5 Tractability of the considered problem in terms of the input dimension

In this section we discuss the *tractability* (in terms of the input dimension n) of the considered problem, i.e., the dependence of the approximation error on n . We show a novel result stating that, assuming a continuous weight selection, the problem of approximating C^k -functions is *intractable*, i.e., that the number of neurons that is required to achieve a non-trivial approximation error is necessarily exponential in n . In the literature this is referred to as the *curse of dimensionality*. The proof of the theorem combines ideas from [19] and [35] and is contained in Appendix E.2.

Theorem 5.1. *Let $s \in \mathbb{N}$. With $\|\cdot\|_{C^k([-1, 1]^s; \mathbb{R})}$ defined similarly to (2.1), we write*

$$\|f\|_{C^\infty([-1, 1]^s; \mathbb{R})} := \sup_{k \in \mathbb{N}} \|f\|_{C^k([-1, 1]^s; \mathbb{R})} \in [0, \infty]$$

for any function $f \in C^\infty([-1, 1]^s; \mathbb{R})$ and denote by $C^{\infty, *, s}$ the set of all $f \in C^\infty([-1, 1]^s; \mathbb{R})$ for which this expression is finite. Let $\bar{a} : C^{\infty, *, s} \rightarrow \mathbb{R}^{2^s - 1}$ be continuous with respect to some norm on $C^{\infty, *, s}$ and moreover, let $M : \mathbb{R}^{2^s - 1} \rightarrow C([-1, 1]^s; \mathbb{R})$ be an arbitrary map. Then it holds

$$\sup_{\substack{f \in C^{\infty, *, s} \\ \|f\|_{C^\infty([-1, 1]^s; \mathbb{R})} \leq 1}} \|f - M(\bar{a}(f))\|_{L^\infty([-1, 1]^s; \mathbb{R})} \geq 1.$$

Note that Theorem 5.1 is formulated for real-valued functions but can be transferred to the complex-valued setting (see Corollary E.6). We decided to include the real-valued statement because it is expected to be of greater interest in the community than the complex-valued analog. Moreover, we stress that Theorem 5.1 is not limited to the class of shallow neural networks but refers to any function class that is parametrizable using finitely many parameters (in particular, e.g., the class of neural networks with possibly more than one hidden layer).

We now examine in what way the constant c appearing in Theorem 3.2 suffers from the curse of dimensionality. To this end, it is convenient to rewrite the result from Theorem 3.2 as

$$\sup_{\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1} \inf_{\Phi \in \mathcal{F}_{n, m, \sigma}^\Phi, \sigma \in \mathbb{C}^m} \|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \leq (\tilde{c} \cdot m)^{-k/(2n)}$$

where the constant $\tilde{c} = \tilde{c}(n, k) > 0$ is independent of m . Writing the result in that way, one sees immediately that, if one seeks to have a worst-case approximation error of less than $\varepsilon > 0$, it is

sufficient to take $m = \lceil \frac{1}{\tilde{c}} \cdot \varepsilon^{-(2n)/k} \rceil$ neurons in the hidden layer of the network. Corollary E.7 shows that it *necessarily* holds $\tilde{c} \leq 16 \cdot 2^{-n}$ and therefore, the constant $\tilde{c}$ unavoidably suffers from the curse of dimensionality. An analysis of the constant (where we refer to Appendices C.1 to C.3 for the details) shows that in our case we can establish the bound $\tilde{c}(n, k) \geq \exp(-C \cdot n^2) \cdot k^{-4n}$ with an *absolute* constant $C > 0$. We remark that, since the constant suffers from the curse of dimensionality in any case, we have not put much effort into optimizing the constant; there is therefore probably ample room for improvement.

6 Limitations

To conduct a comprehensive evaluation of machine learning algorithms, one must analyze the questions of approximation, generalization, and optimization through training algorithms. The present paper, however, only focuses on the aspect of approximation. Analyzing if the proven approximation rate can be attained with learning algorithms such as (stochastic) gradient descent falls outside the scope of this paper. Furthermore, the examination of approximation rates under possibly discontinuous weight assignment is not yet fully resolved by our results. It is an open question which rate is optimally achievable in that case, depending on the choice of the activation function, and specifically in distinguishing between shallow and deep networks. We want to mention the two following points which indicate that this is a quite subtle question:

1. For deep NNs (with more than two hidden layers) with general smooth activation function, it is *not possible* to derive any non-trivial lower bounds in the setting of unrestricted weight assignment, since there exists an activation function with the property that NNs of *constant size* using this activation function can approximate any continuous function to arbitrary precision (see [31, Theorem 4]). Note that [31] considers real-valued NNs, but the results can be transferred to CVNNs with a suitable choice of the activation function.
2. In the real-valued case, fully general lower bounds for the approximation capabilities of shallow NNs have been derived by using results from [22] regarding the approximation properties of so-called ridge functions, i.e., functions of the form $\sum_{j=1}^m \phi_j(\langle a_j, x \rangle)$ with $a_j \in \mathbb{R}^d$ and each $\phi_j : \mathbb{R} \rightarrow \mathbb{R}$. It is an interesting problem to generalize these results to higher-dimensional ridge functions of the form $\sum_{j=1}^m \phi_j(A_j x)$, where each $\phi_j : \mathbb{R}^s \rightarrow \mathbb{R}$ and $A_j \in \mathbb{R}^{s \times d}$. This would imply lower bounds for shallow CVNNs. However, such a generalization seems to be highly non-trivial and is outside the scope of the present paper.

7 Conclusion

This paper analyzes error bounds for the approximation of complex-valued C^k -functions by means of complex-valued neural networks with smooth and non-polyharmonic activation functions. It is demonstrated that complex-valued neural networks with these activation functions achieve the *identical* approximation rate as real-valued networks that employ smooth and non-polynomial activation functions. This is an important theoretical finding, since CVNNs are on the one hand more restrictive than real-valued neural networks (since the mappings between layers should be $\mathbb{C}$ -linear and not just $\mathbb{R}$ -linear), but on the other hand more versatile, since the activation function is a mapping from $\mathbb{C}$ to $\mathbb{C}$ (i.e., from $\mathbb{R}^2 \rightarrow \mathbb{R}^2$) rather than from $\mathbb{R}$ to $\mathbb{R}$ as in the real case. Additionally, it is established that the proven approximation rate is optimal if one assumes a continuous weight selection. In summary, if one focuses on the approximation rate for C^k -functions, CVNNs have the same excellent approximation properties as real-valued networks.

The behavior of the approximation rate for unrestricted weight selection is more subtle. It is shown that a rate of $m^{-k/(2n-1)}$ can be achieved for certain activation functions (Theorem 4.2) but in general, one cannot improve on the rate that is attainable for continuous weight selection (Theorem 4.3).

While the proven approximation *rate* is optimal under the assumption of continuous weight selection, the involved constants suffer from the curse of dimensionality. Section 5, however, shows that this is inevitable in the given setting.

Such theoretical approximation results contribute to the mathematical understanding of Deep Learning. The remarkable approximation-theoretical properties of neural networks can be seen as one reason why neural networks provide outstanding results in many applications.

Acknowledgments. PG and FV acknowledge support by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1. FV acknowledges support by the Hightech Agenda Bavaria.

References

- [1] M. Abramowitz and I. A. Stegun, eds. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. 9. Dover print. Dover books on mathematics. New York, NY: Dover Publ, 2013. ISBN: 978-0-486-61272-0.
- [2] B. Adcock, S. Brugiapaglia, and C. G. Webster. *Sparse polynomial approximation of high-dimensional functions*. Computational science and engineering 25. Philadelphia: Society for Industrial and Applied Mathematics, 2022. 292 pp. ISBN: 978-1-61197-687-8.
- [3] C. D. Aliprantis and K. C. Border. *Infinite dimensional analysis: a hitchhiker’s guide*. 3rd ed. Berlin ; New York: Springer, 2006. ISBN: 978-3-540-29586-0.
- [4] M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge ; New York, NY: Cambridge University Press, 1999. ISBN: 978-0-521-57353-5.
- [5] M. Arjovsky, A. Shah, and Y. Bengio. “Unitary evolution recurrent neural networks”. In: *International conference on machine learning*. PMLR. 2016, pp. 1120–1128. URL: <http://proceedings.mlr.press/v48/arjovsky16.pdf>.
- [6] K. E. Atkinson. *An introduction to numerical analysis*. 2nd ed. New York: Wiley, 1989. ISBN: 978-0-471-62489-9.
- [7] M. B. Balk. *Polyanalytic functions*. Mathematical research 63. Berlin: Akademie Verlag, 1991. ISBN: 978-3-05-501292-1.
- [8] J. A. Barrachina, C. Ren, C. Morisseau, G. Vieillard, and J.-P. Ovarlez. “Complex-valued vs. real-valued neural networks for classification perspectives: An example on non-circular data”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 2990–2994.
- [9] J. A. Barrachina, C. Ren, C. Morisseau, G. Vieillard, and J.-P. Ovarlez. “Comparison between equivalent architectures of complex-valued and real-valued neural networks - Application on polarimetric SAR image segmentation”. In: *Journal of Signal Processing Systems* 95.1 (2023), pp. 57–66.
- [10] A. R. Barron. “Universal approximation bounds for superpositions of a sigmoidal function”. In: *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945. DOI: 10.1109/18.256500.
- [11] H. Bauer. *Measure and integration theory*. De Gruyter studies in mathematics 26. Berlin ; New York: W. de Gruyter, 2001. 230 pp. ISBN: 978-3-11-016719-1.
- [12] A. Benjamin and J. J. Quinn. *Proofs that really count: the art of combinatorial proof*. Dolciani mathematical expositions. Washington, DC: Mathematical Association of America, 2003. ISBN: 978-0-88385-333-7.
- [13] D. Berend and T. Tassa. “Improved bounds on Bell numbers and on moments of sums of random variables”. In: *Probability and Mathematical Statistics* 30.2 (2010), pp. 185–205.
- [14] A. Caragea, D. G. Lee, J. Maly, G. Pfander, and F. Voigtlaender. “Quantitative Approximation Results for Complex-Valued Neural Networks”. In: *SIAM Journal on Mathematics of Data Science* 4.2 (2022), pp. 553–580. DOI: 10.1137/21M1429540.
- [15] H. H. Çevik, Y. E. Acar, and M. Çunkaş. “Day Ahead Wind Power Forecasting Using Complex Valued Neural Network”. In: *2018 International Conference on Smart Energy Systems and Technologies (SEST)*. 2018, pp. 1–6. DOI: 10.1109/SEST.2018.8495637.
- [16] E. Cole, J. Cheng, J. Pauly, and S. Vasanawala. “Analysis of deep complex-valued convolutional neural networks for MRI reconstruction and phase-focused applications”. In: *Magnetic resonance in medicine* 86.2 (2021), pp. 1093–1109.
- [17] G. Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/BF02551274.
- [18] K. Deimling. *Nonlinear Functional Analysis*. Dover books on mathematics. Dover Publications, 2013. ISBN: 978-3-662-00549-1.
- [19] R. A. DeVore, R. Howard, and C. Micchelli. “Optimal nonlinear approximation”. In: *Manuscripta Mathematica* 63.4 (Dec. 1989), pp. 469–478. DOI: 10.1007/BF01171759.
- [20] W. F. Donoghue. *Distributions and Fourier transforms*. OCLC: 45151. New York: Academic Press, 1969. ISBN: 978-0-12-220650-4.
- [21] G. B. Folland. *Real analysis: modern techniques and their applications*. 2nd ed. Pure and applied mathematics. New York: Wiley, 1999. ISBN: 978-0-471-31716-6.
- [22] Y. Gordon, V. Maiorov, M. Meyer, and S. Reisner. “On the Best Approximation by Ridge Functions in the Uniform Norm”. In: *Constructive Approximation* 18.1 (Jan. 1, 2001), pp. 61–85. DOI: 10.1007/s00365-001-0009-5.

- [23] A. Graves, A.-R. Mohamed, and G. Hinton. “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.
- [24] W. P. Johnson. “The Curious History of Faa di Bruno’s Formula”. In: *The American Mathematical Monthly* 109.3 (Mar. 2002), p. 217. DOI: 10.2307/2695352.
- [25] P. C. Kainen, V. Kůrková, and A. Vogt. “Approximation by neural networks is not continuous”. In: *Neurocomputing* 29.1 (1999), pp. 47–56. DOI: [https://doi.org/10.1016/S0925-2312\(99\)00111-3](https://doi.org/10.1016/S0925-2312(99)00111-3).
- [26] L. Kaup and B. Kaup. *Holomorphic Functions of Several Variables: An Introduction to the Fundamental Theory*. De Gruyter, Dec. 1983. ISBN: 978-3-11-004150-7. DOI: 10.1515/9783110838350.
- [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. DOI: 10.1145/3065386.
- [28] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function”. In: *Neural Networks* 6.6 (Jan. 1993), pp. 861–867. DOI: 10.1016/S0893-6080(05)80131-5.
- [29] G. Lorentz. *Approximation of Functions*. Approximation of Functions. Holt, Rinehart and Winston, 1966.
- [30] V. Maiorov. “On Best Approximation by Ridge Functions”. In: *Journal of Approximation Theory* 99.1 (July 1999), pp. 68–94. DOI: 10.1006/jath.1998.3304.
- [31] V. Maiorov and A. Pinkus. “Lower bounds for approximation by MLP neural networks”. In: *Neurocomputing* 25.1 (1999), pp. 81–91. DOI: [https://doi.org/10.1016/S0925-2312\(98\)00111-8](https://doi.org/10.1016/S0925-2312(98)00111-8).
- [32] A. Marset and F. Sahin. “Application of complex-valued convolutional neural network for next generation wireless networks”. In: *2017 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. 2017, pp. 1–5. DOI: 10.1109/WNYIPW.2017.8356260.
- [33] H. N. Mhaskar. “Neural Networks for Optimal Approximation of Smooth and Analytic Functions”. In: *Neural Computation* 8.1 (Jan. 1996), pp. 164–177. DOI: 10.1162/neco.1996.8.1.164.
- [34] C. Muscalu and W. Schlag. *Classical and multilinear harmonic analysis. Volume 1*. OCLC: 887742576. Cambridge, UK: Cambridge University Press, 2013. ISBN: 978-0-521-88245-3.
- [35] E. Novak and H. Woźniakowski. “Approximation of infinitely differentiable multivariate functions is intractable”. In: *Journal of Complexity* 25.4 (2009), pp. 398–404. DOI: <https://doi.org/10.1016/j.jco.2008.11.002>.
- [36] A. Pinkus. *Ridge functions*. Cambridge: Cambridge University Press, 2016. ISBN: 978-1-107-12439-4.
- [37] Y.-D. Qu, R.-Q. Zhang, S.-Q. Shen, J. Yu, and M. Li. “Entanglement Detection with Complex-Valued Neural Networks”. In: *International Journal of Theoretical Physics* 62.9 (2023), p. 206.
- [38] H. El-Rewaidy, U. Neisius, J. Mancio, S. Kucukseymen, J. Rodriguez, A. Paskavitz, B. Menze, and R. Nezafat. “Deep complex convolutional network for fast reconstruction of 3D late gadolinium enhancement cardiac MRI”. In: *NMR in Biomedicine* 33.7 (2020), e4312.
- [39] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: from theory to algorithms*. New York, NY, USA: Cambridge University Press, 2014. 397 pp. ISBN: 978-1-107-05713-5.
- [40] Z. Shen, H. Yang, and S. Zhang. “Deep Network Approximation Characterized by Number of Neurons”. In: *Communications in Computational Physics* 28.5 (2020), pp. 1768–1811. DOI: <https://doi.org/10.4208/cicp.0A-2020-0149>.
- [41] J. W. Siegel and J. Xu. “Approximation rates for neural networks with general activation functions”. In: *Neural Networks* 128 (2020), pp. 313–321. DOI: <https://doi.org/10.1016/j.neunet.2020.05.019>.
- [42] P. Stein. “A Note on the Volume of a Simplex”. In: *The American Mathematical Monthly* 73.3 (Mar. 1966), p. 299. DOI: 10.2307/2315353.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [44] P. Virtue, S. X. Yu, and M. Lustig. “Better than real: Complex-valued neural nets for MRI fingerprinting”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. Beijing: IEEE, Sept. 2017, pp. 3953–3957. ISBN: 978-1-5090-2175-8. DOI: 10.1109/ICIP.2017.8297024.
- [45] F. Voigtlaender. “The universal approximation theorem for complex-valued neural networks”. In: *Applied and Computational Harmonic Analysis* (Dec. 2022), S1063520322001014. DOI: 10.1016/j.acha.2022.12.002.
- [46] D. Yarotsky. “Error bounds for approximations with deep ReLU networks”. In: *Neural Networks* 94 (Oct. 2017), pp. 103–114. DOI: 10.1016/j.neunet.2017.07.002.

- [47] D. Yarotsky and A. Zhevnerchuk. “The phase diagram of approximation rates for deep neural networks”. In: *Advances in neural information processing systems* 33 (2020), pp. 13005–13015. URL: <https://proceedings.neurips.cc/paper/2020/file/979a3f14bae523dc5101c52120c535e9-Paper.pdf>.
- [48] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin. “Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.12 (2017), pp. 7177–7188. DOI: 10.1109/TGRS.2017.2743222.

A Notation, Wirtinger derivatives and special activation functions

A.1 Notation and Wirtinger derivatives

Throughout the paper, multi-indices (i.e., elements of $\mathbb{N}_0^n$) are denoted using boldface. For $\mathbf{m} \in \mathbb{N}_0^n$ we have the usual notation $|\mathbf{m}| = \sum_{j=1}^n \mathbf{m}_j$. For a number $m \in \mathbb{N}_0$ and another multi-index $\mathbf{p} \in \mathbb{N}_0^n$ we write $\mathbf{m} \leq m$ iff $\mathbf{m}_j \leq m$ for every $j \in \{1, \dots, n\}$ and $\mathbf{m} \leq \mathbf{p}$ iff $\mathbf{m}_j \leq \mathbf{p}_j$ for every $j \in \{1, \dots, n\}$. Furthermore we write

$$\binom{\mathbf{p}}{\mathbf{r}} := \prod_{j=1}^n \binom{\mathbf{p}_j}{\mathbf{r}_j}$$

for two multi-indices $\mathbf{p}, \mathbf{r} \in \mathbb{N}_0^n$ with $\mathbf{r} \leq \mathbf{p}$. For a complex vector $z \in \mathbb{C}^n$ we write

$$z^{\mathbf{m}} := \prod_{j=1}^n z_j^{\mathbf{m}_j} \quad \text{and} \quad \bar{z}^{\mathbf{m}} := \prod_{j=1}^n \bar{z}_j^{\mathbf{m}_j}.$$

For a point $x \in \mathbb{R}^n$ and $r > 0$ we define

$$B_r(x) := \{y \in \mathbb{R}^n : \|x - y\| < r\}$$

with $\|\cdot\|$ denoting the usual Euclidean distance. This definition is analogously transferred to $\mathbb{C}^n$.

$\mathbb{C}^n$ is canonically isomorphic to $\mathbb{R}^{2n}$ by virtue of the isomorphism

$$\varphi_n : \mathbb{R}^{2n} \rightarrow \mathbb{C}^n, \quad (x_1, \dots, x_n, y_1, \dots, y_n) \mapsto (x_1 + iy_1, \dots, x_n + iy_n). \quad (\text{A.1})$$

The Wirtinger derivatives defined in Section 2 have the following properties that we are going to use, which can be found for example in [26, E.1a]. Here, we assume that $U \subseteq \mathbb{C}$ is open and $f \in C^1(U; \mathbb{C})$.

1. ∂_{wirt} and $\bar{\partial}_{\text{wirt}}$ are both $\mathbb{C}$ -linear operators on the set $C^1(U; \mathbb{C})$.
2. f is complex-differentiable in $z \in U$ iff $\bar{\partial}_{\text{wirt}} f(z) = 0$ and in this case the equality

$$\partial_{\text{wirt}} f(z) = f'(z)$$

holds true, with $f'(z)$ denoting the complex derivative of f at z .

3. We have the conjugation rules

$$\overline{\partial_{\text{wirt}} f} = \bar{\partial}_{\text{wirt}} \bar{f} \quad \text{and} \quad \overline{\bar{\partial}_{\text{wirt}} f} = \partial_{\text{wirt}} \bar{f}.$$

4. If $g \in C^1(U; \mathbb{C})$, the following product rules for Wirtinger derivatives hold for every $z \in U$:

$$\begin{aligned} \partial_{\text{wirt}}(f \cdot g)(z) &= \partial_{\text{wirt}} f(z) \cdot g(z) + f(z) \cdot \partial_{\text{wirt}} g(z), \\ \bar{\partial}_{\text{wirt}}(f \cdot g)(z) &= \bar{\partial}_{\text{wirt}} f(z) \cdot g(z) + f(z) \cdot \bar{\partial}_{\text{wirt}} g(z). \end{aligned}$$

This product rule is not explicitly stated in [26] but follows easily from the product rule for $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$.

5. If $V \subseteq \mathbb{C}$ is an open set and $g \in C^1(V; \mathbb{C})$ with $g(V) \subseteq U$, then the following chain rules for Wirtinger derivatives hold true:

$$\begin{aligned} \partial_{\text{wirt}}(f \circ g) &= [(\partial_{\text{wirt}} f) \circ g] \cdot \partial_{\text{wirt}} g + [(\bar{\partial}_{\text{wirt}} f) \circ g] \cdot \partial_{\text{wirt}} \bar{g}, \\ \bar{\partial}_{\text{wirt}}(f \circ g) &= [(\partial_{\text{wirt}} f) \circ g] \cdot \bar{\partial}_{\text{wirt}} g + [(\bar{\partial}_{\text{wirt}} f) \circ g] \cdot \bar{\partial}_{\text{wirt}} \bar{g}. \end{aligned}$$

6. If $f \in C^2(U; \mathbb{C})$ then we have

$$\Delta f(z) = 4 (\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} f)(z)$$

for every $z \in U$, with Δ denoting the usual Laplace-Operator $\Delta = \partial^{(2,0)} + \partial^{(0,2)}$ (cf. [7, equation (1.7)]).

For an open set $U \subseteq \mathbb{C}^n$, a function $f \in C^k(U; \mathbb{C})$ and a multi-index $\ell \in \mathbb{N}_0^n$ with $|\ell| \leq k$ we write $\partial_{\text{wirt}}^\ell f$ and $\bar{\partial}_{\text{wirt}}^\ell f$ for the iterated Wirtinger derivatives according to the multi-index ℓ .

Proposition A.1. Let $U \subseteq \mathbb{C}^n$ be an open set and $f \in C^k(U; \mathbb{C})$. Then, identifying f with the function $f \circ \varphi_n$ with φ_n as in (A.1), for any multi-indices $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $|\mathbf{m} + \ell| \leq k$ we have the representation

$$\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} f(a) = \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \mathbf{m} + \ell}} b_{\mathbf{p}} (\partial^{\mathbf{p}} f)(a) \quad \forall a \in U$$

with complex numbers $b_{\mathbf{p}} \in \mathbb{C}$ only depending on $\mathbf{p}, \mathbf{m}$ and ℓ and writing $\mathbf{p} = (\mathbf{p}', \mathbf{p}'')$ for the concatenation of the multi-indices $\mathbf{p}'$ and $\mathbf{p}'' \in \mathbb{N}_0^n$. In particular, the coefficients do not depend on f .

Proof. The proof is by induction over $\mathbf{m}$ and ℓ . We first assume $\mathbf{m} = 0$ and show the claim for all $\ell \in \mathbb{N}_0^n$ with $|\ell| \leq k$. In the case $\ell = 0$ there is nothing to show, so we assume the claim to be true for fixed $\ell \in \mathbb{N}_0^n$ with $|\ell| < k$ and take $j \in \{1, \dots, n\}$ arbitrarily. Then we get

$$\begin{aligned} \bar{\partial}_{\text{wirt}}^{\ell+e_j} f(a) &= \bar{\partial}_{\text{wirt}}^{e_j} \bar{\partial}_{\text{wirt}}^{\ell} f(a) \stackrel{\text{IH}}{=} \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \ell}} b_{\mathbf{p}} \bar{\partial}_{\text{wirt}}^{e_j} (\partial^{\mathbf{p}} f)(a) \\ &= \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \ell}} \frac{b_{\mathbf{p}}}{2} \left(\partial^{(\mathbf{p}' + e_j, \mathbf{p}'')} f \right)(a) + \frac{ib_{\mathbf{p}}}{2} \left(\partial^{(\mathbf{p}', \mathbf{p}'' + e_j)} f \right)(a) \\ &=: \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \ell + e_j}} b_{\mathbf{p}} (\partial^{\mathbf{p}} f)(a), \end{aligned}$$

as was to be shown.

Since we have shown the case $\mathbf{m} = 0$ we may assume the claim to be true for fixed $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $|\mathbf{m} + \ell| < k$. Then we deduce

$$\begin{aligned} \partial_{\text{wirt}}^{\mathbf{m}+e_j} \bar{\partial}_{\text{wirt}}^{\ell} f(a) &= \partial_{\text{wirt}}^{e_j} \partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} f(a) \stackrel{\text{IH}}{=} \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \mathbf{m} + \ell}} b_{\mathbf{p}} \partial_{\text{wirt}}^{e_j} (\partial^{\mathbf{p}} f)(a) \\ &= \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \mathbf{m} + \ell}} \frac{b_{\mathbf{p}}}{2} \left(\partial^{(\mathbf{p}' + e_j, \mathbf{p}'')} f \right)(a) - \frac{ib_{\mathbf{p}}}{2} \left(\partial^{(\mathbf{p}', \mathbf{p}'' + e_j)} f \right)(a) \\ &=: \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \mathbf{m} + \ell + e_j}} b_{\mathbf{p}} (\partial^{\mathbf{p}} f)(a). \end{aligned}$$

Hence the claim follows using the principle of mathematical induction. $\square$

Proposition A.1 is technical but crucial: In the course of the paper we will need to approximate Wirtinger derivatives of certain functions. In fact, however, we will approximate real derivatives and take advantage of the fact that Wirtinger derivatives are linear combinations of these.

A.2 Concrete examples of activation functions

In this section we analyze concrete activation functions that are commonly used in practical applications of complex-valued neural networks. We are going to show that those activation functions are smooth and non-polyharmonic on some non-empty open subset of $\mathbb{C}$. Our first result analyzes a family of “real activation functions interpreted as complex activation functions”.

Proposition A.2. Let $\rho \in C^\infty(\mathbb{R}; \mathbb{R})$ be non-polynomial and let $\psi : \mathbb{C} \rightarrow \mathbb{C}$ be defined as

$$\psi(z) := \rho(\text{Re}(z)) \quad \text{resp.} \quad \psi(z) := \rho(\text{Im}(z)).$$

Then ψ is smooth and non-polyharmonic.

Proof. Since ψ depends only on the real-, resp. imaginary part of the input, we see directly from the definition of the Wirtinger derivatives that

$$\partial_{\text{wirt}} \psi(z) = \bar{\partial}_{\text{wirt}} \psi(z) = \frac{1}{2} \rho'(\text{Re}(z)) \quad \text{resp.} \quad \partial_{\text{wirt}} \psi(z) = -\bar{\partial}_{\text{wirt}} \psi(z) = -\frac{i}{2} \rho'(\text{Im}(z)).$$

Hence we see for arbitrary $m, \ell \in \mathbb{N}_0$ that

$$\left| \partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \psi(z) \right| = \frac{1}{2^{m+\ell}} \left| \rho^{(m+\ell)}(\text{Re}(z)) \right| \quad \text{resp.} \quad \left| \partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \psi(z) \right| = \frac{1}{2^{m+\ell}} \left| \rho^{(m+\ell)}(\text{Im}(z)) \right|.$$

Since ρ is non-polynomial we can choose a real number x , such that $\rho^{(n)}(x) \neq 0$ for all $n \in \mathbb{N}_0$ (cf. for instance [20, p. 53]). Thus, $\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \psi(z) \neq 0$ for all $m, \ell \in \mathbb{N}_0$, whenever $z \in \mathbb{C}$ with $\text{Re}(z) = x$, or $\text{Im}(z) = y$, respectively. $\square$

Next, we consider the modReLU which was defined in Section 2.

Theorem A.3. *Let $b \in (-\infty, 0)$. Writing $\sigma = \sigma_{\text{modReLU}, b}$, one has for every $z \in \mathbb{C}$ with $|z| + b > 0$ the identity*

$$\left(\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \sigma \right) (z) = \begin{cases} z + b \frac{z}{|z|}, & m = \ell = 0, \\ 1 + \frac{b}{2} \cdot \frac{1}{|z|}, & m = 1, \ell = 0, \\ b \cdot q_{m,\ell} \cdot \frac{z^{\ell-m+1}}{|z|^{2\ell+1}}, & m \leq \ell + 1, \ell \geq 1, \\ b \cdot q_{m,\ell} \cdot \frac{\bar{z}^{m-\ell-1}}{|z|^{2m-1}}, & m \geq \ell + 1, m \geq 2 \end{cases}$$

for every $m \in \mathbb{N}_0$ and $\ell \in \mathbb{N}_0$. Here, the numbers $q_{m,\ell}$ are non-zero and rational. Furthermore, note that all cases for choices of m and ℓ are covered, by observing that we can either have the case $m \geq \ell + 1$ (where either $m \geq 2$ or $m = 1, \ell = 0$) or $m \leq \ell + 1$, where the latter case is again split into $\ell = 0$ and $\ell \geq 1$.

Proof. We first calculate certain Wirtinger derivatives for $z \neq 0$. First note

$$\begin{aligned} \bar{\partial}_{\text{wirt}} \left(\frac{1}{|z|^m} \right) &= \frac{1}{2} \left(\partial^{(1,0)} \left(\frac{1}{|z|^m} \right) + i \cdot \partial^{(0,1)} \left(\frac{1}{|z|^m} \right) \right) \\ &= \frac{1}{2} \left(\left(-\frac{m}{2} \right) \frac{2 \text{Re}(z) + i \cdot 2 \text{Im}(z)}{|z|^{m+2}} \right) \\ &= -\frac{m}{2} \cdot \frac{z}{|z|^{m+2}} \end{aligned}$$

and similarly

$$\partial_{\text{wirt}} \left(\frac{1}{|z|^m} \right) = -\frac{m}{2} \cdot \frac{\bar{z}}{|z|^{m+2}}$$

for any $m \in \mathbb{N}$. Using the product rule for Wirtinger derivatives (see Appendix A.1), we see

$$\bar{\partial}_{\text{wirt}} \left(\frac{z^\ell}{|z|^m} \right) = \underbrace{\bar{\partial}_{\text{wirt}} (z^\ell)}_{=0} \cdot \frac{1}{|z|^m} + z^\ell \cdot \bar{\partial}_{\text{wirt}} \left(\frac{1}{|z|^m} \right) = -\frac{m}{2} \cdot \frac{z^{\ell+1}}{|z|^{m+2}} \quad (\text{A.2})$$

for any $m \in \mathbb{N}$ and $\ell \in \mathbb{N}_0$ and furthermore

$$\begin{aligned} \partial_{\text{wirt}} \left(\frac{z^\ell}{|z|^m} \right) &= \partial_{\text{wirt}} (z^\ell) \cdot \frac{1}{|z|^m} + z^\ell \cdot \partial_{\text{wirt}} \left(\frac{1}{|z|^m} \right) \\ &= \ell \cdot z^{\ell-1} \cdot \frac{1}{|z|^m} - z^\ell \cdot \frac{m}{2} \cdot \frac{\bar{z}}{|z|^{m+2}} \\ &= \left(\ell - \frac{m}{2} \right) \cdot \frac{z^{\ell-1}}{|z|^m} \end{aligned} \quad (\text{A.3})$$

for $m, \ell \in \mathbb{N}$, and finally

$$\partial_{\text{wirt}} \left(\frac{\bar{z}^\ell}{|z|^m} \right) = \underbrace{\partial_{\text{wirt}} (\bar{z}^\ell)}_{=0} \cdot \frac{1}{|z|^m} + \bar{z}^\ell \cdot \partial_{\text{wirt}} \left(\frac{1}{|z|^m} \right) = -\frac{m}{2} \cdot \frac{\bar{z}^{\ell+1}}{|z|^{m+2}} \quad (\text{A.4})$$

for $m \in \mathbb{N}$ and $\ell \in \mathbb{N}_0$.

Having proven those three identities, we can start with the actual computation. We first fix $m = 0$ and prove the claimed identity by induction over ℓ . The case $\ell = 0$ follows directly from the definition of the modReLU and for $\ell = 1$, we note

$$\bar{\partial}_{\text{wirt}}\sigma(z) = \underbrace{\bar{\partial}_{\text{wirt}}(z)}_{=0} + b \cdot \bar{\partial}_{\text{wirt}}\left(\frac{z}{|z|}\right) \stackrel{(A.2)}{=} b \cdot \left(-\frac{1}{2}\right) \frac{z^2}{|z|^3},$$

which is the claimed form. Then, using induction, we compute

$$\bar{\partial}_{\text{wirt}}^{\ell+1}\sigma(z) = \bar{\partial}_{\text{wirt}}\left(b \cdot q_{0,\ell} \cdot \frac{z^{\ell+1}}{|z|^{2\ell+1}}\right) \stackrel{(A.2)}{=} \underbrace{b \cdot q_{0,\ell} \cdot \left(-\frac{2\ell+1}{2}\right)}_{=:q_{0,\ell+1}} \cdot \frac{z^{\ell+2}}{|z|^{2\ell+3}},$$

so that the case $m = 0$ is complete.

Now we deal with the case $m \leq \ell + 1$. The case $\ell = 0$ and $m = 1$ is proven by computing

$$\partial_{\text{wirt}}\sigma(z) = \partial_{\text{wirt}}(z) + b \cdot \partial_{\text{wirt}}\left(\frac{z}{|z|}\right) \stackrel{(A.3)}{=} 1 + b \cdot \frac{1}{2} \cdot \frac{1}{|z|}$$

so we can assume $\ell > 0$. Since we already dealt with the case $m = 0$, we can inductively assume the claim to be true for a fixed $m \leq \ell$. Then we compute

$$\left(\partial_{\text{wirt}}^{m+1}\bar{\partial}_{\text{wirt}}^\ell\sigma\right)(z) = \partial_{\text{wirt}}\left(b \cdot q_{m,\ell} \cdot \frac{z^{\ell-m+1}}{|z|^{2\ell+1}}\right) \stackrel{(A.3)}{=} \underbrace{b \cdot q_{m,\ell} \cdot \left(-m + \frac{1}{2}\right)}_{=:q_{m+1,\ell}} \cdot \frac{z^{\ell-m}}{|z|^{2\ell+1}},$$

which is of the desired form. Note that (A.3) is indeed applicable because $\ell - m + 1 \geq 1$.

Finally, we consider the case where $m \geq \ell + 1$ and $m \geq 2$. The case $m = \ell + 1$ has already been shown. Using induction, we see

$$\begin{aligned} \left(\partial_{\text{wirt}}^{m+1}\bar{\partial}_{\text{wirt}}^\ell\sigma\right)(z) &= \partial_{\text{wirt}}\left(\delta_{(m,\ell)=(1,0)} + b \cdot q_{m,\ell} \cdot \frac{\bar{z}^{m-\ell-1}}{|z|^{2m-1}}\right) \\ &\stackrel{(A.4)}{=} \underbrace{b \cdot q_{m,\ell} \cdot \left(-m + \frac{1}{2}\right)}_{=:q_{m+1,\ell}} \cdot \frac{\bar{z}^{m-\ell}}{|z|^{2m+1}}, \end{aligned}$$

so the proof is complete. $\square$

From Theorem A.3 we can now deduce that the modReLU is smooth and non-polyharmonic on some non-empty open subset of $\mathbb{C}$.

Corollary A.4. *Let $b \in (-\infty, 0)$ and $z \in \mathbb{C}$ with $|z| > -b$. Then we have*

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \sigma_{\text{modReLU},b}(z) \neq 0$$

for all $m, \ell \in \mathbb{N}_0$. In particular, $\sigma_{\text{modReLU},b}$ is smooth and non-polyharmonic on the set

$$\{z \in \mathbb{C} : |z| > -b\}.$$

Furthermore, $\sigma_{\text{modReLU},b}$ is continuous on the entire complex plane.

Proof. The first part follows from Theorem A.3 by noting that if $|z| > -b$ we have in particular $|z| > -b/2$ and $|z| > 0$. For the second part, we first note that $\sigma_{\text{modReLU},b}$ is trivially continuous in $z \in \mathbb{C}$ if $|z| \neq -b$. Hence, we assume $|z| = -b$. Take any sequence $(z_j)_{j \in \mathbb{N}}$ with $z_j \rightarrow z$ as $j \rightarrow \infty$ and note

$$|\sigma_{\text{modReLU},b}(z_j) - \sigma_{\text{modReLU},b}(z)| = |\sigma_{\text{modReLU},b}(z_j)| \rightarrow \begin{cases} 0 \rightarrow 0, & \text{if } |z_j| < b, \\ |z_j| + b \rightarrow |z| + b = 0, & \text{if } |z_j| \geq b. \end{cases}$$

This shows the continuity of $\sigma_{\text{modReLU},b}$. $\square$

Next we analyze the complex cardioid, which was defined in Section 2.

Theorem A.5. *For any $z \in \mathbb{C}$ with $z \neq 0$ and any $m, \ell \in \mathbb{N}_0$ we have*

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \text{card}(z) = \begin{cases} \frac{1}{2}z + \frac{1}{4}\frac{z^2}{|z|} + \frac{|z|}{4}, & m = \ell = 0, \\ a_{m,\ell} \frac{z^{\ell-m}}{|z|^{2\ell-1}} + b_{m,\ell} \frac{z^{\ell+2-m}}{|z|^{2\ell+1}}, & m \leq \ell \neq 0, \\ \frac{1}{2} + \frac{1}{8} \cdot \frac{\bar{z}}{|z|} + \frac{3}{8} \cdot \frac{z}{|z|}, & m = \ell + 1 = 1, \\ a_{m,\ell} \frac{\bar{z}}{|z|^{2\ell+1}} + b_{m,\ell} \frac{z}{|z|^{2\ell+1}}, & m = \ell + 1 > 1, \\ a_{m,\ell} \frac{\bar{z}^{m-\ell}}{|z|^{2m-1}} + b_{m,\ell} \frac{\bar{z}^{m-\ell-2}}{|z|^{2m-3}}, & m \geq \ell + 2. \end{cases}$$

Here, the numbers $a_{m,\ell}$ and $b_{m,\ell}$ are non-zero and rational. Furthermore, note that all cases for possible choices of m and ℓ are covered: The case $m \leq \ell$ is split into $\ell = 0$ and $\ell \neq 0$. The case $m = \ell + 1$ is split into $m = 1$ and $m > 1$. Then, the case $m \geq \ell + 2$ remains.

Proof. For the following we always assume $z \in \mathbb{C}$ with $z \neq 0$. Then we can apply the identity $\cos(\angle z) = \frac{\text{Re}(z)}{|z|}$, so we can rewrite

$$\text{card}(z) = \frac{1}{2} \left(1 + \frac{\text{Re}(z)}{|z|} \right) z = \frac{1}{2}z + \frac{1}{4} \frac{(z + \bar{z})z}{|z|} = \frac{1}{2}z + \frac{1}{4} \frac{z^2}{|z|} + \frac{|z|}{4}. \quad (\text{A.5})$$

This establishes the case $m = \ell = 0$. Next, we compute

$$\bar{\partial}_{\text{wirt}}(|z|) = \frac{1}{2} \left(\frac{1}{2} \frac{2 \text{Re}(z)}{|z|} + \frac{i}{2} \frac{2 \text{Im}(z)}{|z|} \right) = \frac{1}{2} \frac{z}{|z|} \quad (\text{A.6})$$

and similarly

$$\partial_{\text{wirt}}(|z|) = \frac{1}{2} \frac{\bar{z}}{|z|}. \quad (\text{A.7})$$

We deduce

$$\bar{\partial}_{\text{wirt}} \text{card}(z) \stackrel{(\text{A.2}), (\text{A.6})}{=} \underbrace{\frac{1}{4} \cdot \left(-\frac{1}{2} \right)}_{=: b_{0,1}} \cdot \frac{z^3}{|z|^3} + \underbrace{\frac{1}{8}}_{=: a_{0,1}} \cdot \frac{z}{|z|}.$$

Using induction, we derive

$$\begin{aligned} \bar{\partial}_{\text{wirt}}^{\ell+1} \text{card}(z) &= a_{0,\ell} \bar{\partial}_{\text{wirt}} \left(\frac{z^\ell}{|z|^{2\ell-1}} \right) + b_{0,\ell} \bar{\partial}_{\text{wirt}} \left(\frac{z^{\ell+2}}{|z|^{2\ell+1}} \right) \\ &\stackrel{(\text{A.2})}{=} \underbrace{a_{0,\ell} \cdot \left(-\frac{2\ell-1}{2} \right)}_{=: a_{0,\ell+1}} \cdot \frac{z^{\ell+1}}{|z|^{2\ell+1}} + \underbrace{b_{0,\ell} \cdot \left(-\frac{2\ell+1}{2} \right)}_{=: b_{0,\ell+1}} \cdot \frac{z^{\ell+3}}{|z|^{2\ell+3}}, \end{aligned}$$

so the claim has been shown if $m = 0$. If we now fix any $\ell \in \mathbb{N}$ and assume that the claim holds true for some $m \in \mathbb{N}_0$ with $m < \ell$, we get

$$\begin{aligned} \partial_{\text{wirt}}^{m+1} \bar{\partial}_{\text{wirt}}^\ell \text{card}(z) &= a_{m,\ell} \partial_{\text{wirt}} \left(\frac{z^{\ell-m}}{|z|^{2\ell-1}} \right) + b_{m,\ell} \partial_{\text{wirt}} \left(\frac{z^{\ell+2-m}}{|z|^{2\ell+1}} \right) \\ &\stackrel{(\text{A.3})}{=} \underbrace{a_{m,\ell} \cdot \left(\frac{1}{2} - m \right)}_{=: a_{m+1,\ell}} \cdot \frac{z^{\ell-m-1}}{|z|^{2\ell-1}} + \underbrace{b_{m,\ell} \cdot \left(\frac{3}{2} - m \right)}_{=: b_{m+1,\ell}} \cdot \frac{z^{\ell+2-m-1}}{|z|^{2\ell+1}}, \end{aligned}$$

so the claim holds true if $m \leq \ell$.

The case $m = \ell + 1$ is split into the case $m = 1$ and $m > 1$. If $m = 1$, then $\ell = 0$ and we compute

$$\partial_{\text{wirt}} \text{card}(z) \stackrel{(\text{A.3}), (\text{A.5}), (\text{A.7})}{=} \frac{1}{2} + \frac{1}{4} \left(2 - \frac{1}{2} \right) \cdot \frac{z}{|z|} + \frac{1}{8} \cdot \frac{\bar{z}}{|z|}.$$

If $m > 1$, we get

$$\begin{aligned} \partial_{\text{wirt}}^{\ell+1} \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) &= a_{\ell,\ell} \partial_{\text{wirt}} \left(\frac{1}{|z|^{2\ell-1}} \right) + b_{\ell,\ell} \cdot \partial_{\text{wirt}} \left(\frac{z^2}{|z|^{2\ell+1}} \right) \\ &\stackrel{(A.3),(A.4)}{=} \underbrace{a_{\ell,\ell} \cdot \left(-\frac{2\ell-1}{2} \right)}_{=:a_{\ell+1,\ell}} \cdot \frac{\bar{z}}{|z|^{2\ell+1}} + \underbrace{b_{\ell,\ell} \cdot \left(2 - \frac{2\ell+1}{2} \right)}_{=:b_{\ell+1,\ell}} \cdot \frac{z}{|z|^{2\ell+1}}. \end{aligned}$$

Next, we deal with the case $m = \ell + 2$: Here, we see

$$\begin{aligned} \partial_{\text{wirt}}^{\ell+2} \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) &= \partial_{\text{wirt}} \left(\frac{1}{2} \delta_{\ell=0} + a_{\ell+1,\ell} \frac{\bar{z}}{|z|^{2\ell+1}} + b_{\ell+1,\ell} \frac{z}{|z|^{2\ell+1}} \right) \\ &\stackrel{(A.3),(A.4)}{=} \underbrace{a_{\ell+1,\ell} \cdot \left(-\frac{2\ell+1}{2} \right)}_{=:a_{\ell+2,\ell}} \cdot \frac{\bar{z}^2}{|z|^{2\ell+3}} + \underbrace{b_{\ell+1,\ell} \cdot \left(1 - \frac{2\ell+1}{2} \right)}_{=:b_{\ell+2,\ell}} \cdot \frac{1}{|z|^{2\ell+1}}. \end{aligned}$$

If we assume the claim to be true for a fixed $m \geq \ell + 2$, we get

$$\begin{aligned} \partial_{\text{wirt}}^{m+1} \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) &= a_{m,\ell} \cdot \partial_{\text{wirt}} \left(\frac{\bar{z}^{m-\ell}}{|z|^{2m-1}} \right) + b_{m,\ell} \cdot \partial_{\text{wirt}} \left(\frac{\bar{z}^{m-\ell-2}}{|z|^{2m-3}} \right) \\ &\stackrel{(A.4)}{=} \underbrace{a_{m,\ell} \cdot \left(-\frac{2m-1}{2} \right)}_{=:a_{m+1,\ell}} \cdot \frac{\bar{z}^{m+1-\ell}}{|z|^{2m+1}} + \underbrace{b_{m,\ell} \cdot \left(-\frac{2m-3}{2} \right)}_{=:b_{m+1,\ell}} \cdot \frac{\bar{z}^{m-\ell-1}}{|z|^{2m-1}}. \end{aligned}$$

Hence, using induction, we have proven the claimed identity. $\square$

The statement regarding the non-polyharmonicity of the complex cardioid is formulated in the following corollary.

Corollary A.6. *For every $z \in \mathbb{C}$ with $z \notin \mathbb{R} \cup i\mathbb{R}$ and every $m, \ell \in \mathbb{N}_0$ we have*

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) \neq 0.$$

In particular, card is smooth and non-polyharmonic on $\mathbb{C} \setminus (\mathbb{R} \cup i\mathbb{R})$. Furthermore, card is continuous on the entire complex plane.

Proof. We start with the first part of the corollary. For the following, let $z \in \mathbb{C}$ with $z \notin \mathbb{R} \cup i\mathbb{R}$, i.e., $\text{Re}(z) \neq 0$ and $\text{Im}(z) \neq 0$. Using the definition of card, we see

$$\text{card}(z) = 0 \iff z = 0 \quad \text{or} \quad \cos(\angle z) = -1 \iff z \in \mathbb{R}_{\leq 0},$$

and thus, $\text{card}(z) \neq 0$ since $\text{Im}(z) \neq 0$. In the case $m \leq \ell \neq 0$ we use Theorem A.5 to get the following chain of implications:

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) = 0 \Leftrightarrow a_{m,\ell} + b_{m,\ell} \frac{z^2}{|z|^2} = 0 \Leftrightarrow z^2 = -|z|^2 \cdot \frac{a_{m,\ell}}{b_{m,\ell}} \Rightarrow z^2 \in \mathbb{R} \Leftrightarrow z \in \mathbb{R} \cup i\mathbb{R}$$

holds, and thus $\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) \neq 0$ for $z \notin \mathbb{R} \cup i\mathbb{R}$.

For the case $m = \ell + 1 = 1$, note

$$\partial_{\text{wirt}} \text{card}(z) = 0 \Leftrightarrow \frac{1}{8} \cdot \bar{z} + \frac{3}{8} \cdot z = -\frac{|z|}{2} \Rightarrow \frac{3}{8} \text{Im}(z) - \frac{1}{8} \text{Im}(z) = 0 \Leftrightarrow \text{Im}(z) = 0,$$

and thus $\partial_{\text{wirt}} \text{card}(z) \neq 0$ since $z \notin \mathbb{R}$.

For $m = \ell + 1 > 1$, we see by considering the real- and imaginary parts that

$$\begin{aligned} \partial_{\text{wirt}}^{\ell+1} \bar{\partial}_{\text{wirt}}^{\ell} \text{card}(z) = 0 &\Leftrightarrow a_{m,\ell} \bar{z} + b_{m,\ell} z = 0 \\ &\stackrel{\text{Re}(z), \text{Im}(z) \neq 0}{\Leftrightarrow} a_{m,\ell} + b_{m,\ell} = 0 \quad \text{and} \quad -a_{m,\ell} + b_{m,\ell} = 0 \\ &\Leftrightarrow a_{m,\ell} = b_{m,\ell} = 0, \end{aligned}$$

and hence $\partial_{\text{wirt}}^{\ell+1} \bar{\partial}_{\text{wirt}}^\ell \text{card}(z) \neq 0$, since $a_{m,\ell} \neq 0 \neq b_{m,\ell}$ by Theorem A.5.

Therefore, it remains to consider the case $m \geq \ell + 2$. But here we easily see

$$\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \text{card}(z) = 0 \quad \Leftrightarrow \quad a_{m,\ell} \frac{\bar{z}^2}{|z|^2} + b_{m,\ell} = 0 \quad \Rightarrow \quad \bar{z}^2 \in \mathbb{R} \quad \Leftrightarrow \quad z \in \mathbb{R} \cup i\mathbb{R},$$

and hence $\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \text{card}(z) \neq 0$. Since all cases have been considered, the claim follows.

Regarding the second part of the corollary, first note that card is trivially continuous on $\mathbb{C} \setminus \{0\}$. Further note that

$$|\text{card}(z)| = \left| \frac{1}{2} \left(1 + \frac{\text{Re}(z)}{|z|} \right) z \right| \leq |z| \rightarrow 0$$

as $z \rightarrow 0$, showing the continuity of card on the entire complex plane $\mathbb{C}$. $\square$

B Postponed proofs concerning the approximation of polynomials

B.1 Divided Differences

Divided differences are well-known in numerical mathematics as they are for example used to calculate the coefficients of an interpolation polynomial in its Newton representation. In our case, we are interested in divided differences since they can be used to obtain a generalization of the classical mean-value theorem for differentiable functions: Given an interval $I \subseteq \mathbb{R}$ and $n + 1$ pairwise distinct data points $x_0, \dots, x_n \in I$ as well as an n -times differentiable real-valued function $f : I \rightarrow \mathbb{R}$, there exists $\xi \in (\min \{x_0, \dots, x_n\}, \max \{x_0, \dots, x_n\})$ such that

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!},$$

where the left-hand side is a divided difference of f (defined below). The classical mean-value theorem is obtained in the case $n = 1$. Our goal in this section is to generalize this result to a multivariate setting by considering divided differences in multiple variables. Such a generalization is probably well-known, but since we could not locate a convenient reference and to make the paper more self-contained, we provide a proof.

Let us first define divided differences formally. Given $n + 1$ data points $(x_0, y_0), \dots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$ with pairwise distinct x_k , we define the associated divided differences recursively via

$$\begin{aligned} [y_k] &:= y_k, \quad k \in \{0, \dots, n\}, \\ [y_k, \dots, y_{k+j}] &:= \frac{[y_{k+1}, \dots, y_{k+j}] - [y_k, \dots, y_{k+j-1}]}{x_{k+j} - x_k}, \quad j \in \{1, \dots, n\}, \quad k \in \{0, \dots, n-j\}. \end{aligned}$$

If the data points are defined by a function f (i.e. $y_k = f(x_k)$ for all $k \in \{0, \dots, n\}$), we write

$$[x_k, \dots, x_{k+j}] f := [y_k, \dots, y_{k+j}].$$

We first consider an alternative representation of divided differences, the so-called *Hermite-Genocchi-Formula*. To state it, we introduce the notation Σ^k to denote a certain k -dimensional simplex.

Definition B.1. Let $s \in \mathbb{N}$. Then we define

$$\Sigma^s := \left\{ x \in \mathbb{R}^s : x_\ell \geq 0 \text{ for all } \ell \text{ and } \sum_{\ell=1}^s x_\ell \leq 1 \right\}.$$

We further set $\Sigma^0 := \{0\} \subseteq \mathbb{R}$. Denoting by λ^s the s -dimensional Lebesgue-measure (and by λ^0 the counting measure), the identity $\lambda^s(\Sigma^s) = \frac{1}{s!}$ holds true; see for instance [42] and note that the case $s = 0$ can be seen directly.

In the following, we consider integrals over the sets Σ^s . These integrals are always formed with respect to the measure λ^s . However, we only write, e.g., $\int_{\Sigma^s} f(x) dx$, with the implicit understanding that the integral is formed with respect to the measure λ^s . Using this convention, we can now consider the alternative representation of divided differences.

Lemma B.2 (Hermite-Genocchi-Formula). *Let $k \in \mathbb{N}_0$. For real numbers $a, b \in \mathbb{R}$ with $a < b$, a function $f \in C^k([a, b]; \mathbb{R})$ (where $C^0([a, b]; \mathbb{R})$ denotes the set of continuous functions) and pairwise distinct $x_0, \dots, x_k \in [a, b]$, the divided difference of f at the data points $x_0, \dots, x_k$ satisfies the identity*

$$[x_0, \dots, x_k] f = \int_{\Sigma^k} f^{(k)} \left(x_0 + \sum_{\ell=1}^k s_\ell (x_\ell - x_0) \right) ds. \quad (\text{B.1})$$

Proof. The case $k = 0$ follows directly from $[x_0]f = f(x_0)$. For the case $k > 0$, see [6, Theorem 3.3]. $\square$

We will make use of the above formula by combining it with the following generalization of the mean-value theorem for integrals.

Lemma B.3. *Let D be a connected and compact topological space. Let $\mathcal{A}$ be some σ -algebra over D and let $\mu : \mathcal{A} \rightarrow [0, \infty)$ be a finite measure on D with $\mu(D) > 0$. Let $f : D \rightarrow \mathbb{R}$ be a continuous function with respect to the standard topology on $\mathbb{R}$ and the topology on D . Moreover, let f be measurable with respect to $\mathcal{A}$ and the Borel σ -algebra on $\mathbb{R}$. Then there exists $\xi \in D$ such that*

$$f(\xi) = \frac{1}{\mu(D)} \cdot \int_D f(x) d\mu(x).$$

Proof. Since D is compact and connected and f continuous, it follows that $f(D) \subset \mathbb{R}$ is compact and connected, and hence $f(D)$ is a compact interval in $\mathbb{R}$. Therefore, there exist $x_{\min} \in D$ and $x_{\max} \in D$ satisfying

$$f(x_{\min}) \leq f(x) \leq f(x_{\max})$$

for all $x \in D$. Thus, one gets

$$\begin{aligned} f(x_{\min}) &= \frac{1}{\mu(D)} \int_D f(x_{\min}) d\mu(x) \leq \frac{1}{\mu(D)} \int_D f(x) d\mu(x) \\ &\leq \frac{1}{\mu(D)} \int_D f(x_{\max}) d\mu(x) = f(x_{\max}). \end{aligned}$$

The claim now follows from the fact that $f(D)$ is an interval. $\square$

We also get a convenient representation of divided differences for the case of equidistant data points.

Lemma B.4. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}$, $k \in \mathbb{N}_0$ and $h > 0$. We consider the case of equidistant data points, meaning $x_j := x_0 + jh$ for all $j = 1, \dots, k$. In this case, we have the formula*

$$[x_0, \dots, x_k] f = \frac{1}{k!h^k} \cdot \sum_{r=0}^k (-1)^{k-r} \binom{k}{r} f(x_r). \quad (\text{B.2})$$

Proof. We prove the result via induction over the number j of considered data points, meaning the following: For all $j \in \{0, \dots, k\}$ we have

$$[x_\ell, \dots, x_{\ell+j}] f = \frac{1}{j!h^j} \cdot \sum_{r=0}^j (-1)^{j-r} \binom{j}{r} f(x_{\ell+r})$$

for all $\ell \in \{0, \dots, k\}$ satisfying $\ell + j \leq k$. The case $j = 0$ is trivial. Therefore, we assume the claim to be true for a fixed $j \in \{0, \dots, k-1\}$, and let $\ell \in \{0, \dots, k\}$ be arbitrary with $\ell + j + 1 \leq k$. We then get

$$\begin{aligned} [x_\ell, \dots, x_{\ell+j+1}] f &= \frac{[x_{\ell+1}, \dots, x_{\ell+j+1}] f - [x_\ell, \dots, x_{\ell+j}] f}{x_{\ell+j+1} - x_\ell} \\ &\stackrel{\text{I.H.}}{=} \frac{1}{j!h^j} \cdot \frac{\sum_{r=0}^j (-1)^{j-r} \binom{j}{r} (f(x_{\ell+r+1}) - f(x_{\ell+r}))}{(j+1)h} \\ &= \frac{1}{(j+1)!h^{j+1}} \sum_{r=0}^j (-1)^{j-r} \binom{j}{r} (f(x_{\ell+r+1}) - f(x_{\ell+r})). \end{aligned}$$

Using an index shift, we deduce

$$\begin{aligned}
& \sum_{r=0}^j (-1)^{j-r} \binom{j}{r} f(x_{\ell+r+1}) - \sum_{r=0}^j (-1)^{j-r} \binom{j}{r} f(x_{\ell+r}) \\
&= \sum_{r=1}^{j+1} (-1)^{j+1-r} \binom{j}{r-1} f(x_{\ell+r}) + \sum_{r=0}^j (-1)^{j+1-r} \binom{j}{r} f(x_{\ell+r}) \\
&= (-1)^{j+1} f(x_{\ell}) + \sum_{r=1}^j \left((-1)^{j+1-r} f(x_{\ell+r}) \left[\binom{j}{r-1} + \binom{j}{r} \right] \right) + f(x_{\ell+j+1}) \\
&= \sum_{r=0}^{j+1} (-1)^{j+1-r} \binom{j+1}{r} f(x_{\ell+r}),
\end{aligned}$$

which yields the claim. $\square$

The final result for divided differences, which is the result that is actually used in the proof of Theorem 3.1 in Appendix B.3, reads as follows:

Theorem B.5. *Let $f : \mathbb{R}^s \rightarrow \mathbb{R}$ and $k \in \mathbb{N}_0, r > 0$, such that $f|_{(-r,r)^s} \in C^k((-r,r)^s; \mathbb{R})$. For $\mathbf{p} \in \mathbb{N}_0^s$ with $|\mathbf{p}| \leq k$ and $h > 0$ let*

$$f_{\mathbf{p},h} := (2h)^{-|\mathbf{p}|} \sum_{0 \leq \mathbf{r} \leq \mathbf{p}} (-1)^{|\mathbf{p}|-|\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} f(h(2\mathbf{r} - \mathbf{p})).$$

Let $m := \max_j \mathbf{p}_j$. Then, for $0 < h < \frac{r}{\max\{1,m\}}$ there exists $\xi \in h[-m, m]^s$ satisfying

$$f_{\mathbf{p},h} = \partial^{\mathbf{p}} f(\xi).$$

Proof. We may assume $m > 0$, since $m = 0$ implies $\mathbf{p} = 0$ and thus $f_{\mathbf{p},h} = f(0)$, so that the claim holds for $\xi = 0$.

We prove via induction over $s \in \mathbb{N}$ that the formula

$$f_{\mathbf{p},h} = \mathbf{p}! \int_{\Sigma^{\mathbf{p}_s}} \int_{\Sigma^{\mathbf{p}_{s-1}}} \cdots \int_{\Sigma^{\mathbf{p}_1}} \partial^{\mathbf{p}} f \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \sigma_{\ell}^{(1)}, \dots, -h\mathbf{p}_s + 2h \sum_{\ell=1}^{\mathbf{p}_s} \ell \sigma_{\ell}^{(s)} \right) d\sigma^{(1)} \cdots d\sigma^{(s)} \quad (\text{B.3})$$

holds for all $\mathbf{p} \in \mathbb{N}_0^s$ with $1 \leq |\mathbf{p}| \leq k$ and all $0 < h < \frac{r}{m}$. The case $s = 1$ is exactly the Hermite-Genocchi-Formula (B.1), combined with (B.2) applied to the data points

$$-hp, -hp + 2h, \dots, hp - 2h, hp.$$

By induction, assume that the claim holds for some $s \in \mathbb{N}$. For a fixed point $y \in (-r, r)$, let

$$f_y : (-r, r)^s \rightarrow \mathbb{R}, \quad x \mapsto f(x, y).$$

For $\mathbf{p} \in \mathbb{N}_0^{s+1}$ with $|\mathbf{p}| \leq k$ and $\mathbf{p}' := (p_1, \dots, p_s)$, we define

$$\Gamma : (-r, r) \rightarrow \mathbb{R}, \quad y \mapsto (f_y)_{\mathbf{p}',h} = (2h)^{-|\mathbf{p}'|} \sum_{0 \leq \mathbf{r}' \leq \mathbf{p}'} (-1)^{|\mathbf{p}'|-|\mathbf{r}'|} \binom{\mathbf{p}'}{\mathbf{r}'} f(h(2\mathbf{r}' - \mathbf{p}'), y).$$

Using the induction hypothesis, we get

$$\begin{aligned}
& \Gamma(y) \\
&= \mathbf{p}'! \int_{\Sigma^{\mathbf{p}_s}} \int_{\Sigma^{\mathbf{p}_{s-1}}} \cdots \int_{\Sigma^{\mathbf{p}_1}} \partial^{(\mathbf{p}',0)} f \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \sigma_{\ell}^{(1)}, \dots, -h\mathbf{p}_s + 2h \sum_{\ell=1}^{\mathbf{p}_s} \ell \sigma_{\ell}^{(s)}, y \right) d\sigma^{(1)} \cdots d\sigma^{(s)}
\end{aligned}$$

for all $y \in (-r, r)$. Furthermore, we calculate

$$\begin{aligned}
& \mathbf{p}_{s+1}! \cdot [-h \cdot \mathbf{p}_{s+1}, -h \cdot \mathbf{p}_{s+1} + 2h, \dots, h \cdot \mathbf{p}_{s+1}] \Gamma \\
& \stackrel{(B.2)}{=} (2h)^{-\mathbf{p}_{s+1}} \sum_{r'=0}^{\mathbf{p}_{s+1}} (-1)^{\mathbf{p}_{s+1}-r'} \binom{\mathbf{p}_{s+1}}{r'} \Gamma(h(2r' - \mathbf{p}_{s+1})) \\
& = (2h)^{-\mathbf{p}_{s+1}} \left[\sum_{r'=0}^{\mathbf{p}_{s+1}} (-1)^{\mathbf{p}_{s+1}-r'} \binom{\mathbf{p}_{s+1}}{r'} (2h)^{-|\mathbf{p}'|} \right. \\
& \quad \left. \times \sum_{0 \leq \mathbf{r}' \leq \mathbf{p}'} (-1)^{|\mathbf{p}'|-|\mathbf{r}'|} \binom{\mathbf{p}'}{\mathbf{r}'} f(h(2\mathbf{r}' - \mathbf{p}'), h(2r' - \mathbf{p}_{s+1})) \right] \\
& = (2h)^{-|\mathbf{p}|} \sum_{0 \leq \mathbf{r} \leq \mathbf{p}} (-1)^{|\mathbf{p}|-|\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} f(h(2\mathbf{r} - \mathbf{p})) \\
& = f_{\mathbf{p},h}.
\end{aligned}$$

On the other hand, we get

$$\begin{aligned}
& [-h \cdot \mathbf{p}_{s+1}, -h \cdot \mathbf{p}_{s+1} + 2h, \dots, h \cdot \mathbf{p}_{s+1}] \Gamma \\
& \stackrel{(B.1)}{=} \int_{\Sigma^{\mathbf{p}_{s+1}}} \Gamma^{(\mathbf{p}_{s+1})} \left(-h\mathbf{p}_{s+1} + 2h \sum_{\ell=1}^{\mathbf{p}_{s+1}} \ell \sigma_{\ell}^{(s+1)} \right) d\sigma^{(s+1)} \\
& = \mathbf{p}'! \int_{\Sigma^{\mathbf{p}_{s+1}}} \dots \int_{\Sigma^{\mathbf{p}_1}} \partial^{\mathbf{p}} f \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \sigma_{\ell}^{(1)}, \dots, -h\mathbf{p}_{s+1} + 2h \sum_{\ell=1}^{\mathbf{p}_{s+1}} \ell \sigma_{\ell}^{(s+1)} \right) d\sigma^{(1)} \dots d\sigma^{(s+1)}.
\end{aligned}$$

Interchanging the order of integration and derivative is possible, since we integrate on compact sets and only consider continuously differentiable functions (see, e.g., [11, Lemma 16.2]).

We have thus proven (B.3) using the principle of induction. The claim of the theorem then follows directly using the mean-value theorem for integrals (Lemma B.3) applied to the topological space

$$D := \Sigma^{\mathbf{p}_1} \times \dots \times \Sigma^{\mathbf{p}_{s+1}}$$

equipped with the product topology where each factor is endowed with the standard topology on $\Sigma^{\mathbf{p}_\ell}$ (where the standard topology on Σ^0 is the discrete topology), the measure

$$\mu := \lambda^{\mathbf{p}_1} \otimes \dots \otimes \lambda^{\mathbf{p}_{s+1}}$$

defined on the product of the Borel σ -algebras on $\Sigma^{\mathbf{p}_\ell}$ (and the σ -algebra of $\{0\}$ is its power set) and to the function

$$(\sigma^{(1)}, \dots, \sigma^{(s+1)}) \mapsto \partial^{\mathbf{p}} f \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \sigma_{\ell}^{(1)}, \dots, -h\mathbf{p}_{s+1} + 2h \sum_{\ell=1}^{\mathbf{p}_{s+1}} \ell \sigma_{\ell}^{(s+1)} \right).$$

Moreover, note that all the simplices $\Sigma^{\mathbf{p}_\ell}$ are compact and connected (in fact convex) with

$$\lambda^{\mathbf{p}_1}(\Sigma^{\mathbf{p}_1}) \dots \lambda^{\mathbf{p}_{s+1}}(\Sigma^{\mathbf{p}_{s+1}}) = \frac{1}{\mathbf{p}'!},$$

see Definition B.1. Therefore, Lemma B.3 yields the existence of a certain $(\xi^{(1)}, \dots, \xi^{(s+1)}) \in D$ with

$$f_{\mathbf{p},h} = \partial^{\mathbf{p}} f \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \xi_{\ell}^{(1)}, \dots, -h\mathbf{p}_{s+1} + 2h \sum_{\ell=1}^{\mathbf{p}_{s+1}} \ell \xi_{\ell}^{(s+1)} \right).$$

Hence, the claim follows letting

$$\xi := \left(-h\mathbf{p}_1 + 2h \sum_{\ell=1}^{\mathbf{p}_1} \ell \xi_{\ell}^{(1)}, \dots, -h\mathbf{p}_{s+1} + 2h \sum_{\ell=1}^{\mathbf{p}_{s+1}} \ell \xi_{\ell}^{(s+1)} \right) \in h[-m, m]^s. \quad \square$$

B.2 Proof of Proposition 2.1

Proof of Proposition 2.1. Let $M \in \mathbb{N}_0$. Since ϕ is not polyharmonic we can pick $z \in U$ with $\Delta^M \phi(z) \neq 0$. By continuity we can choose $\delta > 0$ with $B_\delta(z) \subseteq U$ and $\Delta^M \phi(w) \neq 0$ for all $w \in B_\delta(z)$. For all $m, \ell \in \mathbb{N}_0$ let

$$A_{m,\ell} := \left\{ w \in B_\delta(z) : \partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \phi(w) = 0 \right\}$$

and assume towards a contradiction that

$$\bigcup_{m,\ell \leq M} A_{m,\ell} = B_\delta(z).$$

By [3, Corollary 3.35], $B_\delta(z)$ with its usual topology is completely metrizable. By continuity of $\partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \phi$ the sets $A_{m,\ell}$ are closed in $B_\delta(z)$. Hence, using the Baire category theorem [3, Theorems 3.46 and 3.47], there are $m, \ell \in \mathbb{N}_0$ with $m, \ell \leq M$, $z' \in A_{m,\ell}$ and $\varepsilon > 0$ such that

$$B_\varepsilon(z') \subseteq A_{m,\ell} \subseteq B_\delta(z).$$

But thanks to $\Delta^M \phi = 4^M \partial_{\text{wirt}}^M \bar{\partial}_{\text{wirt}}^M \phi = 4^M \partial_{\text{wirt}}^{M-\ell} \bar{\partial}_{\text{wirt}}^{M-m} \partial_{\text{wirt}}^\ell \bar{\partial}_{\text{wirt}}^m \phi$ (see property 6 in Appendix A.1), this directly implies $\Delta^M \phi \equiv 0$ on $B_\varepsilon(z')$ in contradiction to the choice of $B_\delta(z)$. $\square$

B.3 Proof of Theorem 3.1

The following section is dedicated to proving Theorem 3.1. We are going to show that it is possible to approximate complex polynomials in z and $\bar{z}$ arbitrarily well on Ω_n using shallow complex-valued neural networks. To do so, we follow the proof sketch given after the statement of Theorem 3.1, starting with the following lemma.

Lemma B.6. *Let $\phi : \mathbb{C} \rightarrow \mathbb{C}$ and $\delta > 0$, $b \in \mathbb{C}$, $k \in \mathbb{N}_0$, such that $\phi|_{B_\delta(b)} \in C^k(B_\delta(b); \mathbb{C})$. For fixed $z \in \Omega_n$, where we recall that $\Omega_n = [-1, 1]^n + i[-1, 1]^n$, we consider the map*

$$\phi_z : B_{\frac{\delta}{\sqrt{2n}}}(0) \rightarrow \mathbb{C}, \quad w \mapsto \phi(w^T z + b),$$

which is in C^k since for $w \in B_{\frac{\delta}{\sqrt{2n}}}(0) \subseteq \mathbb{C}^n$ we have

$$|w^T z| \leq \|w\|_2 \cdot \|z\|_2 < \frac{\delta}{\sqrt{2n}} \cdot \sqrt{2n} = \delta.$$

For all multi-indices $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $|\mathbf{m} + \ell| \leq k$ we have

$$\partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^\ell \phi_z(w) = z^{\mathbf{m}} \bar{z}^\ell \cdot \left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b)$$

for all $w \in B_{\frac{\delta}{\sqrt{2n}}}(0)$.

Proof. First we prove the statement

$$\bar{\partial}_{\text{wirt}}^\ell \phi_z(w) = \bar{z}^\ell \cdot \left(\bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b) \quad \text{for all } w \in B_{\frac{\delta}{\sqrt{2n}}}(0) \quad (\text{B.4})$$

by induction over $0 \leq |\ell| \leq k$. The case $\ell = 0$ is trivial. Assuming that (B.4) holds for fixed $\ell \in \mathbb{N}_0^n$ with $|\ell| < k$, we want to show

$$\bar{\partial}_{\text{wirt}}^{\ell+e_j} \phi_z(w) = \bar{z}^{\ell+e_j} \cdot \left(\bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi \right) (w^T z + b) \quad (\text{B.5})$$

for all $w \in B_{\frac{\delta}{\sqrt{2n}}}(0)$, where $j \in \{1, \dots, n\}$ is chosen arbitrarily. To this end, first note

$$\begin{aligned} \bar{\partial}_{\text{wirt}}^{\ell+e_j} \phi_z(w) &= \bar{\partial}_{\text{wirt}}^{e_j} \bar{\partial}_{\text{wirt}}^\ell \phi_z(w) \stackrel{\text{induction}}{=} \bar{\partial}_{\text{wirt}}^{e_j} \left[w \mapsto \bar{z}^\ell \cdot \left(\bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b) \right] \\ &= \bar{z}^\ell \bar{\partial}_{\text{wirt}}^{e_j} \left[w \mapsto \left(\bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (w^T z + b) \right]. \end{aligned}$$

Applying the chain rule for Wirtinger derivatives and using that

$$\bar{\partial}_{\text{wirt}}^{e_j} [w \mapsto w^T z + b] = 0$$

since $w \mapsto w^T z + b$ is holomorphic in every variable, we see

$$\begin{aligned} \bar{\partial}_{\text{wirt}}^{e_j} [w \mapsto (\bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b)] &= (\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b) \cdot \bar{\partial}_{\text{wirt}}^{e_j} [w \mapsto w^T z + b] \\ &\quad + (\bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi)(w^T z + b) \cdot \bar{\partial}_{\text{wirt}}^{e_j} [w \mapsto \overline{w^T z + b}] \\ &= (\bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi)(w^T z + b) \cdot \overline{\partial_{\text{wirt}}^{e_j} [w \mapsto w^T z + b]} \\ &= \bar{z}^{e_j} \cdot (\bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi)(w^T z + b), \end{aligned}$$

using the fact that $w_j \mapsto w^T z + b$ is holomorphic and hence

$$\bar{\partial}_{\text{wirt}}^{e_j} [w \mapsto w^T z + b] = 0 \quad \text{and} \quad \partial_{\text{wirt}}^{e_j} [w \mapsto w^T z + b] = z_j.$$

Thus, we have proven (B.5) and induction yields (B.4).

It remains to show the full claim. We use induction over $|\mathbf{m}|$ and note that the case $\mathbf{m} = 0$ has just been shown. We assume that the claim holds true for fixed $\mathbf{m} \in \mathbb{N}_0^n$ with $|\mathbf{m} + \ell| < k$ and choose $j \in \{1, \dots, n\}$. Thus, we get

$$\begin{aligned} \partial_{\text{wirt}}^{\mathbf{m}+e_j} \bar{\partial}_{\text{wirt}}^\ell \phi_z(w) &= \partial_{\text{wirt}}^{e_j} \partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^\ell \phi_z(w) \stackrel{\text{IH}}{=} \partial_{\text{wirt}}^{e_j} \left(w \mapsto z^{\mathbf{m}} \bar{z}^\ell \cdot (\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b) \right) \\ &= z^{\mathbf{m}} \bar{z}^\ell \cdot \partial_{\text{wirt}}^{e_j} [w \mapsto (\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b)]. \end{aligned}$$

Using the chain rule again, we calculate

$$\begin{aligned} &\partial_{\text{wirt}}^{e_j} [w \mapsto (\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b)] \\ &= (\partial_{\text{wirt}}^{|\mathbf{m}|+1} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b) \cdot \partial_{\text{wirt}}^{e_j} [w \mapsto w^T z + b] \\ &\quad + (\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi)(w^T z + b) \cdot \partial_{\text{wirt}}^{e_j} [w \mapsto \overline{w^T z + b}] \\ &= z^{e_j} \cdot (\partial_{\text{wirt}}^{|\mathbf{m}|+1} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b) + (\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|+1} \phi)(w^T z + b) \cdot \overline{\partial_{\text{wirt}}^{e_j} [w \mapsto w^T z + b]} \\ &= z^{e_j} \cdot (\partial_{\text{wirt}}^{|\mathbf{m}|+1} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi)(w^T z + b). \end{aligned}$$

By induction, this proves the claim. $\square$

As the last preparation for the proof of Theorem 3.1, we need the following lemma.

Lemma B.7. *Let $\phi : \mathbb{C} \rightarrow \mathbb{C}$ and $\delta > 0$, $b \in \mathbb{C}$, $k \in \mathbb{N}_0$, such that $\phi|_{B_\delta(b)} \in C^k(B_\delta(b); \mathbb{C})$. Let $m, n \in \mathbb{N}$ and $\varepsilon > 0$. For $\mathbf{p} \in \mathbb{N}_0^{2n}$, $h > 0$ and $z \in \Omega_n$ we write*

$$\begin{aligned} \Phi_{\mathbf{p},h}(z) &:= (2h)^{-|\mathbf{p}|} \sum_{0 \leq \mathbf{r} \leq \mathbf{p}} (-1)^{|\mathbf{p}|-|\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} (\phi_z \circ \varphi_n)(h(2\mathbf{r} - \mathbf{p})) \\ &= (2h)^{-|\mathbf{p}|} \sum_{0 \leq \mathbf{r} \leq \mathbf{p}} (-1)^{|\mathbf{p}|-|\mathbf{r}|} \binom{\mathbf{p}}{\mathbf{r}} \phi \left([\varphi_n(h(2\mathbf{r} - \mathbf{p}))]^T \cdot z + b \right), \end{aligned}$$

where ϕ_z is as introduced in Lemma B.6 and φ_n is as in (A.1). Furthermore, let

$$\phi_{\mathbf{p}} : \Omega_n \times B_{\frac{\delta}{\sqrt{2n}}}(0) \rightarrow \mathbb{C}, \quad (z, w) \mapsto \partial^{\mathbf{p}} \phi_z(w).$$

Then there exists $h^* > 0$ such that

$$\|\Phi_{\mathbf{p},h} - \phi_{\mathbf{p}}(\cdot, 0)\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \varepsilon$$

for all $\mathbf{p} \in \mathbb{N}_0^{2n}$ with $|\mathbf{p}| \leq k$ and $\mathbf{p} \leq m$ and $h \in (0, h^*)$.

Proof. Fix $\mathbf{p} \in \mathbb{N}_0^{2n}$ with $|\mathbf{p}| \leq k$ and $\mathbf{p} \leq m$. The map

$$B_{\sqrt{2n+1}}(0) \times B_{\delta/(\sqrt{2n+1})}(0) \rightarrow \mathbb{C}, \quad (z, w) \mapsto \phi(w^T z + b)$$

is in C^k since

$$|w^T z| \leq \|w\| \cdot \|z\| < \frac{\delta}{\sqrt{2n+1}} \cdot (\sqrt{2n+1}) = \delta.$$

Therefore, the map

$$B_{\sqrt{2n+1}}(0) \times B_{\delta/(\sqrt{2n+1})}(0) \rightarrow \mathbb{C}, \quad (z, w) \mapsto \partial^{\mathbf{p}} \phi_z(w)$$

is continuous and in particular uniformly continuous on the compact set

$$\Omega_n \times \overline{B_{\delta/(3n)}(0)} \subseteq B_{\sqrt{2n+1}}(0) \times B_{\delta/(\sqrt{2n+1})}(0).$$

Here, we employed $\sqrt{2n+1} < 3n$ for every $n \in \mathbb{N}$. Hence, there exists $h_{\mathbf{p}} \in (0, \frac{\delta}{3n \cdot \sqrt{2n+1}})$, such that

$$|\phi_{\mathbf{p}}(z, \xi) - \phi_{\mathbf{p}}(z, 0)| \leq \frac{\varepsilon}{\sqrt{2}}$$

for all $\xi \in \varphi_n(h \cdot [-m, m]^{2n})$, $h \in (0, h_{\mathbf{p}})$ and $z \in \Omega_n$. Now fix such an $h \in (0, h_{\mathbf{p}})$ and $z \in \Omega_n$. Applying Theorem B.5 to both components of $(\varphi_1^{-1} \circ \Phi_{\mathbf{p}, h})(z)$ and $\varphi_1^{-1} \circ \phi_z \circ \varphi_n|_{(-\frac{\delta}{3n}, \frac{\delta}{3n})^{2n}}$ separately yields the existence of two real vectors $\xi_1, \xi_2 \in h \cdot [-m, m]^{2n}$, such that

$$\begin{aligned} (\varphi_1^{-1} \circ \Phi_{\mathbf{p}, h}(z))_1 &= [\partial^{\mathbf{p}} (\varphi_1^{-1} \circ \phi_z \circ \varphi_n)(\xi_1)]_1 \\ \text{and } (\varphi_1^{-1} \circ \Phi_{\mathbf{p}, h}(z))_2 &= [\partial^{\mathbf{p}} (\varphi_1^{-1} \circ \phi_z \circ \varphi_n)(\xi_2)]_2. \end{aligned}$$

Rewriting this yields

$$\operatorname{Re}(\Phi_{\mathbf{p}, h}(z)) = \operatorname{Re}(\phi_{\mathbf{p}}(z, \varphi_n(\xi_1))) \quad \text{and} \quad \operatorname{Im}(\Phi_{\mathbf{p}, h}(z)) = \operatorname{Im}(\phi_{\mathbf{p}}(z, \varphi_n(\xi_2))).$$

Using this property, we deduce

$$|\operatorname{Re}(\Phi_{\mathbf{p}, h}(z) - \phi_{\mathbf{p}}(z, 0))| = |\operatorname{Re}(\phi_{\mathbf{p}}(z, \varphi_n(\xi_1)) - \phi_{\mathbf{p}}(z, 0))| \leq |\phi_{\mathbf{p}}(z, \varphi_n(\xi_1)) - \phi_{\mathbf{p}}(z, 0)| \leq \frac{\varepsilon}{\sqrt{2}}$$

and analogously for the imaginary part. Since $z \in \Omega_n$ and $h \in (0, h_{\mathbf{p}})$ have been chosen arbitrarily we get the claim by choosing

$$h^* := \min \{h_{\mathbf{p}} : \mathbf{p} \in \mathbb{N}_0^{2n} \text{ with } |\mathbf{p}| \leq k \text{ and } \mathbf{p} \leq m\}.$$

□

Using the previous two lemmas and the results from Appendix B.1 and Appendix B.2, we can now prove Theorem 3.1.

Proof of Theorem 3.1. Let $b \in U$ satisfy

$$\partial_{\text{wirt}}^{\ell_1} \bar{\partial}_{\text{wirt}}^{\ell_2} \phi(b) \neq 0 \quad \text{for all } \ell_1, \ell_2 \in \mathbb{N}_0 \text{ with } \ell_1, \ell_2 \leq mn.$$

Such a point b exists according to Proposition 2.1. Let $p \in \mathcal{P}'$ and fix $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $\mathbf{m}, \ell \leq m$. For each $z \in \Omega_n$, using Lemma B.6, we then have

$$\begin{aligned} z^{\mathbf{m}} \bar{z}^{\ell} &= \left[\left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b) \right]^{-1} \partial_{\text{wirt}}^{\mathbf{m}} \bar{\partial}_{\text{wirt}}^{\ell} \phi_z(0) \\ &\stackrel{\text{Prop. A.1}}{=} \left[\left(\partial_{\text{wirt}}^{|\mathbf{m}|} \bar{\partial}_{\text{wirt}}^{|\ell|} \phi \right) (b) \right]^{-1} \cdot \sum_{\substack{\mathbf{p}=(\mathbf{p}', \mathbf{p}'') \in \mathbb{N}_0^{2n} \\ \mathbf{p}' + \mathbf{p}'' = \mathbf{m} + \ell}} b_{\mathbf{p}', \mathbf{p}''} \partial^{(\mathbf{p}', \mathbf{p}'')} \phi_z(0) \end{aligned} \quad (\text{B.6})$$

with suitably chosen complex coefficients $b_{\mathbf{p}', \mathbf{p}''} \in \mathbb{C}$ depending only on $\mathbf{p}', \mathbf{p}'', \mathbf{m}$ and ℓ . Here we used that $|\mathbf{m}|, |\ell| \leq mn$. Since $\mathcal{P}' \subseteq \mathcal{P}_m^n$ is bounded and $p \in \mathcal{P}'$, we can write

$$p(z) = \sum_{\substack{\mathbf{m}, \ell \in \mathbb{N}_0^n \\ \mathbf{m}, \ell \leq m}} a_{\mathbf{m}, \ell} z^{\mathbf{m}} \bar{z}^{\ell}$$

with $|a_{\mathbf{m},\ell}| \leq c$ for some constant $c = c(\mathcal{P}') > 0$. In combination with (B.6), this easily implies that we can rewrite p as

$$p(z) = \sum_{\substack{\mathbf{p} \in \mathbb{N}_0^{2n} \\ \mathbf{p} \leq 2m}} c_{\mathbf{p}}(p) \partial^{\mathbf{p}} \phi_z(0) \quad (\text{B.7})$$

with coefficients $c_{\mathbf{p}}(p) \in \mathbb{C}$ satisfying $|c_{\mathbf{p}}(p)| \leq c'$ for some constant $c' = c'(\phi, b, \mathcal{P}', m, n)$. By Lemma B.7, we choose $h^* > 0$, such that

$$|\Phi_{\mathbf{p},h^*}(z) - \partial^{\mathbf{p}} \phi_z(0)| \leq \frac{\varepsilon}{\sum_{\mathbf{q} \in \mathbb{N}_0^{2n}, \mathbf{q} \leq 2m} c'}$$

for all $z \in \Omega_n$ and $\mathbf{p} \in \mathbb{N}_0^{2n}$ with $\mathbf{p} \leq 2m$. Furthermore, we can rewrite each function $\Phi_{\mathbf{p},h^*}$ as

$$\Phi_{\mathbf{p},h^*}(z) = \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{Z}^{2n} \\ |\boldsymbol{\alpha}_j| \leq 2m \ \forall j}} \lambda_{\boldsymbol{\alpha},\mathbf{p}} \phi([\varphi_n(h^* \boldsymbol{\alpha})]^T \cdot z + b)$$

with suitable coefficients $\lambda_{\boldsymbol{\alpha},\mathbf{p}} \in \mathbb{C}$. Since the cardinality of the set

$$\{\boldsymbol{\alpha} \in \mathbb{Z}^{2n} : |\boldsymbol{\alpha}_j| \leq 2m \ \forall j\}$$

is $(4m+1)^{2n}$, this can be converted to

$$\Phi_{\mathbf{p},h^*}(z) = \sum_{j=1}^N \lambda_{j,\mathbf{p}} \phi(\rho_j^T \cdot z + b).$$

For p as in (B.7), we then define

$$\theta(z) := \sum_{\substack{\mathbf{p} \in \mathbb{N}_0^{2n} \\ \mathbf{p} \leq 2m}} c_{\mathbf{p}}(p) \cdot \Phi_{\mathbf{p},h^*}(z) = \sum_{j=1}^N \left[\left(\sum_{\substack{\mathbf{p} \in \mathbb{N}_0^{2n} \\ \mathbf{p} \leq 2m}} c_{\mathbf{p}}(p) \lambda_{j,\mathbf{p}} \right) \phi(\rho_j^T \cdot z + b) \right]$$

and note

$$|\theta(z) - p(z)| \leq \sum_{\mathbf{p} \leq 2m} |c_{\mathbf{p}}(p)| \cdot |\Phi_{\mathbf{p},h^*}(z) - \partial^{\mathbf{p}} \phi_z(0)| \leq \varepsilon.$$

Since the coefficients ρ_j have been chosen independently of the polynomial p , we can rewrite θ in the desired form. $\square$

C Postponed proofs for the approximation of C^k -functions

C.1 Prerequisites from Fourier Analysis

This section is dedicated to reviewing some notations and results from Fourier Analysis. In the end, a quantitative result for the approximation of $C^k([-1, 1]^s; \mathbb{R})$ -functions using linear combinations of multivariate Chebyshev polynomials is derived; see Theorem C.15.

We start by recalling several notations and concepts from Fourier Analysis.

Definition C.1. Let $s \in \mathbb{N}$ and $k \in \mathbb{N}_0$. We define

$$C_{2\pi}^k(\mathbb{R}^s; \mathbb{C}) := \{f \in C^k(\mathbb{R}^s; \mathbb{C}) : \forall \mathbf{p} \in \mathbb{Z}^s \ \forall x \in \mathbb{R}^s : f(x + 2\pi \mathbf{p}) = f(x)\}.$$

and $C_{2\pi}(\mathbb{R}^s; \mathbb{C}) := C_{2\pi}^0(\mathbb{R}^s; \mathbb{C})$. For a function $f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$ we write

$$\|f\|_{C^k([- \pi, \pi]^s; \mathbb{C})} := \max_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ |\mathbf{k}| \leq k}} \|\partial^{\mathbf{k}} f\|_{L^\infty([- \pi, \pi]^s; \mathbb{C})} \quad \text{and}$$

$$\|f\|_{L^p([- \pi, \pi]^s; \mathbb{C})} := \left(\frac{1}{(2\pi)^s} \cdot \int_{[- \pi, \pi]^s} |f(x)|^p dx \right)^{1/p} \quad \text{for } p \in [1, \infty).$$

Moreover, we set $\|f\|_{L^\infty([- \pi, \pi]^s; \mathbb{R})} := \|f\|_{C^0([- \pi, \pi]^s; \mathbb{C})}$.

Definition C.2. For any $s \in \mathbb{N}$ and $\mathbf{k} \in \mathbb{Z}^s$, we write

$$e_{\mathbf{k}} : \mathbb{R}^s \rightarrow \mathbb{C}, \quad e_{\mathbf{k}}(x) = e^{i\langle \mathbf{k}, x \rangle}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product of two vectors. For any $f \in C_{2\pi}(\mathbb{R}^s; \mathbb{C})$ we define the $\mathbf{k}$ -th Fourier coefficient of f to be

$$\hat{f}(\mathbf{k}) := \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} f(x) \overline{e_{\mathbf{k}}(x)} dx.$$

Definition C.3. For two functions $f, g \in C_{2\pi}(\mathbb{R}^s; \mathbb{C})$, we define their convolution as

$$f * g : \mathbb{R}^s \rightarrow \mathbb{C}, \quad (f * g)(x) := \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} f(t) g(x - t) dt.$$

In the following we define several so-called kernels.

Definition C.4. Let $m \in \mathbb{N}_0$ be arbitrary.

1. The m -th one-dimensional *Dirichlet-kernel* is defined as

$$D_m := \sum_{h=-m}^m e_h.$$

2. The m -th one-dimensional *Fejér-kernel* is defined as

$$F_m := \frac{1}{m} \sum_{h=0}^{m-1} D_h.$$

3. The m -th one-dimensional *de-la-Vallée-Poussin-kernel* is defined as

$$V_m := (1 + e_m + e_{-m}) \cdot F_m.$$

4. Let $s \in \mathbb{N}$. We extend the above definitions to dimension s by letting

$$\begin{aligned} D_m^s(x_1, \dots, x_s) &:= \prod_{p=1}^s D_m(x_p), \\ F_m^s(x_1, \dots, x_s) &:= \prod_{p=1}^s F_m(x_p), \\ V_m^s(x_1, \dots, x_s) &:= \prod_{p=1}^s V_m(x_p). \end{aligned}$$

We need the following property of the multivariate extension of the de-la-Vallée-Poussin-kernel.

Proposition C.5. Let $m, s \in \mathbb{N}$. Then one has $\|V_m^s\|_{L^1([-\pi, \pi]^s; \mathbb{C})} \leq 3^s$.

Proof. From [34, Exercise 1.3 and Lemma 1.4] it follows $\|F_m\|_{L^1([-\pi, \pi]; \mathbb{C})} = 1$ and hence using the triangle inequality $\|V_m\|_{L^1([-\pi, \pi]; \mathbb{C})} \leq 3$. The claim then follows using Tonelli's theorem. $\square$

The following definition introduces the term of trigonometric polynomial.

Definition C.6. For any $s \in \mathbb{N}$ and $m \in \mathbb{N}_0$ we call a function of the form

$$\mathbb{R}^s \rightarrow \mathbb{C}, \quad x \mapsto \sum_{\substack{\mathbf{k} \in \mathbb{Z}_0^s \\ -m \leq \mathbf{k} \leq m}} a_{\mathbf{k}} e^{i\langle \mathbf{k}, x \rangle}$$

with coefficients $a_{\mathbf{k}} \in \mathbb{C}$ a *trigonometric polynomial of coordinatewise degree at most m* and denote the space of all those functions with $\mathbb{H}_m^s$. Here, we consider the sum over all $\mathbf{k} \in \mathbb{Z}^s$ with $-m \leq \mathbf{k}_j \leq m$ for all $j \in \{1, \dots, s\}$. We then write

$$E_m^s(f) := \min_{T \in \mathbb{H}_m^s} \|f - T\|_{L^\infty(\mathbb{R}^s; \mathbb{C})} \quad (\text{C.1})$$

for any function $f \in C_{2\pi}(\mathbb{R}^s; \mathbb{C})$.

The following proposition shows that convolving with the Fejèr kernel produces a trigonometric polynomial of degree at most $2m - 1$, while reproducing trigonometric polynomials of degree m . Furthermore, the norm of the convolution operator is bounded uniformly in m . These properties will be useful for our proof of Theorem C.15.

Proposition C.7. *Let $s, m \in \mathbb{N}$ and $k \in \mathbb{N}_0$. The map*

$$v_m : C_{2\pi}(\mathbb{R}^s; \mathbb{C}) \rightarrow \mathbb{H}_{2m-1}^s, \quad f \mapsto f * V_m^s$$

is well-defined and satisfies

$$v_m(T) = T \quad \text{for all } T \in \mathbb{H}_m^s. \quad (\text{C.2})$$

Furthermore, there exists a constant $c = c(s) > 0$ (independent of m), such that

$$\begin{aligned} \|v_m(f)\|_{C^k([- \pi, \pi]^s; \mathbb{C})} &\leq c \cdot \|f\|_{C^k([- \pi, \pi]^s; \mathbb{C})} \quad \forall f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{C}), \\ \|v_m(f)\|_{L^\infty([- \pi, \pi]^s; \mathbb{C})} &\leq c \cdot \|f\|_{L^\infty([- \pi, \pi]^s; \mathbb{C})} \quad \forall f \in C_{2\pi}(\mathbb{R}^s; \mathbb{C}). \end{aligned} \quad (\text{C.3})$$

In fact, it holds $c(s) \leq \exp(C \cdot s)$ with an absolute constant $C > 0$.

Proof. A direct computation shows that $f * e_{\mathbf{k}} = \hat{f}(\mathbf{k}) \cdot e_{\mathbf{k}}$. This implies that v_m is well-defined since V_m^s is a trigonometric polynomial of coordinatewise degree at most $2m - 1$.

The operator is bounded on $C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$ and $C_{2\pi}(\mathbb{R}^s; \mathbb{C})$ with norm at most $c = 3^s$, as follows from Young's inequality [34, Lemma 1.1 (ii)], Proposition C.5, and the fact that one has for all $\mathbf{k} \in \mathbb{N}_0^s$ with $|\mathbf{k}| \leq k$ the identity

$$\partial^{\mathbf{k}}(f * V_m^s) = (\partial^{\mathbf{k}}f) * V_m^s \quad \text{for } f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{C}).$$

It remains to show that v_m is the identity on $\mathbb{H}_m^s$. We first prove that

$$e_k * V_m = e_k \quad (\text{C.4})$$

holds for all $k \in \mathbb{Z}$ with $|k| \leq m$. First note that

$$e_k * V_m = e_k * F_m + e_k * (e_m \cdot F_m) + e_k * (e_{-m} \cdot F_m).$$

We then compute

$$e_k * F_m = \frac{1}{m} \sum_{\ell=0}^{m-1} D_\ell * e_k = \frac{1}{m} \sum_{\ell=0}^{m-1} \sum_{h=-\ell}^{\ell} \underbrace{e_h * e_k}_{=\delta_{k,h} \cdot e_k} = \frac{1}{m} \sum_{\ell=|k|}^{m-1} e_k = \frac{m - |k|}{m} \cdot e_k.$$

Similarly, we get

$$\begin{aligned} e_k * (e_m \cdot F_m) &= \frac{1}{m} \sum_{\ell=0}^{m-1} (e_m D_\ell) * e_k = \frac{1}{m} \sum_{\ell=0}^{m-1} \sum_{h=-\ell}^{\ell} \underbrace{e_{h+m} * e_k}_{=\delta_{k,h+m} \cdot e_k} \\ &= \frac{1}{m} \sum_{\substack{0 \leq \ell \leq m-1 \\ \ell \geq m-k}} e_k = \delta_{k \geq 1} \cdot \frac{k}{m} \cdot e_k \end{aligned}$$

and

$$\begin{aligned} e_k * (e_{-m} \cdot F_m) &= \frac{1}{m} \sum_{\ell=0}^{m-1} (e_{-m} D_\ell) * e_k = \frac{1}{m} \sum_{\ell=0}^{m-1} \sum_{h=-\ell}^{\ell} \underbrace{e_{h-m} * e_k}_{=\delta_{k,h-m} \cdot e_k} \\ &= \frac{1}{m} \sum_{\substack{0 \leq \ell \leq m-1 \\ \ell \geq k+m}} e_k = \delta_{k \leq -1} \cdot \frac{-k}{m} \cdot e_k. \end{aligned}$$

Adding up those three identities yields (C.4).

To finally prove (C.2), it clearly suffices to show that

$$e_{\mathbf{k}} * V_m^s = e_{\mathbf{k}}$$

for all $\mathbf{k} \in \mathbb{Z}^s$ with $-m \leq \mathbf{k} \leq m$. But for such $\mathbf{k}$, using $e_{\mathbf{k}}(x) = \prod_{j=1}^s e_{\mathbf{k}_j}(x_j)$, one obtains

$$\begin{aligned} (e_{\mathbf{k}} * V_m^s)(x) &= \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} \prod_{j=1}^s e_{\mathbf{k}_j}(t_j) \cdot V_m(x_j - t_j) dt \\ &\stackrel{\text{Fubini}}{=} \prod_{j=1}^s (e_{\mathbf{k}_j} * V_m)(x_j) \stackrel{(\text{C.4})}{=} \prod_{j=1}^s e_{\mathbf{k}_j}(x_j) = e_{\mathbf{k}}(x) \end{aligned}$$

for any $x \in \mathbb{R}^s$, as was to be shown. $\square$

The following result follows from a theorem in [29].

Proposition C.8. *Let $s, k \in \mathbb{N}$. Then there exists a constant $c = c(s, k) > 0$, such that, for E_m^s as defined in (C.1),*

$$E_m^s(f) \leq \frac{c}{m^k} \cdot \|f\|_{C^k([- \pi, \pi]^s; \mathbb{R})}$$

for all $m \in \mathbb{N}$ and $f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{R})$.

In fact, it holds $c(s, k) \leq \exp(C \cdot ks) \cdot k^k$ with an absolute constant $C > 0$.

Proof. We apply [29, Theorem 6.6] with $n_i = m$ and $p_i = k$, which yields the existence of a constant $c_1 = c_1(s, k) > 0$, such that

$$E_m^s(f) \leq c_1 \cdot \sum_{\ell=1}^s \frac{1}{m^k} \cdot \omega_{\ell}\left(f, \frac{1}{m}\right)$$

for all $m \in \mathbb{N}$ and $f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{R})$, where $\omega_{\ell}(f, \bullet)$ denotes the modulus of continuity of $\frac{\partial^k f}{\partial x_{\ell}^k}$ with respect to x_{ℓ} , where we have the trivial bound

$$\omega_{\ell}\left(f, \frac{1}{m}\right) \leq 2 \cdot \|f\|_{C^k([- \pi, \pi]^s; \mathbb{R})}.$$

Hence, we get

$$E_m^s(f) \leq c_1 \cdot s \cdot 2 \cdot \|f\|_{C^k([- \pi, \pi]^s; \mathbb{R})} \frac{1}{m^k},$$

so the claim follows by choosing $c := 2s \cdot c_1$.

We refer to Appendix C.2 (see Theorem C.18) for a proof of the claimed bound on the constant $c(s, k)$. $\square$

The above proposition bounds the best possible error of approximating f by trigonometric polynomials of coordinatewise degree at most m , but this is in general non-constructive. Our next result shows that a similar bound holds for approximating f by $v_m(f)$.

Theorem C.9. *Let $s \in \mathbb{N}$. Then there exists a constant $c = c(s) > 0$, such that the operator v_m from Proposition C.7 satisfies*

$$\|f - v_m(f)\|_{L^{\infty}(\mathbb{R}^s)} \leq c \cdot E_m^s(f)$$

for any $m \in \mathbb{N}$ and $f \in C_{2\pi}(\mathbb{R}^s; \mathbb{C})$.

In fact, it holds $c(s) \leq \exp(C \cdot s)$ with an absolute constant $C > 0$.

Proof. For any $T \in \mathbb{H}_m^s$ one has

$$\|f - v_m(f)\|_{L^{\infty}(\mathbb{R}^s)} \stackrel{(\text{C.2})}{\leq} \|f - T\|_{L^{\infty}(\mathbb{R}^s)} + \|v_m(T) - v_m(f)\|_{L^{\infty}(\mathbb{R}^s)} \stackrel{(\text{C.3})}{\leq} (c+1) \|f - T\|_{L^{\infty}(\mathbb{R}^s)}.$$

Taking the infimum over all $T \in \mathbb{H}_m^s$ yields the claim. $\square$

By combining Proposition C.8 and Theorem C.9, we immediately get the following bound.

Corollary C.10. *Let $s, k \in \mathbb{N}_0$. Then there exists a constant $c = c(s, k) > 0$, such that*

$$\|f - v_m(f)\|_{L^\infty(\mathbb{R}^s)} \leq \frac{c}{m^k} \cdot \|f\|_{C^k([- \pi, \pi]^s; \mathbb{R})}$$

for every $m \in \mathbb{N}$ and $f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{R})$.

In fact, we have $c(s, k) \leq \exp(C \cdot ks) \cdot k^k$ with an absolute constant $C > 0$.

Up to now, we have studied the approximation of periodic functions by trigonometric polynomials, but our actual goal is to approximate non-periodic functions by algebraic polynomials. The next lemma establishes a connection between the two settings.

Lemma C.11. *Let $k \in \mathbb{N}_0$ and $s \in \mathbb{N}$. For any function $f \in C^k([-1, 1]^s; \mathbb{C})$, we define the corresponding periodic function via*

$$f^* : \mathbb{R}^s \rightarrow \mathbb{C}, \quad f^*(x_1, \dots, x_s) = f(\cos(x_1), \dots, \cos(x_s))$$

and note $f^* \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$. The map

$$C^k([-1, 1]^s; \mathbb{C}) \rightarrow C_{2\pi}^k(\mathbb{R}^s; \mathbb{C}), \quad f \mapsto f^*$$

is a continuous linear operator with respect to the C^k -norms on $C^k([-1, 1]^s; \mathbb{C})$ and $C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$. The operator norm can be bounded from above by k^k .

Proof. The map is well-defined since $\cos$ is a smooth function and 2π -periodic. The linearity of the operator is obvious, so it remains to show its continuity.

The goal is to apply the closed graph theorem [21, Theorem 5.12]. By definition of f^* , and since $\cos : [-\pi, \pi] \rightarrow [-1, 1]$ is surjective, we have the equality $\|f\|_{L^\infty([-1, 1]^s; \mathbb{C})} = \|f^*\|_{L^\infty([- \pi, \pi]^s; \mathbb{C})}$. Let then $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions $f_n \in C^k([-1, 1]^s; \mathbb{C})$ and $g^* \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$ such that $f_n \rightarrow f$ in $C^k([-1, 1]^s; \mathbb{C})$ and $f_n^* \rightarrow g^*$ in $C_{2\pi}^k(\mathbb{R}^s; \mathbb{C})$. We then have

$$\begin{aligned} \|f^* - g^*\|_{L^\infty([- \pi, \pi]^s)} &\leq \|f^* - f_n^*\|_{L^\infty([- \pi, \pi]^s)} + \|f_n^* - g^*\|_{L^\infty([- \pi, \pi]^s)} \\ &= \|f - f_n\|_{L^\infty([-1, 1]^s; \mathbb{C})} + \|f_n^* - g^*\|_{L^\infty([- \pi, \pi]^s)} \\ &\leq \|f - f_n\|_{C^k([-1, 1]^s; \mathbb{C})} + \|f_n^* - g^*\|_{C^k([- \pi, \pi]^s; \mathbb{C})} \rightarrow 0 \quad (n \rightarrow \infty). \end{aligned}$$

It follows $f^* = g^*$ and the closed graph theorem yields the desired continuity.

We refer to Appendix C.2 (see Theorem C.19 and Remark C.20) for a proof of the claimed bound on the operator norm. $\square$

For a function $f \in C^k([-1, 1]^s; \mathbb{C})$ we want to express $v_m(f^*)$ in a convenient way, involving a product of cosines. To this end, we make use of the following identity, which is a generalization of the well-known product-to-sum formula for $\cos$.

Lemma C.12. *Let $s \in \mathbb{N}$. Then it holds for any $x \in \mathbb{R}^s$ that*

$$\prod_{j=1}^s \cos(x_j) = \frac{1}{2^s} \sum_{\sigma \in \{-1, 1\}^s} \cos(\langle \sigma, x \rangle).$$

Proof. This is an inductive generalization of the product-to-sum formula

$$2 \cos(x) \cos(y) = \cos(x - y) + \cos(x + y) \tag{C.5}$$

for $x, y \in \mathbb{R}$, which can be found for instance in [1, Eq. 4.3.32]. The case $s = 1$ holds since $\cos$ is an even function. Assume that the claim holds for a fixed $s \in \mathbb{N}$ and take $x \in \mathbb{R}^{s+1}$. Writing

$x' = (x_1, \dots, x_s)$, we derive

$$\begin{aligned}
\prod_{j=1}^{s+1} \cos(x_j) &= \left(\frac{1}{2^s} \sum_{\sigma \in \{-1,1\}^s} \cos(\langle \sigma, x' \rangle) \right) \cdot \cos(x_{s+1}) \\
&= \frac{1}{2^s} \sum_{\sigma \in \{-1,1\}^s} \cos(\langle \sigma, x' \rangle) \cos(x_{s+1}) \\
&\stackrel{(C.5)}{=} \frac{1}{2^{s+1}} \sum_{\sigma \in \{-1,1\}^s} [\cos(\langle \sigma, x' \rangle + x_{s+1}) + \cos(\langle \sigma, x' \rangle - x_{s+1})] \\
&= \frac{1}{2^{s+1}} \sum_{\sigma \in \{-1,1\}^{s+1}} \cos(\langle \sigma, x \rangle),
\end{aligned}$$

as was to be shown. $\square$

The following proposition states that $v_m(f^*)$ can be expressed as a linear combination of products of cosines. This representation is useful since these cosines can be interpolated by Chebyshev polynomials which in the end leads to the desired approximation result.

Proposition C.13. *Let $s \in \mathbb{N}$ and $k \in \mathbb{N}_0$. For any $f \in C^k([-1, 1]^s; \mathbb{C})$ and $m \in \mathbb{N}$ the de-la-Vallée-Poussin operator given as $f \mapsto v_m(f^*)$ with v_m as in Proposition C.7 and $f \mapsto f^*$ as in Lemma C.11 has a representation*

$$v_m(f^*)(x_1, \dots, x_s) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} \mathcal{V}_{\mathbf{k}}^m(f) \prod_{j=1}^s \cos(\mathbf{k}_j x_j)$$

for continuous linear functionals

$$\mathcal{V}_{\mathbf{k}}^m : C^k([-1, 1]^s; \mathbb{C}) \rightarrow \mathbb{C}, \quad f \mapsto 2^{\|\mathbf{k}\|_0} \cdot a_{\mathbf{k}}^m \cdot \widehat{f^*}(\mathbf{k}),$$

where $\|\mathbf{k}\|_0 = \#\{j \in \{1, \dots, s\} : \mathbf{k}_j \neq 0\}$ and $a_{\mathbf{k}}^m = \widehat{V_m^s}(\mathbf{k})$. Furthermore, if $f \in C^k([-1, 1]^s; \mathbb{R})$, then $\mathcal{V}_{\mathbf{k}}^m(f) \in \mathbb{R}$ for every $\mathbf{k} \in \mathbb{N}_0^s$ with $\mathbf{k} \leq 2m-1$.

Proof. First of all, it is easy to see that $v_m(f^*)$ is even in each variable, which follows directly from the fact that f^* and V_m^s are both even in each variable. Furthermore, if we write

$$V_m^s = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} a_{\mathbf{k}}^m e_{\mathbf{k}}$$

with appropriately chosen coefficients $a_{\mathbf{k}}^m \in \mathbb{R}$, we easily see

$$v_m(f^*) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) e_{\mathbf{k}}.$$

Using Euler's identity and the fact that $v_m(f^*)$ is an even function, we get the representation

$$v_m(f^*)(x) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) \cos(\langle \mathbf{k}, x \rangle)$$

for all $x \in \mathbb{R}^s$. Using $\odot$ to denote the componentwise product of two vectors of the same size, i.e., $x \odot y = (x_i \cdot y_i)_i$, and using the identity $\langle \mathbf{k}, \sigma \odot x \rangle = \langle \sigma, \mathbf{k} \odot x \rangle$, we see since $v_m(f^*)$ is even in

every variable that

$$\begin{aligned}
v_m(f^*)(x) &= \frac{1}{2^s} \cdot \sum_{\sigma \in \{-1,1\}^s} v_m(f^*)(\sigma \odot x) \\
&= \frac{1}{2^s} \cdot \sum_{\sigma \in \{-1,1\}^s} \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) \cos(\langle \mathbf{k}, \sigma \odot x \rangle) \\
&= \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} \left(a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) \frac{1}{2^s} \sum_{\sigma \in \{-1,1\}^s} \cos(\langle \sigma, \mathbf{k} \odot x \rangle) \right) \\
&\stackrel{\text{Lemma C.12}}{=} \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ -(2m-1) \leq \mathbf{k} \leq 2m-1}} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) \prod_{j=1}^s \cos(\mathbf{k}_j x_j) \\
&= \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} 2^{\|\mathbf{k}\|_0} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}) \prod_{j=1}^s \cos(\mathbf{k}_j x_j)
\end{aligned}$$

with

$$\|\mathbf{k}\|_0 := \#\{j \in \{1, \dots, s\} : \mathbf{k}_j \neq 0\}.$$

In the last step we again used that $\cos$ is an even function and that

$$\widehat{f^*}(\mathbf{k}) = \widehat{f^*}(\sigma \odot \mathbf{k})$$

for all $\sigma \in \{-1, 1\}^s$, which also follows easily since f^* and V_m^s are even in every component. Letting

$$\mathcal{V}_{\mathbf{k}}^m(f) := 2^{\|\mathbf{k}\|_0} a_{\mathbf{k}}^m \widehat{f^*}(\mathbf{k}),$$

we have the desired form. The fact that $\mathcal{V}_{\mathbf{k}}^m$ is a continuous linear functional on $C_{2\pi}^k([-1, 1]^s; \mathbb{C})$ follows directly since $f \mapsto \widehat{f^*}(\mathbf{k})$ is a continuous linear functional for every $\mathbf{k}$. If f is real-valued, so is $\widehat{f^*}(\mathbf{k})$ for every $\mathbf{k} \in \mathbb{N}_0^s$ with $\mathbf{k} \leq 2m-1$, since f^* is real-valued and even in every component. $\square$

Our main approximation result involves linear combinations of Chebyshev polynomials where the coefficients in this linear combination are given as $\mathcal{V}_{\mathbf{k}}^m(f)$. It is therefore important to be able to bound the sum of the absolute values $|\mathcal{V}_{\mathbf{k}}^m(f)|$.

Lemma C.14. *Let $s \in \mathbb{N}$. There exists a constant $c = c(s) > 0$, such that the inequality*

$$\sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} |\mathcal{V}_{\mathbf{k}}^m(f)| \leq c \cdot m^{s/2} \cdot \|f\|_{L^\infty([-1, 1]^s; \mathbb{C})}$$

holds for all $m \in \mathbb{N}$ and $f \in C([-1, 1]^s; \mathbb{C})$, where $\mathcal{V}_{\mathbf{k}}^m$ is as in Proposition C.13. In fact, we have $c(s) \leq \exp(C \cdot s)$ with an absolute constant $C > 0$.

Proof. Let $f \in C([-1, 1]^s; \mathbb{C})$ and $m \in \mathbb{N}$. For any multi-index $\ell \in \mathbb{N}_0^s$, it follows from Proposition C.13 that

$$\widehat{v_m(f^*)}(\ell) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} \mathcal{V}_{\mathbf{k}}^m(f) \widehat{g}_{\mathbf{k}}(\ell),$$

with

$$g_{\mathbf{k}} : \mathbb{R}^s \rightarrow \mathbb{R}, \quad (x_1, \dots, x_s) \mapsto \prod_{j=1}^s \cos(\mathbf{k}_j x_j).$$

Now, a calculation using Fubini's theorem and using

$$g_{\mathbf{k}} = \prod_{j=1}^s \frac{1}{2} (e_{\mathbf{k}_j} + e_{-\mathbf{k}_j}) = \prod_{\substack{1 \leq j \leq s \\ \mathbf{k}_j \neq 0}} \frac{1}{2} (e_{\mathbf{k}_j} + e_{-\mathbf{k}_j})$$

for any number $k \in \mathbb{N}_0$ shows

$$\widehat{g}_{\mathbf{k}}(\ell) = \begin{cases} \frac{1}{2^{\|\mathbf{k}\|_0}}, & \mathbf{k} = \ell, \\ 0, & \text{otherwise} \end{cases} \quad \text{for } \mathbf{k}, \ell \in \mathbb{N}_0^s.$$

Therefore, we have the bound $|\mathcal{V}_{\ell}^m(f)| \leq 2^s \cdot \left| \widehat{v_m(f^*)}(\ell) \right|$ for $\ell \in \mathbb{N}_0^s$ with $|\ell| \leq 2m - 1$. Using the Cauchy-Schwarz and the Parseval inequality, we therefore see

$$\begin{aligned} \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} |\mathcal{V}_{\mathbf{k}}^m(f)| &\leq 2^s \cdot \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} \left| \widehat{v_m(f^*)}(\mathbf{k}) \right| \stackrel{\text{CS}}{\leq} 2^s \cdot (2m)^{s/2} \cdot \left(\sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} \left| \widehat{v_m(f^*)}(\mathbf{k}) \right|^2 \right)^{1/2} \\ &\stackrel{\text{Parseval}}{\leq} 2^s \cdot 2^{s/2} \cdot m^{s/2} \cdot \|v_m(f^*)\|_{L^2([-\pi, \pi]^s; \mathbb{C})} \\ &\leq \underbrace{2^s \cdot 2^{s/2}}_{=: c_1(s)} \cdot m^{s/2} \cdot \|v_m(f^*)\|_{L^\infty([-\pi, \pi]^s; \mathbb{C})}. \end{aligned}$$

Using Proposition C.7, we get a constant $c_2(s) \leq \exp(C_0 \cdot s)$ such that

$$\sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} |\mathcal{V}_{\mathbf{k}}^m(f)| \leq c_1(s) \cdot c_2(s) \cdot m^{s/2} \cdot \|f^*\|_{L^\infty([-\pi, \pi]^s; \mathbb{C})} = c(s) \cdot m^{s/2} \cdot \|f\|_{L^\infty([-1, 1]^s; \mathbb{C})},$$

as claimed. $\square$

For any natural number $\ell \in \mathbb{N}_0$, we denote by T_ℓ the ℓ -th Chebyshev polynomial, satisfying

$$T_\ell(\cos(x)) = \cos(\ell x), \quad x \in \mathbb{R}.$$

One can show that T_ℓ is in fact a polynomial of degree ℓ . For a multi-index $\mathbf{k} \in \mathbb{N}_0^s$, we define

$$T_{\mathbf{k}}(x) := \prod_{j=1}^s T_{k_j}(x_j), \quad x \in \mathbb{R}^s.$$

We then get the following approximation result about approximating (non-periodic) C^k -functions by linear combinations of Chebyshev polynomials.

Theorem C.15. *Let $k, s, m \in \mathbb{N}$. Then there exists a constant $c = c(s, k) > 0$ with the following property: For any $f \in C^k([-1, 1]^s; \mathbb{R})$ the polynomial P defined as*

$$P(x) := \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} \mathcal{V}_{\mathbf{k}}^m(f) \cdot T_{\mathbf{k}}(x),$$

with $\mathcal{V}_{\mathbf{k}}^m$ as in Proposition C.13, satisfies

$$\|f - P\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq \frac{c}{m^k} \cdot \|f\|_{C^k([-1, 1]^s; \mathbb{R})}.$$

Here, the maps

$$C([-1, 1]^s; \mathbb{R}) \rightarrow \mathbb{R}, \quad f \mapsto \mathcal{V}_{\mathbf{k}}^m(f)$$

are continuous and linear functionals with respect to the L^∞ -norm. Furthermore, there exists a constant $\tilde{c} = \tilde{c}(s) > 0$, such that the inequality

$$\sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq 2m-1}} |\mathcal{V}_{\mathbf{k}}^m(f)| \leq \tilde{c} \cdot m^{s/2} \cdot \|f\|_{L^\infty([-1, 1]^s; \mathbb{R})}$$

holds for all $f \in C([-1, 1]^s; \mathbb{R})$.

Moreover, we have $c(s, k) \leq \exp(C \cdot ks) \cdot k^{2k}$ and $\tilde{c}(s) \leq \exp(C \cdot s)$ with an absolute constant $C > 0$.

Proof. We choose the constant $c_0 = c_0(s, k)$ according to Corollary C.10. Let $f \in C^k([-1, 1]^s; \mathbb{R})$ be arbitrary. Then we define the corresponding function $f^* \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{R})$ as above. Let P be defined as in the statement of the theorem. Then it follows from the definition of the Chebyshev polynomials T_k , the definition of P , and the formula for $v_m(f^*)$ from Proposition C.13 that

$$P^*(x) = v_m(f^*)(x)$$

is satisfied, where P^* is the corresponding function to P defined similarly to f^* . Overall, we get the bound

$$\|f - P\|_{L^\infty([-1, 1]^s; \mathbb{R})} = \|f^* - P^*\|_{L^\infty([-\pi, \pi]^s; \mathbb{R})} \stackrel{\text{Cor. C.10}}{\leq} \frac{c_0}{m^k} \cdot \|f^*\|_{C^k([-\pi, \pi]^s; \mathbb{R})}.$$

The first claim then follows using the continuity of the map $f \mapsto f^*$ as proven in Lemma C.11. The second part of the theorem has already been proven in Lemma C.14. $\square$

C.2 Details on bounding the constant c in Theorem 3.2

In this appendix we provide details on the bound of the constant c that appears in the formulation of Theorem 3.2. Specifically, we perform a careful investigation of several results from [29] to get an upper bound for the constant appearing in Proposition C.8. Moreover, we analyze the operator norm of the operator

$$C^k([-1, 1]^s; \mathbb{C}) \rightarrow C_{2\pi}^k(\mathbb{R}^s; \mathbb{C}), \quad f \mapsto f^* \quad \text{with} \quad f^*(x) := f(\cos(x_1), \dots, \cos(x_s))$$

appearing in Lemma C.11 and show that it is bounded from above by k^k .

We start with the analysis of some bounds in [29, Chapter 4.3]. Here, a generalization of *Jackson's kernel* is defined for any $m, r \in \mathbb{N}$ as

$$L_{m,r}(t) := \lambda_{m,r}^{-1} \cdot \left(\frac{\sin(mt/2)}{\sin(t/2)} \right)^{2r}, \quad t \in \mathbb{R},$$

where $\lambda_{m,r}$ is chosen such that

$$\int_{[-\pi, \pi]} L_{m,r}(t) dt = 1.$$

The first two important bounds are provided in the following proposition.

Proposition C.16. *Let $m, r \in \mathbb{N}$. Then it holds*

$$\lambda_{m,r}^{-1} \leq \exp(C \cdot r) \cdot m^{1-2r} \quad \text{and} \quad \int_{[0, \pi]} t^k L_{m,r}(t) dt \leq \exp(C \cdot r) \cdot m^{-k}$$

for any $k \leq 2r - 2$, with an absolute constant $C > 0$.

Proof. Since $L_{m,r} \geq 0$ and since $\sin(t/2) \leq t/2$ for $t \in [0, \pi]$, we get

$$\begin{aligned} \lambda_{m,r} &\geq \int_{[0, \pi]} \left(\frac{\sin(mt/2)}{t/2} \right)^{2r} dt = 2^{2r} \cdot \int_{[0, \pi]} \left(\frac{\sin(mt/2)}{t} \right)^{2r} dt \\ &= 2^{2r} \cdot \int_{[0, \pi m/2]} \left(\frac{\sin(u)}{(2u)/m} \right)^{2r} du \cdot \frac{2}{m} \geq m^{2r-1} \cdot \int_{[0, \pi m/2]} \left(\frac{\sin(u)}{u} \right)^{2r} du \\ &\geq m^{2r-1} \cdot \int_{[0, \pi/2]} \left(\frac{\sin(u)}{u} \right)^{2r} du \geq m^{2r-1} \cdot \int_{[0, \pi/2]} \left(\frac{2u}{\pi \cdot u} \right)^{2r} du \geq \left(\frac{2}{\pi} \right)^{2r} \cdot m^{2r-1}. \end{aligned}$$

Here, we employed the inequality $\sin(u) \geq \frac{2}{\pi}u$ for $u \in [0, \pi/2]$ in the penultimate step.¹ This shows the first part of the claim.

¹To see that this inequality holds, note that $\sin''(u) = -\sin(u) \leq 0$ for $u \in [0, \pi/2]$, so that $\sin$ is concave on that interval, and hence $\sin(u) = \sin((1 - \frac{2u}{\pi}) \cdot 0 + \frac{2u}{\pi} \cdot \frac{\pi}{2}) \geq \frac{2u}{\pi} \sin(\frac{\pi}{2}) = \frac{2u}{\pi}$.

For the second part, we again use the estimate $\sin(u) \geq \frac{2}{\pi}u$ for $u \in [0, \pi/2]$ to compute

$$\begin{aligned}
\int_{[0,\pi]} t^k L_{m,r}(t) dt &= \lambda_{m,r}^{-1} \cdot \int_{[0,\pi]} t^k \left(\frac{\sin(mt/2)}{\sin(t/2)} \right)^{2r} dt \leq \lambda_{m,r}^{-1} \cdot \int_{[0,\pi]} t^k \left(\frac{\sin(mt/2)}{t/\pi} \right)^{2r} dt \\
&= \lambda_{m,r}^{-1} \cdot \pi^{2r} \cdot \int_{[0,\pi]} t^{k-2r} \sin(mt/2)^{2r} dt \\
&= \lambda_{m,r}^{-1} \cdot \pi^{2r} \cdot \int_{[0,\pi m/2]} \left(\frac{2u}{m} \right)^{k-2r} \sin(u)^{2r} du \cdot \frac{2}{m} \\
&\leq \lambda_{m,r}^{-1} \cdot \pi^{2r} \cdot m^{2r-1-k} \int_{[0,\pi m/2]} u^{k-2r} \sin(u)^{2r} du \\
&\leq \exp(C_1 \cdot r) \cdot \int_{[0,\infty)} u^{k-2r} \cdot \sin(u)^{2r} du \cdot m^{-k}
\end{aligned}$$

with an absolute constant $C_1 > 0$. Here, we employed the first part of this proposition. It remains to bound the integral. This is done via

$$\begin{aligned}
\int_{[0,\infty)} u^{k-2r} \cdot \sin(u)^{2r} du &= \int_{[0,1]} u^{k-2r} \cdot \sin(u)^{2r} du + \int_{[1,\infty)} u^{k-2r} \cdot \sin(u)^{2r} du \\
&\leq \int_{[0,1]} \underbrace{u^k \cdot \left(\frac{\sin(u)}{u} \right)^{2r}}_{\leq 1} du + \int_{[1,\infty)} u^{-2} du \leq C_2
\end{aligned}$$

with an absolute constant $C_2 > 0$. This proves the claim. $\square$

The proof in [29] proceeds by defining

$$K_{m,r}(t) := L_{m',r}(t), \quad m' = \left\lfloor \frac{m}{r} \right\rfloor + 1.$$

Proposition C.16 shows for $k \leq 2r - 2$ that

$$\int_{[0,\pi]} t^k K_{m,r}(t) dt \leq \exp(C \cdot r) \cdot (m')^{-k}.$$

Since $m' \geq \frac{m}{r}$ we infer

$$\int_{[0,\pi]} t^k K_{m,r}(t) dt \leq \exp(C \cdot r) \cdot \left(\frac{r}{m} \right)^k \leq \exp(C \cdot r) \cdot r^k \cdot m^{-k} \quad (\text{C.6})$$

with an absolute constant $C > 0$.

We can now quantify the constant appearing in [29, Theorem 4.3].

Theorem C.17 (cf. [29, Theorem 4.3]). *Let $k, m \in \mathbb{N}$ and $f \in C_{2\pi}^k(\mathbb{R}; \mathbb{R})$. Let*

$$\omega(f^{(k)}, 1/m) := \max_{x \in \mathbb{R}, |t| \leq 1/m} |f^{(k)}(x+t) - f^{(k)}(x)|.$$

Then it holds

$$E_m^1(f) \leq (\exp(C \cdot k) \cdot k^k) \cdot m^{-k} \cdot \omega(f^{(k)}, 1/m).$$

Here, we recall that $E_m^1(f)$ denotes the best possible approximation error when approximating f using trigonometric polynomials of degree m ; see Equation (C.1).

Proof. We follow the proof of [29, Theorem 4.3]. Take $r = k + 1$ and define

$$I_m(x) := - \int_{[-\pi,\pi]} K_{m,r}(t) \sum_{\ell=1}^{k+1} (-1)^\ell \binom{k+1}{\ell} f(x + \ell t) dt.$$

Then it is shown in the proof of [29, Theorem 4.3] that I_m is a trigonometric polynomial of degree at most m and that

$$|f(x) - I_m(x)| \leq 2 \cdot \omega_{k+1}(f, 1/m) \cdot \int_{[0,\pi]} (mt + 1)^{k+1} K_{m,r}(t) dt.$$

Here, $\omega_{k+1}(f, 1/m)$ denotes the *modulus of smoothness* of f as defined on [29, p. 47]. The integral can be bounded via

$$\begin{aligned}
& \int_{[0, \pi]} (mt+1)^{k+1} K_{m,r}(t) dt \\
&= \int_{[0, 1/m]} \underbrace{(mt+1)^{k+1}}_{\leq 2} K_{m,r}(t) dt + \int_{[1/m, \pi]} \underbrace{(mt+1)^{k+1}}_{\leq 2mt} K_{m,r}(t) dt \\
&\leq 2^{k+1} \cdot \underbrace{\int_{[-\pi, \pi]} K_{m,r}(t) dt}_{=1} + 2^{k+1} m^{k+1} \cdot \int_{[0, \pi]} t^{k+1} K_{m,r}(t) dt \\
&\stackrel{(C.6)}{\leq} 2^{k+1} + 2^{k+1} m^{k+1} \exp(C_1 \cdot r) \cdot (k+1)^{k+1} \cdot m^{-(k+1)} \stackrel{r \leq 2k}{\leq} \exp(C \cdot k) \cdot k^k
\end{aligned}$$

with absolute constants $C, C_1 > 0$. Since $\omega_{k+1}(f, 1/m) \leq m^{-k} \cdot \omega(f^{(k)}, 1/m)$ follows from [29, Equation 3.6(5)], the claim is shown. $\square$

Therefore, we can bound the constant appearing in [29, Theorem 4.3] by $\exp(C \cdot k) \cdot k^k$. It remains to deal with the approximation of *multivariate* periodic functions by *multivariate* trigonometric polynomials which is contained in [29, Theorem 6.6].

Theorem C.18 (cf. [29, Theorem 6.6]). *Let $s, k \in \mathbb{N}$ and $f \in C_{2\pi}^k(\mathbb{R}^s; \mathbb{R})$. Let ω_j denote the modulus of continuity of $\frac{\partial^k f}{\partial x_j^k}$ for $j = 1, \dots, s$. Then, with E_m^s as introduced in Equation (C.1), it holds*

$$E_m^s(f) \leq \exp(C \cdot ks) \cdot k^k \cdot m^{-k} \sum_{j=1}^s \omega_j(1/m),$$

with an absolute constant $C > 0$.

Proof. We follow the proof of [29, Theorem 6.6] with $p_j = k$ and $n_j = m$ for every index $j = 1, \dots, s$. For $j = 1, \dots, s+1$ define the set $\mathcal{T}_j$ consisting of all functions $g \in C_{2\pi}(\mathbb{R}^s; \mathbb{R})$ that are a trigonometric polynomial in x_ℓ of degree at most m for $\ell < j$; in x_ℓ for $\ell \geq j$ they should have continuous partial derivatives $\frac{\partial^p g}{\partial x_\ell^p}$ for $0 \leq p \leq k$; the modulus of continuity of $\frac{\partial^k g}{\partial x_\ell^k}$ should not exceed $2^{K_j} \omega_j$, where

$$K_j = (j-1)(k+1) \leq 2ks \quad \text{for } j > 1 \text{ and } K_1 = 1 \leq 2ks.$$

Then it is shown that if $j \in \{1, \dots, s\}$ and $f_j \in \mathcal{T}_j$ there exists a function $f_{j+1} \in \mathcal{T}_{j+1}$ for which

$$\|f_j - f_{j+1}\|_{L^\infty(\mathbb{R}^s; \mathbb{R})} \leq \exp(C_1 \cdot k) \cdot k^k \cdot m^{-k} \cdot 2^{2ks} \cdot \omega_j(1/m) \leq \exp(C_2 \cdot ks) \cdot k^k \cdot m^{-k} \cdot \omega_j(1/m)$$

for absolute constants $C_1, C_2 > 0$. This is an application of Theorem C.17. Hence, defining $f_1 := f$, we see

$$\|f - f_{s+1}\|_{L^\infty(\mathbb{R}^s; \mathbb{R})} \leq \sum_{j=1}^s \|f_j - f_{j+1}\|_{L^\infty(\mathbb{R}^s; \mathbb{R})} \leq \exp(C_2 \cdot ks) \cdot k^k \cdot m^{-k} \cdot \sum_{j=1}^s \omega_j(1/m). \quad \square$$

Therefore, we have shown that the constant appearing in [29, Theorem 6.6] can be bounded from above by $\exp(C \cdot ks) \cdot k^k$.

In the rest of this section, we discuss the operator norm of the operator defined in Lemma C.11. In Lemma C.11 the closed graph theorem is used to show that the operator is bounded. However, the closed graph theorem does not provide any bound on the norm of the operator. Therefore, in order to quantify the operator norm, we need to apply a different technique, which is *Faa di Bruno's formula*. This formula is a generalization of the chain rule to higher order derivatives.

Theorem C.19. *Let $s, k \in \mathbb{N}$. We define the operator*

$$T : C^k([-1, 1]^s; \mathbb{C}) \rightarrow C_{2\pi}^k([- \pi, \pi]^s; \mathbb{C}), \quad (Tf)(x_1, \dots, x_s) := f(\cos(x_1), \dots, \cos(x_s)).$$

Let $\alpha \in \mathbb{N}_0^s$ with $|\alpha| \leq k$. Then, for any $f \in C^k([-1, 1]^s; \mathbb{C})$, we have

$$\|\partial^\alpha (Tf)\|_{L^\infty([- \pi, \pi]^s; \mathbb{C})} \leq \prod_{j=1}^s \alpha_j^{\alpha_j} \cdot \|f\|_{C^{|\alpha|}([-1, 1]^s)}.$$

Proof. The proof is by induction over s . The case $s = 1$ is an application of Faa di Bruno's formula: We can write $Tf = f \circ g$ with $g(x) = \cos(x)$. We then take $\ell \in \mathbb{N}_0$ with $\ell \leq k$ and some $x \in [-\pi, \pi]$. The set partition version of Faa di Bruno's formula (see for instance [24, p. 219]) then yields

$$\left| (f \circ g)^{(\ell)}(x) \right| \leq \sum_{\pi \in \Pi_\ell} \left(\left| f^{(|\pi|)}(g(x)) \right| \cdot \prod_{B \in \pi} \left| g^{(|B|)}(x) \right| \right).$$

Here, Π_ℓ denotes the set of all partitions of the set $\{1, \dots, \ell\}$. Since all derivatives of g are bounded by 1 in absolute value and $|\pi| \leq \ell$ for every partition $\pi \in \Pi_\ell$ we get

$$\|(f \circ g)^{(\ell)}\|_{L^\infty([-\pi, \pi]; \mathbb{C})} \leq |\Pi_\ell| \cdot \|f\|_{C^\ell([-1, 1]; \mathbb{C})}.$$

The number $|\Pi_\ell|$ is the number of possible partitions of the set $\{1, \dots, \ell\}$ and is the so-called ℓ -th *Bell number*. It can be bounded from above by ℓ^ℓ (see [13, Theorem 2.1]). This proves the case $s = 1$.

We now assume that the claim holds for an arbitrary but fixed $s \in \mathbb{N}$. Take $\alpha \in \mathbb{N}_0^{s+1}$ with $|\alpha| \leq k$. We decompose $\alpha = (\alpha', \alpha_{s+1})$ with $\alpha' \in \mathbb{N}_0^s$. For a fixed variable $y_{s+1} \in [-1, 1]$, we define

$$f_{y_{s+1}}(y_1, \dots, y_s) := f(y_1, \dots, y_s, y_{s+1}) \quad \text{for } (y_1, \dots, y_s) \in [-1, 1]^s.$$

We denote $g(x_1, \dots, x_{s+1}) := (\cos(x_1), \dots, \cos(x_{s+1}))$, $g_s(x_1, \dots, x_s) := (\cos(x_1), \dots, \cos(x_s))$ and $\theta(x_{s+1}) := \cos(x_{s+1})$. For every $(x_1, \dots, x_{s+1}) \in [-\pi, \pi]^{s+1}$ it then holds

$$(f \circ g)(x_1, \dots, x_{s+1}) = (f_{\theta(x_{s+1})} \circ g_s)(x_1, \dots, x_s).$$

We now differentiate $f \circ g$ with respect to the multiindex α and get

$$\begin{aligned} [\partial^\alpha (f \circ g)](x_1, \dots, x_{s+1}) &= \frac{\partial^{\alpha_{s+1}}}{\partial x_{s+1}^{\alpha_{s+1}}} \left[\partial^{\alpha'} (f_{\theta(x_{s+1})} \circ g_s)(x_1, \dots, x_s) \right] \\ &= (h_{x_1, \dots, x_s} \circ \theta)^{(\alpha_{s+1})}(x_{s+1}) \end{aligned}$$

where we define

$$h_{x_1, \dots, x_s}(y_{s+1}) := \partial^{\alpha'} (f_{y_{s+1}} \circ g_s)(x_1, \dots, x_s) \quad \text{for } (x_1, \dots, x_s) \in [-\pi, \pi]^s \text{ and } y_{s+1} \in [-1, 1].$$

Using the case $s = 1$, we get

$$|[\partial^\alpha (f \circ g)](x_1, \dots, x_{s+1})| = \left| (h_{x_1, \dots, x_s} \circ \theta)^{(\alpha_{s+1})}(x_{s+1}) \right| \leq \alpha_{s+1}^{\alpha_{s+1}} \cdot \|h_{x_1, \dots, x_s}\|_{C^{\alpha_{s+1}}([-1, 1]; \mathbb{C})}$$

for any fixed $(x_1, \dots, x_s) \in [-\pi, \pi]^s$.

It remains to bound $\|h_{x_1, \dots, x_s}\|_{C^{\alpha_{s+1}}([-1, 1]; \mathbb{C})}$. To this end, we fix $\ell \in \mathbb{N}_0$ with $\ell \leq \alpha_{s+1}$. We further denote

$$F_{y_1, \dots, y_s}(y_{s+1}) := f(y_1, \dots, y_s, y_{s+1}) \quad \text{for } (y_1, \dots, y_{s+1}) \in [-1, 1]^{s+1}.$$

For arbitrary $(x_1, \dots, x_s) \in [-\pi, \pi]^s$ and $y_{s+1} \in [-1, 1]$ we then see

$$h_{x_1, \dots, x_s}^{(\ell)}(y_{s+1}) = \partial^{\alpha'} \left[(x_1, \dots, x_s) \mapsto F_{g_s(x_1, \dots, x_s)}^{(\ell)}(y_{s+1}) \right] = \partial^{\alpha'} [H_{y_{s+1}} \circ g_s](x_1, \dots, x_s)$$

where

$$H_{y_{s+1}}(y_1, \dots, y_s) := F_{y_1, \dots, y_s}^{(\ell)}(y_{s+1}) \quad \text{for } (y_1, \dots, y_s) \in [-1, 1]^s.$$

Hence, we see by induction that

$$\begin{aligned} \left| h_{x_1, \dots, x_s}^{(\ell)}(y_{s+1}) \right| &= \left| \partial^{\alpha'} [H_{y_{s+1}} \circ g_s](x_1, \dots, x_s) \right| \stackrel{\text{IH}}{\leq} \prod_{j=1}^s \alpha_j^{\alpha_j} \cdot \|H_{y_{s+1}}\|_{C^{|\alpha'|}([-1, 1]^s; \mathbb{C})} \\ &\leq \prod_{j=1}^s \alpha_j^{\alpha_j} \cdot \|f\|_{C^{|\alpha|}([-1, 1]^{s+1}; \mathbb{C})} \end{aligned}$$

as was to be shown. □

Remark C.20. For a multiindex $\alpha \in \mathbb{N}_0^s$ with $|\alpha| \leq k$ we see

$$\prod_{j=1}^s \alpha_j^{\alpha_j} \leq k^{\sum_{j=1}^s \alpha_j} \leq k^k.$$

Hence, the norm of the operator introduced in Lemma C.11 can be bounded from above by k^k .

C.3 Proof of Theorem 3.2

For any natural number $\ell \in \mathbb{N}_0$, we denote by T_ℓ the ℓ -th Chebyshev polynomial, satisfying

$$T_\ell(\cos(x)) = \cos(\ell x), \quad x \in \mathbb{R}.$$

For a multi-index $\mathbf{k} \in \mathbb{N}_0^s$ we define

$$T_{\mathbf{k}}(x) := \prod_{j=1}^s T_{k_j}(x_j), \quad x \in [-1, 1]^s.$$

The proof of Theorem 3.2 relies on the fact that C^k -functions can be approximated at a certain rate using linear combinations of the $T_{\mathbf{k}}$ (see Theorem C.15). We also refer to Figure 4 for an illustration of the overall proof strategy of Theorem 3.2.

Proof of Theorem 3.2. Choose $M \in \mathbb{N}$ as the largest integer for which $(16M - 7)^{2n} \leq m$, where we assume without loss of generality that $9^{2n} \leq m$, which can be done by choosing $\sigma_j = 0$ for all $j \in \{1, \dots, m\}$ for $m < 9^{2n}$, at the cost of possibly enlarging c . First we note that by the choice of M the inequality

$$m \leq (16M + 9)^{2n}$$

holds true. Since $16M + 9 \leq 25M$, we get $m \leq 25^{2n} \cdot M^{2n}$ or equivalently

$$\frac{m^{1/2n}}{25} \leq M. \quad (\text{C.7})$$

According to Theorem C.15 we choose a constant $c_1 = c_1(n, k)$ with the property that for any function $f \in C^k([-1, 1]^{2n}; \mathbb{R})$ there exists a polynomial

$$P = \sum_{0 \leq \mathbf{k} \leq 2M-1} \mathcal{V}_{\mathbf{k}}^M(f) \cdot T_{\mathbf{k}}$$

of coordinatewise degree at most $2M - 1$ satisfying

$$\|f - P\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \leq \frac{c_1}{M^k} \cdot \|f\|_{C^k([-1, 1]^{2n}; \mathbb{R})}.$$

Furthermore, according to Theorem C.15, we choose a constant $c_2 = c_2(n)$, such that the inequality

$$\sum_{0 \leq \mathbf{k} \leq 2M-1} |\mathcal{V}_{\mathbf{k}}^M(f)| \leq c_2 \cdot M^n \cdot \|f\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \leq c_2 \cdot M^n \cdot \|f\|_{C^k([-1, 1]^{2n}; \mathbb{R})}$$

holds for all $f \in C^k([-1, 1]^{2n}; \mathbb{R})$. The final constant is then defined to be

$$c = c(n, k) := \sqrt{2} \cdot 25^k \cdot (c_1 + c_2).$$

Fix $\mathbf{k} \leq 2M - 1$. Since $T_{\mathbf{k}}$ is a polynomial of componentwise degree less or equal to $2M - 1$ with φ_n as in (A.1), we have a representation

$$(T_{\mathbf{k}} \circ \varphi_n^{-1})(z) = \sum_{\substack{\ell^1, \ell^2 \in \mathbb{N}_0^n \\ \ell^1, \ell^2 \leq 2M-1}} a_{\ell^1, \ell^2}^{\mathbf{k}} \prod_{t=1}^n \text{Re}(z_t)^{\ell_t^1} \text{Im}(z_t)^{\ell_t^2}$$

with suitably chosen coefficients $a_{\ell^1, \ell^2}^{\mathbf{k}} \in \mathbb{C}$. By using the identities $\text{Re}(z_t) = \frac{1}{2}(z_t + \bar{z}_t)$ and also $\text{Im}(z_t) = \frac{1}{2i}(z_t - \bar{z}_t)$ we can rewrite $T_{\mathbf{k}} \circ \varphi_n^{-1}$ into a complex polynomial in z and $\bar{z}$, i.e.,

$$(T_{\mathbf{k}} \circ \varphi_n^{-1})(z) = \sum_{\substack{\ell^1, \ell^2 \in \mathbb{N}_0^n \\ \ell^1, \ell^2 \leq 4M-2}} b_{\ell^1, \ell^2}^{\mathbf{k}} z^{\ell^1} \bar{z}^{\ell^2}$$

with complex coefficients $b_{\ell^1, \ell^2}^{\mathbf{k}} \in \mathbb{C}$. Using Theorem 3.1, we choose $\rho_1, \dots, \rho_m \in \mathbb{C}^n$ and $b \in \mathbb{C}$, such that for any polynomial $P \in \{T_{\mathbf{k}} \circ \varphi_n^{-1} : \mathbf{k} \leq 2M - 1\} \subseteq \mathcal{P}_{4M-2}^n$ there are coefficients $\sigma_1(P), \dots, \sigma_m(P) \in \mathbb{C}$, such that

$$\|g_P - P\|_{L^\infty(\Omega_n; \mathbb{C})} \leq M^{-k-n}, \quad (\text{C.8})$$

where

$$g_P := \sum_{t=1}^m \sigma_t(P) \phi(\rho_t^T z + b).$$

Note that here we implicitly use the bound $(4 \cdot (4M - 2) + 1)^{2n} \leq m$. We are now going to show that the chosen constant and the chosen vectors ρ_t have the desired property.

Let $f \in C^k(\Omega_n; \mathbb{C})$. By splitting f into real and imaginary part, we write $f = f_1 + i \cdot f_2$ with $f_1, f_2 \in C^k(\Omega_n; \mathbb{R})$. For the following, fix $j \in \{1, 2\}$ and note that $f_j \circ \varphi_n \in C^k([-1, 1]^{2n}; \mathbb{R})$. By choice of c_1 , there exists a polynomial P with the property

$$\|f_j \circ \varphi_n - P\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \leq \frac{c_1}{M^k} \cdot \|f_j \circ \varphi_n\|_{C^k([-1, 1]^{2n}; \mathbb{R})}$$

or equivalently

$$\|f_j - P \circ \varphi_n^{-1}\|_{L^\infty(\Omega_n; \mathbb{R})} \leq \frac{c_1}{M^k} \cdot \|f_j\|_{C^k(\Omega_n; \mathbb{R})}, \quad (\text{C.9})$$

where $P \circ \varphi_n^{-1}$ can be written in the form

$$(P \circ \varphi_n^{-1})(z) = \sum_{0 \leq \mathbf{k} \leq 2M-1} \mathcal{V}_{\mathbf{k}}^M(f_j \circ \varphi_n) \cdot (T_{\mathbf{k}} \circ \varphi_n^{-1})(z).$$

We choose the function $g_{T_{\mathbf{k}} \circ \varphi_n^{-1}}$ according to (C.8). Thus, writing

$$g_j := \sum_{0 \leq \mathbf{k} \leq 2M-1} \mathcal{V}_{\mathbf{k}}^M(f_j \circ \varphi_n) \cdot g_{T_{\mathbf{k}} \circ \varphi_n^{-1}},$$

we obtain

$$\begin{aligned} \|P \circ \varphi_n^{-1} - g_j\|_{L^\infty(\Omega_n; \mathbb{R})} &\leq \sum_{0 \leq \mathbf{k} \leq 2M-1} |\mathcal{V}_{\mathbf{k}}^M(f_j \circ \varphi_n)| \cdot \underbrace{\|T_{\mathbf{k}} \circ \varphi_n^{-1} - g_{T_{\mathbf{k}} \circ \varphi_n^{-1}}\|_{L^\infty(\Omega_n; \mathbb{R})}}_{\leq M^{-k-n}} \\ &\leq M^{-k-n} \cdot \sum_{0 \leq \mathbf{k} \leq 2M-1} |\mathcal{V}_{\mathbf{k}}^M(f_j \circ \varphi_n)| \\ &\leq \frac{c_2}{M^k} \|f_j \circ \varphi_n\|_{C^k([-1, 1]^{2n}; \mathbb{R})} = \frac{c_2}{M^k} \|f_j\|_{C^k(\Omega_n; \mathbb{R})}. \end{aligned} \quad (\text{C.10})$$

Combining (C.9) and (C.10), we see

$$\|f_j - g_j\|_{L^\infty(\Omega_n; \mathbb{R})} \leq \frac{c_1 + c_2}{M^k} \cdot \|f_j\|_{C^k(\Omega_n; \mathbb{R})} \leq \frac{c_1 + c_2}{M^k} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}.$$

In the end, define

$$g := g_1 + i \cdot g_2.$$

Since the vectors ρ_t have been chosen fixed, it is clear that, after rearranging, g has the desired form, i.e., $g = \sigma^T \Phi$ where $\Phi(z) = (\phi(\rho_t z + b))_{t=1}^m$. Furthermore, one obtains the bound

$$\begin{aligned} \|f - g\|_{L^\infty(\Omega_n; \mathbb{C})} &\leq \sqrt{\|f_1 - g_1\|_{L^\infty(\Omega_n; \mathbb{R})}^2 + \|f_2 - g_2\|_{L^\infty(\Omega_n; \mathbb{R})}^2} \\ &\leq \frac{c_1 + c_2}{M^k} \cdot \sqrt{\|f\|_{C^k(\Omega_n; \mathbb{C})}^2 + \|f\|_{C^k(\Omega_n; \mathbb{C})}^2} \\ &\leq \frac{\sqrt{2} \cdot (c_1 + c_2)}{M^k} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}. \end{aligned}$$

Using (C.7), we see

$$\|f - g\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \frac{\sqrt{2} \cdot 25^k \cdot (c_1 + c_2)}{m^{k/2n}} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})},$$

as desired.

The linearity and continuity of the maps $f \mapsto \sigma_j(f)$ (with respect to the $\|\cdot\|_{L^\infty}$ -norm) follow easily from the fact that the map $f \mapsto \mathcal{V}_{\mathbf{k}}^M(f)$ is a continuous linear functional for every multiindex $0 \leq \mathbf{k} \leq 2M - 1$. $\square$

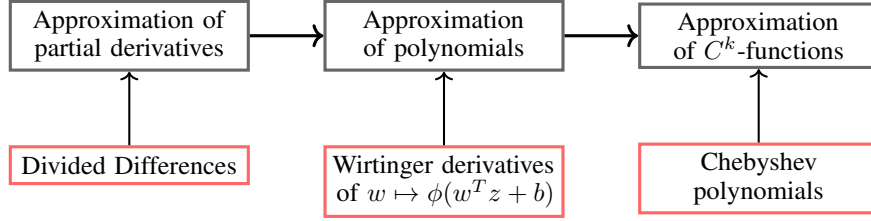


Figure 4: Schematic for the proof of the main result (Theorem 3.2). The first row shows the different steps of the proof and the second row indicates the main tools used.

D Approximation of holomorphically extendable functions

In this appendix we provide the proofs for the statements contained in Remark 3.3. The proofs mainly rely on results about sparse polynomial approximation [2].

Definition D.1 (cf. [2, Assumption 2.3]). Let $s \in \mathbb{N}$ and $\nu \in (1, \infty)^s$. For every $j \in \{1, \dots, s\}$ let

$$\mathcal{E}_{\nu_j} := \left\{ \frac{z + z^{-1}}{2} : z \in \mathbb{C}, 1 \leq |z| \leq \nu_j \right\}.$$

We then define the (filled-in) *Bernstein polyellipse* of parameter ν as

$$\mathcal{E}_\nu := \mathcal{E}_{\nu_1} \times \dots \times \mathcal{E}_{\nu_s} \subseteq \mathbb{C}^s$$

and observe that $[-1, 1]^s \subseteq \mathcal{E}_\nu$ when we interpret $[-1, 1]^s$ as a subset of $\mathbb{C}^s$. We further define

$$\mathcal{V}_s(\nu) := \left\{ f : [-1, 1]^s \rightarrow \mathbb{C} : \exists \text{ open } U \supseteq \mathcal{E}_\nu \text{ and } \tilde{f} : U \rightarrow \mathbb{C} \text{ holomorphic with } \tilde{f}|_{[-1, 1]^s} = f \right\}.$$

Here, $[-1, 1]^s$ is again interpreted as a subset of $\mathbb{C}^s$. Moreover, note that such an extension $\tilde{f}$ is, if existent, unique, as follows from the identity theorem for holomorphic functions. Hence, the expression

$$\|f\|_{\mathcal{V}_s(\nu)} := \|\tilde{f}\|_{L^\infty(\mathcal{E}_\nu; \mathbb{C})}$$

is well-defined.

Definition D.2 (cf. [2, pp. 25, 28 ff.]). Let $s \in \mathbb{N}$. We define a probability measure on $[-1, 1]^s$ via

$$d\mu_s := \prod_{j=1}^s \frac{1}{\pi \sqrt{1 - x_j^2}} dx.$$

We define the *normalized* Chebyshev polynomials for $\mathbf{k} \in \mathbb{N}_0^s$ as

$$\widetilde{T}_{\mathbf{k}}(x) := 2^{\|\mathbf{k}\|_0/2} \prod_{j=1}^s \cos(\mathbf{k}_j \arccos(x_j)),$$

where $\|\mathbf{k}\|_0 := \#\{1 \leq j \leq s : \mathbf{k}_j \neq 0\}$. Note that this definition differs slightly from the notion used in Appendices C.1 and C.3.

The following lemma is crucial for deriving the approximation rate in Theorem D.5. The proofs can be found in [2].

Lemma D.3 (cf. [2, Remark 2.15 and Theorem 3.2]). Let $s \in \mathbb{N}$, $\mathbf{k} \in \mathbb{N}_0^s$, $\nu \in (1, \infty)^s$ and $f \in \mathcal{V}_s(\nu)$. Then it holds

1. $\|\widetilde{T}_{\mathbf{k}}\|_{L^\infty([-1, 1]^s; \mathbb{R})} = 2^{\|\mathbf{k}\|_0/2}$, where $\|\mathbf{k}\|_0 = \#\{1 \leq j \leq s : \mathbf{k}_j \neq 0\}$;
2. $|\langle \widetilde{T}_{\mathbf{k}}, f \rangle_{\mu_s}| \leq \nu^{-\mathbf{k}} \cdot 2^{\|\mathbf{k}\|_0/2} \cdot \|f\|_{\mathcal{V}_s(\nu)}.$

It is a well-known fact that the Chebyshev polynomials $\widetilde{T}_{\mathbf{k}}$ form an *orthonormal basis* of $L^2_{\mu_s}([-1, 1]^s; \mathbb{C})$. The following proposition states that functions from $\mathcal{V}_s(\nu)$ can even be approximated uniformly by linear combinations of the $\widetilde{T}_{\mathbf{k}}$ at a certain rate. The proof follows essentially by applying Lemma D.3.

Proposition D.4. Let $s, m \in \mathbb{N}$ and $\boldsymbol{\nu} \in (1, \infty)^s$. Let $\nu := \min_{j=1, \dots, s} \nu_j$. Then there exists a constant $c = c(s, \boldsymbol{\nu}) > 0$ with the following property: For every $f \in \mathcal{V}_s(\boldsymbol{\nu})$, defining

$$P_m := \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \leq m}} \langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s} \cdot \widetilde{T}_{\mathbf{k}},$$

it holds

$$\|f - P_m\|_{L^\infty([-1, 1]^s; \mathbb{C})} \leq c \cdot \nu^{-m} \cdot \|f\|_{\mathcal{V}_s(\boldsymbol{\nu})}.$$

Proof. Let $f \in \mathcal{V}_s(\boldsymbol{\nu})$. Since the $\widetilde{T}_{\mathbf{k}}$ form an orthonormal basis of $L_{\mu_s}^2([-1, 1]^s; \mathbb{C})$ and since μ_s is a probability measure, so that $f \in C([-1, 1]^s; \mathbb{C}) \subseteq L_{\mu_s}^2([-1, 1]^s; \mathbb{C})$, it follows that

$$f = \sum_{\mathbf{k} \in \mathbb{N}_0^s} \langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s} \cdot \widetilde{T}_{\mathbf{k}}$$

with unconditional convergence in $L_{\mu_s}^2$. For $x \in [-1, 1]^s$, note that

$$\sum_{\mathbf{k} \in \mathbb{N}_0^s} |\langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s}| \cdot |\widetilde{T}_{\mathbf{k}}(x)| \leq \sum_{\mathbf{k} \in \mathbb{N}_0^s} |\langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s}| \cdot \|\widetilde{T}_{\mathbf{k}}\|_{L^\infty([-1, 1]^s; \mathbb{C})} \leq 2^s \|f\|_{\mathcal{V}_s(\boldsymbol{\nu})} \cdot \sum_{\mathbf{k} \in \mathbb{N}_0^s} \nu^{-\mathbf{k}}.$$

Here, we employed Lemma D.3 at the last inequality. From $\nu > 1$ it follows that

$$\sum_{\mathbf{k} \in \mathbb{N}_0^s} \langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s} \cdot \widetilde{T}_{\mathbf{k}}$$

converges *pointwise* (even uniformly) and, since all the involved functions are continuous, this pointwise limit then has to coincide with the $L_{\mu_s}^2$ -limit f . Hence, it holds

$$\begin{aligned} \|f - P_m\|_{L^\infty([-1, 1]^s; \mathbb{C})} &\leq \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k} \not\leq m}} |\langle f, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_s}| \cdot \|\widetilde{T}_{\mathbf{k}}\|_{L^\infty([-1, 1]^s; \mathbb{C})} \\ &\leq 2^s \|f\|_{\mathcal{V}_s(\boldsymbol{\nu})} \cdot \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ |\mathbf{k}| \not\leq m}} \nu^{-\mathbf{k}}, \end{aligned}$$

where we again used Lemma D.3. To complete the proof we compute

$$\begin{aligned} \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ |\mathbf{k}| \not\leq m}} \nu^{-\mathbf{k}} &\leq \sum_{j=1}^s \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^s \\ \mathbf{k}_j > m}} \nu^{-\mathbf{k}} = \sum_{j=1}^s \left(\prod_{\substack{1 \leq \ell \leq s \\ \ell \neq j}} \left(\sum_{k=0}^{\infty} \nu_\ell^{-k} \right) \cdot \sum_{k=m+1}^{\infty} \nu_j^{-k} \right) \\ &\leq \sum_{j=1}^s \left(\nu_j^{-(m+1)} \prod_{1 \leq \ell \leq s} \left(\sum_{k=0}^{\infty} \nu_\ell^{-k} \right) \right) \\ &= \prod_{1 \leq \ell \leq s} \left(\sum_{k=0}^{\infty} \nu_\ell^{-k} \right) \cdot \sum_{j=1}^s \nu_j^{-(m+1)} \\ &= \prod_{1 \leq \ell \leq s} \left(\frac{\nu_\ell}{\nu_\ell - 1} \right) \cdot s \cdot \nu^{-(m+1)} \end{aligned}$$

and define $c(s, \boldsymbol{\nu}) := 2^s \cdot s \cdot \prod_{\ell=1}^s \frac{\nu_\ell}{\nu_\ell - 1} \cdot \nu^{-1}$. $\square$

To formulate the result for the approximation of holomorphically extendable functions using CVNNs we need to transfer the definition of $\mathcal{V}_s(\boldsymbol{\nu})$ to the complex setting. For $\boldsymbol{\nu} \in (1, \infty)^{2n}$ and with φ_n as in Equation (A.1), we hence write

$$\mathcal{W}_n(\boldsymbol{\nu}) := \left\{ f : \Omega_n \rightarrow \mathbb{C} : f \circ \varphi_n|_{[-1, 1]^{2n}} \in \mathcal{V}_{2n}(\boldsymbol{\nu}) \right\}$$

For $f \in \mathcal{W}_n(\nu)$ we define

$$\|f\|_{\mathcal{W}_n(\nu)} := \|f \circ \varphi_n\|_{\mathcal{V}_{2n}(\nu)}.$$

Thus, $\mathcal{W}_n(\nu)$ consists of all complex-valued functions defined on Ω_n that can be holomorphically extended onto some polyellipse in $\mathbb{C}^{2n}$, where $\Omega_n \subseteq \mathbb{C}^n$ is interpreted as a subset of $\mathbb{R}^{2n}$ and then as a subset of $\mathbb{C}^{2n}$. The final approximation result then reads as follows.

Theorem D.5. *Let $n \in \mathbb{N}$ and $\nu \in (1, \infty)^{2n}$. Set*

$$\nu := \min_{1 \leq j \leq 2n} \nu_j.$$

Then there exists a constant $c = c(n, \nu) > 0$ with the following property: For every function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ that is smooth and non-polyharmonic on some open set $\emptyset \neq U \subset \mathbb{C}$ and for every $m \in \mathbb{N}$ there exists a first layer $\Phi \in \mathcal{F}_{n,m}^\phi$ with the property that for every $f \in \mathcal{W}_n(\nu)$ there exist coefficients $\sigma = \sigma(f) \in \mathbb{C}^m$ such that

$$\|f - \sigma^T \Phi\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c \cdot \nu^{-m^{1/(2n)}/17} \cdot \|f\|_{\mathcal{W}_n(\nu)}.$$

Moreover, the map $f \mapsto \sigma(f)$ is a continuous linear functional with respect to the L^∞ -norm.

Proof. Choose $M \in \mathbb{N}$ as the largest integer satisfying $(8M + 1)^{2n} \leq m$, where we assume without loss of generality that $9^{2n} \leq m$. This can be done by choosing $\sigma = 0$ for $m < 9^{2n}$ at the cost of possibly enlarging c . Note that the maximality of M implies $(8M + 9)^{2n} > m$. By using $8M + 9 \leq 17M$, this gives us

$$M > \frac{1}{17} \cdot m^{1/(2n)}. \quad (\text{D.1})$$

Choose the constant $c_1 = c_1(2n, \nu)$ according to Proposition D.4.

Fix $\mathbf{k} \in \mathbb{N}_0^{2n}$ with $\mathbf{k} \leq M$. Since $\widetilde{T}_{\mathbf{k}}$ is a polynomial of componentwise degree at most M , we have a representation

$$\left(\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}\right)(z) = \sum_{\substack{\ell^1, \ell^2 \in \mathbb{N}_0^n \\ \ell^1, \ell^2 \leq M}} a_{\ell^1, \ell^2}^{\mathbf{k}} \prod_{t=1}^n \operatorname{Re}(z_t)^{\ell_t^1} \operatorname{Im}(z_t)^{\ell_t^2}$$

with suitably chosen coefficients $a_{\ell^1, \ell^2}^{\mathbf{k}} \in \mathbb{C}$. By using the identities $\operatorname{Re}(z_t) = \frac{1}{2}(z_t + \bar{z}_t)$ and also $\operatorname{Im}(z_t) = \frac{1}{2i}(z_t - \bar{z}_t)$, we can rewrite $\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}$ into a complex polynomial in z and $\bar{z}$, i.e.,

$$\left(\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}\right)(z) = \sum_{\substack{\ell^1, \ell^2 \in \mathbb{N}_0^n \\ \ell^1, \ell^2 \leq 2M}} b_{\ell^1, \ell^2}^{\mathbf{k}} z^{\ell^1} \bar{z}^{\ell^2}$$

with complex coefficients $b_{\ell^1, \ell^2}^{\mathbf{k}} \in \mathbb{C}$. Using Theorem 3.1, we choose $\rho_1, \dots, \rho_m \in \mathbb{C}^n$ and $b \in \mathbb{C}$, such that for any polynomial $P \in \left\{\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1} : \mathbf{k} \leq M\right\} \subseteq \mathcal{P}_{2M}^n$ there exist coefficients $\sigma_1(P), \dots, \sigma_m(P) \in \mathbb{C}$, such that

$$\|g_P - P\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \left(\sum_{\mathbf{k} \in \mathbb{N}_0^{2n}} \nu^{-\mathbf{k}} \right)^{-1} \cdot \nu^{-(M+1)}, \quad (\text{D.2})$$

where

$$g_P := \sum_{t=1}^m \sigma_t(P) \phi(\rho_t^T z + b).$$

Note that here we implicitly use the bound $(4 \cdot (2M) + 1)^{2n} \leq m$. We are now going to show that the first layer $\Phi \in \mathcal{F}_{n,m}^\phi$ defined using the ρ_t and b (i.e., $\Phi(z) = (\phi(\rho_t^T z + b))_{t=1}^m$) has the desired property.

To this end, take an arbitrary function $f \in \mathcal{W}_n(\nu)$. Proposition D.4 tells us that

$$\|f - P_M \circ \varphi_n^{-1}\|_{L^\infty(\Omega_n; \mathbb{C})} = \|f \circ \varphi_n - P_M\|_{L^\infty([-1,1]^{2n}; \mathbb{C})} \leq c_1 \cdot \nu^{-(M+1)} \cdot \|f \circ \varphi_n\|_{\mathcal{V}_{2n}(\nu)}, \quad (\text{D.3})$$

where

$$P_M \circ \varphi_n^{-1} = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq M}} \langle f \circ \varphi_n, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_{2n}} \cdot (\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}).$$

We then define the approximating network

$$g := \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq M}} \langle f \circ \varphi_n, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_{2n}} \cdot g_{\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}}.$$

From the definition of the functions $g_{\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}}$ it follows directly that g is a shallow CVNN with first layer $\Phi \in \mathcal{F}_{n,m}^\phi$. Furthermore, it holds

$$\begin{aligned} \|P_M \circ \varphi_n^{-1} - g\|_{L^\infty(\Omega_n; \mathbb{C})} &\leq \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq M}} |\langle f \circ \varphi_n, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_{2n}}| \cdot \|\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1} - g_{\widetilde{T}_{\mathbf{k}} \circ \varphi_n^{-1}}\|_{L^\infty(\Omega_n; \mathbb{C})} \\ &\stackrel{(\text{D.2}), \text{Lem. D.3}}{\leq} 2^n \cdot \|f \circ \varphi_n\|_{\mathcal{V}_{2n}(\nu)} \cdot \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq M}} \nu^{-\mathbf{k}} \cdot \left(\sum_{\mathbf{k} \in \mathbb{N}_0^{2n}} \nu^{-\mathbf{k}} \right)^{-1} \cdot \nu^{-(M+1)} \\ &\leq 2^n \cdot \|f\|_{\mathcal{W}_n(\nu)} \cdot \nu^{-(M+1)}. \end{aligned}$$

Combining this estimate with Equation (D.3) and applying the triangle inequality gives us

$$\|f - g\|_{L^\infty(\Omega_n; \mathbb{C})} \leq (c_1 + 2^n) \cdot \nu^{-(M+1)} \cdot \|f\|_{L^\infty(\mathcal{E}_\nu; \mathbb{C})}.$$

Hence, the claim follows taking $c := c_1 + 2^n$ and using (D.1).

The continuity of the linear map $f \mapsto \sigma(f)$ with respect to the L^∞ -norm follows directly from the fact that $f \mapsto \langle f \circ \varphi_n, \widetilde{T}_{\mathbf{k}} \rangle_{\mu_{2n}}$ is trivially continuous with respect to the $L^2_{\mu_{2n}}$ -norm and hence also continuous with respect to the L^∞ -norm since μ_{2n} is a probability measure. $\square$

E Postponed proofs for the optimality results in the case of continuous weight selection

In this section we provide the proofs for the optimality results derived if a continuous weight selection is assumed. Specifically, we prove Theorems 4.1 and 5.1. The proofs of both results rely on a very general result about the approximation in normed spaces by subsets that are parametrizable with m parameters [19]. We decided to include a detailed proof for this approximation result in this paper since the nature of the continuity assumption in [19] is not completely clear.

Proposition E.1 ([19, Theorem 3.1]). *Let $(X, \|\cdot\|_X)$ be a normed space, $\emptyset \neq K \subseteq X$ a subset and $V \subseteq X$ a linear, not necessarily closed subspace of X containing K . Let $m \in \mathbb{N}$, let $\bar{a} : K \rightarrow \mathbb{R}^m$ be a map which is continuous with respect to some norm $\|\cdot\|_V$ on V and $M : \mathbb{R}^m \rightarrow X$ some arbitrary map. Let*

$$b_m(K)_X := \sup_{X_{m+1}} \sup \{ \varrho \geq 0 : U_\varrho(X_{m+1}) \subseteq K \}, \quad (\text{E.1})$$

where the first supremum is taken over all $(m+1)$ -dimensional linear subspaces X_{m+1} of X and

$$U_\varrho(X_{m+1}) := \{y \in X_{m+1} : \|y\|_X \leq \varrho\}.$$

Further, we set $b_m(K)_X := 0$ if the supremum in (E.1) is not well-defined as a quantity in $[0, \infty]$. Then it holds

$$\sup_{x \in K} \|x - M(\bar{a}(x))\|_X \geq b_m(K)_X.$$

Proof. The claim is trivial if $b_m(K)_X = 0$. Thus, assume $b_m(K)_X > 0$. Let $0 < \varrho \leq b_m(K)_X$ be any number such that there exists an $(m+1)$ -dimensional subspace X_{m+1} of X with $U_\varrho(X_{m+1}) \subseteq K$. It follows $U_\varrho(X_{m+1}) \subseteq V$, hence $X_{m+1} \subseteq V$, so $\|\cdot\|_V$ defines a norm on X_{m+1} . Thus, the

restriction of $\bar{a}$ to $\partial U_\varrho(X_{m+1})$ is a continuous mapping to $\mathbb{R}^m$ with respect to $\|\cdot\|_V$. Since all norms are equivalent on the finite-dimensional space X_{m+1} , the Borsuk-Ulam-Theorem [18, Corollary 4.2] yields the existence of a point $x_0 \in \partial U_\varrho(X_{m+1})$ with $\bar{a}(x_0) = \bar{a}(-x_0)$. We then see

$$\begin{aligned} 2\varrho &= 2\|x_0\|_X \leq \|x_0 - M(\bar{a}(x_0))\|_X + \|x_0 + M(\bar{a}(-x_0))\|_X \\ &= \|x_0 - M(\bar{a}(x_0))\|_X + \|-x_0 - M(\bar{a}(-x_0))\|_X, \end{aligned}$$

and hence at least one of the two summands on the right has to be larger than or equal to ϱ . $\square$

E.1 Proof of Theorem 4.1

Using Proposition E.1, we can deduce our lower bound in the context of C^k -spaces. The proof is in fact almost identical to what is done in [19, Theorem 4.2]. However, we decided to include a detailed proof in this paper, since [19] considers Sobolev functions and not C^k -functions.

Theorem E.2. *Let $s, k \in \mathbb{N}$. Then there exists a constant $c = c(s, k) > 0$ with the following property: For any $m \in \mathbb{N}$ and any map $\bar{a} : C^k([-1, 1]^s; \mathbb{R}) \rightarrow \mathbb{R}^m$ that is continuous with respect to some norm on $C^k([-1, 1]^s; \mathbb{R})$ and any (possibly discontinuous) map $M : \mathbb{R}^m \rightarrow C([-1, 1]^s; \mathbb{R})$, we have*

$$\sup_{\substack{f \in C^k([-1, 1]^s; \mathbb{R}) \\ \|f\|_{C^k([-1, 1]^s; \mathbb{R})} \leq 1}} \|f - M(\bar{a}(f))\|_{L^\infty([-1, 1]^s; \mathbb{R})} \geq c \cdot m^{-k/s}.$$

Proof. The idea is to apply Proposition E.1 to $X := C([-1, 1]^s; \mathbb{R})$, $V := C^k([-1, 1]^s; \mathbb{R})$ and the set $K := \{f \in C^k([-1, 1]^s; \mathbb{R}) : \|f\|_{C^k([-1, 1]^s; \mathbb{R})} \leq 1\}$.

Assume in the beginning that $m = n^s$ with an integer $n > 1$. Pick $\phi \in C^\infty(\mathbb{R}^s)$ with $\phi \equiv 1$ on $[-3/4, 3/4]^s$ and $\phi \equiv 0$ outside of $[-1, 1]^s$. Fix $c_0 = c_0(s, k) > 0$ with

$$1 \leq \|\phi\|_{C^k([-1, 1]^s; \mathbb{R})} \leq c_0.$$

Let $Q_1, \dots, Q_m$ be the partition (disjoint up to null-sets) of $[-1, 1]^s$ into closed cubes of sidelength $2/n$. For every $j \in \{1, \dots, m\}$ we write $Q_j = \times_{\ell=1}^s [a_\ell^{(j)} - 1/n, a_\ell^{(j)} + 1/n]$ with an appropriately chosen vector $a = (a_1^{(j)}, \dots, a_s^{(j)}) \in [-1, 1]^s$ and let

$$\phi_j(x) := \phi(n(x - a^{(j)})) \text{ for } x \in \mathbb{R}^s.$$

By choice of ϕ , the maps ϕ_j are supported on a proper subset of Q_j for every $j \in \{1, \dots, m\}$ and an inductive argument shows

$$\partial^{\mathbf{k}} \phi_j(x) = n^{|\mathbf{k}|} \cdot (\partial^{\mathbf{k}} \phi)(n(x - a^{(j)})) \text{ for every } \mathbf{k} \in \mathbb{N}_0^s \text{ and } x \in \mathbb{R}^s$$

and hence in particular

$$\|\phi_j\|_{C^k([-1, 1]^s; \mathbb{R})} \leq n^{|\mathbf{k}|} \cdot c_0. \quad (\text{E.2})$$

Let $X_m := \text{span}\{\phi_1, \dots, \phi_m\}$ and $S \in U_1(X_m) = \{f \in X_m : \|f\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq 1\}$. Then we can write S in the form $S = \sum_{j=1}^m c_j \phi_j$ with real numbers $c_1, \dots, c_m \in \mathbb{R}$. Suppose there exists $j^* \in \{1, \dots, m\}$ with $|c_{j^*}| > 1$. Then we have

$$\|S\|_{L^\infty([-1, 1]^s; \mathbb{R})} \geq |S(a^{(j^*)})| \geq |c_{j^*}| > 1,$$

since the functions ϕ_j have disjoint support and $\phi_j(a^{(j)}) = 1$. This is a contradiction to $S \in U_1(X_m)$ and we can thus infer that $\max_j |c_j| \leq 1$. Furthermore, we see again because the functions ϕ_j have disjoint support that

$$\|\partial^{\mathbf{k}} S\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq \max_j |c_j| \cdot \|\partial^{\mathbf{k}} \phi_j\|_{L^\infty([-1, 1]^s; \mathbb{R})} \stackrel{(\text{E.2})}{\leq} n^{|\mathbf{k}|} \cdot c_0 \leq c_0 \cdot n^k = c_0 \cdot m^{k/s}$$

for every $\mathbf{k} \in \mathbb{N}_0^s$ with $|\mathbf{k}| \leq k$ and hence

$$\|S\|_{C^k([-1, 1]^s; \mathbb{R})} \leq c_0 \cdot m^{k/s}.$$

Thus, letting $\varrho := c_0^{-1} \cdot m^{-k/s}$ yields $U_\varrho(X_m) \subseteq K$, so we see by Proposition E.1 that

$$\sup_{f \in K} \|f - M_{m-1}(\bar{a}(f))\|_{L^\infty([-1,1]^s; \mathbb{R})} \geq \varrho = c_1 \cdot m^{-k/s}$$

with $c_1 = c_0^{-1}$ for every map $\bar{a} : X \rightarrow \mathbb{R}^{m-1}$ which is continuous with respect to some norm on V and any map $M_{m-1} : \mathbb{R}^{m-1} \rightarrow X$. Using the inequality $m \leq 2(m-1)$ (note $m > 1$) we get

$$\sup_{f \in K} \|f - M_{m-1}(\bar{a}(f))\|_{L^\infty([-1,1]^s; \mathbb{R})} \geq c_1 \cdot m^{-k/s} \geq c_1 \cdot (2(m-1))^{-k/s} \geq c_2 \cdot (m-1)^{-k/s}$$

with $c_2 = c_1 \cdot 2^{-k/s}$. Hence, the claim has been shown for all numbers m of the form $n^s - 1$ with an integer $n > 1$.

In the end, let $m \in \mathbb{N}$ be arbitrary and pick $n \in \mathbb{N}$ with $n^s \leq m < (n+1)^s$. For given maps $\bar{a} : V \rightarrow \mathbb{R}^m$ and $M : \mathbb{R}^m \rightarrow X$ with $\bar{a}$ continuous with respect to some norm on V , let

$$\tilde{a} : V \rightarrow \mathbb{R}^{(n+1)^s-1}, \quad f \mapsto (\bar{a}(f), 0) \quad \text{and}$$

$$M_{(n+1)^s-1} : \mathbb{R}^{(n+1)^s-1} \rightarrow X, \quad (x, y) \mapsto M(x),$$

where $x \in \mathbb{R}^m$, $y \in \mathbb{R}^{(n+1)^s-1-m}$. Then we get

$$\begin{aligned} \sup_{f \in K} \|f - M(\bar{a}(f))\|_{L^\infty([-1,1]^s; \mathbb{R})} &= \sup_{f \in K} \|f - M_{(n+1)^s-1}(\tilde{a}(f))\|_{L^\infty([-1,1]^s; \mathbb{R})} \\ &\geq c_2 \cdot ((n+1)^s - 1)^{-k/s} \geq c_2 \cdot (2^s n^s)^{-k/s} \geq c_3 \cdot m^{-k/s} \end{aligned}$$

with $c_3 = c_2 \cdot 2^{-k}$. Here we used the bound $(n+1)^s - 1 \leq (2n)^s$. This proves the full claim. $\square$

Using this theorem, we can now prove Theorem 4.1.

Proof of Theorem 4.1. Let $\bar{a} : C^k(\Omega_n; \mathbb{C}) \rightarrow \mathbb{C}^m$ be any map that is continuous with respect to some norm $\|\cdot\|_V$ on $C^k(\Omega_n; \mathbb{C})$, and let $M : \mathbb{C}^m \rightarrow C(\Omega_n; \mathbb{C})$ be arbitrary. With φ_n, φ_m defined as in Equation (A.1), let

$$\tilde{a} : C^k([-1,1]^{2n}; \mathbb{R}) \rightarrow \mathbb{R}^{2m}, \quad \tilde{f} \mapsto \varphi_m^{-1} \left(\bar{a} \left(\tilde{f} \circ \varphi_n^{-1} \big|_{\Omega_n} \right) \right).$$

Clearly, $\tilde{a}$ is continuous on $C^k([-1,1]^{2n}; \mathbb{R})$ with respect to the norm $\|\cdot\|_{\tilde{V}}$ on $C^k([-1,1]^{2n}; \mathbb{R})$ defined as

$$\|\tilde{f}\|_{\tilde{V}} := \left\| \tilde{f} \circ \varphi_n^{-1} \big|_{\Omega_n} \right\|_V \quad \text{for } \tilde{f} \in C^k([-1,1]^{2n}; \mathbb{R}).$$

Let

$$\widetilde{M} : \mathbb{R}^{2m} \rightarrow C([-1,1]^{2n}; \mathbb{R}), \quad \widetilde{M}(x) := \operatorname{Re}(M(\varphi_m(x))) \circ \varphi_n \big|_{[-1,1]^{2n}}.$$

Then it holds

$$\begin{aligned} &\sup_{f \in C^k(\Omega_n; \mathbb{C})} \|f - M(\bar{a}(f))\|_{L^\infty(\Omega_n; \mathbb{C})} \\ &\leq \sup_{\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1} \|f - \operatorname{Re}(M(\bar{a}(f)))\|_{L^\infty(\Omega_n; \mathbb{R})} \\ &= \sup_{\substack{\tilde{f} \in C^k([-1,1]^{2n}; \mathbb{R}) \\ \|\tilde{f}\|_{C^k([-1,1]^{2n}; \mathbb{R})} \leq 1}} \left\| \tilde{f} \circ \varphi_n^{-1} - \operatorname{Re} \left(M \left(\bar{a} \left(\tilde{f} \circ \varphi_n^{-1} \big|_{\Omega_n} \right) \right) \right) \right\|_{L^\infty(\Omega_n; \mathbb{R})} \\ &= \sup_{\substack{\tilde{f} \in C^k([-1,1]^{2n}; \mathbb{R}) \\ \|\tilde{f}\|_{C^k([-1,1]^{2n}; \mathbb{R})} \leq 1}} \left\| \tilde{f} - \operatorname{Re} \left(M \left(\varphi_m \left(\varphi_m^{-1} \left(\bar{a} \left(\tilde{f} \circ \varphi_n^{-1} \big|_{\Omega_n} \right) \right) \right) \right) \right) \circ \varphi_n \right\|_{L^\infty([-1,1]^{2n}; \mathbb{R})} \\ &= \sup_{\substack{\tilde{f} \in C^k([-1,1]^{2n}; \mathbb{R}) \\ \|\tilde{f}\|_{C^k([-1,1]^{2n}; \mathbb{R})} \leq 1}} \left\| \tilde{f} - \widetilde{M}(\tilde{a}(\tilde{f})) \right\|_{L^\infty([-1,1]^{2n}; \mathbb{R})} \geq \tilde{c} \cdot (2m)^{-k/(2n)}, \end{aligned}$$

with a constant $\tilde{c} = \tilde{c}(n, k)$ provided by Theorem E.2. Hence, the claim follows by choosing $c = c(n, k) := 2^{-k/(2n)} \cdot \tilde{c}$. $\square$

As a corollary, we formulate a special case of Theorem 4.1 for the case of shallow complex-valued neural networks.

Corollary E.3. *Let $n, k \in \mathbb{N}$. Then there exists a constant $c = c(n, k) > 0$ with the following property: For any $m \in \mathbb{N}$, $\phi \in C(\mathbb{C}; \mathbb{C})$ and any map*

$$\eta : C^k(\Omega_n; \mathbb{C}) \rightarrow (\mathbb{C}^n)^m \times \mathbb{C}^m \times \mathbb{C}^m, \quad g \mapsto (\eta_1(g), \eta_2(g), \eta_3(g))$$

which is continuous with respect to some norm on $C^k(\Omega_n; \mathbb{C})$, there exists $f \in C^k(\Omega_n; \mathbb{C})$ satisfying $\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1$ and

$$\|f - \Psi(f)\|_{L^\infty(\Omega_n; \mathbb{C})} \geq c \cdot m^{-k/(2n)},$$

where $\Psi(f) \in C(\Omega_n; \mathbb{C})$ is given by

$$\Psi(f)(z) := \sum_{j=1}^m (\eta_3(f))_j \phi \left([\eta_1(f)]_j^T z + (\eta_2(f))_j \right).$$

Proof. Using Theorem 4.1, we deduce that there exists $f \in C^k(\Omega_n; \mathbb{C})$ satisfying $\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1$ and

$$\|f - \Psi(f)\|_{L^\infty(\Omega_n; \mathbb{C})} \geq c' \cdot (m(n+2))^{-k/(2n)}$$

for a constant $c' = c'(n, k) > 0$. Hence, the claim follows by letting $c := c' \cdot (n+2)^{-k/(2n)}$. $\square$

E.2 Proof of Theorem 5.1

We can use Proposition E.1 not only to show that the rate of convergence established in this paper is optimal (which is done in Appendix E.1) but also to show that the problem of approximating C^k -functions using a set of functions that can be parametrized with finitely many parameters is intractable in the sense that it suffers from the curse of dimensionality, provided that the map which assigns to each C^k -function the parameters of the approximating function is continuous. This is the subject of this section.

In [35] a certain space of polynomials was used to show the intractability in the case of *linear* approximation methods. We are also going to use this class of polynomials, but combine it with Proposition E.1 to infer intractability in the case of *continuous* approximation methods. We start with a lemma discussing an important property of this space of polynomials. This property is stated as part of a proof in [35], but no complete proof is provided.

Lemma E.4. *Let $s \in \mathbb{N}$ and consider a function $f \in C^\infty([-1, 1]^s; \mathbb{R})$ which is given via*

$$f(x) = \sum_{\mathbf{k} \in \{0,1\}^s} a_{\mathbf{k}} x^{\mathbf{k}} \tag{E.3}$$

with coefficients $a_{\mathbf{k}} \in \mathbb{R}$ for every $\mathbf{k} \in \{0,1\}^s$. Then it holds

$$\|f\|_{C^k([-1,1]^s; \mathbb{R})} = \|f\|_{L^\infty([-1,1]^s; \mathbb{R})}$$

for every $k \in \mathbb{N}$.

Proof. The proof is by induction over s . We start with the case $s = 1$ and note that we can write $f(x) = ax + b$ with $a, b \in \mathbb{R}$ in that case. Switching to $-f$ if necessary, we can assume $a \geq 0$. Clearly, $\|f\|_{L^\infty([-1,1]; \mathbb{R})} \leq |a| + |b|$. Conversely, if $b \geq 0$ then $|f(1)| = |a + b| = |a| + |b|$. If otherwise $b < 0$ then $|f(-1)| = |b - a| = |a - b| = |a| + |b|$. Thus, $\|f\|_{L^\infty([-1,1]; \mathbb{R})} = |a| + |b|$. For the derivatives, we have $\|f'\|_{L^\infty([-1,1]; \mathbb{R})} = |a|$ and $\|f^{(k)}\|_{L^\infty([-1,1]; \mathbb{R})} = 0$ for $k \geq 2$. This proves the claim in the case $s = 1$.

We now assume that the claim holds for some arbitrary but fixed $s \in \mathbb{N}$. We further let $\alpha \in \mathbb{N}_0^{s+1}$ and fix a point $(x_1, \dots, x_{s+1}) \in [-1, 1]^{s+1}$. We decompose $\alpha = (\alpha', \alpha_{s+1})$ with $\alpha' \in \mathbb{N}_0^s$. Let

$$\tilde{f} : [-1, 1] \rightarrow \mathbb{R}, \quad y_{s+1} \mapsto \partial^{(\alpha', 0)} f(x_1, \dots, x_s, y_{s+1})$$

and note

$$\partial^\alpha f(x_1, \dots, x_{s+1}) = \tilde{f}^{(\alpha_{s+1})}(x_{s+1}).$$

Note that f is affine-linear with respect to each variable (with all other variables hold fixed). Hence, $\tilde{f}$ is an affine function and we can thus apply the case $s = 1$ to $\tilde{f}$ and get

$$\|\tilde{f}^{(\alpha_{s+1})}\|_{L^\infty([-1,1];\mathbb{R})} \leq \|\tilde{f}\|_{L^\infty([-1,1];\mathbb{R})}.$$

Putting this together, we infer

$$|\partial^\alpha f(x_1, \dots, x_{s+1})| \leq \|\tilde{f}\|_{L^\infty([-1,1];\mathbb{R})} = \sup_{y_{s+1} \in [-1,1]} |\partial^{(\alpha',0)} f(x_1, \dots, x_s, y_{s+1})|. \quad (\text{E.4})$$

We now fix an arbitrary point $y_{s+1} \in [-1, 1]$ and consider

$$\hat{f}: [-1, 1]^s \rightarrow \mathbb{R}, \quad (y_1, \dots, y_s) \mapsto f(y_1, \dots, y_s, y_{s+1}).$$

Then it holds

$$\partial^{(\alpha',0)} f(x_1, \dots, x_s, y_{s+1}) = \partial^{\alpha'} \hat{f}(x_1, \dots, x_s).$$

Applying the induction hypothesis to $\hat{f}$ (which is easily seen to be of the form (E.3)) we get

$$|\partial^{\alpha'} \hat{f}(x_1, \dots, x_s)| \leq \|\hat{f}\|_{L^\infty([-1,1]^s;\mathbb{R})} \leq \|f\|_{L^\infty([-1,1]^{s+1};\mathbb{R})}. \quad (\text{E.5})$$

Combining (E.4) and (E.5) yields

$$|\partial^\alpha f(x_1, \dots, x_{s+1})| \leq \|f\|_{L^\infty([-1,1]^{s+1};\mathbb{R})}.$$

Since $\alpha \in \mathbb{N}_0^{s+1}$ was arbitrary, we get the claim by noting that

$$\|f\|_{C^k([-1,1]^s;\mathbb{R})} \geq \|f\|_{L^\infty([-1,1]^s;\mathbb{R})}$$

holds trivially for every $k \in \mathbb{N}$. □

Using the above lemma, we can now deduce that the approximation of smooth functions using continuous approximation methods is intractable in terms of the input dimension.

Proof of Theorem 5.1. We apply Proposition E.1 to $X := C([-1, 1]^s; \mathbb{R})$, $V := C^{\infty,*s}$ and to the set $K := \{f \in C^{\infty,*s} : \|f\|_{C^\infty([-1,1]^s;\mathbb{R})} \leq 1\}$ and $m := 2^s - 1$. The space

$$X_{m+1} := \left\{ [-1, 1]^s \ni x \mapsto \sum_{\mathbf{k} \in \{0,1\}^s} a_{\mathbf{k}} x^{\mathbf{k}} : a_{\mathbf{k}} \in \mathbb{R} \right\}$$

consisting of all functions considered in the previous Lemma E.4 is an $(m + 1)$ -dimensional subspace of $C([-1, 1]^s; \mathbb{R})$. For every $f \in X_{m+1}$ with $\|f\|_{L^\infty([-1,1]^s;\mathbb{R})} \leq 1$, Lemma E.4 tells us $\|f\|_{C^\infty([-1,1]^s;\mathbb{R})} \leq 1$. Hence, $U_1(X_{m+1}) \subseteq K$ and Proposition E.1 then yields the claim. □

Remark E.5. The statement of Theorem 5.1 also holds if the functions satisfy $\bar{a} : C^{\infty,*s} \rightarrow \mathbb{R}^m$ and $M : \mathbb{R}^m \rightarrow C([-1, 1]^s; \mathbb{R})$ with $m \leq 2^s - 1$. This can be seen by defining

$$\tilde{a} : C^{\infty,*s} \rightarrow \mathbb{R}^{2^s-1}, \quad f \mapsto (\bar{a}(f), 0, \dots, 0)$$

and

$$\tilde{M} : \mathbb{R}^{2^s-1} \rightarrow C([-1, 1]^s; \mathbb{R}), \quad (a, b) \mapsto M(a)$$

with $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^{2^s-1-m}$.

The following Corollary E.6 transfers Theorem 5.1 to the complex-valued setting.

Corollary E.6. *Let $n \in \mathbb{N}$. For any function $f \in C^\infty(\Omega_n; \mathbb{C})$ we write*

$$\|f\|_{C^\infty(\Omega_n;\mathbb{C})} := \sup_{k \in \mathbb{N}} \|f\|_{C^k(\Omega_n;\mathbb{C})}$$

and let $C_{\mathbb{C}}^{\infty,*n}$ denote the space consisting of all functions for which this expression is finite. Let $\bar{a} : C_{\mathbb{C}}^{\infty,*n} \rightarrow \mathbb{C}^{2^{2n-1}-1}$ be continuous with respect to some norm on $C_{\mathbb{C}}^{\infty,*n}$ and moreover, let $M : \mathbb{C}^{2^{2n-1}-1} \rightarrow C(\Omega_n; \mathbb{C})$ be an arbitrary map. Then it holds

$$\sup_{\substack{f \in C_{\mathbb{C}}^{\infty,*n} \\ \|f\|_{C^\infty(\Omega_n;\mathbb{C})} \leq 1}} \|f - M(\bar{a}(f))\|_{L^\infty(\Omega_n;\mathbb{C})} \geq 1.$$

Proof. The transfer to the complex-valued setting works in the same manner as the proof of Theorem 4.1 (see Appendix E.1). We write $m := 2^{2n-1} - 1$ and note $2m = 2^{2n} - 2 \leq 2^{2n} - 1$. We define $\tilde{a} : C^{\infty,*,2n}_{\mathbb{C}} \rightarrow \mathbb{R}^{2m}$ and $\tilde{M} : \mathbb{R}^{2m} \rightarrow C^k([-1, 1]^{2n}; \mathbb{R})$ in the same way as in the proof of Theorem 4.1. Using again the same technique as in the proof of Theorem 4.1, we get

$$\sup_{\substack{f \in C^{\infty,*,n}_{\mathbb{C}} \\ \|f\|_{C^{\infty}(\Omega_n; \mathbb{C})} \leq 1}} \|f - M(\tilde{a}(f))\|_{L^{\infty}(\Omega_n; \mathbb{C})} \geq \sup_{\substack{\tilde{f} \in C^{\infty,*,2n} \\ \|\tilde{f}\|_{C^{\infty}([-1, 1]^{2n}; \mathbb{R})} \leq 1}} \|\tilde{f} - \tilde{M}(\tilde{a}(\tilde{f}))\|_{L^{\infty}([-1, 1]^{2n}; \mathbb{R})} \geq 1,$$

applying Theorem 5.1 in the last inequality, using $2m \leq 2^{2n} - 1$. $\square$

We conclude this appendix by adding a note on the constant appearing in our main approximation bound.

Corollary E.7. *Let $n \in \mathbb{N}$ with $n \geq 2$ and $\alpha > 0$ and let $\phi \in C(\mathbb{C}; \mathbb{C})$. Let $\tilde{c} = \tilde{c}(n, \alpha) > 0$ be such that for every $m \in \mathbb{N}$ there exists a mapping*

$$\eta : C^{\infty,*,n}_{\mathbb{C}} \rightarrow (\mathbb{C}^n)^m \times \mathbb{C}^m \times \mathbb{C}^m, \quad g \mapsto (\eta_1(g), \eta_2(g), \eta_3(g))$$

that is continuous with respect to any norm on $C^{\infty,*,n}_{\mathbb{C}}$ and such that

$$\|f - \Psi(f)\|_{L^{\infty}(\Omega_n; \mathbb{C})} \leq (\tilde{c} \cdot m)^{-\alpha} \cdot \|f\|_{C^{\infty}(\Omega_n; \mathbb{C})},$$

for every $f \in C^{\infty,*,n}_{\mathbb{C}}$. Here, $\Psi(f) \in C(\Omega_n; \mathbb{C})$ is given by

$$\Psi(f)(z) := \sum_{j=1}^m (\eta_3(f))_j \phi \left([\eta_1(f)]_j^T z + (\eta_2(f))_j \right).$$

Then it necessarily holds $\tilde{c} \leq 16 \cdot 2^{-n}$.

Proof. We first assume $n \geq 4$. We take $m = \left\lfloor \frac{2^{2n-1}-1}{n+2} \right\rfloor$ and note that then $m(n+2) \leq 2^{2n-1} - 1$. Therefore, Corollary E.6 applies and we infer that for each $\varepsilon \in (0, 1)$, there exists $f = f_{\varepsilon} \in C^{\infty,*,n}_{\mathbb{C}}$ with $\|f\|_{C^{\infty}(\Omega_n; \mathbb{C})} \leq 1$ and such that

$$1 - \varepsilon \leq \|f - \Psi(f)\|_{L^{\infty}(\Omega_n; \mathbb{C})} \leq (\tilde{c} \cdot m)^{-\alpha} \cdot \|f\|_{C^{\infty}(\Omega_n; \mathbb{C})} \leq (\tilde{c} \cdot m)^{-\alpha}.$$

This then necessarily implies $\tilde{c} \cdot m \leq 1$ or equivalently $\tilde{c} \leq 1/m$. It therefore suffices to derive a lower bound for m . Firstly, we note

$$2^{2n-1} = 2^{n-3} \cdot 2^{n+2} = 2^{n-3} \cdot (1+1)^{n+2} \geq 2^{n-3}(n+3),$$

where we applied Bernoulli's inequality. Because of $n \geq 4 \geq 3$, this yields

$$2^{2n-1} - 1 \geq 2^{n-3}(n+3) - 2^{n-3} = 2^{n-3}(n+2).$$

Hence, we get

$$m \geq \frac{2^{2n-1} - 1}{n+2} - 1 \geq 2^{n-3} - 1 = 2^{n-3}(1 - 2^{3-n}) \geq 2^{n-4} = \frac{2^n}{16}.$$

Here, we used $n \geq 4$ in the last inequality. An explicit computation shows that the same bounds also holds in the cases $n = 2$ and $n = 3$. This proves the claim. $\square$

F Postponed proofs for the optimality results in the case of unrestricted weight selection

F.1 Approximation using Ridge Functions

In this section we prove for $s \in \mathbb{N}_{\geq 2}$ that every function in $C^k([-1, 1]^s; \mathbb{R})$ can be uniformly approximated with an error of the order $m^{-k/(s-1)}$ using a linear combination of m so-called *ridge functions*. In fact, we only consider ridge *polynomials*, meaning functions of the form

$$\mathbb{R}^s \rightarrow \mathbb{R}, \quad x \mapsto p(a^T x)$$

for a fixed vector $a \in \mathbb{R}^s$ and a polynomial $p : \mathbb{R} \rightarrow \mathbb{R}$. Note that this result has already been obtained in a slightly different form in [30, Theorem 1]; namely, it is shown there that the rate of approximation $m^{-k/(s-1)}$ can be achieved by functions of the form $\sum_{j=1}^m f_j(a_j^T x)$ with $a_j \in \mathbb{R}^s$ and $f_j \in L_{\text{loc}}^1(\mathbb{R})$. We will need the fact that the f_j can actually be chosen as polynomials and that the vectors $a_1, \dots, a_m$ can be chosen independently from the particular function f . This is shown in the proof of [30], but not stated explicitly. For this reason, and in order to clarify the proof itself and to make the paper more self-contained, we decided to present the proof in this appendix.

Lemma F.1. *Let $m, s \in \mathbb{N}$. Then we denote by*

$$P_m^s := \left\{ \mathbb{R}^s \rightarrow \mathbb{R}, \quad x \mapsto \sum_{\substack{k \in \mathbb{N}_0^s \\ |k| \leq m}} a_k x^k : a_k \in \mathbb{R} \right\}$$

the set of real polynomials of degree at most m . The subset of homogeneous polynomials of degree m is defined as

$$H_m^s := \left\{ \mathbb{R}^s \rightarrow \mathbb{R}, \quad x \mapsto \sum_{\substack{k \in \mathbb{N}_0^s \\ |k| = m}} a_k x^k : a_k \in \mathbb{R} \right\}.$$

Then there exists a constant $c = c(s) > 0$ satisfying

$$\dim(H_m^s) \leq c \cdot m^{s-1} \quad \forall m \in \mathbb{N}.$$

Proof. It is immediate that the set

$$\{\mathbb{R}^s \rightarrow \mathbb{R}, \quad x \mapsto x^{\mathbf{k}} : \mathbf{k} \in \mathbb{N}_0^s, |\mathbf{k}| = m\}$$

forms a basis of H_m^s , hence

$$\dim(H_m^s) = \#\{\mathbf{k} \in \mathbb{N}_0^s : |\mathbf{k}| = m\}.$$

This quantity clearly equals the number of possibilities for drawing m times from a set with s elements with replacement. Hence, we see

$$\dim(H_m^s) = \binom{s+m-1}{m},$$

see for instance [12, Identity 143]. A further estimation shows

$$\binom{s+m-1}{m} = \prod_{j=1}^{s-1} \frac{m+j}{j} = \prod_{j=1}^{s-1} \left(1 + \frac{m}{j}\right) \leq (1+m)^{s-1} \leq 2^{s-1} \cdot m^{s-1}.$$

Hence, the claim follows with $c(s) = 2^{s-1}$. $\square$

A combination of results from [36] together with the fact that it is possible to approximate C^k -functions using polynomials of degree at most m with an error of the order m^{-k} , as shown in Theorem C.15, yields the desired result.

Theorem F.2. *Let $s, k \in \mathbb{N}$ with $s \geq 2$ and $r > 0$. Then there exists a constant $c = c(s, k) > 0$ with the following property: For every $m \in \mathbb{N}$ there exist $a_1, \dots, a_m \in \mathbb{R}^s \setminus \{0\}$ with $\|a_j\|_2 = r$, such that for every function $f \in C^k([-1, 1]^s; \mathbb{R})$ there exist polynomials $p_1, \dots, p_m \in P_m^1$ satisfying*

$$\left\| f(x) - \sum_{j=1}^m p_j(a_j^T x) \right\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq c \cdot m^{-k/(s-1)} \cdot \|f\|_{C^k([-1, 1]^s; \mathbb{R})}.$$

Proof. We first pick the constant $c_1 = c_1(s)$ according to Lemma F.1. Then we define the constant $c_2 = c_2(s) := (2s)^{s-1} \cdot c_1(s)$ and let $M \in \mathbb{N}$ be the largest integer satisfying

$$c_2 \cdot M^{s-1} \leq m.$$

Here, we assume without loss of generality that $m \geq c_2$, which can be justified by choosing $p_j = 0$ for every $j \in \{1, \dots, m\}$ if $m < c_2$, at the cost of possibly enlarging c . Note that the choice of M implies $c_2 \cdot (2M)^{s-1} \geq c_2 \cdot (M+1)^{s-1} > m$, and thus

$$M \geq \frac{1}{2} \cdot c_2^{-1/(s-1)} \cdot m^{1/(s-1)} = c_3 \cdot m^{1/(s-1)} \quad (\text{F.1})$$

with $c_3 = c_3(s) := 1/2 \cdot c_2^{-1/(s-1)}$.

Using [36, Proposition 5.9] and Lemma F.1 we can pick $a_1, \dots, a_m \in \mathbb{R}^s \setminus \{0\}$ satisfying

$$H_{s(2M-1)}^s = \text{span} \left\{ x \mapsto (a_j^T x)^{s(2M-1)} : j \in \{1, \dots, m\} \right\}, \quad (\text{F.2})$$

where we used that

$$c_1 \cdot (s(2M-1))^{s-1} \leq c_1 \cdot (2s)^{s-1} \cdot M^{s-1} = c_2 \cdot M^{s-1} \leq m.$$

Here we can assume $\|a_j\|_2 = r$ for every $j \in \{1, \dots, m\}$ since multiplying each a_j with a positive constant does not change the span in (F.2). From [36, Corollary 5.12] we infer that

$$P_{s(2M-1)}^s = \text{span} \left\{ x \mapsto (a_j^T x)^r : j \in \{1, \dots, m\}, 0 \leq r \leq s(2M-1) \right\}. \quad (\text{F.3})$$

Let $f \in C^k([-1, 1]^s; \mathbb{R})$. Then, according to Theorem C.15, there exists a polynomial $P : \mathbb{R}^s \rightarrow \mathbb{R}$ of *coordinatewise* degree at most $2M-1$ satisfying

$$\|f - P\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq c_4 \cdot M^{-k} \cdot \|f\|_{C^k([-1, 1]^s; \mathbb{R})},$$

where $c_4 = c_4(s, k) > 0$. Note that by construction it holds $P \in P_{s(2M-1)}^s$. Using (F.3) we deduce the existence of polynomials $p_1, \dots, p_m : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$P(x) = \sum_{j=1}^m p_j(a_j^T x) \quad \text{for all } x \in \mathbb{R}^s.$$

Combining the previously shown bounds, we get

$$\begin{aligned} \left\| f(x) - \sum_{j=1}^m p_j(a_j^T x) \right\|_{L^\infty([-1, 1]^s; \mathbb{R})} &= \|f(x) - P(x)\|_{L^\infty([-1, 1]^s; \mathbb{R})} \leq c_4 \cdot M^{-k} \cdot \|f\|_{C^k([-1, 1]^s; \mathbb{R})} \\ &\stackrel{(\text{F.1})}{\leq} c \cdot m^{-k/(s-1)} \cdot \|f\|_{C^k([-1, 1]^s; \mathbb{R})}, \end{aligned}$$

as desired. Here, we defined $c = c(s, k) := c_4 \cdot c_3^{-k}$. □

F.2 Proof of Theorem 4.2

Using Theorem F.2, we can prove the following statement for complex-valued C^k -functions, which will play an important role in the proof of Theorem 4.2.

Proposition F.3. *Let $n, k \in \mathbb{N}$. Then there exists a constant $c = c(n, k) > 0$ with the following property: For any $m \in \mathbb{N}$ there exist complex vectors $b_1, \dots, b_m \in \mathbb{C}^n$ with $\|b_j\|_2 = 1/\sqrt{2n}$ for $j = 1, \dots, m$ and with the property that for any function $f \in C^k(\Omega_n; \mathbb{C})$ there exist functions $g_1, \dots, g_m \in C(\Omega_1; \mathbb{C})$ such that*

$$\left\| f(z) - \sum_{j=1}^m g_j(b_j^T \cdot z) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}.$$

Note that the vectors $b_1, \dots, b_m$ can be chosen independently from the considered function f , whereas $g_1, \dots, g_m$ do depend on f .

Proof. Theorem F.2 yields the existence of a constant $c_1 = c_1(n, k) > 0$ with the property that for any $m \in \mathbb{N}$ there exist real vectors $a_1, \dots, a_m \in \mathbb{R}^{2n}$ with $\|a_j\|_2 = 1/\sqrt{2n}$ such that for any function $\tilde{f} \in C^k([-1, 1]^{2n}; \mathbb{R})$ there exist functions $\tilde{g}_1, \dots, \tilde{g}_m \in C([-1, 1]; \mathbb{R})$ satisfying

$$\left\| \tilde{f}(x) - \sum_{j=1}^m \tilde{g}_j(a_j^T x) \right\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \leq c_1 \cdot m^{-k/(2n-1)} \cdot \|\tilde{f}\|_{C^k([-1, 1]^{2n}; \mathbb{R})}.$$

We then define the vectors $b_1, \dots, b_m \in \mathbb{C}^n$ componentwise via

$$(b_j)_\ell := (a_j)_\ell - i \cdot (a_j)_{n+\ell}, \quad \ell \in \{1, \dots, n\}, \quad j \in \{1, \dots, m\}.$$

First we see $\|b_j\|_2 = \|a_j\|_2 = 1/\sqrt{2n}$. We first consider real-valued functions, i.e., $f \in C^k(\Omega_n; \mathbb{R})$. Let φ_n be defined as in (A.1). By the choice of the constant c_1 we can find continuous functions $\tilde{g}_1, \dots, \tilde{g}_m \in C([-1, 1]; \mathbb{R})$ such that

$$\left\| (f \circ \varphi_n)(x) - \sum_{j=1}^m \tilde{g}_j(a_j^T x) \right\|_{L^\infty([-1, 1]^{2n})} \leq c_1 \cdot m^{-k/(2n-1)} \cdot \|f \circ \varphi_n\|_{C^k([-1, 1]^{2n}; \mathbb{R})}.$$

We then define $g_j \in C(\Omega_1; \mathbb{R})$ by $g_j(z) := \tilde{g}_j(\operatorname{Re}(z))$ for any $j \in \{1, \dots, m\}$. For $z \in \Omega_n$ we then have

$$\begin{aligned} g_j((b_j)^T z) &= \tilde{g}_j\left(\operatorname{Re}\left(\sum_{\ell=1}^n (b_j)_\ell \cdot z_\ell\right)\right) \\ &= \tilde{g}_j\left(\operatorname{Re}\left(\sum_{\ell=1}^n \left((a_j)_\ell - i \cdot (a_j)_{n+\ell}\right) (\varphi_n^{-1}(z)_\ell + i \cdot \varphi_n^{-1}(z)_{n+\ell})\right)\right) \\ &= \tilde{g}_j\left(\sum_{\ell=1}^n \left[(a_j)_\ell \varphi_n^{-1}(z)_\ell + (a_j)_{n+\ell} \varphi_n^{-1}(z)_{n+\ell}\right]\right) \\ &= \tilde{g}_j\left((a_j)^T \cdot \varphi_n^{-1}(z)\right). \end{aligned} \tag{F.4}$$

Therefore,

$$\begin{aligned} \left\| f(z) - \sum_{j=1}^m g_j(b_j^T z) \right\|_{L^\infty(\Omega_n; \mathbb{R})} &= \left\| (f \circ \varphi_n)(x) - \sum_{j=1}^m g_j(b_j^T \cdot \varphi_n(x)) \right\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \\ &\stackrel{(F.4)}{=} \left\| (f \circ \varphi_n)(x) - \sum_{j=1}^m \tilde{g}_j(a_j^T \cdot x) \right\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \\ &\leq c_1 \cdot m^{-k/(2n-1)} \cdot \|f \circ \varphi_n\|_{C^k([-1, 1]^{2n}; \mathbb{R})}. \end{aligned}$$

By the above, for $f \in C^k(\Omega_n; \mathbb{C})$ we can pick functions $g_1^{\operatorname{Re}}, \dots, g_m^{\operatorname{Re}}, g_1^{\operatorname{Im}}, \dots, g_m^{\operatorname{Im}} \in C(\Omega_1; \mathbb{R})$ satisfying

$$\begin{aligned} \left\| \operatorname{Re}(f(z)) - \sum_{j=1}^m g_j^{\operatorname{Re}}(b_j^T z) \right\|_{L^\infty(\Omega_n; \mathbb{R})} &\leq c_1 \cdot m^{-k/(2n-1)} \cdot \|\operatorname{Re}(f \circ \varphi_n)\|_{C^k([-1, 1]^{2n}; \mathbb{R})}, \\ \left\| \operatorname{Im}(f(z)) - \sum_{j=1}^m g_j^{\operatorname{Im}}(b_j^T z) \right\|_{L^\infty(\Omega_n; \mathbb{R})} &\leq c_1 \cdot m^{-k/(2n-1)} \cdot \|\operatorname{Im}(f \circ \varphi_n)\|_{C^k([-1, 1]^{2n}; \mathbb{R})}. \end{aligned}$$

Defining $g_j := g_j^{\operatorname{Re}} + i \cdot g_j^{\operatorname{Im}}$ yields

$$\left\| f(z) - \sum_{j=1}^m g_j(b_j^T z) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \leq c_1 \cdot \sqrt{2} \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})},$$

completing the proof. $\square$

The special activation function that yields the improved approximation rate of $m^{-k/(2n-1)}$ (see Theorem 4.2) is constructed in the following lemma.

Lemma F.4. *Let $\{u_\ell\}_{\ell=1}^\infty$ be an enumeration of the set of complex polynomials in z and $\bar{z}$ with coefficients in $\mathbb{Q} + i\mathbb{Q}$. Then there exists a function $\phi \in C^\infty(\mathbb{C}; \mathbb{C})$ with the following properties:*

1. *For every $\ell \in \mathbb{N}$ and $z \in \Omega_1$ one has*

$$\phi(z + 3\ell) = u_\ell(z).$$

2. *ϕ is non-polyharmonic.*

Proof. Let $\psi \in C^\infty(\mathbb{C}; \mathbb{R})$ with $0 \leq \psi \leq 1$ and

$$\psi|_{\Omega_1} \equiv 1, \quad \text{supp}(\psi) \subseteq \tilde{\Omega},$$

where $\tilde{\Omega} := \{z \in \mathbb{C} : |\text{Re}(z)|, |\text{Im}(z)| < \frac{3}{2}\}$. We then define

$$\phi := f \cdot \psi + \sum_{\ell=1}^{\infty} u_\ell(\bullet - 3\ell) \cdot \psi(\bullet - 3\ell),$$

where $f(z) = e^{\text{Re}(z)}$. Note that ϕ is smooth since it is a locally finite sum of smooth functions. Furthermore, ϕ is non-polyharmonic on the interior of Ω_1 , since the calculation in the proof of Proposition A.2 shows for z in the interior of Ω_1 and $\rho : \mathbb{R} \rightarrow \mathbb{R}$, $t \mapsto e^t$ that

$$\left| \partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell \phi(z) \right| = \left| \partial_{\text{wirt}}^m \bar{\partial}_{\text{wirt}}^\ell f(z) \right| = \frac{1}{2^{m+\ell}} \left| \rho^{(m+\ell)}(\text{Re}(z)) \right| > 0$$

for arbitrary $m, \ell \in \mathbb{N}_0$. Finally, property (1) follows directly by construction of ϕ because

$$(\tilde{\Omega} + 3\ell) \cap (\tilde{\Omega} + 3\ell') = \emptyset$$

for $\ell \neq \ell'$. □

Using the properties of the special activation function constructed in Lemma F.4 and applying the approximation result from Proposition F.3 we can now prove Theorem 4.2.

Proof of Theorem 4.2. Let ϕ be the activation function constructed in Lemma F.4. We choose the constant c according to Proposition F.3. Let $m \in \mathbb{N}$ and $f \in C^k(\Omega_n; \mathbb{C})$. We can without loss of generality assume that $f \not\equiv 0$. Again, according to Proposition F.3, we can choose $\rho_1, \dots, \rho_m \in \mathbb{C}^n$ with $\|2\rho_j\| = 1/\sqrt{2n}$ and $g_1, \dots, g_m \in C(\Omega; \mathbb{C})$ with the property

$$\left\| f(z) - \sum_{j=1}^m g_j(\rho_j^T z) \right\|_{L^\infty(\Omega_n)} \leq c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}.$$

Recall from Lemma F.4 that $\{u_\ell\}_{\ell=1}^\infty$ is an enumeration of the set of complex polynomials in z and $\bar{z}$. Hence, using the complex version of the Stone-Weierstraß-Theorem (see for instance [21, Theorem 4.51]), we can pick $\ell_1, \dots, \ell_m \in \mathbb{N}$ such that

$$\|g_j - u_{\ell_j}\|_{L^\infty(\Omega_1; \mathbb{C})} \leq m^{-1-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})} \quad (\text{F.5})$$

for every $j \in \{1, \dots, m\}$. Since $\phi(\bullet + 3\ell) = u_\ell$ on Ω_1 for each $\ell \in \mathbb{N}$, and since $\rho_j^T z \in \Omega_1$ for $j \in \{1, \dots, m\}$ and $z \in \Omega_n$, we estimate

$$\begin{aligned}
& \left\| f(z) - \sum_{j=1}^m \phi(\rho_j^T z + 3\ell_j) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \\
& \leq \left\| f(z) - \sum_{j=1}^m g_j(\rho_j^T z) \right\|_{L^\infty(\Omega_n; \mathbb{C})} + \sum_{j=1}^m \|g_j(\rho_j^T z) - \phi(\rho_j^T \cdot z + 3\ell_j)\|_{L^\infty(\Omega_n; \mathbb{C})} \\
& \leq c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})} + \sum_{j=1}^m \|g_j(z) - u_{\ell_j}(z)\|_{L^\infty(\Omega_1; \mathbb{C})} \\
& \stackrel{(F.5)}{\leq} c \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})} + m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})} \\
& = (c+1) \cdot m^{-k/(2n-1)} \cdot \|f\|_{C^k(\Omega_n; \mathbb{C})}. \quad \square
\end{aligned}$$

F.3 Proof of Theorem 4.3

As a preparation for the proof of Theorem 4.3, we first prove a similar result in the real-valued setting. We remark that the proof idea is inspired by the proof of [46, Theorem 4].

Theorem F.5. *Let $n, k \in \mathbb{N}$ and*

$$\phi : \mathbb{R} \rightarrow \mathbb{R}, \quad \phi(x) := \frac{1}{1 + e^{-x}}$$

be the sigmoid function. Then there exists a constant $c = c(n, k) > 0$ with the following property: If the numbers $\varepsilon \in (0, \frac{1}{2})$ and $m \in \mathbb{N}$ are such that for every function $f \in C^k([-1, 1]^n; \mathbb{R})$ with $\|f\|_{C^k([-1, 1]^n; \mathbb{R})} \leq 1$ there exist coefficients $\rho_1, \dots, \rho_m \in \mathbb{R}^n$, $\eta_1, \dots, \eta_m \in \mathbb{R}$ and $\sigma_1, \dots, \sigma_m \in \mathbb{R}$ satisfying

$$\left\| f(x) - \sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T x + \eta_j) \right\|_{L^\infty([-1, 1]^n; \mathbb{R})} \leq \varepsilon,$$

then necessarily

$$m \geq c \cdot \frac{\varepsilon^{-n/k}}{\ln(1/\varepsilon)}.$$

Proof. We first pick a function $\psi \in C^\infty(\mathbb{R}^n; \mathbb{R})$ with the property that $\psi(0) = 1$ and $\psi(x) = 0$ for every $x \in \mathbb{R}^n$ with $\|x\|_2 > \frac{1}{4}$. We then choose

$$c_1 = c_1(n, k) := \left(\|\psi\|_{C^k([-1, 1]^n; \mathbb{R})} \right)^{-1}.$$

Now, let $\varepsilon \in (0, \frac{1}{2})$ and $m \in \mathbb{N}$ be arbitrary with the property stated in the formulation of the theorem. If $\varepsilon > \frac{c_1}{2} \cdot \frac{1}{6^k}$, then $m \geq c \cdot \frac{\varepsilon^{-n/k}}{\ln(1/\varepsilon)}$ trivially holds (as long as $c = c(n, k) > 0$ is sufficiently small). Hence, we can assume that $\varepsilon \leq \frac{c_1}{2} \cdot \frac{1}{6^k}$. Now, let N be the smallest integer with $N \geq 2$, for which

$$\frac{c_1}{2^{k+1}} \cdot N^{-k} \leq \varepsilon.$$

Note that this implies

$$N^k \geq \frac{c_1}{\varepsilon} \cdot \frac{1}{2^{k+1}} \geq \frac{c_1}{2^{k+1}} \cdot \frac{2}{c_1} \cdot 6^k = 3^k$$

and hence $N \geq 3$, whence $N - 1 \geq 2$. Therefore, by minimality of N , and since $\frac{N}{2} \leq N - 1$ because of $N \geq 2$, it follows that

$$\varepsilon < \frac{c_1}{2^{k+1}} \cdot (N - 1)^{-k} \leq \frac{c_1}{2^{k+1}} 2^k \cdot N^{-k} = \frac{c_1}{2} \cdot N^{-k}. \quad (F.6)$$

Now, for every $\alpha \in \{-N, \dots, N\}^n$ pick $z_\alpha \in \{0, 1\}$ arbitrary and let $y_\alpha := z_\alpha c_1 N^{-k}$. Define the function

$$f(x) := \sum_{\alpha \in \{-N, \dots, N\}^n} y_\alpha \cdot \psi(Nx - \alpha), \quad x \in \mathbb{R}^n.$$

Clearly, $f \in C^\infty(\mathbb{R}^n; \mathbb{R})$. Furthermore, since the supports of the functions $\psi(\bullet - \alpha)$, $\alpha \in \mathbb{Z}^n$ are pairwise disjoint, we see for any multi-index $\mathbf{k} \in \mathbb{N}_0^n$ with $|\mathbf{k}| \leq k$ that

$$\begin{aligned} \|\partial^{\mathbf{k}} f\|_{L^\infty([-1, 1]^n; \mathbb{R})} &\leq N^{|\mathbf{k}|} \cdot \max_{\alpha} |y_\alpha| \cdot \|\partial^{\mathbf{k}} \psi\|_{L^\infty([-1, 1]^n; \mathbb{R})} \\ &\leq N^k \cdot \max_{\alpha} |y_\alpha| \cdot \|\psi\|_{C^k([-1, 1]^n; \mathbb{R})} \leq 1, \end{aligned}$$

so we conclude that $\|f\|_{C^k([-1, 1]^n; \mathbb{R})} \leq 1$. Additionally, for any fixed $\beta \in \{-N, \dots, N\}^n$ we see

$$f\left(\frac{\beta}{N}\right) = y_\beta,$$

by choice of ψ . See also Figure 5 for an illustration of the function f .

By assumption, we can choose suitable coefficients $\rho_1, \dots, \rho_m \in \mathbb{R}^n$, $\eta_1, \dots, \eta_m \in \mathbb{R}$ and furthermore $\sigma_1, \dots, \sigma_m \in \mathbb{R}$ such that

$$\|f - g\|_{L^\infty([-1, 1]^n; \mathbb{R})} \leq \varepsilon$$

for

$$g := \sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T \cdot \bullet + \eta_j).$$

Letting

$$\tilde{g} := g(\bullet/N) = \sum_{j=1}^m \sigma_j \cdot \phi\left(\frac{\rho_j^T}{N} \cdot \bullet + \eta_j\right), \quad (\text{F.7})$$

we see for every $\alpha \in \{-N, \dots, N\}^n$ that

$$\tilde{g}(\alpha) = g\left(\frac{\alpha}{N}\right) \begin{cases} \geq y_\alpha - \varepsilon = c_1 N^{-k} - \varepsilon \stackrel{(\text{F.6})}{>} (c_1/2) N^{-k}, & \text{if } z_\alpha = 1, \\ \leq y_\alpha + \varepsilon \stackrel{(\text{F.6})}{<} (c_1/2) N^{-k}, & \text{if } z_\alpha = 0. \end{cases}$$

Therefore, we get $\mathbb{1}(\tilde{g} > (c_1/2)N^{-k})(\alpha) = z_\alpha$ for any $\alpha \in \{-N, \dots, N\}^n$. Since the choice of z_α has been arbitrary, it follows that the set

$$H := \left\{ \mathbb{1}(\tilde{g} > (c_1/2)N^{-k}) \big|_{\{-N, \dots, N\}^n} : \tilde{g} \text{ of form (F.7)} \right\}$$

shatters the whole set $\{-N, \dots, N\}^n$. Therefore, we conclude that

$$\text{VC}(H) \geq (2N + 1)^n \geq N^n.$$

On the other hand, [4, Theorem 8.11] shows that

$$\text{VC}(H) \leq 2m(n + 2) \log_2(60n \cdot N) \leq c_3 \cdot m \cdot \ln(N)$$

with a suitably chosen constant $c_3 = c_3(n)$. Here we used that $N \geq 3$ so that $\ln(N) \geq \ln(3) > 0$. Combining those two inequalities yields

$$m \geq \frac{N^n}{c_3 \cdot \ln(N)}.$$

Using that $N \geq c_4 \cdot \varepsilon^{-1/k}$ with $c_4 := c_4(n, k) = \left(\frac{c_1}{2^{k+1}}\right)^{1/k}$ and $N \leq c_5 \cdot \varepsilon^{-1/k}$ with the definition $c_5 := c_5(n, k) = \left(\frac{c_1}{2}\right)^{1/k}$, we see that

$$m \geq \frac{c_4^n \cdot \varepsilon^{-n/k}}{c_3 \cdot \ln(c_5 \cdot \varepsilon^{-1/k})} \geq c_6 \cdot \frac{\varepsilon^{-n/k}}{\ln(1/\varepsilon)}$$

with $c_6 = c_6(n, k) > 0$ chosen appropriately. $\square$

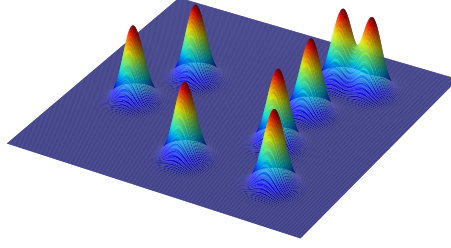


Figure 5: Illustration of the function f considered in the proof of Theorem F.5.

As a corollary, we get a similar result for complex-valued neural networks.

Corollary F.6. *Let $n, k \in \mathbb{N}$ and*

$$\phi : \mathbb{C} \rightarrow \mathbb{C}, \quad \phi(z) := \frac{1}{1 + e^{-\operatorname{Re}(z)}}.$$

Then there exists a constant $c = c(n, k) > 0$ with the following property: If $\varepsilon \in (0, \frac{1}{2})$ and $m \in \mathbb{N}$ are such that for every function $f \in C^k(\Omega_n; \mathbb{C})$ with $\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1$ there exist coefficients $\rho_1, \dots, \rho_m \in \mathbb{C}^n$, $\eta_1, \dots, \eta_m \in \mathbb{C}$ and $\sigma_1, \dots, \sigma_m \in \mathbb{C}$ satisfying

$$\left\| f(z) - \sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T z + \eta_j) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \varepsilon,$$

then necessarily

$$m \geq c \cdot \frac{\varepsilon^{-2n/k}}{\ln(1/\varepsilon)}.$$

Proof. We choose the constant $c = c(2n, k)$ according to the previous Theorem F.5 and let φ_n as in (A.1). Then, let $\varepsilon \in (0, \frac{1}{2})$ and $m \in \mathbb{N}$ with the properties assumed in the statement of the corollary. If we then take an arbitrary function $f \in C^k([-1, 1]^{2n}; \mathbb{R})$ with $\|f\|_{C^k([-1, 1]^{2n}; \mathbb{R})} \leq 1$, we deduce the existence of $\rho_1, \dots, \rho_m \in \mathbb{C}^n$, $\eta_1, \dots, \eta_m \in \mathbb{C}$ and $\sigma_1, \dots, \sigma_m \in \mathbb{C}$, such that

$$\begin{aligned} & \left\| f(x) - \operatorname{Re} \left(\sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T \cdot \varphi_n(x) + \eta_j) \right) \right\|_{L^\infty([-1, 1]^{2n}; \mathbb{R})} \\ & \leq \left\| (f \circ \varphi_n^{-1})(z) - \sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T z + \eta_j) \right\|_{L^\infty(\Omega_n; \mathbb{C})} \leq \varepsilon. \end{aligned}$$

In the next step, we show that

$$\mathbb{R}^{2n} \ni x \mapsto \operatorname{Re} \left(\sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T \cdot \varphi_n(x) + \eta_j) \right)$$

is a real-valued shallow neural network with m neurons in the hidden layer and the real sigmoid function as activation function. Then the claim follows using Theorem F.5.

For every $j \in \{1, \dots, m\}$ we pick a matrix $\tilde{\rho}_j \in \mathbb{R}^{2n \times 2}$ with the property that one has

$$\tilde{\rho}_j^T \cdot \varphi_n^{-1}(z) = \varphi_1^{-1}(\rho_j^T \cdot z)$$

for every $z \in \mathbb{C}^n$. This is possible, since this is equivalent to

$$\tilde{\rho}_j^T v = \varphi_1^{-1}(\rho_j^T \varphi_n(v))$$

for all $v \in \mathbb{R}^{2n}$, where the right-hand side is an $\mathbb{R}$ -linear map $\mathbb{R}^{2n} \rightarrow \mathbb{R}^2$. Denoting the first column of $\hat{\rho}_j$ by $\hat{\rho}_j^T$, we get

$$\hat{\rho}_j^T \cdot \varphi_n^{-1}(z) = \operatorname{Re}(\rho_j^T \cdot z) \quad \text{for all } z \in \mathbb{C}^n.$$

Writing γ for the classical real-valued sigmoid function (i.e. $\gamma(x) = \frac{1}{1+e^{-x}}$), we deduce for arbitrary $x \in \mathbb{R}^{2n}$ that

$$\begin{aligned} \operatorname{Re} \left(\sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T \cdot \varphi_n(x) + \eta_j) \right) &= \operatorname{Re} \left(\sum_{j=1}^m \sigma_j \cdot \gamma(\operatorname{Re}(\rho_j^T \cdot \varphi_n(x) + \eta_j)) \right) \\ &= \operatorname{Re} \left(\sum_{j=1}^m \sigma_j \cdot \gamma(\hat{\rho}_j^T x + \operatorname{Re}(\eta_j)) \right) \\ &= \sum_{j=1}^m \operatorname{Re}(\sigma_j) \cdot \gamma(\hat{\rho}_j^T x + \operatorname{Re}(\eta_j)), \end{aligned}$$

where in the last step we used that γ is real-valued. As noted above, this completes the proof. $\square$

Now, we can finally prove Theorem 4.3.

Proof of Theorem 4.3. Let $\alpha = \frac{2n}{k}$ and choose $c_2 = c_2(\alpha) = c_2(n, k) > 0$ such that the inequality $\ln(x) \leq c_2 \cdot x^{\alpha/2}$ holds for all $x \geq 1$. Furthermore, let $c_1 = c_1(n, k) > 0$ as in Corollary F.6. By choosing $c = c(n, k) > 0$ sufficiently small, we can ensure that $\varepsilon_m := c \cdot (m \cdot \ln(m))^{-k/(2n)}$ satisfies

$$\ln\left(\frac{c_1}{c_2}\right) + \frac{\alpha}{2} \ln(1/\varepsilon_m) \geq \frac{\alpha}{4} \cdot \ln(1/\varepsilon_m) \quad \text{for all } m \in \mathbb{N}_{\geq 2}.$$

By further shrinking $c = c(n, k) > 0$ if necessary, we may assume

$$c \cdot (2 \cdot \ln(2))^{-k/(2n)} < \frac{1}{2}$$

and hence $c \cdot (m \cdot \ln(m))^{-k/(2n)} < \frac{1}{2}$ for all $m \in \mathbb{N}_{\geq 2}$. Finally, setting $c_3 := \frac{\alpha}{4}$ and shrinking c even further, we can arrange that $c^\alpha < c_1 \cdot c_3$. Now, assume towards a contradiction that for every $f \in C^k(\Omega_n; \mathbb{C})$ with $\|f\|_{C^k(\Omega_n; \mathbb{C})} \leq 1$ there are coefficients $\rho_1, \dots, \rho_m \in \mathbb{C}^n$, $\sigma_1, \dots, \sigma_m \in \mathbb{C}$ and $\eta_1, \dots, \eta_m \in \mathbb{C}$ such that

$$\left\| f(z) - \sum_{j=1}^m \sigma_j \cdot \phi(\rho_j^T z + \eta_j) \right\|_{L^\infty(\Omega_n; \mathbb{C})} < c \cdot (m \cdot \ln(m))^{-k/(2n)}.$$

Applying Corollary F.6, we then get

$$m \geq c_1 \cdot \frac{\varepsilon^{-2n/k}}{\ln(1/\varepsilon)}$$

with $\varepsilon = c \cdot (m \cdot \ln m)^{-k/(2n)} \in (0, \frac{1}{2})$ and $c_1 = c_1(n, k) > 0$. Recall from the beginning of the proof that $\alpha := 2n/k$ and that $\ln(x) \leq c_2 x^{\alpha/2}$ for every $x \geq 1$. We observe

$$m \geq c_1 \cdot \frac{\varepsilon^{-\alpha}}{\ln(1/\varepsilon)} \geq \frac{c_1}{c_2} \varepsilon^{-\alpha/2},$$

which implies

$$\ln(m) \geq \ln\left(\frac{c_1}{c_2}\right) + \frac{\alpha}{2} \cdot \ln(1/\varepsilon) \geq \frac{\alpha}{4} \cdot \ln(1/\varepsilon) = c_3 \cdot \ln(1/\varepsilon).$$

Overall we then get

$$m \geq c_1 \cdot \frac{\varepsilon^{-\alpha}}{\ln(1/\varepsilon)} = c_1 \cdot c^{-\alpha} \cdot \frac{m \cdot \ln(m)}{\ln(1/\varepsilon)} \geq c_1 \cdot c_3 \cdot c^{-\alpha} \cdot \frac{m \cdot \ln(m)}{\ln(m)} = c_1 \cdot c_3 \cdot c^{-\alpha} \cdot m.$$

Since we chose c such that $c^\alpha < c_1 \cdot c_3$ we get the desired contradiction. $\square$

B.4. On best approximation using multivariate ridge functions with applications to generalized translation networks



On best approximation by multivariate ridge functions with applications to generalized translation networks

Paul Geuchen¹ · Palina Salanevich² · Olov Schavemaker² · Felix Voigtlaender¹

Received: 13 March 2025 / Accepted: 12 May 2025
© The Author(s) 2025

Abstract

In this paper, we prove sharp upper and lower bounds for the approximation of Sobolev functions by sums of *multivariate ridge functions*, i.e., functions of the form $\mathbb{R}^d \ni x \mapsto \sum_{k=1}^n \varrho_k(A_k x) \in \mathbb{R}$ with $\varrho_k : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and $A_k \in \mathbb{R}^{\ell \times d}$. We show that the order of approximation asymptotically behaves as $n^{-r/(d-\ell)}$, where r is the regularity (order of differentiability) of the Sobolev functions to be approximated. Our lower bound even holds when approximating L^∞ -Sobolev functions of regularity r with error measured in L^1 , while our upper bound applies to the approximation of L^p -Sobolev functions in L^p for any $1 \leq p \leq \infty$. These bounds generalize well-known results regarding the approximation properties of *univariate* ridge functions to the multivariate case. We use our results to obtain sharp asymptotic bounds for the approximation of Sobolev functions using generalized translation networks and complex-valued neural networks.

Keywords Multivariate ridge functions · Sobolev functions · Shallow neural networks · Complex-valued neural networks · Approximation rates

Mathematics Subject Classification 41A30 · 41A25 · 41A63 · 46E35 · 68T07

✉ Felix Voigtlaender
felix.voigtlaender@ku.de

Paul Geuchen
paul.geuchen@ku.de

Palina Salanevich
p.salanevich@uu.nl

Olov Schavemaker
o.p.schavemaker@uu.nl

¹ Mathematical Institute for Machine Learning and Data Science (MIDS), Catholic University of Eichstätt–Ingolstadt (KU), Auf der Schanz 49, 85049 Ingolstadt, Germany

² Mathematical Institute, Utrecht University, Budapestlaan 6, 3584 CD Utrecht, Netherlands

1 Introduction

Ridge functions — both univariate and multivariate — and their linear combinations have received significant attention in the mathematical literature [3, 11, 15, 26], in particular due to their applications to neural networks and related areas [7, 19]. A (multivariate) ridge function on $\mathbb{R}^d$ is defined as a composition of the form $\varrho \circ A$ with $\varrho : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and $A \in \mathbb{R}^{\ell \times d}$, where $\ell < d$. In the case when $\ell = 1$, we refer to such a ridge function as univariate. A topic of particular interest is the approximation capacity of linear combinations of ridge functions. Although this question has been extensively studied in the case of *univariate* ridge functions [10, 12, 17, 18, 25], the case of *multivariate* ridge functions has largely remained unexplored until now.

In this paper, we establish asymptotically tight bounds for the approximation of Sobolev functions by sums of multivariate ridge functions, more precisely by sums of ℓ -variate ridge functions for any fixed $1 \leq \ell < d$. Specifically, we show that for every L^p -Sobolev function with smoothness (order of differentiability) r defined on the unit ball in $\mathbb{R}^d$, there exists a sum of n multivariate ridge functions such that the error of approximation in the L^p -norm is asymptotically $n^{-r/(d-\ell)}$, where $p \in [1, \infty]$. We furthermore prove that this rate is *optimal*.

Moreover, we show that our results for multivariate ridge functions can be used to derive tight asymptotic bounds for the approximation of Sobolev functions by *generalized translation networks* (GTNs) and *complex-valued* neural networks (CVNNs). A (shallow) generalized translation network with activation dimension ℓ is a function of the form

$$\mathbb{R}^d \ni x \mapsto \sum_{i=1}^n c_i \tau(A_i x + v_i) \in \mathbb{R}, \quad (1.1)$$

with matrices $A_i \in \mathbb{R}^{\ell \times d}$, bias vectors $v_i \in \mathbb{R}^\ell$, coefficients $c_i \in \mathbb{R}$ and a fixed activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$. Note that one obtains classical shallow neural networks in the case $\ell = 1$. Both generalized translation networks and complex-valued neural networks recently have received increased attention from both theoretical and applied perspectives; see [2, 14, 20, 22] and the references therein.

Our results demonstrate that shallow GTNs with higher activation dimensions achieve a strictly better optimal approximation rate than GTNs with a smaller activation dimension. Moreover, we show that there exists a (nice) complex activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ such that shallow CVNNs can approximate Sobolev functions of smoothness r on the unit ball of $\mathbb{C}^d$ with rate $n^{-r/(2d-2)}$, whereas the best rate that can be achieved by shallow *real-valued* neural networks is $n^{-r/(2d-1)}$. We also show that the rate of $n^{-r/(2d-2)}$ is optimal in the class of all shallow CVNNs.

1.1 Related work

In this subsection, we discuss how our results relate to the existing literature, both regarding ridge functions and regarding various classes of neural networks.

1.1.1 Approximation by sums of univariate ridge functions

The approximation of Sobolev functions by sums of *univariate* ridge functions was extensively studied by V. Maiorov and his co-authors, see [10, 17, 18]. As a first result in this direction, the sharp bound of $n^{-r/(d-1)}$ was shown for the approximation of L^2 -Sobolev functions of regularity r on the d -dimensional unit ball, with error measured in the L^2 -norm [17]. Two years later, this result was extended to the approximation of L^p -Sobolev functions of regularity r , with error measured in the L^q -norm for arbitrary $p \geq q \geq 2$; see [10]. Finally, in 2010 the paper [18] showed the same sharp bound of $n^{-r/(d-1)}$ for every $p \geq q \geq 1$.

The techniques used in the present paper rely to a large extent on the techniques used in [18], combined with a generalization of a result from [17]. In the process of writing the present paper, we discovered that the proof presented in [18] contains a gap, which we discuss in detail in Appendix A. The present work resolves the issue in the proof in [18] by replacing an orthogonal projection onto a certain space of polynomials by a *quasi-projection* defined in Section 2.3.

1.1.2 Shallow neural networks with general activation function

In [19], it is shown that there exists an analytic, sigmoidal (but quite bespoke and somewhat pathological) activation function $\tau : \mathbb{R} \rightarrow \mathbb{R}$ with the property that every sum of n univariate ridge functions can be approximated up to arbitrary precision using shallow neural networks with activation function τ and $3n$ neurons. Combining this with the results presented in [10, 18] then readily implies that shallow neural networks with n neurons and the bespoke activation function τ can approximate Sobolev functions of regularity r at a rate of $n^{-r/(d-1)}$. In this work, we extend this statement to the case of generalized translation networks (as in Equation (1.1)); see Theorem 1.3.

1.1.3 Shallow neural networks and continuous weight selection

In [21], Mhaskar proves an upper bound of $n^{-r/d}$ for approximating Sobolev functions of regularity r using shallow neural networks with *any* smooth but non-polynomial activation function. In fact, [21] does not only consider classical shallow neural networks but also generalized translation networks (see Equation (1.1)). It is then shown that, for a fixed activation dimension ℓ , the rate of $n^{-r/d}$ is *optimal* under the assumption of *continuous weight selection*, i.e., assuming that the map that assigns to a Sobolev function f the coefficients of the approximating network is continuous. Specifically, this shows that under the restriction of continuous weight selection, there is *no* benefit (in terms of the approximation rate) in choosing a generalized translation network instead of a classical shallow network. Our work shows that, if one drops this continuity assumption, one can indeed *improve* the approximation rate from $n^{-r/(d-1)}$ (for classical shallow neural networks) to $n^{-r/(d-\ell)}$ using a generalized translation network with activation dimension ℓ , at least for a specific choice of the activation function; see Theorem 1.3.

1.1.4 Deep neural networks with general activation function

With respect to neural networks, the results derived in the present paper can be used to determine the optimal rates of approximating Sobolev functions of regularity r , where optimality is considered with respect to the class of all *shallow* (real- or complex-valued) neural networks *with general (continuous) activation function*. This is a novel result for the case of complex-valued neural networks, and recovers known results for the case of real-valued networks.

In the setting of arbitrary (continuous) activation functions considered here, the question of optimal approximation rates is only meaningful for *shallow* networks, since as a consequence of the so-called *Kolmogorov-Arnold representation theorem* (see [19, Theorem 5]) one can show that using the analytic, sigmoidal (but bespoke and pathological) activation function τ mentioned above, *every* continuous function can be approximated *arbitrarily well* by neural networks *of constant size* with two hidden layers. It is easy to see that an analogous result then also holds for networks with more than two hidden layers. Recently, it was shown that similar results (for networks with 11 hidden layers) also hold when using a slightly less "contrived" (but still impractical) activation function [35]. These results thus show that no meaningful lower bounds can be derived in the class of *deep* neural networks with *general* activation function.

1.1.5 Deep neural networks with ReLU activation function

In modern deep learning, the most commonly used activation function is the ReLU (rectified linear unit), which has thus been the focus of intensive research. For this special activation function, one can bound the VC dimension of the class of all ReLU networks of a given size [1], and this can be used to derive lower bounds for the approximation of Sobolev functions in terms of the number of weights or neurons, even for arbitrarily deep networks and without assuming continuous weight selection [34]. Thus, the above phenomenon of "infinite approximation power" of networks with two hidden layers does *not* occur for the ReLU. Moreover, for the case where $r/d \gg 1$, it has been shown in [24, 28] that *shallow* ReLU networks cannot achieve the rate of $n^{-r/d}$ for approximating Sobolev functions of regularity r . This is an instance of the general phenomenon that deeper ReLU networks can achieve better approximation rates than shallow networks [31], at least when the functions to be approximated are very smooth [24, 28].

Optimal upper bounds that match the lower bounds (in some cases up to log factors) have been derived in [29, 30, 34]. In particular, it has been shown in [34] that for arbitrarily deep ReLU networks with unconstrained weights one can obtain a strictly better approximation rate than using networks with constrained weights (i.e., assuming continuous weight selection, or assuming that the magnitude of the weights is polynomially bounded in terms of the number of neurons of the network).

Overall, these results show that in order to fully understand the "landscape of approximation results" for special, elementary activation functions such as the ReLU, an in-depth analysis is needed for each activation function. Our bound in the class of shallow networks with *general* activation function then serves as a benchmark that can be realized using some activation functions (such as those in [19, 35]), but not

generically with functions such as the ReLU. Similarly, not every complex activation function can achieve the corresponding approximation bound of $n^{-r/(2d-2)}$; see, e.g., [9, Theorem 4.3]. This helps explain why the activation function that we construct for attaining the optimal rate in the case of shallow complex-valued networks (see Theorem 5.9) is relatively impractical.

1.1.6 Complex-valued neural networks

While the previously discussed results deal with real-valued neural networks, in recent years there has been a growing interest in establishing approximation theoretical results for complex-valued neural networks (CVNNs). Classical statements about the universality of real-valued neural networks were generalized to the complex-valued setting in [8, 33]. Furthermore, the quantitative statement for real-valued neural networks from [21] was generalized to the complex-valued setting in [9]. The same paper also proves the existence of a complex-valued activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ with the property that shallow complex-valued neural networks with that activation function achieve an approximation rate of $n^{-r/(2d-1)}$ for Sobolev functions of smoothness r on the unit ball in $\mathbb{C}^d$. In the present work, we prove that the same activation function in fact achieves the optimal approximation rate of $n^{-r/(2d-2)}$, see Theorem 1.4. Note that the optimal rate achievable by real-valued networks is $n^{-r/(2d-1)}$, as follows by identifying $\mathbb{C}^d \cong \mathbb{R}^{2d}$. Thus, while the quantitative approximation bounds established for CVNNs so far showed *similar* approximation theoretical properties for complex-valued and real-valued neural networks, the result derived in this paper indicates a *superior* expressivity of CVNNs.

1.2 Notation

By $\mathbb{Z}$ we denote the set of integers. We use $\mathbb{N}$ to denote the positive integers excluding 0 and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $x \in \mathbb{R}^d$ and $p \in [1, \infty)$, we denote by $\|x\|_{\ell^p} := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ the ℓ^p -norm of the vector x and by $\|x\|_{\ell^\infty} := \max_{j=1,\dots,d} |x_j|$ the ℓ^∞ -norm of x . We write $B^d := \{x \in \mathbb{R}^d : \|x\|_{\ell^2} \leq 1\}$ for the closed d -dimensional Euclidean unit ball. Moreover, we let $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_{\ell^2} = 1\}$ denote the unit sphere in $\mathbb{R}^d$. We write $\mathbb{A}^d$ for the d -dimensional Lebesgue measure. The symbol $I_{d \times d}$ denotes the d -dimensional identity matrix. We denote the set of all (real) orthogonal $d \times d$ matrices by $O(d)$. For measurable functions $f, g : B^d \rightarrow \mathbb{R}$, we let

$$\langle f, g \rangle := \int_{B^d} f(x) \cdot g(x) \, dx, \quad (1.2)$$

if the expression on the right-hand side is well-defined, i.e., if $f \cdot g \in L^1(B^d)$. By $\mathcal{B}(W_d^{r,p})$ we denote the set of all *Sobolev functions* $f : B^d \rightarrow \mathbb{R}$ for which $\|f\|_{W_d^{r,p}} \leq 1$, where $\|\cdot\|_{W_d^{r,p}}$ is defined as in (2.4). We refer to Section 2.2 for more details on the class of Sobolev functions.

Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. For a measurable subset $K \subseteq \mathbb{F}^d$, a measurable function $f : K \rightarrow \mathbb{F}$ and a number $p \in [1, \infty)$, we define

$$\|f\|_{L^p(K)} := \left(\int_K |f(x)|^p \, dx \right)^{1/p} \quad \text{and} \quad \|f\|_{L^\infty(K)} := \operatorname{ess\,sup}_{x \in K} |f(x)|,$$

where $\operatorname{ess\,sup}$ denotes the *essential supremum*. For $p \in [1, \infty]$ and a measurable subset $K \subseteq \mathbb{F}^d$, we let

$$L^p(K; \mathbb{F}) := \{f : K \rightarrow \mathbb{F} : f \text{ measurable with } \|f\|_{L^p(K)} < \infty\}.$$

We define

$$L^p_{\operatorname{loc}}(\mathbb{F}^d; \mathbb{F}) := \left\{ f : \mathbb{F}^d \rightarrow \mathbb{F} : f \text{ measurable and } f|_K \in L^p(K; \mathbb{F}) \text{ for every compact } K \subseteq \mathbb{F}^d \right\}.$$

Moreover, we set

$$\mathcal{M}_b(\mathbb{F}^d; \mathbb{F}) := \left\{ f : \mathbb{F}^d \rightarrow \mathbb{F} : f \text{ measurable and bounded on every compact } K \subseteq \mathbb{F}^d \right\}.$$

Note here that the definition of $\mathcal{M}_b(\mathbb{F}^d; \mathbb{F})$ does not involve *essential* boundedness but genuine boundedness on every compact set. In the case $\mathbb{F} = \mathbb{R}$, we often omit the second argument in $L^p(K; \mathbb{F})$, $L^p_{\operatorname{loc}}(\mathbb{F}^d; \mathbb{F})$ and $\mathcal{M}_b(\mathbb{F}^d; \mathbb{F})$. For a subset $A \subseteq \mathbb{F}^d$, we denote by A° its interior with respect to the usual Euclidean topology.

For a multiindex $\mathbf{k} \in \mathbb{N}_0^d$ and a vector $x \in \mathbb{F}^d$, we define

$$|\mathbf{k}| := \sum_{j=1}^d k_j, \quad \mathbf{k}! := \prod_{j=1}^d k_j! \quad \text{and} \quad x^{\mathbf{k}} := \prod_{j=1}^d x_j^{k_j}.$$

For a natural number $s \in \mathbb{N}_0$ and a set $K \subseteq \mathbb{R}^d$, we let

$$\mathcal{P}_s(K) := \left\{ K \ni x \mapsto \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^d \\ |\mathbf{k}| \leq s}} a_{\mathbf{k}} x^{\mathbf{k}} : a_{\mathbf{k}} \in \mathbb{R} \right\}$$

denote the set of (real) polynomials on K of degree at most s and write $\mathcal{P}(K) := \bigcup_{s \in \mathbb{N}_0} \mathcal{P}_s(K)$ for the set of polynomials on K .

1.3 Main results

In this section, we present the main results derived in the present paper. We start by stating the lower bound for the approximation of Sobolev functions by sums of multivariate ridge functions.

Theorem 1.1 *Let $d, \ell, r \in \mathbb{N}$ with $d > \ell$ and $p, q \in [1, \infty]$. Then there exists a positive constant $c = c(d, \ell, p, q, r) > 0$ with the following property: For every $n \in \mathbb{N}$ there exists a Sobolev function $f \in \mathcal{B}(W_d^{r,p})$ such that for every choice of functions $\varrho_1, \dots, \varrho_n : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ with the property that $\varrho_k(A_k \cdot \bullet) \in L^1(B^d)$ for every $k \in \{1, \dots, n\}$, we have*

$$\left\| f(x) - \sum_{k=1}^n \varrho_k(A_k x) \right\|_{L^q(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

We refer to Section 3 and specifically to Corollary 3.10 for the proof. Remarkably, this lower bound holds under the weak assumption that $\varrho_k(A_k \cdot \bullet) \in L^1(B^d)$ for every $k \in \{1, \dots, n\}$. This is, for example, satisfied whenever A_k is full-rank and $\varrho_k \in L^1_{\text{loc}}(\mathbb{R}^\ell)$; see Proposition 2.1.

The complementary upper bound can already be achieved by only considering polynomial functions and a *fixed* choice of matrices $A_k \in \mathbb{R}^{\ell \times d}$.

Theorem 1.2 *Let $d, \ell, r \in \mathbb{N}$ with $d > \ell$ and $1 \leq q \leq p \leq \infty$. Then there exists a positive constant $C = C(d, \ell, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ such that for any $f \in \mathcal{B}(W_d^{r,p})$ there exist polynomials $P_1, \dots, P_n \in \mathcal{P}(\mathbb{R}^\ell)$ with*

$$\left\| f(x) - \sum_{k=1}^n P_k(A_k x) \right\|_{L^q(B^d)} \leq C \cdot n^{-r/(d-\ell)}.$$

The proof of this statement can be found in Section 4 (Theorem 4.2).

We use these results to obtain sharp asymptotic bounds for the approximation error using generalized translation networks and complex-valued neural networks. We start by stating a lower and upper bound for the approximation error using generalized translation networks.

Theorem 1.3 *Let $\ell \in \mathbb{N}$. Then the following two statements hold:*

- (1) *For any $d, r \in \mathbb{N}$ with $d > \ell$ and any $p, q \in [1, \infty]$, there exists a constant $c = c(d, \ell, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exists a function $f \in \mathcal{B}(W_d^{r,p})$ such that for any choice of $\tau \in L^1_{\text{loc}}(\mathbb{R}^\ell)$, matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ with $\text{rank}(A_k) \in \{0, \ell\}$, biases $b_1, \dots, b_n \in \mathbb{R}^\ell$ and coefficients $c_1, \dots, c_n \in \mathbb{R}$ we have*

$$\left\| f(x) - \sum_{k=1}^n c_k \tau(A_k x + b_k) \right\|_{L^q(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

If $\tau \in \mathcal{M}_b(\mathbb{R}^\ell)$, the assumption $\text{rank}(A_k) \in \{0, \ell\}$ is not needed.

- (2) *There exists a smooth activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ with the property that for any choice of $d, r \in \mathbb{N}$ with $d > \ell$ and any $1 \leq q \leq p \leq \infty$ there exists a constant $C = C(d, \ell, p, q, r) > 0$ satisfying the following: For any $n \in \mathbb{N}$ there*

exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ such that for any function $f \in \mathcal{B}(W_d^{r,p})$ there exist biases $b_1, \dots, b_n \in \mathbb{R}^\ell$ and coefficients $c_1, \dots, c_n \in \mathbb{R}$ with

$$\left\| f(x) - \sum_{k=1}^n c_k \tau(A_k x + b_k) \right\|_{L^q(B^d)} \leq C \cdot n^{-r/(d-\ell)}.$$

The proofs of these statements are easy consequences of Theorems 1.1 and 1.2; see Section 5.1.

We furthermore establish similar bounds for the approximation error using complex-valued neural networks (CVNNs). While the lower bound for CVNNs is an immediate consequence of Theorem 1.1 and might be viewed as a special case of Theorem 1.3(1) (when taking $\ell = 2$), the upper bound is *not* immediately obtained as a special case of Theorem 1.3(2); see the discussion at the beginning of Section 5. To express the result in a convenient way, we use the identification $\mathbb{C}^d \cong \mathbb{R}^{2d}$ so that we can consider $B^{2d} \subseteq \mathbb{R}^{2d}$ as a subset of $\mathbb{C}^d$.

Theorem 1.4 *The following two statements hold:*

- (1) *For any $d, r \in \mathbb{N}$ with $d \geq 2$ and any $p, q \in [1, \infty]$, there exists a constant $c = c(d, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exists a function $f : \mathbb{C}^d \rightarrow \mathbb{C}$ that satisfies $\operatorname{Re}(f), \operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p})$ such that for any choice of a function $\phi \in L^1_{\text{loc}}(\mathbb{C}; \mathbb{C})$, complex vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$, biases $\beta_1, \dots, \beta_n \in \mathbb{C}$ and coefficients $\gamma_1, \dots, \gamma_n \in \mathbb{C}$ we have*

$$\left\| f(z) - \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \right\|_{L^q(B^{2d})} \geq c \cdot n^{-r/(2d-2)}.$$

- (2) *There exists a smooth¹ activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ with the property that for any choice of $d, r \in \mathbb{N}$ with $d \geq 2$ and any $1 \leq q \leq p \leq \infty$ there exists a constant $C = C(d, p, q, r) > 0$ satisfying the following: For any $n \in \mathbb{N}$ there exist complex vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$ such that for any function $f : \mathbb{C}^d \rightarrow \mathbb{C}$ with $\operatorname{Re}(f), \operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p})$ there exist biases $\beta_1, \dots, \beta_n \in \mathbb{C}$ and coefficients $\gamma_1, \dots, \gamma_n \in \mathbb{C}$ with*

$$\left\| f(z) - \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \right\|_{L^q(B^{2d})} \leq C \cdot n^{-r/(2d-2)}.$$

We refer to Section 5.2 for the proof of this theorem (see in particular Corollary 5.3 and Theorem 5.9).

1.4 Organization of the paper

In Section 2, we introduce the central objects of the present work and discuss preliminary statements which are important ingredients for our proofs. Section 3 is devoted

¹ Smoothness is understood in the sense of real variables here, i.e., identifying $\mathbb{C} \cong \mathbb{R}^2$.

to proving the lower bound stated in Theorem 1.1. The complementary upper bound from Theorem 1.2 is proven in Section 4. In Section 5, these bounds are used to obtain the results for generalized translation networks and complex-valued neural networks stated in Theorems 1.3 and 1.4. Lastly, the appendices contain postponed proofs and a discussion of the gap in the proof of the lower bound in [18].

2 Preliminaries

2.1 Multivariate ridge functions

In this work, we study the approximation using sums of n arbitrary multivariate (more precisely, ℓ -variate) ridge functions. Here, for a given $d \in \mathbb{N}$ and a natural number $\ell \in \{1, \dots, d-1\}$, an ℓ -variate ridge function is a function $\varrho^* : \mathbb{R}^d \rightarrow \mathbb{R}$ of the form $\varrho^*(x) := \varrho(Ax)$, where $\varrho : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and $A \in \mathbb{R}^{\ell \times d}$. Note that we get a classical ridge function in the case $\ell = 1$. In order to derive approximation bounds, one needs to impose certain (mild) assumptions on the functions ϱ or ϱ^* which we discuss here.

Recall that B^d denotes the closed unit ball in $\mathbb{R}^d$. We let

$$\mathcal{R}_{d,\ell} := \left\{ \varrho^* : B^d \rightarrow \mathbb{R}, x \mapsto \varrho(Ax) : A \in \mathbb{R}^{\ell \times d}, \right. \\ \left. \varrho : \mathbb{R}^\ell \rightarrow \mathbb{R} \text{ measurable with } \varrho^* \in L^1(B^d) \right\}$$

denote the set of all ℓ -variate ridge functions $B^d \rightarrow \mathbb{R}$ that belong to $L^1(B^d)$. Moreover, we define

$$\mathcal{R}_{d,\ell}^* := \left\{ \varrho^* : B^d \rightarrow \mathbb{R}, x \mapsto \varrho(Ax) : A \in \mathbb{R}^{\ell \times d}, \varrho \in \mathcal{M}_b(\mathbb{R}^\ell) \right\},$$

where we recall that $\mathcal{M}_b(\mathbb{R}^\ell)$ denotes the set of all locally bounded measurable functions. Here, it is important to note the subtle difference that in the definition of $\mathcal{R}_{d,\ell}$ we require the composed map ϱ^* to be in $L^1(B^d)$, whereas in the definition of $\mathcal{R}_{d,\ell}^*$ we require the low-dimensional map ϱ to be in $\mathcal{M}_b(\mathbb{R}^\ell)$.

The following proposition provides a sufficient condition for a function to belong to $\mathcal{R}_{d,\ell}$.

Proposition 2.1 *Let $d \in \mathbb{N}$ and $\ell \in \{1, \dots, d-1\}$. Let $\varrho \in L^1_{\text{loc}}(\mathbb{R}^\ell)$ and $A \in \mathbb{R}^{\ell \times d}$ with $\text{rank}(A) = \ell$. Then we have $\varrho^* \in \mathcal{R}_{d,\ell}$ where $\varrho^*(x) := \varrho(Ax)$ for $x \in B^d$.*

Proof According to the definition of the set $\mathcal{R}_{d,\ell}$, it suffices to show $\varrho^* \in L^1(B^d)$. Let $A = U\Sigma V^T$ be the singular value decomposition of A with orthogonal matrices $U \in \mathbb{R}^{\ell \times \ell}$, $V \in \mathbb{R}^{d \times d}$ and a “diagonal” matrix $\Sigma \in \mathbb{R}^{\ell \times d}$ storing the singular values $\sigma_1 > \dots > \sigma_\ell > 0$ of A (note that $\text{rank}(A) = \ell$). Using the rotation invariance of the Lebesgue measure, we obtain

$$\int_{B^d} |\varrho^*(x)| \, dx = \int_{B^d} |\varrho(U\Sigma V^T x)| \, dx \stackrel{y=V^T x}{=} \int_{B^d} |\varrho(U\Sigma y)| \, dy.$$

We decompose a vector $y \in \mathbb{R}^d$ as $y = (y', y'')$ with $y' \in \mathbb{R}^\ell$ and $y'' \in \mathbb{R}^{d-\ell}$. By letting $\tilde{\Sigma} \in \mathbb{R}^{\ell \times \ell}$ be the matrix that arises from Σ by deleting the last $d - \ell$ columns (i.e., $\tilde{\Sigma}$ is the diagonal matrix with the singular values $\sigma_1, \dots, \sigma_\ell$ on the diagonal), we obtain

$$\varrho(U\Sigma y) = \tilde{\varrho}(y') \quad \text{for every } y = (y', y'') \in \mathbb{R}^\ell \times \mathbb{R}^{d-\ell} \cong \mathbb{R}^d,$$

with $\tilde{\varrho}(y') := \varrho(U\tilde{\Sigma}y')$. Thus, we get

$$\int_{B^d} |\varrho(U\Sigma y)| \, dy \leq \int_{[-1,1]^d} |\varrho(U\Sigma y)| \, dy = \int_{[-1,1]^{d-\ell}} \int_{[-1,1]^\ell} |\tilde{\varrho}(y')| \, dy' \, dy'',$$

where we used Tonelli's theorem. We compute

$$\begin{aligned} \int_{[-1,1]^\ell} |\tilde{\varrho}(y')| \, dy' &= \int_{[-1,1]^\ell} |\varrho(U\tilde{\Sigma}y')| \, dy' \stackrel{z=U\tilde{\Sigma}y'}{=} \int_{U\tilde{\Sigma}[-1,1]^\ell} |\varrho(z)| \, dz \cdot \frac{1}{\det(\tilde{\Sigma})}. \\ &\int_{U\tilde{\Sigma}[-1,1]^\ell} |\varrho(z)| \, dz =: \alpha < \infty. \end{aligned}$$

Here, we used the behavior of the Lebesgue measure under a linear change of variables, the fact that $U\tilde{\Sigma}[-1, 1]^\ell \subseteq \mathbb{R}^\ell$ is compact and that $\varrho \in L^1_{\text{loc}}(\mathbb{R}^\ell)$. Overall, we get

$$\int_{B^d} |\varrho^*(x)| \, dx \leq \alpha \cdot \int_{[-1,1]^{d-\ell}} dy'' = \alpha \cdot 2^{d-\ell} < \infty,$$

which yields the claim. $\square$

Remark 2.2 We remark that the assumption that A is full-rank in Proposition 2.1 cannot simply be omitted. To see this, take $d \in \mathbb{N}$ with $d \geq 3$ and any measurable function $h : \mathbb{R} \rightarrow \mathbb{R}$ that satisfies

$$\int_{[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]} |h(x)| \, dx = \infty.$$

For arbitrary $\ell \in \mathbb{N}$ with $d > \ell \geq 2$, we let

$$\varrho : \mathbb{R}^\ell \rightarrow \mathbb{R}, \quad \varrho(x_1, \dots, x_\ell) = \begin{cases} h(x_1) & \text{if } x_2 = \dots = x_\ell = 0, \\ 0 & \text{else.} \end{cases}$$

Since $\varrho \equiv 0$ almost everywhere, we infer $\varrho \in L^1_{\text{loc}}(\mathbb{R}^\ell)$. However, we can construct the matrix

$$A := \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{\ell \times d}.$$

For $x \in B^d$, we then get $\varrho(Ax) = h(x_1)$. This yields

$$\begin{aligned} \int_{B^d} |\varrho(Ax)| \, dx &\geq \int_{\left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]^d} |\varrho(Ax)| \, dx = \int_{\left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]^d} |h(x_1)| \, dx \\ &= \left(\frac{2}{\sqrt{d}}\right)^{d-1} \cdot \int_{\left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]} |h(x_1)| \, dx_1 = \infty. \end{aligned}$$

Thus, $\varrho(A \cdot \bullet) \notin L^1(B^d)$.

As we aim to study the approximation properties of sums of multivariate ridge functions, we define

$$\mathcal{R}_{n,d,\ell} := \left\{ \sum_{j=1}^n \varrho_j : \varrho_1, \dots, \varrho_n \in \mathcal{R}_{d,\ell} \right\} \quad (2.1)$$

and

$$\mathcal{R}_{n,d,\ell}^* := \left\{ \sum_{j=1}^n \varrho_j : \varrho_1, \dots, \varrho_n \in \mathcal{R}_{d,\ell}^* \right\}.$$

Next, we note that, given a fixed measurable function $f : B^d \rightarrow \mathbb{R}$, the approximation error when approximating f using elements of $\mathcal{R}_{n,d,\ell}$ in the $\|\cdot\|_{L^1(B^d)}$ -norm remains the same when replacing $\mathcal{R}_{n,d,\ell}$ by $\mathcal{R}_{n,d,\ell}^*$. This will allow us to focus on functions in $\mathcal{R}_{n,d,\ell}^*$ for deriving our lower bounds.

Proposition 2.3 *Let $n, d \in \mathbb{N}$ and $\ell \in \{1, \dots, d-1\}$. Moreover, let $f : B^d \rightarrow \mathbb{R}$ be a measurable function. Then we have*

$$\inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} = \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)}.$$

Proof In order to show

$$\inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} \geq \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)},$$

we may without loss of generality assume that

$$\inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} < \infty.$$

In this case, for a given $\varepsilon > 0$, there exists a function $\bar{R} \in \mathcal{R}_{n,d,\ell}$ satisfying

$$\|f - \bar{R}\|_{L^1(B^d)} \leq \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} + \varepsilon/2.$$

By definition of $\mathcal{R}_{n,d,\ell}$, we can choose $\varrho_1, \dots, \varrho_n : \mathbb{R}^\ell \rightarrow \mathbb{R}$ and $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ such that

$$\overline{R}(x) = \sum_{j=1}^n \varrho_j(A_j x) \quad \text{for } x \in \mathbb{R}^d,$$

where $(B^d \ni x \mapsto \varrho_j(A_j x)) \in L^1(B^d)$ for $j \in \{1, \dots, n\}$. For $N \in \mathbb{N}$ and $j \in \{1, \dots, n\}$, we define

$$\varrho_j^N(x) := \begin{cases} \varrho_j(x), & \text{if } |\varrho_j(x)| \leq N, \\ 0, & \text{otherwise.} \end{cases}$$

Then we have $\varrho_j^N \in \mathcal{M}_b(\mathbb{R}^\ell)$ (in fact, ϱ_j^N is even globally bounded) for every $j \in \{1, \dots, n\}$ and $N \in \mathbb{N}$. We define

$$R^N : B^d \rightarrow \mathbb{R}, \quad R^N(x) := \sum_{j=1}^n \varrho_j^N(A_j x)$$

and note $R^N \in \mathcal{R}_{n,d,\ell}^*$.

Clearly, we have $R^N \rightarrow \overline{R}$ pointwise on B^d as $N \rightarrow \infty$. Moreover, we observe

$$|R^N(x)| \leq \sum_{j=1}^n |\varrho_j^N(A_j x)| \leq \sum_{j=1}^n |\varrho_j(A_j x)|$$

for every $x \in B^d$, where the right-hand side belongs to $L^1(B^d)$. Therefore, by dominated convergence we get

$$\|R^N - \overline{R}\|_{L^1(B^d)} \rightarrow 0 \quad (N \rightarrow \infty).$$

Hence, we can pick $N \in \mathbb{N}$ large enough such that

$$\|R^N - \overline{R}\|_{L^1(B^d)} \leq \varepsilon/2.$$

We thus get

$$\begin{aligned} \|f - R^N\|_{L^1(B^d)} &\leq \|f - \overline{R}\|_{L^1(B^d)} + \|R^N - \overline{R}\|_{L^1(B^d)} \\ &\leq \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we get

$$\inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} \geq \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)}. \quad (2.2)$$

Lastly, note that

$$\mathcal{R}_{n,d,\ell}^* \subseteq \mathcal{R}_{n,d,\ell},$$

which implies

$$\inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)} \geq \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)}. \quad (2.3)$$

Equations (2.2) and (2.3) together yield the claim. $\square$

In the following proposition, we note that the matrices in the definition of the set $\mathcal{R}_{d,\ell}^*$ (and therefore also in the definition of the set $\mathcal{R}_{n,d,\ell}^*$) can be replaced by matrices with *orthonormal rows*. Here, we say that a matrix $A \in \mathbb{R}^{\ell \times d}$ has orthonormal rows if and only if $AA^T = I_{\ell \times \ell}$, which is equivalent to the rows of A being an orthonormal system in $\mathbb{R}^d$. Completing the columns of A^T to an orthonormal basis of $\mathbb{R}^d$, we obtain in that case the existence of an orthogonal matrix $\sigma \in \mathbb{R}^{d \times d}$ with

$$A\sigma = I_{\ell \times d},$$

where

$$I_{\ell \times d} := (I_{\ell \times \ell} | 0) \in \mathbb{R}^{\ell \times d}.$$

Proposition 2.4 *Let $d, \ell \in \mathbb{N}$ with $\ell < d$. Then we have*

$$\mathcal{R}_{d,\ell}^* = \left\{ B^d \rightarrow \mathbb{R}, x \mapsto \varrho(Ax) : \varrho \in \mathcal{M}_b(\mathbb{R}^\ell), A \in \mathbb{R}^{\ell \times d} \text{ with } AA^T = I_{\ell \times \ell} \right\}.$$

Proof The inclusion “ $\supseteq$ ” is trivial. To prove “ $\subseteq$ ”, let $\varrho \in \mathcal{M}_b(\mathbb{R}^\ell)$ and $A \in \mathbb{R}^{\ell \times d}$ be arbitrary. Let $A = U\Sigma V^T$ be the compact singular value decomposition of A with (semi-)orthogonal matrices $U \in \mathbb{R}^{\ell \times \ell}$, $V \in \mathbb{R}^{d \times \ell}$ (i.e., $V^T V = U U^T = I_{\ell \times \ell}$) and a diagonal matrix $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_\ell) \in \mathbb{R}^{\ell \times \ell}$ with $\sigma_1 \geq \dots \geq \sigma_\ell \geq 0$. We define

$$\tilde{\varrho} : \mathbb{R}^\ell \rightarrow \mathbb{R}, \quad \tilde{\varrho}(x) = \varrho(U\Sigma x).$$

Clearly, it holds that $\tilde{\varrho} \in \mathcal{M}_b(\mathbb{R}^\ell)$. Moreover, for arbitrary $x \in \mathbb{R}^d$ we get

$$\varrho(Ax) = \varrho(U\Sigma V^T x) = \tilde{\varrho}(V^T x)$$

where $V^T \in \mathbb{R}^{\ell \times d}$ is a matrix with orthonormal rows. This proves the claim. $\square$

2.2 Sobolev functions on B^d

In this work, we study the approximation of Sobolev functions by sums of multivariate ridge functions. Therefore, we present our notation regarding Sobolev functions in this paragraph.

We identify $W^{r,p}(B^d)$ with $W^{r,p}((B^d)^\circ)$, where $(B^d)^\circ$ is the open unit ball in $\mathbb{R}^d$. Thus, for arbitrary $p \in [1, \infty]$, we call a function $f \in L^p(B^d)$ an L^p -Sobolev function of regularity $r \in \mathbb{N}$ if for every multiindex $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq r$ the derivative $\partial^{\mathbf{k}} f$ exists in the weak sense on $(B^d)^\circ$ and is itself contained in $L^p(B^d)$. For such a function $f : B^d \rightarrow \mathbb{R}$, we define

$$\|f\|_{W_d^{r,p}} := \begin{cases} \left(\sum_{\mathbf{k} \in \mathbb{N}_0^d, |\mathbf{k}| \leq r} \|\partial^{\mathbf{k}} f\|_{L^p(B^d)}^p \right)^{1/p}, & p < \infty, \\ \max_{\mathbf{k} \in \mathbb{N}_0^d, |\mathbf{k}| \leq r} \|\partial^{\mathbf{k}} f\|_{L^\infty(B^d)}, & p = \infty. \end{cases} \quad (2.4)$$

We then write $\mathcal{B}(W_d^{r,p})$ for the unit ball in the L^p -Sobolev space of regularity r , i.e., for the set of functions $f \in W^{r,p}(B^d)$ for which $\|f\|_{W_d^{r,p}} \leq 1$.

The following result (a Jackson-type bound for Sobolev functions) is folklore and is essential for deriving the approximation bounds in this paper.

Proposition 2.5 (cf. [21, Equation (2.10)]) *Let $d, r \in \mathbb{N}$ and $1 \leq q \leq p \leq \infty$. Then there exists a constant $C = C(d, p, q, r) > 0$ with the following property: For any $s \in \mathbb{N}$ and any $f \in \mathcal{B}(W_d^{r,p})$ there exists a polynomial $P \in \mathcal{P}_s(B^d)$ such that*

$$\|f - P\|_{L^q(B^d)} \leq C \cdot s^{-r}.$$

2.3 Quasi-projection operator onto $\mathcal{P}_s(B^d)$

The main goal of this subsection is to set up certain “quasi-projection” operators $\text{Pr}_s : L^1(B^d) \rightarrow \mathcal{P}_{2s-1}(B^d)$ for $s \in \mathbb{N}$ such that $\text{Pr}_s(P) = P$ for all $P \in \mathcal{P}_s(B^d)$ and such that $\sup_{s \in \mathbb{N}} \|\text{Pr}_s\|_{L^1(B^d) \rightarrow L^1(B^d)} < \infty$.

The space $L^2(B^d)$ together with the inner product defined in (1.2) forms a Hilbert space. Since the set $\mathcal{P}(B^d)$ is dense in $L^2(B^d)$ and the set

$$\mathcal{B} := \{x^{\mathbf{k}} : \mathbf{k} \in \mathbb{N}_0^d\}$$

forms a basis of $\mathcal{P}(B^d)$, we conclude that the linear span of $\mathcal{B}$ is dense in $L^2(B^d)$. Let the map $\varphi : \mathbb{N} \rightarrow \mathbb{N}_0^d$ be a bijection with the property that $n \mapsto |\varphi(n)|$ is non-decreasing. By applying the Gram-Schmidt algorithm to $\mathcal{B}$ with respect to the inner product $\langle \cdot, \cdot \rangle$ on $L^2(B^d)$ in the order that is given by the enumeration φ , we obtain a countable set $\Pi^d = \{P_i\}_{i \in \mathbb{N}} \subseteq \mathcal{P}(B^d)$ with the following properties:

- (1) Π^d is an orthonormal basis of $L^2(B^d)$.
- (2) For each $s \in \mathbb{N}$ the set $\Pi_s^d := \{P_i\}_{i \in I_s}$ forms an orthonormal basis of $\mathcal{P}_s(B^d)$, where we define $I_s := \{i \in \mathbb{N} : \deg(P_i) \leq s\}$.

For a given $s \in \mathbb{N}_0$, we let $J_s := \{i \in \mathbb{N} : \deg(P_i) = s\}$. For $x, y \in B^d$, we then set

$$Q_s(x, y) := \sum_{i \in J_s} P_i(x) P_i(y), \quad L_s(x, y) := \sum_{k=0}^{\infty} \eta\left(\frac{k}{s}\right) Q_k(x, y) \quad \text{for } s \in \mathbb{N},$$

where $\eta : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth function with $\eta(x) = 1$ for $x \in [-1, 1]$ and $\eta(x) = 0$ for $|x| \geq 2$. Lastly, following [4, Def. 11.5.1, Def. 11.1.1, bottom of p. 268], for $f \in L^1(B^d)$, we define $\text{Pr}_s f : B^d \rightarrow \mathbb{R}$ via

$$(\text{Pr}_s f)(x) := \langle f, L_s(x, \bullet) \rangle.$$

For $f \in L^1(B^d)$, a computation shows

$$\begin{aligned} (\text{Pr}_s f)(x) &= \langle f, L_s(x, \bullet) \rangle = \sum_{k=0}^{2s-1} \eta(k/s) \cdot \langle f, Q_k(x, \bullet) \rangle = \sum_{k=0}^{2s-1} \sum_{i \in J_k} \eta(k/s) \cdot P_i(x) \cdot \langle f, P_i \rangle \\ &= \sum_{i \in I_{2s-1}} \eta(\deg(P_i)/s) \cdot \langle f, P_i \rangle \cdot P_i(x) = \sum_{i \in I_{2s-1}} a_{i,s} \cdot \langle f, P_i \rangle \cdot P_i(x), \end{aligned} \quad (2.5)$$

where we denote $a_{i,s} := \eta(\deg(P_i)/s)$.

Alternatively, motivated by [4, Equation (11.1.12)], if we let $\text{proj}_k(f) := \sum_{j \in J_k} \langle f, P_j \rangle \cdot P_j$ denote the orthogonal projection onto the subspace spanned by $\{P_j\}_{j \in J_k}$, we may write

$$\text{Pr}_s f = \sum_{k=0}^{\infty} \eta(k/s) \cdot \text{proj}_k(f).$$

In order to show that Pr_s satisfies the properties stated at the beginning of this subsection it will be helpful to define, for $k, \sigma \in \mathbb{N}_0$,

$$S_k^\sigma(f) := \frac{1}{\binom{k+\sigma}{\sigma}} \cdot \sum_{j=0}^k \binom{k-j+\sigma}{\sigma} \cdot \text{proj}_j(f),$$

see [4, Equation (11.2.8) and (A.4.2)]. The sequence $(S_k^\sigma(f))_{k \in \mathbb{N}_0}$ is called the sequence of σ -Cesaro means of the sequence $(\text{proj}_k(f))_{k \in \mathbb{N}_0}$.

Moreover, for a function $g : \mathbb{R} \rightarrow \mathbb{R}$, we let $(\Delta g)(x) := g(x) - g(x+1)$ and recursively define $\Delta^{\sigma+1} g := \Delta(\Delta^\sigma(g))$ for any $\sigma \in \mathbb{N}_0$. The following technical lemma provides a useful identity.

Lemma 2.6 *For $s \in \mathbb{N}$ and $\sigma \in \mathbb{N}_0$, let everything be defined as above. We let $\eta^*(x) := \eta(x/s)$. Then*

$$\text{Pr}_s f = \sum_{k=0}^{\infty} \left(\Delta^{\sigma+1} \eta^* \right)(k) \cdot \binom{k+\sigma}{\sigma} \cdot S_k^\sigma(f).$$

The proof of Lemma 2.6 can be found in Appendix B.1.

We can now deduce three properties of the operator Pr_s that are central for the present work.

Proposition 2.7 (cf. [4, Theorem 11.5.2])

- (1) $\text{Pr}_s : L^1(B^d) \rightarrow \mathcal{P}_{2s-1}(B^d)$ is a well-defined linear operator.
- (2) For every $P \in \mathcal{P}_s(B^d)$ we have $\text{Pr}_s P = P$.
- (3) There exists a constant $C = C(d) > 0$ such that $\|\text{Pr}_s f\|_{L^1(B^d)} \leq C \cdot \|f\|_{L^1(B^d)}$ for every $f \in L^1(B^d)$ and $s \in \mathbb{N}$.

Proposition 2.7 is stated in [4], but the proof is omitted, since it is similar to the proof of a different result in [4]. Since Proposition 2.7 is essential for our argument and to make our paper more self-contained, we give a proof.

Proof of Proposition 2.7 (1) Follows from Equation (2.5).

- (2) Let $P \in \mathcal{P}_s(B^d)$. By orthogonality, we have $\langle P, P_i \rangle = 0$ if $\deg(P_i) > s$. By (2.5), we hence get

$$\begin{aligned} \text{Pr}_s P &= \sum_{i \in I_{2s-1}} \eta(\deg(P_i)/s) \cdot \langle P, P_i \rangle \cdot P_i \\ &= \sum_{i \in I_s} \underbrace{\eta(\deg(P_i)/s)}_{\in [0,1]} \cdot \langle P, P_i \rangle \cdot P_i = \sum_{i \in I_s} \langle P, P_i \rangle \cdot P_i = P, \end{aligned}$$

where the last step uses that $(P_i)_{i \in I_s}$ is an orthonormal basis of $\mathcal{P}_s(B^d)$.

- (3) By Lemma 2.6, we have

$$\text{Pr}_s f = \sum_{k=0}^{\infty} \left(\left(\Delta^{\sigma+1} \eta^* \right) (k) \cdot \binom{k+\sigma}{\sigma} \cdot S_k^{\sigma}(f) \right) \quad \text{for every } \sigma \in \mathbb{N}_0$$

with $\eta^*(x) := \eta(x/s)$. Note that the choice of η and the definition of $\Delta^{\sigma+1}$ imply that $\Delta^{\sigma+1} \eta^*(k) = 0$ for $k \geq 2s$. We hence get

$$\|\text{Pr}_s f\|_{L^1(B^d)} \leq \sum_{k=0}^{2s-1} \left| \left(\Delta^{\sigma+1} \eta^* \right) (k) \right| \cdot \binom{k+\sigma}{\sigma} \cdot \|S_k^{\sigma}(f)\|_{L^1(B^d)}.$$

We observe, with $(\eta^*)^{(\sigma+1)}$ denoting the $(\sigma + 1)$ -th derivative of η^* ,

$$\begin{aligned} \left| \left(\Delta^{\sigma+1} \eta^* \right) (k) \right| &\leq \int_{[0,1]^{\sigma+1}} \left| (\eta^*)^{(\sigma+1)} (k + u_1 + \dots + u_{\sigma+1}) \right| du_1 \dots du_{\sigma+1} \\ &\leq \sup_{x \in \mathbb{R}} \left| (\eta^*)^{(\sigma+1)} (x) \right|, \end{aligned}$$

as follows from [4, Proposition A.3.1 (ii)]. Since $(\eta^*)^{(\sigma+1)}(x) = s^{-\sigma-1} \cdot \eta^{(\sigma+1)}(x/s)$ for every $x \in \mathbb{R}$ by the chain rule, we get

$$\left| \left(\Delta^{\sigma+1} \eta^* \right) (k) \right| \leq s^{-\sigma-1} \cdot C_1$$

with $C_1 = C_1(\sigma) := \sup_{x \in \mathbb{R}} |\eta^{(\sigma+1)}(x)| < \infty$.

Secondly, using [32, Exercise 0.0.5], we have

$$\binom{k + \sigma}{\sigma} \leq \left(\frac{e(k + \sigma)}{\sigma} \right)^\sigma = \left(\frac{ek}{\sigma} + e \right)^\sigma \leq e^\sigma \cdot (k + 1)^\sigma = C_2 \cdot (k + 1)^\sigma$$

for $\sigma \geq 1$, with $C_2 = C_2(\sigma) = e^\sigma$.

Lastly, by picking $\sigma > \frac{d}{2}$, we can ensure that

$$\|S_k^\sigma(f)\|_{L^1(B^d)} \leq C_3 \cdot \|f\|_{L^1(B^d)} \quad \text{for all } f \in L^1(B^d)$$

with a constant $C_3 = C_3(d, \sigma)$ which does not depend on k . This follows from [4, Theorem 11.4.1], where we choose $\kappa = (0, \dots, 0, 1/2) \in \mathbb{R}^{d+1}$ in order to obtain the Lebesgue measure.

Hence, by fixing $\sigma > d/2$, we obtain

$$\begin{aligned} \|\text{Pr}_s f\|_{L^1(B^d)} &\leq C_1 \cdot C_2 \cdot C_3 \cdot s^{-\sigma-1} \cdot \sum_{k=0}^{2s-1} (k + 1)^\sigma \cdot \|f\|_{L^1(B^d)} \\ &\leq C_1 \cdot C_2 \cdot C_3 \cdot s^{-\sigma-1} \cdot \sum_{k=0}^{2s-1} (2s)^\sigma \cdot \|f\|_{L^1(B^d)} \\ &\leq C_1 \cdot C_2 \cdot C_3 \cdot s^{-\sigma-1} \cdot 2^{\sigma+1} \cdot s^{\sigma+1} \cdot \|f\|_{L^1(B^d)} = C \cdot \|f\|_{L^1(B^d)}, \end{aligned}$$

where $C = C(d) := C_1 C_2 C_3 \cdot 2^{\sigma+1}$. Note that the choice of σ only depends on d . This proves the claim. $\square$

The properties from Proposition 2.7 justify that we call Pr_s a *quasi-projection* onto $\mathcal{P}_s(B^d)$ with range in $\mathcal{P}_{2s-1}(B^d)$. Crucially, this operator is bounded with respect to the L^1 -norm and the operator norm can be upper bounded independently of s , whereas the same does not hold for an *orthogonal* projection onto $\mathcal{P}_s(B^d)$.

3 Proof of the lower bound

In this section, we prove the asymptotic lower bound of $n^{-r/(d-\ell)}$ for the approximation error of Sobolev functions using linear combinations of n multivariate ridge functions, see Theorem 1.1. We split this section into three parts: Firstly, we provide an overview of the proof strategy in Section 3.1. In Section 3.2, we prove that the cardinality of a certain set, defined in (3.7), can be bounded from above in a suitable way. This upper bound is then used in Section 3.3 to prove Theorem 1.1.

3.1 Proof overview

Since the proof of the lower bound is in large parts quite technical, we provide a proof overview to explain the underlying idea of the proof and its main steps.

First, it is not difficult to see that it suffices to prove Theorem 1.1 for the set $\mathcal{R}_{n,d,\ell}^*$ and for the case $p = \infty$, $q = 1$, i.e., it suffices to show

$$\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)} \gtrsim_{d,\ell,r} n^{-r/(d-\ell)}, \quad (3.1)$$

where the notation “ $\gtrsim_{d,\ell,r}$ ” indicates an inequality up to a multiplicative factor depending only on d , ℓ and r .

As the next core idea of the proof, we note that the infimum over $\mathcal{R}_{n,d,\ell}^*$ in (3.1) can effectively be replaced by an infimum over $\text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)$, where the degree $s \in \mathbb{N}$ needs to be carefully balanced with the number of summands n . Here, we recall that Pr_s denotes the quasi projection onto $\mathcal{P}_s(B^d)$ with range in $\mathcal{P}_{2s-1}(B^d)$ as discussed in Section 2.3. To see that it is enough to consider $\text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)$, we pick a (large) constant $C = C(d, \ell, r) > 0$ and pick $s \in \mathbb{N}$ such that $C \cdot n \leq s^{d-\ell} \leq 2C \cdot n$. Using Jackson’s inequality (see Proposition 2.5) and the properties of Pr_s noted in Proposition 2.7 one can deduce the existence of a constant $C_{\text{proj}} = C_{\text{proj}}(d, r) > 0$ such that

$$\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \|f - \text{Pr}_s(f)\|_{L^1(B^d)} \leq C_{\text{proj}} \cdot s^{-r} \leq C_{\text{proj}} \cdot C^{-r/(d-\ell)} \cdot n^{-r/(d-\ell)}. \quad (3.2)$$

The goal is then to show that

$$\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{P \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)} \|f - P\|_{L^1(B^d)} \gtrsim_{d,\ell,r} C^{(r/d)-(r/(d-\ell))} \cdot n^{-r/(d-\ell)}. \quad (3.3)$$

Indeed, since

$$\begin{aligned} \|f - R\|_{L^1(B^d)} &\gtrsim_d \|\text{Pr}_s(f) - \text{Pr}_s(R)\|_{L^1(B^d)} \\ &\geq \|f - \text{Pr}_s(R)\|_{L^1(B^d)} - \|f - \text{Pr}_s(f)\|_{L^1(B^d)} \end{aligned}$$

for every $R \in \mathcal{R}_{n,d,\ell}^*$, we see that (3.2) and (3.3) together imply (3.1) if C is chosen sufficiently large.

Next, we construct a function $f \in \mathcal{B}(W_d^{r,\infty})$ realizing the lower bound in (3.3) by considering a set of certain sums of smooth bump functions. More precisely, we pick $m \asymp_{d,\ell,r} C^{-1} \cdot s^d$ (where $C = C(d, \ell, r) > 0$ is the constant from above which balances the degree s with n). We then choose $\xi_1, \dots, \xi_m \in B^d$ such that the cubes $Q_i := \xi_i + Q$ are pairwise disjoint subsets of B^d (in fact, the Q_i have distance of order $m^{-1/d}$ from each other), where $Q \subseteq \mathbb{R}^d$ is a cube with $\lambda^d(Q) \asymp_d m^{-1}$, meaning that the side lengths are of order $m^{-1/d}$. For given $\varepsilon \in \{\pm 1\}^m$, we then let $f_\varepsilon \in \mathcal{B}(W_d^{r,\infty})$ be a smooth function satisfying $f_\varepsilon \equiv \varepsilon_i \cdot \frac{\kappa}{m^{r/d}}$ on Q_i for every $i \in \{1, \dots, m\}$. The constant $\kappa = \kappa(d, r) > 0$ is needed to ensure $\|f_\varepsilon\|_{W_d^{r,\infty}} \leq 1$. For arbitrary $P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)$ we then compute

$$\begin{aligned} \|f_\varepsilon - P\|_{L^1(B^d)} &\geq \sum_{i=1}^m \int_{Q_i} |f_\varepsilon(x) - P(x)| \, dx = \sum_{i=1}^m \int_Q |f_\varepsilon(\xi_i + t) - P(\xi_i + t)| \, dt \\ &\geq \int_Q \inf_{\tilde{P} \in \Pr_s(\mathcal{R}_{n,d,\ell}^*), \eta \in \mathbb{R}^d} \sum_{i=1}^m \left| \frac{\kappa}{m^{r/d}} \varepsilon_i - \tilde{P}(\xi_i + \eta) \right| \, dt \\ &\asymp_{d,r} \frac{1}{m^{1+r/d}} \cdot \inf_{\tilde{P} \in \Pr_s(\mathcal{R}_{n,d,\ell}^*), \eta \in \mathbb{R}^d} \left\| \varepsilon - (\tilde{P}(\xi_i + \eta))_{i=1}^m \right\|_{\ell^1}; \end{aligned} \quad (3.4)$$

see Proposition 3.7 for the details. By [18, Lemma 6] it is known that if

$$\left| \left\{ (\text{sgn}(\tilde{P}(\xi_1 + \eta)), \dots, \text{sgn}(\tilde{P}(\xi_m + \eta))) : \tilde{P} \in \Pr_s(\mathcal{R}_{n,d,\ell}^*), \eta \in \mathbb{R}^d \right\} \right| \leq 2^{m/4} \quad (3.5)$$

then there exists $\varepsilon^* \in \{\pm 1\}^m$ with

$$\inf_{\tilde{P} \in \Pr_s(\mathcal{R}_{n,d,\ell}^*), \eta \in \mathbb{R}^d} \left\| \varepsilon^* - (\tilde{P}(\xi_i + \eta))_{i=1}^m \right\|_{\ell^1} \geq am,$$

where $a > 0$ is an absolute constant. Plugging this into (3.4) yields

$$\|f_{\varepsilon^*} - P\|_{L^1(B^d)} \gtrsim_{d,r} m^{-r/d} \asymp_{d,\ell,r} C^{r/d} \cdot s^{-r} \gtrsim_{d,\ell,r} C^{(r/d)-(r/(d-\ell))} \cdot n^{-r/(d-\ell)},$$

as desired.

It remains to show that (3.5) is indeed true. The proof of this fact is contained in Section 3.2 and is based on an application of [17, Lemma 3] together with a careful evaluation of the inner product of a (multivariate) ridge function and a given polynomial using multivariate polar coordinates; see Lemma 3.5.

On a high level, the proof structure follows that of the result for univariate ridge functions in [18], properly modified and adapted to the multivariate setting. In particular, we highlight the following two key differences:

- Central to the proof in [18] is Lemma 2, where it is shown that the inner product (as defined in (1.2)) between a (univariate) ridge function and a polynomial from a

special system of orthogonal polynomials admits a certain separation of variables. We, in contrast, express the inner product between a (multivariate) ridge function and a fixed polynomial in a different way, proving a generalization of [17, Theorem 3]; see Lemma 3.5. Moreover, while the specific choice of the system of orthogonal polynomials is of central importance in [18], we impose (almost) no restrictions on the set of orthogonal polynomials that we consider but can simply use the “naive” system defined in Section 2.3.

- In contrast to the proof in [18], we do not use the *orthogonal* projections onto the space $\mathcal{P}_s(B^d)$ but the quasi-projection operators onto $\mathcal{P}_s(B^d)$ with range in $\mathcal{P}_{2s-1}(B^d)$ defined in Section 2.3. The fact that these operators are uniformly bounded with respect to the L^1 -norm enables us to show the lower bound with respect to the L^1 -norm and therefore close the gap in the proof in [18] which we describe in more detail in Appendix A.

3.2 Sign set cardinality

In this section, we show that (3.5) is satisfied under suitable assumptions. To this end, let $d, \ell, m, n, s \in \mathbb{N}$ with $\ell < d$ be arbitrary and fix² $\vartheta = \vartheta(d, m) \in \mathbb{N}$ satisfying

$$\frac{m^{1/d}}{2} \leq \vartheta \leq m^{1/d}.$$

Consider the lattice

$$\Xi := \left\{ \left(\frac{i_1 + 1/2}{\sqrt{d}\vartheta}, \dots, \frac{i_d + 1/2}{\sqrt{d}\vartheta} \right) : i_1, \dots, i_d \in \mathbb{Z} \cap [-\vartheta, \vartheta - 1] \right\} \subseteq B^d.$$

Then it holds that $|\Xi| = (2\vartheta)^d \geq m$. Let $\{\xi_1, \dots, \xi_m\} \subseteq \Xi$ with $\xi_i \neq \xi_j$ for $i \neq j$. We then set

$$\Pi_{m,s,n,\ell,d} := \left\{ (P(\xi_1 + t), \dots, P(\xi_m + t)) : P \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*), t \in \mathbb{R}^d \right\} \subseteq \mathbb{R}^m. \quad (3.6)$$

Here, Pr_s is as introduced in Section 2.3. Moreover, we define

$$E^m := \{\pm 1\}^m = \{\varepsilon = (\varepsilon_1, \dots, \varepsilon_m) : \varepsilon_i = \pm 1, i = 1, \dots, m\}.$$

We want to show that the set defined in (3.5), i.e.,

$$\text{sgn}(\Pi_{m,s,n,\ell,d}) := \{(\text{sgn}(h_1), \dots, \text{sgn}(h_m)) : h \in \Pi_{m,s,n,\ell,d}\} \subseteq E^m \quad (3.7)$$

(where $\text{sgn}(a) = 1$ if $a \geq 0$ and $\text{sgn}(a) = -1$ if $a < 0$) is “small” relative to E^m . Precisely, in this section, we will show the following statement, which may be seen as a generalization of [18, Lemma 4] to the case of multivariate ridge functions.

² Such a ϑ always exists: If $m^{1/d} \geq 2$ this follows from $m^{1/d} - \frac{m^{1/d}}{2} \geq 1$ and if $m^{1/d} \in [1, 2]$ we can simply pick $\vartheta = 1$.

Lemma 3.1 *Let $d, \ell \in \mathbb{N}$ with $\ell < d$ and $c_1 > 0$ be arbitrary and let $c_0 := 4^{\ell+3}/c_1$. Then there exists a constant $C = C(d, \ell, c_1)$ such that for any choice of $m, s, n \in \mathbb{N}$ with*

$$c_1 s^d \leq m \leq 2c_1 s^d, \quad c_0 n \leq s^{d-\ell} \leq 2c_0 n \quad \text{and} \quad s \geq C \quad (3.8)$$

we have

$$|\text{sgn}(\Pi_{m,s,n,\ell,d})| \leq 2^{m/4}.$$

Here, $\Pi_{m,s,n,\ell,d}$ and $\text{sgn}(\Pi_{m,s,n,\ell,d})$ are as defined in (3.6) and (3.7).

In the case that $|\text{sgn}(\Pi_{m,s,n,\ell,d})| \leq 2^{m/4}$, there exists a vector $\varepsilon^* \in E^m$ that is “far away” from $\Pi_{m,s,n,\ell,d}$ and we will use this vector in Section 3.3 to construct a function that realizes the lower bound for ridge function approximation. This intuition is backed up by the following lemma.

Lemma 3.2 (cf. [18, Lemma 6]) *Let $m \in \mathbb{N}$ and $\Gamma \subseteq \mathbb{R}^m$ with*

$$|\text{sgn}(\Gamma)| \leq 2^{m/4}.$$

Then there exists a vector $\varepsilon^* \in E^m$ such that

$$\inf_{x \in \Gamma} \|\varepsilon^* - x\|_{\ell^1} \geq am$$

for an absolute constant $a > 0$.

The proof of the lemma is given in [18]. However, in order to clarify the proof in [18] and to keep the paper more self-contained, we include the proof in Appendix B.2.

The remaining part of this subsection is dedicated to proving Lemma 3.1. The following lemma provides the central tool in order to obtain a bound on the cardinality of $\text{sgn}(\Pi_{m,s,n,\ell,d})$.

Lemma 3.3 (cf. [17, Lemma 3]) *Let $m, s, N, K \in \mathbb{N}$ be natural numbers such that $N + K \leq m/2$. Moreover, let $\pi_{\alpha,\beta}(z)$ for $\alpha \in \{1, \dots, m\}$ and $\beta \in \{1, \dots, K\}$ be polynomials of degree at most s in $z \in \mathbb{R}^N$, and set*

$$\pi_\alpha : \mathbb{R}^K \times \mathbb{R}^N \rightarrow \mathbb{R}, \quad \pi_\alpha(b, z) := \sum_{\beta=1}^K b_\beta \pi_{\alpha,\beta}(z) \quad \text{for } \alpha \in \{1, \dots, m\}.$$

Then, for

$$\Pi_{m,s,N,K}^* := \left\{ (\pi_1(b, z), \dots, \pi_m(b, z)) : (b, z) \in \mathbb{R}^K \times \mathbb{R}^N \right\} \subseteq \mathbb{R}^m,$$

we have

$$|\text{sgn}(\Pi_{m,s,N,K}^*)| \leq (4s)^N (N + K + 1)^{N+2} \left(\frac{2em}{N + K} \right)^{N+K}.$$

In order to show that the set $\text{sgn}(\Pi_{m,s,n,\ell,d})$ is “small” compared to E^m we need to express the set $\Pi_{m,s,n,\ell,d}$ in a way that fits the setting of Lemma 3.3. We start with a technical auxiliary statement. It is well-known, but for the sake of completeness we include a short proof in Appendix B.3.

Lemma 3.4 *Let $a, b \in \mathbb{N}_0$. Then there exist coefficients $\alpha_0 = \alpha_0(a, b), \dots, \alpha_{a+b} = \alpha_{a+b}(a, b) \in \mathbb{R}$ as well as $\beta_0 = \beta_0(a, b), \dots, \beta_{a+b} = \beta_{a+b}(a, b) \in \mathbb{R}$ with*

$$\cos(\varphi)^a \sin(\varphi)^b = \sum_{h=0}^{a+b} (\alpha_h \cos(h\varphi) + \beta_h \sin(h\varphi)), \quad \varphi \in \mathbb{R}.$$

In the following, we show that the inner product of a fixed polynomial of degree at most s and an arbitrary multivariate ridge function can be expressed in a convenient way. This is in fact a generalization of [17, Theorem 3] to the case of multivariate ridge functions.

Lemma 3.5 *Let $s, d, \ell \in \mathbb{N}$ with $\ell < d$ and set $\mu := 2^\ell(s+d+1)^\ell$. Given $P \in \mathcal{P}_s(\mathbb{R}^d)$, there exist polynomials $Q_1(\bullet; P), \dots, Q_\mu(\bullet; P) \in \mathcal{P}_s(\mathbb{R}^{d^2})$ (where we identify $\mathbb{R}^{d \times d}$ with $\mathbb{R}^{d^2}$) such that, given $A \in \mathbb{R}^{\ell \times d}$ and $\sigma \in O(d)$ with $A\sigma = I_{\ell \times d}$ and $\varrho \in \mathcal{M}_b(\mathbb{R}^\ell)$, writing $\varrho_A(x) = \varrho(Ax)$, we have*

$$\langle \varrho_A, P \rangle = \sum_{h=1}^{\mu} b_h(\varrho) Q_h(\sigma; P),$$

with coefficients $b_h(\varrho) \in \mathbb{R}$ that only depend on h and ϱ .

Proof Let $P \in \mathcal{P}_s(\mathbb{R}^d)$ be an arbitrary polynomial of degree at most s and let $\sigma \in \mathbb{R}^{d \times d}$. Reordering in terms of y , we can write

$$P(\sigma y) = \sum_{|\mathbf{k}| \leq s} P_{\mathbf{k}}(\sigma; P) y^{\mathbf{k}} \quad \text{for every } y \in \mathbb{R}^d, \quad (3.9)$$

where each function $\sigma \mapsto P_{\mathbf{k}}(\sigma; P)$ is a polynomial of degree at most s in the d^2 variables $(\sigma_{i,j})_{i,j=1,\dots,d}$. For a multiindex $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq s$, we set

$$q_{\mathbf{k},d,\ell} := \frac{1}{\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell} \cdot \int_{\mathbb{S}^{d-\ell-1}} \xi_{\ell+1}^{\mathbf{k}_{\ell+1}} \dots \xi_d^{\mathbf{k}_d} d\mathcal{H}^{d-\ell-1}(\xi), \quad (3.10)$$

where $\mathcal{H}^{d-\ell-1}$ denotes the $(d - \ell - 1)$ -dimensional Hausdorff measure. Let further $k < \ell$. Using Lemma 3.4, we can pick coefficients $\alpha_{h,\mathbf{k}}^k, \alpha_{h,\mathbf{k}}^\ell, \beta_{h,\mathbf{k}}^k, \beta_{h,\mathbf{k}}^\ell \in \mathbb{R}$ that only depend on h, d, ℓ, k and $\mathbf{k}$ such that

$$\begin{aligned} & \cos(\varphi)^{\mathbf{k}_{k+1}} \sin(\varphi)^{k-1+\mathbf{k}_1+\dots+\mathbf{k}_k} \\ &= \sum_{h=0}^{k-1+\mathbf{k}_1+\dots+\mathbf{k}_{k+1}} (\alpha_{h,\mathbf{k}}^k \cos(h\varphi) + \beta_{h,\mathbf{k}}^k \sin(h\varphi)) \quad \text{and} \end{aligned}$$

$$\begin{aligned} & \cos(\varphi)^{\mathbf{k}_{\ell+1}+\dots+\mathbf{k}_d+d-\ell+1} \sin(\varphi)^{\ell-1+\mathbf{k}_1+\dots+\mathbf{k}_\ell} \\ &= \sum_{h=0}^{d+\mathbf{k}_1+\dots+\mathbf{k}_d} (\alpha_{h,\mathbf{k}}^\ell \cos(h\varphi) + \beta_{h,\mathbf{k}}^\ell \sin(h\varphi)) \end{aligned}$$

hold for every $\varphi \in \mathbb{R}$. Since $k < \ell < d$ and $|\mathbf{k}| \leq s$, we can add zeros to obtain

$$\begin{aligned} & \cos(\varphi)^{\mathbf{k}_{k+1}} \sin(\varphi)^{k-1+\mathbf{k}_1+\dots+\mathbf{k}_k} \\ &= \sum_{h=0}^{d+s} (\alpha_{h,\mathbf{k}}^k \cos(h\varphi) + \beta_{h,\mathbf{k}}^k \sin(h\varphi)) \quad \text{and} \\ & \cos(\varphi)^{\mathbf{k}_{\ell+1}+\dots+\mathbf{k}_d+d-\ell+1} \sin(\varphi)^{\ell-1+\mathbf{k}_1+\dots+\mathbf{k}_\ell} \\ &= \sum_{h=0}^{d+s} (\alpha_{h,\mathbf{k}}^\ell \cos(h\varphi) + \beta_{h,\mathbf{k}}^\ell \sin(h\varphi)) \end{aligned} \quad (3.11)$$

for every $\varphi \in \mathbb{R}$. Note that each sum $\sum_{h=0}^{d+s} (\alpha_{h,\mathbf{k}}^k \cos(h\varphi) + \beta_{h,\mathbf{k}}^k \sin(h\varphi))$ for $k \in \{1, \dots, \ell\}$ consists in total of $2(d+s+1)$ summands. Therefore, after rearranging, we can write

$$\begin{aligned} & \prod_{k=1}^{\ell} \sum_{h=0}^{d+s} (\alpha_{h,\mathbf{k}}^k \cos(h\varphi_k) + \beta_{h,\mathbf{k}}^k \sin(h\varphi_k)) \\ &= \sum_{h=1}^{2^\ell(d+s+1)^\ell} \tilde{\zeta}_{h,\mathbf{k}} f_h(\varphi_1, \dots, \varphi_\ell) \quad \text{for all } (\varphi_1, \dots, \varphi_\ell) \in \mathbb{R}^\ell, \end{aligned} \quad (3.12)$$

where the $\tilde{\zeta}_{h,\mathbf{k}} \in \mathbb{R}$ are coefficients that depend only on $h, \mathbf{k}, d, \ell$ and each function f_h is of the form

$$f_h(\varphi_1, \dots, \varphi_\ell) = \prod_{k=1}^{\ell} \left\{ \begin{matrix} \cos \\ \sin \end{matrix} \right\}_{h,k} (\tau_{h,k} \varphi_k)$$

with $\tau_{h,k} \in \{0, 1, \dots, d+s\}$ and $\left\{ \begin{matrix} \cos \\ \sin \end{matrix} \right\}_{h,k} \in \{\cos, \sin\}$ for $h \in \{1, \dots, 2^\ell(d+s+1)^\ell\}$ and $k \in \{1, \dots, \ell\}$. Note that f_h only depends on h (and d, ℓ) but *not* on $\mathbf{k}$. We then set

$$\zeta_{h,\mathbf{k}} := q_{\mathbf{k},d,\ell} \tilde{\zeta}_{h,\mathbf{k}}, \quad (3.13)$$

where $q_{\mathbf{k},d,\ell}$ is as defined in (3.10). Lastly, for $h \in \{1, \dots, 2^\ell(s+d+1)^\ell\}$ we define

$$Q_h(\sigma; P) := \sum_{|\mathbf{k}| \leq s} \zeta_{h,\mathbf{k}} \cdot P_{\mathbf{k}}(\sigma; P), \quad (3.14)$$

where the $P_{\mathbf{k}}$ are as defined in (3.9). Since $P_{\mathbf{k}}(\bullet; P) \in \mathcal{P}_s(\mathbb{R}^{d^2})$ for every $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq s$, we infer that the same holds for the functions $Q_h(\bullet; P)$.

We claim that the polynomials $Q_h(\bullet; P)$ have the property stated in the formulation of the lemma. To see this, let $\varrho \in \mathcal{M}_b(\mathbb{R}^\ell)$. Fix a multiindex $\mathbf{k} \in \mathbb{N}_0^d$. By Fubini's theorem,

$$\begin{aligned} & \int_{B^d} \varrho(y_1, \dots, y_\ell) y^{\mathbf{k}} \, dy \\ &= \int_{B^\ell} \varrho(y_1, \dots, y_\ell) y_1^{\mathbf{k}_1} \dots y_\ell^{\mathbf{k}_\ell} \left(\int_{B_{y_1, \dots, y_\ell}} y_{\ell+1}^{\mathbf{k}_{\ell+1}} \dots y_d^{\mathbf{k}_d} \, d(y_{\ell+1}, \dots, y_d) \right) d(y_1, \dots, y_\ell) \end{aligned}$$

with

$$B_{y_1, \dots, y_\ell} := \left\{ (y_{\ell+1}, \dots, y_d) \in \mathbb{R}^{d-\ell} : y_{\ell+1}^2 + \dots + y_d^2 \leq 1 - y_1^2 - \dots - y_\ell^2 \right\}.$$

We first study the inner integral: Transforming to polar coordinates (see, e.g., [5, p. 118]), we get

$$\begin{aligned} & \int_{B_{y_1, \dots, y_\ell}} y_{\ell+1}^{\mathbf{k}_{\ell+1}} \dots y_d^{\mathbf{k}_d} \, d(y_{\ell+1}, \dots, y_d) \\ &= \int_0^{\sqrt{1-y_1^2-\dots-y_\ell^2}} r^{d-\ell-1} \cdot \left(\int_{\mathbb{S}^{d-\ell-1}} (r\xi_{\ell+1})^{\mathbf{k}_{\ell+1}} \dots (r\xi_d)^{\mathbf{k}_d} \, d\mathcal{H}^{d-\ell-1}(\xi) \right) dr \\ &= \int_0^{\sqrt{1-y_1^2-\dots-y_\ell^2}} r^{\mathbf{k}_{\ell+1}+\dots+\mathbf{k}_d+d-\ell-1} \cdot \left(\int_{\mathbb{S}^{d-\ell-1}} \xi_{\ell+1}^{\mathbf{k}_{\ell+1}} \dots \xi_d^{\mathbf{k}_d} \, d\mathcal{H}^{d-\ell-1}(\xi) \right) dr \\ &= q_{\mathbf{k}, d, \ell} \cdot (1 - y_1^2 - \dots - y_\ell^2)^{(\mathbf{k}_{\ell+1}+\dots+\mathbf{k}_d+d-\ell)/2}, \end{aligned}$$

where $q_{\mathbf{k}, d, \ell}$ is as defined in (3.10). This shows

$$\begin{aligned} & \int_{B^d} \varrho(y_1, \dots, y_\ell) y^{\mathbf{k}} \, dy \\ &= q_{\mathbf{k}, d, \ell} \cdot \int_{B^\ell} \varrho(y_1, \dots, y_\ell) y_1^{\mathbf{k}_1} \dots y_\ell^{\mathbf{k}_\ell} (1 - y_1^2 - \dots - y_\ell^2)^{(\mathbf{k}_{\ell+1}+\dots+\mathbf{k}_d+d-\ell)/2} d(y_1, \dots, y_\ell). \end{aligned} \quad (3.15)$$

Let us first assume $\ell \geq 2$. We transform to hyperspherical coordinates (see for instance [23]) in the last integrand. This transformation is given by (see also [23, Definition 2])

$$\begin{aligned} y_1 &= r \prod_{k=1}^{\ell-1} \sin(\varphi_k), \\ y_2 &= r \cos(\varphi_1) \prod_{k=2}^{\ell-1} \sin(\varphi_k), \end{aligned}$$

$$\begin{aligned}
 y_3 &= r \cos(\varphi_2) \prod_{k=3}^{\ell-1} \sin(\varphi_k), \\
 &\vdots \\
 y_{\ell-1} &= r \cos(\varphi_{\ell-2}) \sin(\varphi_{\ell-1}) = r \cos(\varphi_{\ell-2}) \prod_{k=\ell-1}^{\ell-1} \sin(\varphi_k) \\
 y_\ell &= r \cos(\varphi_{\ell-1}) = r \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell-1} \sin(\varphi_k),
 \end{aligned}$$

where $r \in [0, 1]$, $\varphi_1 \in [-\pi, \pi]$ and $\varphi_k \in [0, \pi]$ for $k \in \{2, \dots, \ell-1\}$. One can show that $y_1^2 + \dots + y_\ell^2 = r^2$. The absolute value of the determinant of the Jacobian of that transformation is given by

$$r^{\ell-1} \prod_{k=2}^{\ell-1} \sin^{k-1}(\varphi_k); \quad \text{see [23] for a proof.}$$

Performing this transformation in the integral in (3.15) gives

$$\begin{aligned}
 &\int_{B^\ell} \varrho(y_1, \dots, y_\ell) y_1^{\mathbf{k}_1} \dots y_\ell^{\mathbf{k}_\ell} (1 - y_1^2 - \dots - y_\ell^2)^{(\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell)/2} \mathrm{d}(y_1, \dots, y_\ell) \\
 &= \int_0^1 \int_{-\pi}^\pi \int_{[0, \pi]^{\ell-2}} r^{\ell-1} \prod_{k=2}^{\ell-1} \sin^{k-1}(\varphi_k) \\
 &\quad \cdot \varrho\left(r \prod_{k=1}^{\ell-1} \sin(\varphi_k), r \cos(\varphi_1) \prod_{k=2}^{\ell-1} \sin(\varphi_k), \dots, r \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell-1} \sin(\varphi_k)\right) \\
 &\quad \left(r \prod_{k=1}^{\ell-1} \sin(\varphi_k)\right)^{\mathbf{k}_1} \left(r \cos(\varphi_1) \prod_{k=2}^{\ell-1} \sin(\varphi_k)\right)^{\mathbf{k}_2} \dots \left(r \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell-1} \sin(\varphi_k)\right)^{\mathbf{k}_\ell} \\
 &\quad (1 - r^2)^{(\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell)/2} \mathrm{d}(\varphi_2, \dots, \varphi_{\ell-1}) \mathrm{d}\varphi_1 \mathrm{d}r.
 \end{aligned}$$

Note that in the case $\ell = 2$, the innermost integral (over $[0, \pi]^{\ell-2}$) has to be omitted. We now additionally apply the substitution $r = \sin(\varphi_\ell)$ (with $\varphi_\ell \in [0, \frac{\pi}{2}]$) and get

$$\begin{aligned}
 &\int_{B^\ell} \varrho(y_1, \dots, y_\ell) y_1^{\mathbf{k}_1} \dots y_\ell^{\mathbf{k}_\ell} (1 - y_1^2 - \dots - y_\ell^2)^{(\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell)/2} \mathrm{d}(y_1, \dots, y_\ell) \\
 &= \int_0^{\pi/2} \int_{-\pi}^\pi \int_{[0, \pi]^{\ell-2}} \prod_{k=1}^{\ell} \sin^{k-1}(\varphi_k) \\
 &\quad \cdot \varrho\left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k)\right)
 \end{aligned}$$

$$\begin{aligned}
& \left(\prod_{k=1}^{\ell} \sin(\varphi_k) \right)^{\mathbf{k}_1} \left(\cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k) \right)^{\mathbf{k}_2} \dots \left(\cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k) \right)^{\mathbf{k}_{\ell}} \\
& \cos(\varphi_{\ell})^{\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell + 1} d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell} \\
& = \int_0^{\pi/2} \int_{-\pi}^{\pi} \int_{[0, \pi]^{\ell-2}} \left(\prod_{k=1}^{\ell-1} \cos(\varphi_k)^{\mathbf{k}_{k+1}} \sin(\varphi_k)^{k-1 + \mathbf{k}_1 + \dots + \mathbf{k}_k} \right) \\
& \quad \cdot \cos(\varphi_{\ell})^{\mathbf{k}_{\ell+1} + \dots + \mathbf{k}_d + d - \ell + 1} \\
& \quad \sin(\varphi_{\ell})^{\ell-1 + \mathbf{k}_1 + \dots + \mathbf{k}_{\ell}} \cdot \varrho \left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \sin(\varphi_{\ell}) \right) \\
& \quad d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell}.
\end{aligned}$$

Using (3.15) and recalling our choice of $\alpha_{h,\mathbf{k}}^k$, $\alpha_{h,\mathbf{k}}^{\ell}$, $\beta_{h,\mathbf{k}}^k$ and $\beta_{h,\mathbf{k}}^{\ell}$ in (3.11) as well as our choice of $\tilde{\zeta}_{h,\mathbf{k}}$ and $\zeta_{h,\mathbf{k}}$ in (3.12) and (3.13), we obtain

$$\begin{aligned}
& \int_{B^d} \varrho(y_1, \dots, y_{\ell}) y^{\mathbf{k}} dy \\
& = q_{\mathbf{k},d,\ell} \int_0^{\pi/2} \int_{-\pi}^{\pi} \int_{[0, \pi]^{\ell-2}} \prod_{k=1}^{\ell} \sum_{h=0}^{d+s} (\alpha_{h,\mathbf{k}}^k \cos(h\varphi_k) + \beta_{h,\mathbf{k}}^k \sin(h\varphi_k)) \\
& \quad \varrho \left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k) \right) d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell} \\
& = q_{\mathbf{k},d,\ell} \int_0^{\pi/2} \int_{-\pi}^{\pi} \int_{[0, \pi]^{\ell-2}} \sum_{h=1}^{2^{\ell(d+s+1)}} \tilde{\zeta}_{h,\mathbf{k}} f_h(\varphi_1, \dots, \varphi_{\ell}) \\
& \quad \varrho \left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k) \right) d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell} \\
& = \sum_{h=1}^{2^{\ell(d+s+1)}} \zeta_{h,\mathbf{k}} \int_0^{\pi/2} \int_{-\pi}^{\pi} \int_{[0, \pi]^{\ell-2}} f_h(\varphi_1, \dots, \varphi_{\ell}) \\
& \quad \varrho \left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k) \right) d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell} \\
& = \sum_{h=1}^{2^{\ell(d+s+1)}} \zeta_{h,\mathbf{k}} \cdot b_h(\varrho), \tag{3.16}
\end{aligned}$$

where

$$\begin{aligned}
& b_h(\varrho) \\
& := \int_0^{\pi/2} \int_{-\pi}^{\pi} \int_{[0, \pi]^{\ell-2}} f_h(\varphi_1, \dots, \varphi_{\ell})
\end{aligned}$$

$$\varrho \left(\prod_{k=1}^{\ell} \sin(\varphi_k), \cos(\varphi_1) \prod_{k=2}^{\ell} \sin(\varphi_k), \dots, \cos(\varphi_{\ell-1}) \prod_{k=\ell}^{\ell} \sin(\varphi_k) \right) \\ d(\varphi_2, \dots, \varphi_{\ell-1}) d\varphi_1 d\varphi_{\ell},$$

which only depends on h and ϱ .

In the case $\ell = 1$, the integral in (3.15) evaluates to

$$\begin{aligned} \int_{B^d} g(y_1) y^{\mathbf{k}} dy &= q_{\mathbf{k},d,1} \cdot \int_{-1}^1 \varrho(y_1) y_1^{\mathbf{k}_1} (1 - y_1^2)^{(\mathbf{k}_2 + \dots + \mathbf{k}_d + d - 1)/2} dy_1 \\ &\stackrel{y_1 = \sin(\varphi)}{=} q_{\mathbf{k},d,1} \cdot \int_{-\pi/2}^{\pi/2} \varrho(\sin(\varphi)) \sin^{\mathbf{k}_1}(\varphi) \cos(\varphi)^{\mathbf{k}_2 + \dots + \mathbf{k}_d + d} d\varphi \\ &\stackrel{(3.11)}{=} q_{\mathbf{k},d,1} \cdot \int_{-\pi/2}^{\pi/2} \varrho(\sin(\varphi)) \sum_{h=0}^{d+s} \left(\alpha_{h,\mathbf{k}}^1 \cos(h\varphi) + \beta_{h,\mathbf{k}}^1 \sin(h\varphi) \right) d\varphi \\ &\stackrel{(3.12)}{=} q_{\mathbf{k},d,1} \cdot \sum_{h=1}^{2(d+s+1)} \tilde{\zeta}_{h,\mathbf{k}} \cdot \int_{-\pi/2}^{\pi/2} \varrho(\sin(\varphi)) \cdot f_h(\varphi) d\varphi \\ &\stackrel{(3.13)}{=} \sum_{h=1}^{2(d+s+1)} \zeta_{h,\mathbf{k}} \cdot b_h(\varrho), \end{aligned} \tag{3.17}$$

with

$$b_h(\varrho) := \int_{-\pi/2}^{\pi/2} \varrho(\sin(\varphi)) \cdot f_h(\varphi) d\varphi,$$

which only depends on h and ϱ .

Finally, let $A \in \mathbb{R}^{\ell \times d}$ and $\sigma \in O(d)$ with $A\sigma = I_{\ell \times d}$. Putting everything together, we get

$$\begin{aligned} \langle \varrho_A, P \rangle &= \int_{B^d} \varrho(Ax) P(x) dx \\ &\stackrel{x = \sigma y}{=} \int_{B^d} \varrho(y_1, \dots, y_{\ell}) P(\sigma y) dy \\ &\stackrel{(3.9)}{=} \sum_{|\mathbf{k}| \leq s} P_{\mathbf{k}}(\sigma; P) \int_{B^d} \varrho(y_1, \dots, y_{\ell}) y^{\mathbf{k}} dy \\ &\stackrel{(3.16), (3.17)}{=} \sum_{|\mathbf{k}| \leq s} P_{\mathbf{k}}(\sigma; P) \left(\sum_{h=1}^{2^{\ell}(s+d+1)^{\ell}} \zeta_{h,\mathbf{k}} \cdot b_h(\varrho) \right) \\ &= \sum_{h=1}^{2^{\ell}(s+d+1)^{\ell}} \left(b_h(\varrho) \cdot \underbrace{\sum_{|\mathbf{k}| \leq s} \zeta_{h,\mathbf{k}} \cdot P_{\mathbf{k}}(\sigma; P)}_{=Q_h(\sigma; P); \text{ see (3.14)}} \right), \end{aligned}$$

which concludes the proof. $\square$

We are now ready to apply Lemma 3.3 and prove Lemma 3.1.

Proof (Proof of Lemma 3.1) Let Pr_s be the quasi-projection onto $\mathcal{P}_s(B^d)$ with range in $\mathcal{P}_{2s-1}(B^d)$ as defined in Section 2.3. Recall that, according to Equation (2.5), we have

$$\text{Pr}_s(f) = \sum_{i \in I_{2s-1}} a_{i,s} \cdot \langle f, P_i \rangle P_i \quad \text{for all } f \in L^1(B^d) \quad (3.18)$$

with real coefficients $a_{i,s} \in \mathbb{R}$ and polynomials $P_i \in \mathcal{P}_{2s-1}(B^d)$ ($i \in I_{2s-1}$) that form an orthonormal basis of $\mathcal{P}_{2s-1}(B^d)$.

We now claim that there exist $\{\gamma_i^j\}_{(i,j) \in I_{2s-1} \times I_{2s-1}} \subset \mathcal{P}_{2s-1}(\mathbb{R}^d)$ satisfying

$$a_{i,s} \cdot P_i(\xi + t) = \sum_{j \in I_{2s-1}} \gamma_i^j(t) P_j(\xi) \quad (3.19)$$

for all $i \in I_{2s-1}$ and $\xi \in B^d$, $t \in \mathbb{R}^d$. Indeed, for fixed $t \in \mathbb{R}^d$ and $i \in I_{2s-1}$, the function defined by $g_t(\xi) := a_{i,s} \cdot P_i(\xi + t)$ is a polynomial of degree at most $2s - 1$ in the variable ξ . Since $\{P_j\}_{j \in I_{2s-1}}$ is an orthonormal basis for $\mathcal{P}_{2s-1}(B^d)$, we thus get

$$g_t(\xi) = \sum_{j \in I_{2s-1}} \underbrace{\langle g_t, P_j \rangle}_{=: \gamma_i^j(t)} P_j(\xi) \quad \text{for all } \xi \in B^d.$$

Note that

$$\gamma_i^j(t) = a_{i,s} \cdot \int_{B^d} P_i(\xi + t) P_j(\xi) \, d\xi$$

are polynomials of degree at most $2s - 1$ in t .

Let $\xi_1, \dots, \xi_m \in B^d$ be taken as in the beginning of Section 3.2. Set $K := n2^\ell(2s + d)^\ell$ and let

$$\psi = (\psi_1, \psi_2) : \{1, \dots, K\} \rightarrow \{1, \dots, n\} \times \{1, \dots, 2^\ell(2s + d)^\ell\} \quad (3.20)$$

be a bijection. Further, for $i \in I_{2s-1}$ and $h \in \{1, \dots, 2^\ell(2s + d)^\ell\}$, let $Q_h(\bullet; P_i) \in \mathcal{P}_{2s-1}(\mathbb{R}^{d^2})$ be given according to Lemma 3.5. For $\alpha \in \{1, \dots, m\}$ and $\beta \in \{1, \dots, K\}$, we then set

$$\pi_{\alpha,\beta}(\sigma_1, \dots, \sigma_n, t) := \sum_{i \in I_{2s-1}} \sum_{j \in I_{2s-1}} Q_{\psi_2(\beta)}(\sigma_{\psi_1(\beta)}; P_i) \gamma_i^j(t) P_j(\xi_\alpha), \quad (3.21)$$

for $\sigma_1, \dots, \sigma_n \in \mathbb{R}^{d \times d} \cong \mathbb{R}^{d^2}$ and $t \in \mathbb{R}^d$. By letting $N := nd^2 + d$, we note that $\pi_{\alpha,\beta} \in \mathcal{P}_{4s-2}(\mathbb{R}^N)$. We claim that

$$\Pi_{m,s,n,\ell,d} \subseteq \left\{ (\pi_1(b, z), \dots, \pi_m(b, z)) : (b, z) \in \mathbb{R}^K \times \mathbb{R}^N \right\} \quad (3.22)$$

with $\pi_\alpha(b, z) := \sum_{\beta=1}^K b_\beta \cdot \pi_{\alpha,\beta}(z)$.

To this end, recall from (3.6) that

$$\Pi_{m,s,n,\ell,d} = \left\{ (P(\xi_1 + t), \dots, P(\xi_m + t)) : P \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*), t \in \mathbb{R}^d \right\}.$$

Let $P \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)$. From (3.18), we have

$$P = \sum_{i \in I_{2s-1}} a_{i,s} \cdot \langle R, P_i \rangle P_i,$$

where $R \in \mathcal{R}_{n,d,\ell}^*$. According to Proposition 2.4, we can write $R = \sum_{k=1}^n \varrho_k(A_k \cdot \bullet)$ with $\varrho_k \in \mathcal{M}_b(\mathbb{R}^\ell)$ and $A_k A_k^T = I_{\ell \times \ell}$ for every $k \in \{1, \dots, n\}$, which implies for each k the existence of a matrix $\sigma_k \in O(d)$ with $A_k \sigma_k = I_{\ell \times d}$. Thus, Lemma 3.5 yields that

$$P = \sum_{i \in I_{2s-1}} \sum_{k=1}^n a_{i,s} \cdot \langle \varrho_k(A_k \bullet), P_i \rangle P_i = \sum_{i \in I_{2s-1}} \sum_{k=1}^n \sum_{h=1}^{2^\ell(2s+d)^\ell} a_{i,s} \cdot b_h(\varrho_k) Q_h(\sigma_k; P_i) P_i$$

with coefficients $b_h(\varrho_k) \in \mathbb{R}$ that only depend on ϱ_k and h .

Recalling the choice of $\gamma_i^j(t)$ in (3.19), the choice of ψ in (3.20), and the definition of $\pi_{\alpha,\beta}$ in (3.21) then yields for every $\alpha \in \{1, \dots, m\}$ and $t \in \mathbb{R}^d$ that

$$\begin{aligned} P(\xi_\alpha + t) &= \sum_{i \in I_{2s-1}} \sum_{k=1}^n \sum_{h=1}^{2^\ell(2s+d)^\ell} a_{i,s} \cdot b_h(\varrho_k) Q_h(\sigma_k; P_i) P_i(\xi_\alpha + t) \\ &= \sum_{i \in I_{2s-1}} \sum_{k=1}^n \sum_{h=1}^{2^\ell(2s+d)^\ell} \sum_{j \in I_{2s-1}} b_h(\varrho_k) Q_h(\sigma_k; P_i) \gamma_i^j(t) P_j(\xi_\alpha) \\ &= \sum_{k=1}^n \sum_{h=1}^{2^\ell(2s+d)^\ell} b_h(\varrho_k) \sum_{i \in I_{2s-1}} \sum_{j \in I_{2s-1}} Q_h(\sigma_k; P_i) \gamma_i^j(t) P_j(\xi_\alpha) \\ &= \sum_{\beta=1}^K b_{\psi_2(\beta)}(\varrho_{\psi_1(\beta)}) \pi_{\alpha,\beta}(\sigma_1, \dots, \sigma_n, t). \end{aligned}$$

This proves (3.22).

To apply Lemma 3.3, we need $N + K \leq m/2$ to be satisfied. Once we show this, by substituting the values of N and K and applying Lemma 3.3 with $4s - 2$ in place of s , we would obtain

$$\begin{aligned} |\text{sgn}(\Pi_{m,s,n,\ell,d})| &\leq (16s - 8)^N (N + K + 1)^{N+2} \left(\frac{2em}{N + K} \right)^{N+K} \\ &\leq (16s)^N (N + K + 1)^{N+2} \left(\frac{2em}{N + K} \right)^{N+K} \leq 2^{m/4} \end{aligned}$$

if $N \lg(16s) + (N + 2) \lg(N + K + 1) + (N + K) \lg(2em/(N + K)) \leq m/4$. Here, we write $\lg$ for the logarithm with respect to base 2. Therefore, if we can prove that

$$\begin{aligned} N + K &\leq m/2 \quad \text{and} \\ N \lg(16s) + (N + 2) \lg(N + K + 1) + (N + K) \lg(2em/(N + K)) &\leq m/4 \end{aligned} \quad (3.23)$$

for sufficiently large $s \in \mathbb{N}$ (depending on d, ℓ, c_1), then we are done.

Note that $c_0 > 2 \cdot 4^\ell / c_1$, which directly implies $c_1/2 > 4^\ell / c_0$. Since it suffices to compare leading coefficients (with respect to s), we hence get from (3.8) that

$$\begin{aligned} m/2 &\geq c_1 s^d / 2 > (s^{d-\ell} / c_0) d^2 + d + (s^{d-\ell} / c_0) 2^\ell (2s + d)^\ell \\ &\geq n d^2 + d + n 2^\ell (2s + d)^\ell = N + K \end{aligned} \quad (3.24)$$

for sufficiently large $s \in \mathbb{N}$ (depending on d, ℓ, c_1).

To show that $N \lg(16s) + (N + 2) \lg(N + K + 1) + (N + K) \lg(2em/(N + K)) \leq m/4$ for sufficiently large s (depending on d, ℓ, c_1), we start by showing

$$K \lg(2em/(N + K)) \leq \tilde{c} \cdot m \quad \text{with } \tilde{c} < \frac{1}{4} \quad (3.25)$$

for sufficiently large s (depending on d, ℓ, c_1). Indeed, by (3.8), we get

$$K \lg\left(\frac{2em}{N + K}\right) \leq (s^{d-\ell} / c_0) 2^\ell (2s + d)^\ell \lg\left(\frac{2e(2c_1 s^d)}{(s^{d-\ell} / (2c_0)) d^2 + d + (s^{d-\ell} / (2c_0)) 2^\ell (2s + d)^\ell}\right).$$

We first study the argument of the logarithm. By comparing leading coefficients, we note

$$\frac{2e(2c_1 s^d)}{(s^{d-\ell} / (2c_0)) d^2 + d + (s^{d-\ell} / (2c_0)) 2^\ell (2s + d)^\ell} \rightarrow \frac{4ec_1}{4^\ell / (2c_0)} = \frac{2ec_1 c_0}{4^{\ell-1}} \quad (s \rightarrow \infty),$$

which shows that

$$\frac{2e(2c_1 s^d)}{(s^{d-\ell} / (2c_0)) d^2 + d + (s^{d-\ell} / (2c_0)) 2^\ell (2s + d)^\ell} \leq \frac{ec_1 c_0}{4^{\ell-2}}$$

for sufficiently large s (depending on d, ℓ, c_1). Recalling that $m \geq c_1 s^d$ and by comparing leading coefficients, it suffices to ensure

$$c_1/4 > \frac{4^\ell}{c_0} \lg(ec_1 c_0 / 4^{\ell-2})$$

in order for (3.25) to be satisfied. This is equivalent to

$$c_0 > \frac{4^{\ell+1}}{c_1} \lg(ec_1 c_0 / 4^{\ell-2}).$$

Plugging in $c_0 = 4^{\ell+3}/c_1$, we obtain that this is equivalent to

$$4^2 > \lg(e \cdot 4^5),$$

which is satisfied since $\lg(e \cdot 4^5) = \lg(e) + 10 \leq 2 + 10 = 12 < 16 = 4^2$. Hence, (3.25) is satisfied for s sufficiently large (depending on d, ℓ, c_1).

Further, we note by (3.8) that

$$N \lg(16s) \leq \left(\frac{d^2}{c_0} s^{d-\ell} + d \right) \lg(16s) \ll s^d, \quad (3.26)$$

i.e., this term is of lower order than s^d . Additionally, we get by (3.8) that

$$(N+2) \lg(N+K+1) \leq \left(\frac{d^2}{c_0} s^{d-\ell} + d + 2 \right) \lg \left(\frac{d^2}{c_0} s^{d-\ell} + d + \frac{2^\ell}{c_0} s^{d-\ell} (2s+d)^\ell + 1 \right) \ll s^d. \quad (3.27)$$

Lastly, we observe again by (3.8) that

$$N \lg(2em/(N+K)) \leq N \lg(2em) \leq \left(\frac{d^2}{c_0} + d \right) \cdot s^{d-\ell} \cdot \lg(4ec_1 s^d) \ll s^d. \quad (3.28)$$

Because of (3.8), Equations (3.24), (3.25), (3.26), (3.27) and (3.28) together imply (3.23), which proves the lemma. $\square$

3.3 Concluding the proof of the lower bound

In this subsection, we prove Theorem 1.1. To this end, we start by constructing a subset of $\mathcal{B}(W_d^{r,\infty})$ consisting of sums of smooth bump functions, which we will later show to realize the desired lower bound.

Let $d, r \in \mathbb{N}$ be fixed and set

$$\Omega^d := \left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}} \right]^d \subseteq B^d. \quad (3.29)$$

Let $\tilde{\omega} : \mathbb{R}^d \rightarrow [0, 1]$ be a smooth function with

$$\tilde{\omega}(x) = \begin{cases} 1, & x \in \Omega^d/2, \\ 0, & x \in \mathbb{R}^d \setminus \Omega^d. \end{cases}$$

We scale $\tilde{\omega}$ such that its restriction to B^d belongs to $\mathcal{B}(W_d^{r,\infty})$ and call this normalization ω . Let ϑ and $\xi_1, \dots, \xi_m$ be chosen as at the beginning of Section 3.2. Recall that

$$E^m = \{\varepsilon = (\varepsilon_1, \dots, \varepsilon_m) : \varepsilon_i = \pm 1, i = 1, \dots, m\}.$$

Let

$$f_\varepsilon : B^d \rightarrow \mathbb{R}, \quad f_\varepsilon(x) := (2\vartheta)^{-r} \sum_{i=1}^m \varepsilon_i \cdot \omega(2\vartheta(x - \xi_i))$$

for $\varepsilon \in E^m$ and define

$$\mathcal{F}^{m,r,d} := \{f_\varepsilon : \varepsilon \in E^m\}. \quad (3.30)$$

Proposition 3.6 *For any choice of $m, r, d \in \mathbb{N}$ it holds that $\mathcal{F}^{m,r,d} \subseteq \mathcal{B}(W_d^{r,\infty})$.*

Proof Since ω is smooth, we conclude $\mathcal{F}^{m,r,d} \subseteq C^\infty(B^d)$. Let $\varepsilon \in E^m$. We want to show that $|\partial^{\mathbf{k}} f_\varepsilon(x)| \leq 1$ for all $x \in B^d$ and every multiindex $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq r$. For $i \in \{1, \dots, m\}$ we let $\Omega_i := \xi_i + \Omega^d/(2\vartheta)$ denote the closed cube with center ξ_i and side length $1/(\vartheta\sqrt{d})$. Since $\partial\Omega_i$ is a null set for every i , and f_ε and all of its derivatives are continuous, we may assume $x \in (B^d)^\circ \setminus \bigcup_{i=1}^m \partial\Omega_i$. Moreover, since $\|\xi_i - \xi_j\|_{\ell^\infty} \geq \frac{1}{\sqrt{d}\vartheta}$ for every $i, j \in \{1, \dots, m\}$ with $i \neq j$, we note that there exists at most one $i \in \{1, \dots, m\}$ with $\|x - \xi_i\|_{\ell^\infty} < \frac{1}{2\sqrt{d}\vartheta}$, which is equivalent to $x \in \Omega_i^\circ$.

In the case such i exists, we get $2\vartheta(x - \xi_j) \notin \Omega^d$ for every $j \neq i$, whence, by definition of ω , we have

$$f_\varepsilon(x) = (2\vartheta)^{-r} \cdot \varepsilon_i \cdot \omega(2\vartheta(x - \xi_i)).$$

Since $x \in \Omega_i^\circ$, this identity even holds in an open neighborhood of x . Therefore, for any multiindex $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| \leq r$, we have

$$\partial^{\mathbf{k}} f_\varepsilon(x) = (2\vartheta)^{|\mathbf{k}|-r} \cdot \varepsilon_i \cdot \left(\partial^{\mathbf{k}} \omega\right)(2\vartheta(x - \xi_i)).$$

Using that $2\vartheta \geq 1$ and $\omega \in \mathcal{B}(W_d^{r,\infty})$, we get

$$\left|\partial^{\mathbf{k}} f_\varepsilon(x)\right| = (2\vartheta)^{|\mathbf{k}|-r} \cdot \left|\left(\partial^{\mathbf{k}} \omega\right)(2\vartheta(x - \xi_i))\right| \leq 1.$$

In the case when $x \notin \bigcup_{i=1}^m \Omega_i$, we have $f_\varepsilon \equiv 0$ in an open neighborhood of x . Combining the two cases concludes the proof. $\square$

We can now establish a lower bound for the approximation of functions from $\mathcal{F}^{m,r,d}$ using functions from $\text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)$ in the $\|\cdot\|_{L^1(B^d)}$ -norm. This is in fact a generalization of [18, Lemma 7] to the case of multivariate ridge functions.

Proposition 3.7 *Let $d, \ell, r \in \mathbb{N}$ with $\ell < d$. Then there exists a constant $c = c(d, \ell, r) > 0$ with the following property: For any $c_1 > 0$ and $c_0 := 4^{\ell+3}/c_1$, there exists a constant $C = C(d, \ell, c_1)$ such that for any choice of $m, s, n \in \mathbb{N}$ with*

$$c_1 s^d \leq m \leq 2c_1 s^d, \quad c_0 n \leq s^{d-\ell} \leq 2c_0 n, \quad \text{and} \quad s \geq C,$$

there exists a function $f_{\varepsilon^*} \in \mathcal{F}^{m,r,d}$ satisfying

$$\inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|f_{\varepsilon^*} - P\|_{L^1(B^d)} \geq c \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)}.$$

Here, $\mathcal{F}^{m,r,d}$ is as defined in (3.30).

Proof For a given $c_1 > 0$, we let $C = C(d, \ell, c_1) > 0$ be chosen according to Lemma 3.1. Let $f_{\varepsilon} \in \mathcal{F}^{m,r,d}$ and $P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)$ be arbitrary. We recall that then

$$f_{\varepsilon}(x) = (2\vartheta)^{-r} \sum_{i=1}^m \varepsilon_i \cdot \omega(2\vartheta(x - \xi_i))$$

for given grid points $\xi_1, \dots, \xi_m \in \Omega^d$ with Ω^d as in (3.29). Moreover, recall from the start of Section 3.2 that $\vartheta \in \mathbb{N}$ is chosen such that $\frac{m^{1/d}}{2} \leq \vartheta \leq m^{1/d}$. With the notation $\Omega_i := \xi_i + \Omega^d/(2\vartheta)$, we get $B^d \supseteq \bigcup_{i=1}^m \Omega_i^\circ$, where the union is disjoint. This yields

$$\|f_{\varepsilon} - P\|_{L^1(B^d)} \geq \sum_{i=1}^m \int_{\Omega_i^\circ} |f_{\varepsilon}(x) - P(x)| \, dx = \sum_{i=1}^m \int_{\Omega_i} |f_{\varepsilon}(x) - P(x)| \, dx,$$

where in the last equality we used that $\partial\Omega_i$ is a measure zero set. Applying the change of variables $x := t + \xi_i$, we conclude

$$\begin{aligned} \sum_{i=1}^m \int_{\Omega_i} |f_{\varepsilon}(x) - P(x)| \, dx &= \int_{\Omega^d/(2\vartheta)} \sum_{i=1}^m |f_{\varepsilon}(t + \xi_i) - P(t + \xi_i)| \, dt \\ &\geq \int_{\Omega^d/(4\vartheta)} \sum_{i=1}^m |f_{\varepsilon}(t + \xi_i) - P(t + \xi_i)| \, dt. \end{aligned}$$

For any $t \in \Omega^d/(4\vartheta)$ and fixed $i \in \{1, \dots, m\}$, since $\xi_i + t \in \Omega_i^\circ$ and $2\vartheta t \in \Omega^d/2$, it follows from the choice of ω that

$$f_{\varepsilon}(\xi_i + t) = (2\vartheta)^{-r} \cdot \varepsilon_i \cdot \omega(2\vartheta t) = (2\vartheta)^{-r} \cdot \varepsilon_i \cdot c_2$$

with an absolute constant $c_2 = c_2(d, r) > 0$. Hence, for any $t \in \Omega^d/(4\vartheta)$ we get

$$\begin{aligned} \sum_{i=1}^m |f_{\varepsilon}(t + \xi_i) - P(t + \xi_i)| &\geq \inf_{\substack{P' \in \Pr_s(\mathcal{R}_{n,d,\ell}^*) \\ \tau \in \mathbb{R}^d}} \sum_{i=1}^m |(2\vartheta)^{-r} \cdot \varepsilon_i \cdot c_2 - P'(\xi_i + \tau)| \\ &= c_2 \cdot (2\vartheta)^{-r} \cdot \inf_{\substack{P' \in \Pr_s(\mathcal{R}_{n,d,\ell}^*) \\ \tau \in \mathbb{R}^d}} \sum_{i=1}^m |\varepsilon_i - P'(\xi_i + \tau)|. \end{aligned} \tag{3.31}$$

At the last step, we used that the set $\mathcal{R}_{n,d,\ell}^*$ is invariant under multiplication with nonzero factors and that the map Pr_s is linear, whence the scaling invariance translates to $\text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)$. Using the bound (3.31), we get

$$\begin{aligned} \|f_\varepsilon - P\|_{L^1(B^d)} &\geq \lambda^d (\Omega^d/(4\vartheta)) \cdot c_2 \cdot (2\vartheta)^{-r} \cdot \inf_{\substack{P' \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*) \\ \tau \in \mathbb{R}^d}} \sum_{i=1}^m |\varepsilon_i - P'(\xi_i + \tau)| \\ &= \left(\frac{1}{2\vartheta \cdot \sqrt{d}} \right)^d \cdot c_2 \cdot (2\vartheta)^{-r} \cdot \inf_{\substack{P' \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*) \\ \tau \in \mathbb{R}^d}} \sum_{i=1}^m |\varepsilon_i - P'(\xi_i + \tau)| \\ &\stackrel{\vartheta \leq m^{1/d}}{\geq} c_2 \cdot c_3 \cdot \frac{1}{m \cdot (2\vartheta)^r} \cdot \inf_{x \in \Pi_{m,s,n,\ell,d}} \|x - \varepsilon\|_{\ell^1}, \end{aligned}$$

where we set $c_3 = c_3(d) := \left(\frac{1}{2\sqrt{d}} \right)^d$ and recall that $\Pi_{m,s,n,\ell,d}$ was defined in (3.6). From Lemmas 3.2 and 3.1, we infer, under the constraint that $s \geq C$, the existence of $\varepsilon^* \in E^m$ with

$$\inf_{x \in \Pi_{m,s,n,\ell,d}} \|x - \varepsilon^*\|_{\ell^1} \geq am$$

with an absolute constant $a > 0$. Hence, we get

$$\|f_{\varepsilon^*} - P\|_{L^1(B^d)} \geq a \cdot c_2 \cdot c_3 \cdot (2\vartheta)^{-r} \stackrel{\vartheta \leq m^{1/d}}{\geq} a \cdot c_2 \cdot c_3 \cdot 2^{-r} \cdot m^{-r/d}.$$

By assumption, we have $m \leq 2c_1 \cdot s^d$, which yields

$$\|f_{\varepsilon^*} - P\|_{L^1(B^d)} \geq a \cdot c_2 \cdot c_3 \cdot 2^{-r} \cdot (2c_1)^{-r/d} \cdot s^{-r}.$$

Furthermore, since $s \leq (2c_0)^{1/(d-\ell)} \cdot n^{1/(d-\ell)}$, we get

$$\|f_{\varepsilon^*} - P\|_{L^1(B^d)} \geq a \cdot c_2 \cdot c_3 \cdot 2^{-r} \cdot (2c_1)^{-r/d} \cdot (2c_0)^{-r/(d-\ell)} \cdot n^{-r/(d-\ell)}.$$

Since $c_0 = \frac{4^{\ell+3}}{c_1}$, this translates to

$$\|f_{\varepsilon^*} - P\|_{L^1(B^d)} \geq a \cdot c_2 \cdot c_3 \cdot 2^{-r} \cdot 2^{-r/d} \cdot 2^{-r(2\ell+7)/(d-\ell)} \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)}.$$

Therefore, the claim follows letting $c = c(d, \ell, r) := a \cdot c_2 \cdot c_3 \cdot 2^{-r} \cdot 2^{-r/d} \cdot 2^{-r(2\ell+7)/(d-\ell)}$. $\square$

The following lemma is a central ingredient in order to prove the final approximation bound and follows immediately from the fact that the quasi-projection Pr_s is bounded with respect to the L^1 -norm, with the bound being independent of s .

Lemma 3.8 (corrected version of [18, Lemma 5]) *Let $d \in \mathbb{N}$. Then there exists a positive constant $c = c(d) > 0$ with the following property: For any choice of $\ell, s, n \in \mathbb{N}$ with $\ell < d$ and $P \in \mathcal{P}_s(B^d)$ it holds*

$$\inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|P - R\|_{L^1(B^d)} \geq c \cdot \inf_{P' \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)} \|P - P'\|_{L^1(B^d)}.$$

Proof Let $P \in \mathcal{P}_s(B^d)$ and $R \in \mathcal{R}_{n,d,\ell}^*$ be arbitrary. Since $P - R$ is bounded on B^d , we clearly have $P - R \in L^1(B^d)$. Then we have

$$\|P - R\|_{L^1(B^d)} \geq c \cdot \|\text{Pr}_s(P - R)\|_{L^1(B^d)} = c \cdot \|P - \text{Pr}_s(R)\|_{L^1(B^d)}$$

by the properties of Pr_s from Proposition 2.7. By taking the infimum on the right-hand side, we get

$$\|P - R\|_{L^1(B^d)} \geq c \cdot \inf_{P' \in \text{Pr}_s(\mathcal{R}_{n,d,\ell}^*)} \|P - P'\|_{L^1(B^d)}.$$

Now we can take the infimum over $R \in \mathcal{R}_{n,d,\ell}^*$ on the left-hand side and obtain the claim. $\square$

We now have everything we need to complete our proof of the lower bound.

Theorem 3.9 *Let $d, \ell, r \in \mathbb{N}$ with $\ell < d$. Then there exists a constant $c = c(d, \ell, r) > 0$ with the property that for any $n \in \mathbb{N}$ we have*

$$\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|f - R\|_{L^1(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

Proof Let $c_1 = c_1(d, \ell, r) > 0$ be arbitrary (to be determined later) and set $c_0 := 4^{\ell+3}/c_1$. Let the two constants $\tilde{c} = \tilde{c}(d, \ell, r) > 0$ and $\tilde{C} = \tilde{C}(d, \ell, c_1) > 0$ be given by Proposition 3.7.

For the moment, we assume that $n \geq N$ with a large number $N = N(d, \ell, c_1) \in \mathbb{N}$ such that the conditions

$$n^{1/(d-\ell)} \left((2c_0)^{1/(d-\ell)} - c_0^{1/(d-\ell)} \right) \geq 1, \quad c_0^{1/(d-\ell)} \cdot n^{1/(d-\ell)} \geq \tilde{C}$$

$$\text{and } c_0^{1/(d-\ell)} \cdot n^{1/(d-\ell)} \geq c_1^{-1/d}$$

are satisfied. The first condition ensures that we can pick a natural number $s \in \mathbb{N}$ that satisfies $c_0 n \leq s^{d-\ell} \leq 2c_0 n$. From the second condition, we infer that this implies $s \geq \tilde{C}$. And lastly, the third condition guarantees that

$$s \geq c_0^{1/(d-\ell)} \cdot n^{1/(d-\ell)} \geq c_1^{-1/d} \quad \text{and thus} \quad (2c_1)s^d - c_1s^d \geq 1,$$

whence we can pick a natural number $m \in \mathbb{N}$ with $c_1s^d \leq m \leq (2c_1)s^d$. To summarize, we have

$$c_1s^d \leq m \leq 2c_1s^d, \quad c_0n \leq s^{d-\ell} \leq 2c_0n \quad \text{and} \quad s \geq \tilde{C}.$$

According to Proposition 3.7, we can thus pick a function $f_\varepsilon \in \mathcal{F}^{m,r,d}$ that satisfies

$$\inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|P - f_\varepsilon\|_{L^1(B^d)} \geq \tilde{C} \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)}. \quad (3.32)$$

Moreover, using Propositions 2.5 and 3.6, we pick a polynomial $g_\varepsilon \in \mathcal{P}_s(B^d)$ with

$$\|g_\varepsilon - f_\varepsilon\|_{L^1(B^d)} \leq C_2 \cdot s^{-r},$$

with an absolute constant $C_2 = C_2(d, r) > 0$. We thus get

$$\begin{aligned} \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|R - f_\varepsilon\|_{L^1(B^d)} &\geq \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|g_\varepsilon - R\|_{L^1(B^d)} - \|g_\varepsilon - f_\varepsilon\|_{L^1(B^d)} \\ &\geq c' \cdot \inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|g_\varepsilon - P\|_{L^1(B^d)} - C_2s^{-r}, \end{aligned}$$

where $c' = c'(d) \in (0, 1]$ is provided by Lemma 3.8 (we can assume that $c' \leq 1$ by shrinking it, if necessary). This can be further bounded via

$$\begin{aligned} &c' \cdot \inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|g_\varepsilon - P\|_{L^1(B^d)} - C_2s^{-r} \\ &\geq c' \cdot \left(\inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|f_\varepsilon - P\|_{L^1(B^d)} - \|f_\varepsilon - g_\varepsilon\|_{L^1(B^d)} \right) - C_2s^{-r} \\ &\geq c' \cdot \inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|f_\varepsilon - P\|_{L^1(B^d)} - c'C_2s^{-r} - C_2s^{-r} \\ &\stackrel{c' \leq 1}{\geq} c' \cdot \inf_{P \in \Pr_s(\mathcal{R}_{n,d,\ell}^*)} \|f_\varepsilon - P\|_{L^1(B^d)} - 2C_2s^{-r} \\ &\geq c'\tilde{C} \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)} - 2C_2s^{-r}, \end{aligned}$$

where we used (3.32). Since $s \geq c_0^{1/(d-\ell)} \cdot n^{1/(d-\ell)}$ and $c_0 = 4^{\ell+3}/c_1$, we get

$$\begin{aligned} \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|R - f_\varepsilon\|_{L^1(B^d)} &\geq c' \tilde{c} \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)} - 2C_2 s^{-r} \\ &\geq c' \tilde{c} \cdot c_1^{(r/(d-\ell))-(r/d)} \cdot n^{-r/(d-\ell)} - 2C_2 c_0^{-r/(d-\ell)} n^{-r/(d-\ell)} \\ &= \left(c' \tilde{c} \cdot c_1^{(r/(d-\ell))-(r/d)} - 2C_2 \cdot 4^{-r(\ell+3)/(d-\ell)} \cdot c_1^{r/(d-\ell)} \right) \cdot n^{-r/(d-\ell)}. \end{aligned}$$

We set $C_3 = C_3(d, \ell, r) := 2C_2 \cdot 4^{-r(\ell+3)/(d-\ell)}$. We then see

$$c' \tilde{c} \cdot c_1^{(r/(d-\ell))-(r/d)} - C_3 \cdot c_1^{r/(d-\ell)} > 0 \Leftrightarrow c' \tilde{c} \cdot c_1^{-r/d} > C_3 \Leftrightarrow c_1 < \left(\frac{c' \tilde{c}}{C_3} \right)^{d/r}.$$

Therefore, by picking $c_1 = c_1(d, \ell, r) > 0$ small enough, we can achieve the desideratum by letting

$$c = c(d, \ell, r) := c' \tilde{c} \cdot c_1^{(r/(d-\ell))-(r/d)} - C_3 \cdot c_1^{r/(d-\ell)}$$

and noting that $f_\varepsilon \in \mathcal{B}(W_d^{r,\infty})$ according to Proposition 3.6.

Recall that we assumed that $n \geq N$. In order to obtain the desired claim for any $n \in \mathbb{N}$, we first note that $\mathcal{R}_{n,d,\ell}^* \subseteq \mathcal{R}_{N,d,\ell}^*$ for any $n \leq N$. The above therefore tells us

$$\begin{aligned} &\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}^*} \|R - f\|_{L^1(B^d)} \\ &\geq \sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{N,d,\ell}^*} \|R - f\|_{L^1(B^d)} > 0 \quad \text{for any } n \leq N. \end{aligned}$$

Hence, we obtain the claim by possibly shrinking the size of the constant c . $\square$

It is now straightforward to generalize the claim to the case of approximation of Sobolev functions from $\mathcal{B}(W_d^{r,p})$ with respect to $\|\cdot\|_{L^q}$ for arbitrary $p, q \in [1, \infty]$.

Corollary 3.10 *Let $d, \ell, r \in \mathbb{N}$ with $\ell < d$ and $p, q \in [1, \infty]$. Then there exists a positive constant $c = c(d, \ell, p, q, r) > 0$ with the property that for any $n \in \mathbb{N}$ we have*

$$\sup_{f \in \mathcal{B}(W_d^{r,p})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^q(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

Here, $\mathcal{R}_{n,d,\ell}$ is as defined in Equation (2.1).

Proof Let $c' = c'(d, \ell, r) > 0$ be chosen according to Theorem 3.9. A direct application of Proposition 2.3 then yields

$$\sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} \geq c' \cdot n^{-r/(d-\ell)}.$$

Moreover, according to Hölder's inequality, we may pick a constant $C_1 = C_1(d, p) > 0$ which satisfies

$$\|f\|_{L^p(B^d)} \leq C_1 \cdot \|f\|_{L^\infty(B^d)} \quad \text{for all measurable } f : B^d \rightarrow \mathbb{R}.$$

This implies $c_2 \cdot \mathcal{B}(W_d^{r,\infty}) \subseteq \mathcal{B}(W_d^{r,p})$ for a constant $c_2 = c_2(d, r, p) > 0$, which yields

$$\begin{aligned} \sup_{f \in \mathcal{B}(W_d^{r,p})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} &\geq \sup_{f \in c_2 \cdot \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)} \\ &= \sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|c_2 \cdot f - R\|_{L^1(B^d)} \\ &= c_2 \cdot \sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|f - R\|_{L^1(B^d)}, \end{aligned}$$

where the last equality uses the fact that the set $\mathcal{R}_{n,d,\ell}$ is invariant with respect to scaling. Moreover, there exists another constant $C_3 = C_3(d, q) > 0$ with

$$\|f\|_{L^1(B^d)} \leq C_3 \cdot \|f\|_{L^q(B^d)} \quad \text{for all measurable } f : B^d \rightarrow \mathbb{R},$$

again according to Hölder's inequality. Therefore,

$$\begin{aligned} \sup_{f \in \mathcal{B}(W_d^{r,p})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|R - f\|_{L^q(B^d)} &\geq C_3^{-1} \cdot \sup_{f \in \mathcal{B}(W_d^{r,p})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|R - f\|_{L^1(B^d)} \\ &\geq c_2 \cdot C_3^{-1} \cdot \sup_{f \in \mathcal{B}(W_d^{r,\infty})} \inf_{R \in \mathcal{R}_{n,d,\ell}} \|R - f\|_{L^1(B^d)} \\ &\geq c_2 \cdot C_3^{-1} \cdot c' \cdot n^{-r/(d-\ell)}. \end{aligned}$$

This yields the claim by picking $c = c(d, \ell, p, q, r) := c_2 \cdot C_3^{-1} \cdot c'$. $\square$

4 Proof of the upper bound

In this section, we prove the upper bound $\mathcal{O}(n^{-r/(d-\ell)})$ for the rate of approximating Sobolev functions using sums of n multivariate ridge functions, which is the optimal rate according to Corollary 3.10. Moreover, we show that this optimal rate can even be attained using sums of *polynomial* ridge functions with *fixed* matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$. To this end, for fixed $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$, we introduce the notation

$$\mathcal{R}_{n,d,\ell}^{\text{poly}}(A_1, \dots, A_n) := \left\{ B^d \ni x \mapsto \sum_{j=1}^n P_j(A_j x) : P_j \in \mathcal{P}(\mathbb{R}^\ell) \right\}.$$

Furthermore, we denote by $\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1})$ the space of *homogeneous* polynomials in $d - \ell + 1$ variables of degree s , i.e.,

$$\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1}) := \left\{ \mathbb{R}^{d-\ell+1} \ni x \mapsto \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{d-\ell+1} \\ |\mathbf{k}|=s}} a_{\mathbf{k}} x^{\mathbf{k}} : a_{\mathbf{k}} \in \mathbb{R} \right\}.$$

We start by showing that each (not necessarily homogeneous) polynomial of degree at most s can be written as the sum of n ℓ -variate polynomial ridge functions, if we choose the number n to exceed the dimension of the space $\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1})$. This result is mainly based on results from [26, Section 5].

Proposition 4.1 *Let $d, n, \ell, s \in \mathbb{N}$ with $\ell < d$ and*

$$\dim(\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1})) \leq n.$$

Then there exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ with

$$\mathcal{P}_s(B^d) \subseteq \mathcal{R}_{n,d,\ell}^{\text{poly}}(A_1, \dots, A_n).$$

Proof Using [26, Proposition 5.9] and $\dim(\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1})) \leq n$, we can pick $a_1, \dots, a_n \in \mathbb{R}^{d-\ell+1}$ satisfying

$$\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1}) = \text{span} \left\{ \mathbb{R}^{d-\ell+1} \ni x \mapsto (a_i^T x)^s : 1 \leq i \leq n \right\}.$$

From [26, Corollary 5.12], we infer that

$$\mathcal{P}_s(\mathbb{R}^{d-\ell+1}) = \text{span} \left\{ \mathbb{R}^{d-\ell+1} \ni x \mapsto (a_i^T x)^k : 1 \leq i \leq n, 0 \leq k \leq s \right\}.$$

Let $P \in \mathcal{P}_s(\mathbb{R}^d)$ be arbitrary. With $x \in \mathbb{R}^{d-\ell+1}$ and $y \in \mathbb{R}^{\ell-1}$, we can then write

$$P(x, y) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} P_{\mathbf{k}}(x),$$

with suitably chosen polynomials $P_{\mathbf{k}} \in \mathcal{P}_s(\mathbb{R}^{d-\ell+1})$. Moreover, each $P_{\mathbf{k}}$ can be written as

$$P_{\mathbf{k}}(x) = \sum_{\substack{1 \leq i \leq n \\ 0 \leq k \leq s}} \alpha_{i,k,\mathbf{k}} (a_i^T x)^k = \sum_{i=1}^n \underbrace{\sum_{k=0}^s \alpha_{i,k,\mathbf{k}} (a_i^T x)^k}_{=: Q_{i,\mathbf{k}}(a_i^T x)}$$

with suitable real coefficients $\alpha_{i,k,\mathbf{k}} \in \mathbb{R}$ and with $Q_{i,\mathbf{k}} \in \mathcal{P}_s(\mathbb{R})$ for each $i \in \{1, \dots, n\}$ and $\mathbf{k} \in \mathbb{N}_0^{\ell-1}$ with $|\mathbf{k}| \leq s$.

This gives us

$$P(x, y) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} P_{\mathbf{k}}(x) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} \sum_{i=1}^n Q_{i,\mathbf{k}}(a_i^T x) = \sum_{i=1}^n \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} Q_{i,\mathbf{k}}(a_i^T x).$$

For $i \in \{1, \dots, n\}$ we define $H_i \in \mathcal{P}(\mathbb{R}^\ell)$ as

$$H_i : \mathbb{R}^\ell = \mathbb{R} \times \mathbb{R}^{\ell-1} \rightarrow \mathbb{R}, \quad (t, y) \mapsto \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} Q_{i,\mathbf{k}}(t)$$

and furthermore $A_i \in \mathbb{R}^{\ell \times d}$ as

$$A_i := \left(\begin{array}{c|c} a_i^T & 0_{1 \times (\ell-1)} \\ \hline 0_{(\ell-1) \times (d-\ell+1)} & I_{(\ell-1) \times (\ell-1)} \end{array} \right),$$

whence it holds that

$$H_i \left(A_i \begin{pmatrix} x \\ y \end{pmatrix} \right) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{\ell-1} \\ |\mathbf{k}| \leq s}} y^{\mathbf{k}} Q_{i,\mathbf{k}}(a_i^T x)$$

for $x \in \mathbb{R}^{d-\ell+1}$ and $y \in \mathbb{R}^{\ell-1}$. We thus get

$$P(x, y) = \sum_{i=1}^n H_i \left(A_i \begin{pmatrix} x \\ y \end{pmatrix} \right).$$

Since $P \in \mathcal{P}_s(\mathbb{R}^d)$ was arbitrary, the claim is shown. $\square$

We can now prove the upper bound.

Theorem 4.2 *Let $d, \ell, r \in \mathbb{N}$ with $d > \ell$ and $1 \leq q \leq p \leq \infty$. Then there exists a positive constant $C = C(d, \ell, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ such that for any $f \in \mathcal{B}(W_d^{r,p})$ there exists a function $R \in \mathcal{R}_{n,d,\ell}^{\text{poly}}(A_1, \dots, A_n)$ with*

$$\|f - R\|_{L^q(B^d)} \leq C \cdot n^{-r/(d-\ell)}.$$

Proof Let $C_1 = C_1(d, \ell) > 0$ be a constant satisfying

$$\dim \left(\mathcal{P}_s^h(\mathbb{R}^{d-\ell+1}) \right) \leq C_1 \cdot s^{d-\ell} \quad \text{for every } s \in \mathbb{N};$$

we refer to [9, Lemma F.1] for a proof of the existence of such a constant. We may assume $n \geq C_1$ at the cost of possibly increasing C in the end, similar to the end of the proof of Theorem 3.9. We then pick $s \in \mathbb{N}$ as the maximal number with

$$C_1 \cdot s^{d-\ell} \leq n.$$

Note that this implies

$$C_1 \cdot (2s)^{d-\ell} > n \Leftrightarrow C_1 \cdot 2^{d-\ell} \cdot s^{d-\ell} > n \Leftrightarrow s > c_2 \cdot n^{1/(d-\ell)}$$

with $c_2 = c_2(d, \ell) := C_1^{1/(\ell-d)} \cdot \frac{1}{2}$. Moreover, note that according to Proposition 4.1 the inclusion $\mathcal{P}_s(B^d) \subseteq \mathcal{R}_{n,d,\ell}^{\text{poly}}(A_1, \dots, A_n)$ holds for a specific choice of $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$. Let $f \in \mathcal{B}(W_d^{r,p})$ be arbitrary. According to Proposition 2.5, we can pick $P \in \mathcal{P}_s(B^d) \subseteq \mathcal{R}_{n,d,\ell}^{\text{poly}}(A_1, \dots, A_n)$ with

$$\|f - P\|_{L^q(B^d)} \leq C_3 \cdot s^{-r} \leq C_3 \cdot c_2^{-r} \cdot n^{-r/(d-\ell)}$$

with an absolute constant $C_3 = C_3(d, p, q, r) > 0$. Hence, in the end the claim follows defining the constant $C := C_3 \cdot c_2^{-r}$. $\square$

Remark 4.3 The proof of Theorem 4.2 shows that the only property of Sobolev functions that is actually needed in order to obtain the approximation rate is the fact that these functions have the property stated in Proposition 2.5. Therefore, one can establish the same approximation rate for the set consisting of all L^p -functions $f : B^d \rightarrow \mathbb{R}$ for which

$$\sup_{s \in \mathbb{N}} \inf_{P \in \mathcal{P}_s(B^d)} (\|f - P\|_{L^q(B^d)} \cdot s^r) \leq C,$$

where C is a fixed constant. Note that the final bound depends on the constant C .

5 Application to neural networks

In this section, we apply the results obtained in the previous sections to the case of shallow neural networks. More specifically, we show upper and lower bounds for the approximation of Sobolev functions using *shallow generalized translation networks* and *shallow complex-valued neural networks*. Here, for $d, \ell, n \in \mathbb{N}$ and any (activation) function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ we define the set of *shallow generalized translation networks* with d input neurons, activation dimension ℓ , n hidden-layer neurons and activation function τ as

$$\mathcal{NN}_{d,\ell,n}^\tau := \left\{ \mathbb{R}^d \ni x \mapsto \sum_{k=1}^n c_k \tau(A_k x + b_k) \in \mathbb{R} : A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}, b_1, \dots, b_n \in \mathbb{R}^\ell, c_1, \dots, c_n \in \mathbb{R} \right\}.$$

Note that we obtain a classical shallow neural network in the case $\ell = 1$. For technical reasons, we further introduce the set

$$\mathcal{NN}_{d,\ell,n}^{\tau,*} := \left\{ \mathbb{R}^d \ni x \mapsto \sum_{k=1}^n c_k \tau(A_k x + b_k) \in \mathbb{R} : A_k \in \mathbb{R}^{\ell \times d}, \text{rank}(A_k) \in \{0, \ell\}, b_k \in \mathbb{R}^\ell, c_k \in \mathbb{R} \text{ for } k \in \{1, \dots, n\} \right\}$$

of generalized translation networks where the matrices A_k are limited to full-rank matrices and the zero matrix. Lastly, for fixed matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$, we let

$$\mathcal{NN}_{d,\ell,n}^{\tau}(A_1, \dots, A_n) := \left\{ \mathbb{R}^d \ni x \mapsto \sum_{k=1}^n c_k \tau(A_k x + b_k) \in \mathbb{R} : b_1, \dots, b_n \in \mathbb{R}^\ell, c_1, \dots, c_n \in \mathbb{R} \right\}$$

denote the set of all generalized translation networks with fixed weight matrices $A_1, \dots, A_n$.

Let us now turn to complex-valued neural networks (CVNNs): For any fixed (activation) function $\phi : \mathbb{C} \rightarrow \mathbb{C}$, we write

$$\mathcal{CVNN}_{d,n}^{\phi} := \left\{ \mathbb{C}^d \ni z \mapsto \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \in \mathbb{C} : \alpha_1, \dots, \alpha_n \in \mathbb{C}^d, \beta_1, \dots, \beta_n, \gamma_1, \dots, \gamma_n \in \mathbb{C} \right\}$$

for the set of shallow CVNNs with d input neurons, n hidden-layer neurons and activation function ϕ . Moreover, for fixed weight vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$ we let

$$\mathcal{CVNN}_{d,n}^{\phi}(\alpha_1, \dots, \alpha_n) := \left\{ \mathbb{C}^d \ni z \mapsto \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \in \mathbb{C} : \beta_1, \dots, \beta_n, \gamma_1, \dots, \gamma_n \in \mathbb{C} \right\}.$$

At first glance, one might be tempted to think that one obtains the set $\mathcal{CVNN}_{d,n}^{\phi}$ as a special case of generalized translation networks by replacing d by $2d$, putting $\ell = 2$ and using $\mathbb{C} \cong \mathbb{R}^2$. However, this is not the case, for the following reasons:

- The functions in $\mathcal{NN}_{2d,2,n}^{\tau}$ map to $\mathbb{R}$ whereas CVNNs map to $\mathbb{C}$, which is equivalent to *two* real output neurons.
- The activation function in $\mathcal{NN}_{2d,2,n}^{\tau}$ is a function $\mathbb{R}^2 \rightarrow \mathbb{R}$, whereas the activation function of a CVNN is a function $\mathbb{C} \rightarrow \mathbb{C}$, i.e., $\mathbb{R}^2 \rightarrow \mathbb{R}^2$.
- The matrices A_k in the definition of $\mathcal{NN}_{2d,2,n}^{\tau}$ are arbitrary elements of $\mathbb{R}^{2 \times (2d)}$.

The complex vectors α_k in the definition of $\mathcal{CVNN}_{d,n}^{\phi}$ may be regarded as elements of $\mathbb{R}^{2 \times (2d)}$ but only elements of $\mathbb{R}^{2 \times (2d)}$ with a specific structure arise in this way. That is, when viewed as elements of $\mathbb{R}^{2 \times (2d)}$ they have the block form

$$\left(\begin{array}{cc|cc| \dots | cc} a_1 & b_1 & a_2 & b_2 & \dots & a_d & b_d \\ -b_1 & a_1 & -b_2 & a_2 & \dots & -b_d & a_d \end{array} \right) \in \mathbb{R}^{2 \times (2d)} \quad (5.1)$$

with real numbers $a_1, \dots, a_d, b_1, \dots, b_d \in \mathbb{R}$ that represent the real and imaginary parts of the entries of the weight vector α_k .

5.1 Generalized translation networks

In this section, we provide sharp bounds on the rate of approximation of Sobolev functions by generalized translation networks as defined above. A lower bound can be obtained immediately from Corollary 3.10 by noting that every function from $\mathcal{NN}_{d,\ell,n}^\tau$ is in fact the sum of n ℓ -variate ridge functions.

Corollary 5.1 *Let $d, \ell, r \in \mathbb{N}$ with $\ell < d$ and $p, q \in [1, \infty]$ be arbitrary. Then there exists a positive constant $c = c(d, \ell, p, q, r) > 0$ with the following property: For every $n \in \mathbb{N}$ there exists a function $f \in \mathcal{B}(W_d^{r,p})$ such that for any activation function $\tau \in L_{\text{loc}}^1(\mathbb{R}^\ell)$ we have*

$$\inf_{\mathcal{T} \in \mathcal{NN}_{d,\ell,n}^{\tau,*}} \|f - \mathcal{T}\|_{L^q(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

If $\tau \in \mathcal{M}_b(\mathbb{R}^\ell)$, we furthermore have

$$\inf_{\mathcal{T} \in \mathcal{NN}_{d,\ell,n}^\tau} \|f - \mathcal{T}\|_{L^q(B^d)} \geq c \cdot n^{-r/(d-\ell)}.$$

Proof Let $\mathcal{T} \in \mathcal{NN}_{d,\ell,n}^\tau$ be arbitrary, i.e.,

$$\mathcal{T}(x) = \sum_{k=1}^n c_k \tau(A_k x + b_k).$$

For $k \in \{1, \dots, n\}$, we define $\varrho_k(x) := c_k \cdot \tau(x + b_k)$. In the case of $\tau \in \mathcal{M}_b(\mathbb{R}^\ell)$ we clearly get $\varrho_k \in \mathcal{M}_b(\mathbb{R}^\ell)$ for any $k \in \{1, \dots, n\}$, which implies $\mathcal{T} \in \mathcal{R}_{n,d,\ell}^*$ in that case. On the other hand, if $\tau \in L_{\text{loc}}^1(\mathbb{R}^\ell)$ and $A_k \in \mathbb{R}^{\ell \times d}$ is full-rank, we get that $\varrho_k \in L^1(B^d)$ by Proposition 2.1. If $A_k = 0$, the function ϱ_k is constant and therefore trivially contained in $L^1(B^d)$. The claim then follows from Corollary 3.10. $\square$

Note the subtle difference that we have to restrict to full-rank matrices A_k (or the zero matrix) in the case of an arbitrary L_{loc}^1 -activation function. This restriction is not necessary in the case of a locally bounded activation function. Whether this restriction in the case $\tau \in L_{\text{loc}}^1(\mathbb{R}^\ell)$ is necessary or can be removed remains a question for further investigation.

In the following theorem, we show that for every $\ell \in \mathbb{N}$ there exists a smooth activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ such that the optimal rate of $n^{-r/(d-\ell)}$ when approximating Sobolev functions with networks from $\mathcal{NN}_{d,\ell,n}^\tau$ can indeed be achieved. This activation function is constructed in a “piecewise” manner, such that the shifts of this function, restricted to B^ℓ , form a dense subset of $C(B^\ell)$.

Theorem 5.2 Let $\ell \in \mathbb{N}$. Then there exists a smooth activation function $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ with the following property: For every $d, r \in \mathbb{N}$ with $\ell < d$ and $1 \leq q \leq p \leq \infty$, there exists a positive constant $C = C(d, \ell, p, q, r) > 0$ such that for every $n \in \mathbb{N}$ there exist matrices $A_1, \dots, A_n \in \mathbb{R}^{\ell \times d}$ with the property that for every $f \in \mathcal{B}(W_d^{r,p})$ there exists a shallow generalized translation network $\mathcal{T} \in \mathcal{NN}_{d,\ell,n}^\tau(A_1, \dots, A_n)$ with

$$\|f - \mathcal{T}\|_{L^q(B^d)} \leq C \cdot n^{-r/(d-\ell)}.$$

Proof Let $\{u_m : B^\ell \rightarrow \mathbb{R}\}_{m \in \mathbb{N}}$ be a countable set of smooth functions which is dense in $C(B^\ell)$ with respect to $\|\cdot\|_{L^\infty}$. For instance, we can take the set of polynomials in ℓ variables with rational coefficients. We write e_1 for the first standard basis vector in $\mathbb{R}^\ell$. Let then $\tau : \mathbb{R}^\ell \rightarrow \mathbb{R}$ be a smooth function with

$$\tau(x + 3m \cdot e_1) = u_m(x) \quad \text{for every } x \in B^\ell, m \in \mathbb{N}.$$

We show that τ has the desired property. To this end, let $d, n, r \in \mathbb{N}$ with $d > \ell$, $1 \leq q \leq p \leq \infty$ and $f \in \mathcal{B}(W_d^{r,p})$ be arbitrary. We use Theorem 4.2 and obtain the existence of a positive constant $C_1 = C_1(d, \ell, p, q, r) > 0$, polynomials $Q_1, \dots, Q_n \in \mathcal{P}(\mathbb{R}^\ell)$ and matrices $B_1, \dots, B_n \in \mathbb{R}^{\ell \times d}$ with

$$\left\| f(x) - \sum_{k=1}^n Q_k(B_k x) \right\|_{L^q(B^d)} \leq C_1 \cdot n^{-r/(d-\ell)}.$$

Note that the choice of the matrices B_k does *not* depend on the choice of f according to Theorem 4.2. Let $k \in \{1, \dots, n\}$. If $B_k \neq 0$, we set³ $A_k := B_k / \|B_k\|_{\ell^2 \rightarrow \ell^2}$ and $P_k(x) := Q_k(\|B_k\|_{\ell^2 \rightarrow \ell^2} \cdot x)$. If $B_k = 0$, set $A_k := B_k$ and $P_k := Q_k$. By construction, we have

$$\sum_{k=1}^n Q_k(B_k x) = \sum_{k=1}^n P_k(A_k x).$$

Note that we then have $A_k x \in B^\ell$ for every $x \in B^d$. Due to the density of the u_m , we can for every $k \in \{1, \dots, n\}$ pick a number $m_k \in \mathbb{N}$ such that

³ For a matrix $M \in \mathbb{R}^{\ell \times d}$, we let $\|M\|_{\ell^2 \rightarrow \ell^2} := \sup_{x \in B^d} \|Mx\|_{\ell^2}$.

$$\|P_k - u_{m_k}\|_{L^\infty(B^\ell)} \leq n^{-r/(d-\ell)} \cdot n^{-1} \cdot \left(\lambda^d(B^d)\right)^{-1/q}$$

and hence

$$\|P_k(A_k x) - u_{m_k}(A_k x)\|_{L^q(B^d)} \leq n^{-r/(d-\ell)} \cdot n^{-1}.$$

Then, because of $\tau(A_k x + 3m_k \cdot e_1) = u_{m_k}(A_k x)$ for $x \in B^d$, we get that

$$\begin{aligned} & \left\| f(x) - \sum_{k=1}^n \tau(A_k x + 3m_k \cdot e_1) \right\|_{L^q(B^d)} \\ & \leq \left\| f(x) - \sum_{k=1}^n P_k(A_k x) \right\|_{L^q(B^d)} + \sum_{k=1}^n \|P_k(A_k x) - u_{m_k}(A_k x)\|_{L^q(B^d)} \\ & \leq (C_1 + 1) \cdot n^{-r/(d-\ell)}. \end{aligned}$$

Hence, the claim follows by letting $C := C_1 + 1$. $\square$

5.2 Complex-valued neural networks

The goal of this subsection is to prove the sharp approximation rate of $n^{-r/(2d-2)}$ for the approximation of (complex-valued) Sobolev functions on the unit ball in $\mathbb{C}^d$ using shallow CVNNs with d input neurons, n hidden-layer neurons and locally integrable activation function; see below for precise definitions. Establishing a *lower* bound of $n^{-r/(2d-2)}$ is a direct consequence of Corollary 3.10 and can be found in Corollary 5.3. Showing that there exists an activation function for which the *upper* bound of $n^{-r/(2d-2)}$ can be attained is more difficult and in particular requires the translation of several results from [26, Section 5] to the complex-valued setting; see Theorem 5.9.

In order to establish our approximation bounds for shallow CVNNs, we first introduce some new notation. For $d \in \mathbb{N}$, we let

$$B^d(\mathbb{C}) := \left\{ z \in \mathbb{C}^d : \|z\|_{\ell^2} \leq 1 \right\}, \quad \text{where } \|z\|_{\ell^2} := \sqrt{\sum_{j=1}^d \operatorname{Re}(z_j)^2 + \operatorname{Im}(z_j)^2}.$$

Moreover, for $p \in [1, \infty]$, we let

$$\mathcal{B}(W_d^{r,p}(\mathbb{C})) := \left\{ f : B^d(\mathbb{C}) \rightarrow \mathbb{C} : \operatorname{Re}(f), \operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p}) \right\},$$

where we canonically identify $\mathbb{C}^d \cong \mathbb{R}^{2d}$ (and thus $B^d(\mathbb{C}) \cong B^{2d}$).

We first derive a lower bound for the approximation of Sobolev functions using shallow CVNNs as introduced above. This is a relatively straightforward consequence of Corollary 3.10 and may be viewed as a special case of Corollary 5.1.

Corollary 5.3 *Let $d, r \in \mathbb{N}$ with $d \geq 2$ and $p, q \in [1, \infty]$ be arbitrary. Then there exists a positive constant $c = c(d, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exists a function $f \in \mathcal{B}(W_d^{r,p}(\mathbb{C}))$ such that for any (complex) activation function $\phi \in L_{\text{loc}}^1(\mathbb{C}; \mathbb{C})$ we have*

$$\inf_{\Phi \in \mathcal{CVNN}_{d,n}^{\phi}} \|f - \Phi\|_{L^q(B^d(\mathbb{C}))} \geq c \cdot n^{-r/(2d-2)}.$$

Proof Let $\phi \in L_{\text{loc}}^1(\mathbb{C}; \mathbb{C})$ and $\Phi \in \mathcal{CVNN}_{d,n}^{\phi}$ be given. Then there exist $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$ and $\beta_1, \dots, \beta_n, \gamma_1, \dots, \gamma_n \in \mathbb{C}$ such that

$$\Phi(z) = \sum_{k=1}^n \gamma_k \phi(\alpha_k^T z + \beta_k) \quad \text{for every } z \in \mathbb{C}^d.$$

We note that

$$\begin{aligned} \operatorname{Re}(\Phi(z)) &= \sum_{k=1}^n \operatorname{Re}(\gamma_k \phi(\alpha_k^T z + \beta_k)) \\ &= \sum_{k=1}^n \left[\operatorname{Re}(\gamma_k) \cdot (\operatorname{Re} \phi)(\alpha_k^T z + \beta_k) - \operatorname{Im}(\gamma_k) \cdot (\operatorname{Im} \phi)(\alpha_k^T z + \beta_k) \right]. \end{aligned}$$

The goal is to show $\operatorname{Re}(\Phi) \in \mathcal{R}_{2n,2d,2}$ by identifying $\mathbb{C}^d$ with $\mathbb{R}^{2d}$. According to the computation above, this is satisfied if

$$(\operatorname{Part} \phi)(\alpha_k^T \cdot \bullet + \beta_k) \in L^1(B^d(\mathbb{C})),$$

for any $\operatorname{Part} \in \{\operatorname{Re}, \operatorname{Im}\}$ and $k \in \{1, \dots, n\}$. Since $\phi \in L_{\text{loc}}^1(\mathbb{C})$, we get

$$(\operatorname{Part} \phi)(\bullet + \beta_k) \in L_{\text{loc}}^1(\mathbb{R}^2),$$

by identifying $\mathbb{C}$ with $\mathbb{R}^2$. Moreover, if $\alpha_k \neq 0$, it follows that the associated real-valued matrix $A \in \mathbb{R}^{2 \times 2d}$ (see (5.1)) has full rank 2. Hence, in this case we get

$$(\operatorname{Part} \phi)(\alpha_k^T \cdot \bullet + \beta_k) \in L^1(B^d(\mathbb{C}))$$

by Proposition 2.1. Conversely, if $\alpha_k = 0$, the map $(\operatorname{Part} \phi)(\alpha_k^T \cdot \bullet + \beta_k)$ is constant and therefore trivially contained in $L^1(B^d(\mathbb{C}))$. This proves $\operatorname{Re}(\Phi) \in \mathcal{R}_{2n,2d,2}$.

Hence, according to Corollary 3.10, we conclude the existence of a real-valued function $f \in \mathcal{B}(W_{2d}^{r,p})$ and a constant $\tilde{c} = \tilde{c}(r, d, p, q) > 0$ such that

$$\inf_{\phi \in L_{\text{loc}}^1(\mathbb{C})} \inf_{\Phi \in \mathcal{CVNN}_{d,n}^\phi} \|f - \text{Re}(\Phi)\|_{L^q(B^d(\mathbb{C}))} \geq \tilde{c} \cdot (2n)^{-r/(2d-2)} = c \cdot n^{-r/(2d-2)},$$

by letting $c := 2^{-r/(2d-2)} \cdot \tilde{c}$. The claim is then obtained by noting that

$$\begin{aligned} \|f - \text{Re}(\Phi)\|_{L^q(B^d(\mathbb{C}))} &= \|\text{Re}(f - \Phi)\|_{L^q(B^d(\mathbb{C}))} \\ &\leq \|f - \Phi\|_{L^q(B^d(\mathbb{C}))} \quad \text{for every } \Phi \in \mathcal{CVNN}_{d,n}^\phi. \end{aligned}$$

□

It remains to show that there exists a complex activation function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ for which the rate of $n^{-r/(2d-2)}$ (which is proven to be optimal according to Corollary 5.3) can indeed be attained. The activation function for which we will show that it achieves the desired approximation rate is the same “piecewise” activation function that was already constructed in [9, Lemma F.4], where an approximation rate of $n^{-r/(2d-1)}$ has been proven (see [9, Theorem 4.2]). This function is constructed following the same idea as in Theorem 5.2. The main reason why the approach from Theorem 5.2 cannot be used directly to obtain the desired result was already discussed at the beginning of the section: while in the definition of the set $\mathcal{NN}_{2d,2,n}^\tau$ the matrices A_k are arbitrary, they are restricted to a specific structure when considering CVNNs. Specifically, it is not straightforward to show that in the case $\ell = 2$ one can pick the matrices appearing in the proof of Proposition 4.1 to have the structure considered in (5.1).

Therefore, we translate several results from [26, Section 5] to the case of complex polynomials in z and $\bar{z}$. Here, we make use of the *Wirtinger Calculus* which we briefly discuss here. For a function $f \in C^1(U; \mathbb{C})$ with an open set $U \subseteq \mathbb{C}$ and where C^1 refers to differentiability with respect to real variables, we define for $w = x + iy \in U$ (with $x, y \in \mathbb{R}$) the Wirtinger derivatives at w as

$$\begin{aligned} \partial_{\text{wirt}} f(w) &:= \frac{\partial f}{\partial z}(w) := \frac{1}{2} \left(\frac{\partial f}{\partial x}(w) - i \cdot \frac{\partial f}{\partial y}(w) \right) \\ \bar{\partial}_{\text{wirt}} f(w) &:= \frac{\partial f}{\partial \bar{z}}(w) := \frac{1}{2} \left(\frac{\partial f}{\partial x}(w) + i \cdot \frac{\partial f}{\partial y}(w) \right). \end{aligned}$$

The intuition behind the Wirtinger derivatives is to formally treat z and $\bar{z}$ as *independent* variables and to then take derivatives only with respect to z or $\bar{z}$. For multiindices $\mathbf{k}, \ell \in \mathbb{N}_0^d$ we write $\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell}$ for iterated multivariate Wirtinger derivatives according to the multiindices $\mathbf{k}$ and ℓ . This is well-defined when applied to functions of sufficient regularity, since Wirtinger derivatives commute because they are linear combinations of partial derivatives.

For $s, t \in \mathbb{N}_0$ and $d \in \mathbb{N}$ we define

$$\mathcal{P}_s(\mathbb{C}^d) := \left\{ \mathbb{C}^d \ni z \mapsto \sum_{\substack{\mathbf{k}, \ell \in \mathbb{N}_0^d \\ |\mathbf{k}|, |\ell| \leq s}} a_{\mathbf{k}, \ell} z^{\mathbf{k}} \bar{z}^{\ell} : a_{\mathbf{k}, \ell} \in \mathbb{C} \right\}$$

and

$$\mathcal{P}_{s,t}^h(\mathbb{C}^d) := \left\{ \mathbb{C}^d \ni z \mapsto \sum_{\substack{\mathbf{k}, \ell \in \mathbb{N}_0^d \\ |\mathbf{k}|=s, |\ell|=t}} a_{\mathbf{k}, \ell} z^{\mathbf{k}} \bar{z}^{\ell} : a_{\mathbf{k}, \ell} \in \mathbb{C} \right\}.$$

We remark that the condition $|\mathbf{k}|, |\ell| \leq s$ appearing in the definition of $\mathcal{P}_s(\mathbb{C}^d)$ is different from $|\mathbf{k}| + |\ell| \leq s$, which would be the direct generalization of the definition in the real case to the complex case. Clearly, it holds that

$$\mathcal{P}_s(\mathbb{C}^d) = \bigoplus_{s', t' \leq s} \mathcal{P}_{s', t'}^h(\mathbb{C}^d). \quad (5.2)$$

For a complex polynomial $Q \in \mathcal{P}_s(\mathbb{C}^d)$ with

$$Q(z) = \sum_{\substack{\mathbf{k}, \ell \in \mathbb{N}_0^d \\ |\mathbf{k}|, |\ell| \leq s}} a_{\mathbf{k}, \ell} z^{\mathbf{k}} \bar{z}^{\ell},$$

we define the associated differential operator as

$$Q(D) := \sum_{\substack{\mathbf{k}, \ell \in \mathbb{N}_0^d \\ |\mathbf{k}|, |\ell| \leq s}} a_{\mathbf{k}, \ell} \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell},$$

where the notations ∂_{wirt} and $\bar{\partial}_{\text{wirt}}$ refer to the (multivariate) Wirtinger derivatives mentioned above. A computation shows that for multiindices $\mathbf{k}, \ell, \mathbf{k}', \ell' \in \mathbb{N}_0^d$ with $|\mathbf{k}| = |\mathbf{k}'|$ and $|\ell| = |\ell'|$ we get

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} (z^{\mathbf{k}'} \bar{z}^{\ell'}) = \mathbb{1}_{(\mathbf{k}, \ell) = (\mathbf{k}', \ell')} \cdot \mathbf{k}! \cdot \ell!. \quad (5.3)$$

We refer to Lemma B.1 in Appendix B.4 for a rigorous proof of that identity. Let $\mathcal{L}(\mathcal{P}_{s,t}^h(\mathbb{C}^d); \mathbb{C})$ denote the space of $\mathbb{C}$ -linear maps from $\mathcal{P}_{s,t}^h(\mathbb{C}^d)$ to $\mathbb{C}$. According to (5.3), the set

$$\left\{ \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} : \mathbf{k}, \ell \in \mathbb{N}_0^d \text{ with } |\mathbf{k}| = s \text{ and } |\ell| = t \right\}$$

forms a basis of $\mathcal{L}(\mathcal{P}_{s,t}^h(\mathbb{C}^d); \mathbb{C})$, so we may conclude that

$$\mathcal{L}(\mathcal{P}_{s,t}^h(\mathbb{C}^d); \mathbb{C}) = \{Q(D) : Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)\}. \quad (5.4)$$

Moreover, for a fixed vector $a \in \mathbb{C}^d$ and multiindices $\mathbf{k}, \ell \in \mathbb{N}_0^d$ with $|\mathbf{k}| = s$ and $|\ell| = t$ for $s, t \in \mathbb{N}_0$, we get

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \left((a^T z)^s (\overline{a^T z})^t \right) = s! \cdot t! \cdot a^{\mathbf{k}} \bar{a}^{\ell}. \quad (5.5)$$

Here, we refer to Lemma B.2 for a proof of this fact. Hence, for $Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$, we observe

$$Q(D) \left((a^T z)^s (\overline{a^T z})^t \right) = s! \cdot t! \cdot Q(a) \quad \text{for all } a \in \mathbb{C}^d. \quad (5.6)$$

Moreover, we note that

$$\begin{aligned} (a^T z)^s (\overline{a^T z})^t &= \left(\sum_{j=1}^n a_j z_j \right) \left(\sum_{j=1}^n \overline{a_j z_j} \right) \\ &= \left(\sum_{j_1, \dots, j_s=1}^n a_{j_1} \cdots a_{j_s} \cdot z_{j_1} \cdots z_{j_s} \right) \left(\sum_{j_1, \dots, j_t=1}^n \overline{a_{j_1}} \cdots \overline{a_{j_t}} \cdot \overline{z_{j_1}} \cdots \overline{z_{j_t}} \right) \in \mathcal{P}_{s,t}^h(\mathbb{C}^d) \end{aligned} \quad (5.7)$$

for every $a \in \mathbb{C}^d$. This leads to the following proposition, which is a generalization of [26, Proposition 5.1] to the complex-valued setting.

Proposition 5.4 *Let $\Omega \subseteq \mathbb{C}^d$, $s, t \in \mathbb{N}_0$ and $P \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$. Then we have*

$$P \in \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\}$$

if and only if $Q(D)P = 0$ for every $Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$ that vanishes on Ω .

Proof Let

$$V := \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\} \stackrel{(5.7)}{\subseteq} \mathcal{P}_{s,t}^h(\mathbb{C}^d).$$

Then, from elementary linear algebra, for $P \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$ we have $P \in V$ if and only if $L(P) = 0$ for every $L \in \mathcal{L}(\mathcal{P}_{s,t}^h(\mathbb{C}^d); \mathbb{C})$ with $L|_V \equiv 0$. From (5.4), we infer that this is equivalent to $Q(D)P = 0$ for every $Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$ with $Q(D)|_V \equiv 0$. But from (5.6) we get that $Q(D)|_V \equiv 0$ if and only if $Q|_{\Omega} \equiv 0$. This proves the claim. $\square$

This immediately gives us the following characterization, which is the generalization of [26, Corollary 5.11] to the complex-valued case.

Proposition 5.5 *Let $\Omega \subseteq \mathbb{C}^d$ and $s, t \in \mathbb{N}_0$. Then we have*

$$\mathcal{P}_{s,t}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\}$$

if and only if for every $Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$ we have

$$Q|_{\Omega} \equiv 0 \Rightarrow Q \equiv 0.$$

Proof We again let

$$V := \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\} \stackrel{(5.7)}{\subseteq} \mathcal{P}_{s,t}^h(\mathbb{C}^d).$$

According to Proposition 5.4 we have $\mathcal{P}_{s,t}^h(\mathbb{C}^d) = V$ if and only if

$$\forall P, Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d) : Q|_{\Omega} \equiv 0 \Rightarrow Q(D)P = 0.$$

We reformulate the latter to

$$\forall Q \in \mathcal{P}_{s,t}^h(\mathbb{C}^d) : Q|_{\Omega} \equiv 0 \Rightarrow \left(Q(D)P = 0 \text{ for all } P \in \mathcal{P}_{s,t}^h(\mathbb{C}^d) \right).$$

But $Q(D)P = 0$ for all $P \in \mathcal{P}_{s,t}^h(\mathbb{C}^d)$ holds if and only if $Q \equiv 0$, which follows for instance from (5.6). $\square$

Note that the previous proposition in particular shows the following a priori not entirely obvious statement

$$\mathcal{P}_{s,t}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \mathbb{C}^d \right\},$$

since the only polynomial that vanishes on $\mathbb{C}^d$ is the zero polynomial. Since the space $\mathcal{P}_{s,t}^h(\mathbb{C}^d)$ is finite-dimensional, we in particular infer the existence of a set $\Omega \subseteq \mathbb{C}^d$ with $|\Omega| = \dim_{\mathbb{C}}(\mathcal{P}_{s,t}^h(\mathbb{C}^d))$ and

$$\mathcal{P}_{s,t}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\}. \quad (5.8)$$

The following proposition is crucial for the proof of the upper bound. For its real-valued analogon, we refer to [26, Corollary 5.12].

Proposition 5.6 *Let $\Omega \subseteq \mathbb{C}^d$ and $s, t \in \mathbb{N}_0$. If*

$$\mathcal{P}_{s,t}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\},$$

then also

$$\mathcal{P}_{s',t'}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^{s'} (\overline{a^T z})^{t'} : a \in \Omega \right\}$$

for all $s', t' \in \mathbb{N}_0$ with $s' \leq s$ and $t' \leq t$.

Proof We know that $\mathcal{P}_{s',t'}^h(\mathbb{C}^d) \stackrel{(5.7)}{\supseteq} \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^{s'} (\overline{a^T z})^{t'} : a \in \Omega \right\}$. Suppose that

$$\mathcal{P}_{s',t'}^h(\mathbb{C}^d) \supsetneq \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^{s'} (\overline{a^T z})^{t'} : a \in \Omega \right\}$$

for some $s' \leq s$ and $t' \leq t$. According to Proposition 5.5, we can then pick $Q \in \mathcal{P}_{s',t'}^h(\mathbb{C}^d)$ with $Q \neq 0$ and $Q|_{\Omega} \equiv 0$. Let $\tilde{Q} \in \mathcal{P}_{s-s',t-t'}^h(\mathbb{C}^d) \setminus \{0\}$ be arbitrary. Then we have $Q\tilde{Q} \in \mathcal{P}_{s,t}^h(\mathbb{C}^d) \setminus \{0\}$ and $Q\tilde{Q}|_{\Omega} \equiv 0$, which contradicts Proposition 5.5, since $\mathcal{P}_{s,t}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a^T z)^s (\overline{a^T z})^t : a \in \Omega \right\}$ by assumption of the proposition. $\square$

We can now show that each polynomial from $\mathcal{P}_s(\mathbb{C}^d)$ can be written as the sum of (complex) ridge polynomials, where the number of summands depends on the dimension of the space $\mathcal{P}_{s,s}^h(\mathbb{C}^d)$. This statement is the translation of Proposition 4.1 to the complex-valued setting.

Theorem 5.7 Let $d \in \mathbb{N}$ and $s \in \mathbb{N}_0$ and pick $n \in \mathbb{N}$ with

$$\dim_{\mathbb{C}}(\mathcal{P}_{s,s}^h(\mathbb{C}^d)) \leq n.$$

Then there exist $a_1, \dots, a_n \in \mathbb{C}^d$ with $\|a_j\|_{\ell^2} \leq 1$ for $j \in \{1, \dots, n\}$ with the following property: For every $P \in \mathcal{P}_s(\mathbb{C}^d)$ there exist $P_1, \dots, P_n \in \mathcal{P}_s(\mathbb{C})$ with

$$P(z) = \sum_{j=1}^n P_j(a_j^T z).$$

Proof We pick $a_1, \dots, a_n \in \mathbb{C}^d$ with $\|a_j\|_{\ell^2} \leq 1$ for $j \in \{1, \dots, n\}$ such that

$$\mathcal{P}_{s,s}^h(\mathbb{C}^d) = \text{span}_{\mathbb{C}} \left\{ z \mapsto (a_j^T z)^s (\overline{a_j^T z})^s : j = 1, \dots, n \right\}.$$

This is possible according to (5.8). Note that we can scale the a_j as we want since scaling does not change the span above. Let $P \in \mathcal{P}_s(\mathbb{C}^d)$ be arbitrary. Then we can write

$$P(z) = \sum_{s',t' \leq s} Q_{s',t'}(z)$$

with $Q_{s',t'} \in \mathcal{P}_{s',t'}^h(\mathbb{C}^d)$ for every $s', t' \leq s$. According to Proposition 5.6, we can write

$$Q_{s',t'}(z) = \sum_{j=1}^n a_{s',t',j} \cdot (a_j^T z)^{s'} (\overline{a_j^T z})^{t'}$$

with suitable coefficients $a_{s',t',j} \in \mathbb{C}$. This gives us

$$\begin{aligned} P(z) &= \sum_{s',t' \leq s} \sum_{j=1}^n a_{s',t',j} \cdot (a_j^T z)^{s'} (\overline{a_j^T z})^{t'} = \sum_{j=1}^n \sum_{s',t' \leq s} a_{s',t',j} \cdot (a_j^T z)^{s'} (\overline{a_j^T z})^{t'} \\ &= \sum_{j=1}^n P_j(a_j^T z) \end{aligned}$$

with

$$P_j(z) := \sum_{s',t' \leq s} a_{s',t',j} \cdot z^{s'} \bar{z}^{t'}.$$

Since clearly $P_j \in \mathcal{P}_s(\mathbb{C})$ for every $j \in \{1, \dots, n\}$, the claim is shown. $\square$

The activation function that yields the optimal approximation rate is obtained in the following lemma. We refer to [9] for the proof.

Lemma 5.8 (cf. [9, Lemma F.4]) *Let $\{u_\ell\}_{\ell=1}^\infty$ be an enumeration of the set of complex polynomials in z and $\bar{z}$ with coefficients in $\mathbb{Q} + i\mathbb{Q}$. Then there exists a smooth function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ (where by “smoothness” we refer to smoothness with respect to real variables) with the property that for every $\ell \in \mathbb{N}$ and $z \in [-1, 1] + i \cdot [-1, 1] \subseteq \mathbb{C}$ one has*

$$\phi(z + 3\ell) = u_\ell(z).$$

Note that since $B^1(\mathbb{C}) \subseteq [-1, 1] + i \cdot [-1, 1]$, the function ϕ in particular satisfies

$$\phi(z + 3\ell) = u_\ell(z) \quad \text{for every } z \in B^1(\mathbb{C}). \quad (5.9)$$

We can now state and prove the main result of this section.

Theorem 5.9 *Let $\phi : \mathbb{C} \rightarrow \mathbb{C}$ be the activation function from Lemma 5.8. Moreover, let $d, r \in \mathbb{N}$ with $d \geq 2$ and $1 \leq q \leq p \leq \infty$. Then there exists a constant $C = C(d, p, q, r) > 0$ with the following property: For any $n \in \mathbb{N}$ there exist complex vectors $\alpha_1, \dots, \alpha_n \in \mathbb{C}^d$ such that for every $f \in \mathcal{B}(W_d^{r,p}(\mathbb{C}))$ there exists a shallow CVNN $\Phi \in \mathcal{CVNN}_{d,n}^\phi(\alpha_1, \dots, \alpha_n)$ with*

$$\|f - \Phi\|_{L^q(B^d)} \leq C \cdot n^{-r/(2d-2)}.$$

Proof Let $s \in \mathbb{N}$. Then it is easy to see that

$$\begin{aligned} \dim_{\mathbb{C}}(\mathcal{P}_{s,s}^h(\mathbb{C}^d)) &= \#\left\{z^{\mathbf{k}} \bar{z}^{\ell} : |\mathbf{k}| = |\ell| = s\right\} = \left(\#\left\{\mathbf{k} \in \mathbb{N}_0^d : |\mathbf{k}| = s\right\}\right)^2 \\ &\leq C_1 \cdot \left(s^{d-1}\right)^2 = C_1 \cdot s^{2d-2}, \end{aligned}$$

where $C_1 = C_1(d) > 0$; again, see for instance [9, Lemma F.1] We may assume $n \geq C_1$ at the cost of possibly enlarging the constant C in the end, similar to the end of the proof of Theorem 3.9. We then pick $s \in \mathbb{N}$ as the largest number satisfying $C_1 \cdot s^{2d-2} \leq n$. Note that this implies

$$C_1 \cdot (2s)^{2d-2} > n \Leftrightarrow s > c_2 \cdot n^{1/(2d-2)}$$

for a constant $c_2 = c_2(d) > 0$.

Let $f \in \mathcal{B}(W_d^{r,p}(\mathbb{C}))$ be arbitrary. For a multiindex $\mathbf{k} \in \mathbb{N}_0^{2d}$ and $z \in \mathbb{C}^d$, we write

$$(\operatorname{Re}(z), \operatorname{Im}(z))^{\mathbf{k}} := \prod_{j=1}^d \operatorname{Re}(z_j)^{\mathbf{k}_j} \cdot \operatorname{Im}(z_j)^{\mathbf{k}_{d+j}}.$$

By applying Proposition 2.5 to $\operatorname{Re}(f) \in \mathcal{B}(W_{2d}^{r,p})$ and $\operatorname{Im}(f) \in \mathcal{B}(W_{2d}^{r,p})$, we obtain the existence of a polynomial

$$P(z) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2d} \\ |\mathbf{k}| \leq s}} a_{\mathbf{k}} \cdot (\operatorname{Re}(z), \operatorname{Im}(z))^{\mathbf{k}}$$

with

$$\|f - P\|_{L^q(B^d(\mathbb{C}))} \leq C_3 \cdot s^{-r} \leq C_4 \cdot n^{-r/(2d-2)},$$

where $C_3 = C_3(d, p, q, r) > 0$ and $C_4 = C_4(d, p, q, r) > 0$ are constants and $a_{\mathbf{k}} \in \mathbb{C}$ for $\mathbf{k} \in \mathbb{N}_0^{2d}$ with $|\mathbf{k}| \leq s$. For a fixed $\mathbf{k} \in \mathbb{N}_0^{2d}$, we compute

$$\begin{aligned} (\operatorname{Re}(z), \operatorname{Im}(z))^{\mathbf{k}} &= \prod_{j=1}^d \operatorname{Re}(z_j)^{\mathbf{k}_j} \cdot \operatorname{Im}(z_j)^{\mathbf{k}_{d+j}} \\ &= \prod_{j=1}^d \left[\frac{1}{2^{\mathbf{k}_j} \cdot (2i)^{\mathbf{k}_{d+j}}} \cdot (z_j + \overline{z_j})^{\mathbf{k}_j} \cdot (z_j - \overline{z_j})^{\mathbf{k}_{d+j}} \right]. \end{aligned}$$

For each $j \in \{1, \dots, d\}$ we let $\mathcal{Q}_j(z_j) := (z_j + \overline{z_j})^{\mathbf{k}_j}$ and $\widetilde{\mathcal{Q}}_j(z_j) := (z_j - \overline{z_j})^{\mathbf{k}_{d+j}}$. Note that $\mathcal{Q}_j \in \mathcal{P}_{\mathbf{k}_j}(\mathbb{C})$ and $\widetilde{\mathcal{Q}}_j \in \mathcal{P}_{\mathbf{k}_{d+j}}(\mathbb{C})$, which yields $\mathcal{Q}_j \widetilde{\mathcal{Q}}_j \in \mathcal{P}_{\mathbf{k}_j + \mathbf{k}_{d+j}}(\mathbb{C})$. Overall, this yields

$$\prod_{j=1}^d \operatorname{Re}(z_j)^{\mathbf{k}_j} \cdot \operatorname{Im}(z_j)^{\mathbf{k}_{d+j}} \in \mathcal{P}_{|\mathbf{k}|}(\mathbb{C}^d) \subseteq \mathcal{P}_s(\mathbb{C}^d),$$

which then yields $P \in \mathcal{P}_s(\mathbb{C}^d)$. Since $n \geq \dim_{\mathbb{C}}(\mathcal{P}_{s,s}^h(\mathbb{C}^d))$, we can apply Theorem 5.7 and write

$$P(z) = \sum_{j=1}^n P_j(\alpha_j^T z),$$

where $P_j \in \mathcal{P}_s(\mathbb{C})$ and $\alpha_j \in \mathbb{C}^d$ with $\|\alpha_j\|_{\ell^2} \leq 1$ for $j \in \{1, \dots, n\}$. Note that the choice of the α_j is *independent* of the choice of f (and P) according to Theorem 5.7. Recall from Lemma 5.8 that $\{u_\ell\}_{\ell=1}^\infty$ is an enumeration of the set of complex polynomials in z and $\bar{z}$ with coefficients in $\mathbb{Q} + i\mathbb{Q}$. Since this set is dense in $C(B^1(\mathbb{C}))$ with respect to $\|\cdot\|_{L^\infty}$ and $|\alpha_j^T z| \leq \|\alpha_j\|_{\ell^2} \cdot \|z\|_{\ell^2} \leq 1$ for every $z \in B^d(\mathbb{C})$, we can pick $\ell_1, \dots, \ell_n \in \mathbb{N}$ with

$$\|P_j - u_{\ell_j}\|_{L^\infty(B^1(\mathbb{C}))} \leq n^{-1-\frac{r}{2d-2}} \cdot \left(\lambda^{2d}(B^d(\mathbb{C}))\right)^{-1/q}$$

and hence

$$\|P_j(\alpha_j^T z) - u_{\ell_j}(\alpha_j^T z)\|_{L^q(B^d(\mathbb{C}))} \leq n^{-1-\frac{r}{2d-2}} \quad \text{for every } j \in \{1, \dots, n\}.$$

We then get

$$\begin{aligned} & \left\| f(z) - \sum_{j=1}^n \phi(\alpha_j^T z + 3\ell_j) \right\|_{L^q(B^d(\mathbb{C}))} \\ &= \left\| f(z) - \sum_{j=1}^n u_{\ell_j}(\alpha_j^T z) \right\|_{L^q(B^d(\mathbb{C}))} \\ &\leq \left\| f(z) - \sum_{j=1}^n P_j(\alpha_j^T z) \right\|_{L^q(B^d(\mathbb{C}))} + \sum_{j=1}^n \|P_j(\alpha_j^T z) - u_{\ell_j}(\alpha_j^T z)\|_{L^q(B^d(\mathbb{C}))} \\ &\leq \|f - P\|_{L^q(B^d(\mathbb{C}))} + n^{-r/(2d-2)} \leq (C_4 + 1) \cdot n^{-r/(2d-2)}. \end{aligned}$$

Since $z \mapsto \sum_{j=1}^n \phi(\alpha_j^T z + 3\ell_j) \in \mathcal{CVNN}_{d,n}^\phi(\alpha_1, \dots, \alpha_n)$, the claim follows by letting $C := C_4 + 1$. $\square$

We remark that one can *not* expect the rate of $n^{-r/(2d-2)}$ for *general* (smooth, non-polyharmonic) activation functions ϕ , as follows from [9, Theorem 4.3].

Appendix A: Discussion of an issue in [18]

In this appendix, we discuss an issue in the paper [18] in which a lower bound of $n^{-r/(d-1)}$ for the error of approximating Sobolev functions by sums of n univariate

ridge functions with respect to the L^1 -norm is shown. One of the central ingredients of the proof in that work is [18, Lemma 5], which, in the notation of the present work, can be stated as follows:

Let $d, n, s \in \mathbb{N}$ with $d > 1$ and let $P \in \mathcal{P}_s(B^d)$ be arbitrary. Let $\pi_s : L^2(B^d) \rightarrow \mathcal{P}_s(B^d)$ denote the orthogonal projection onto $\mathcal{P}_s(B^d)$. Then it holds that

$$\inf_{R \in \mathcal{R}_{n,d,1}^*} \|P - R\|_{L^1(B^d)} \geq \inf_{P' \in \pi_s(\mathcal{R}_{n,d,1}^*)} \|P - P'\|_{L^1(B^d)}.$$

This claim is thus essentially identical to a univariate version of Lemma 3.8 (i.e., for $\ell = 1$), with the difference that the quasi-projection Pr_s is replaced by the orthogonal projection π_s and that there is no absolute constant appearing in the inequality.

The proof presented in [18] relies on the fact that the set $\mathcal{R}_{n,d,1}^*$ can be written as a union of subspaces $U_i \subseteq L^2(B^d)$. Then, by showing that

$$\inf_{R \in U_i} \|P - R\|_{L^1(B^d)} \geq \inf_{P' \in \pi_s(U_i)} \|P - P'\|_{L^1(B^d)} \quad (\text{A.1})$$

holds for any i , one easily gets the bound by forming the infimum over i on both sides. No special property of the subspaces U_i is used in [18]. Yet, for arbitrary subspaces of $L^2(B^d)$ the inequality (A.1) in fact *fails to hold*, even if an absolute constant is allowed. We show this in the following proposition.

Proposition A.1 *Let $d \in \mathbb{N}$. Then for any constant $\kappa > 0$ there exist one-dimensional subspaces $U, V \subseteq L^2(B^d)$ and a function $P \in U$ with the property that*

$$\inf_{g \in V} \|P - g\|_{L^1(B^d)} < \kappa \cdot \inf_{\tilde{g} \in \pi_U(V)} \|P - \tilde{g}\|_{L^1(B^d)}.$$

Here, $\pi_U : L^2(B^d) \rightarrow U$ denotes the orthogonal projection onto U .

Proof Let $\kappa > 0$ be an arbitrary constant. For $n \in \mathbb{N}$, let $\psi_n : \mathbb{R} \rightarrow \mathbb{R}$ be continuous with $0 \leq \psi_n \leq 1$, with $\text{supp}(\psi_n) \subseteq [\frac{1}{n}, \infty)$, and with $\psi_n \equiv 1$ on $[\frac{2}{n}, 1]$. Then, writing $x = (x_1, \dots, x_d)$ for $x \in \mathbb{R}^d$, define the functions

$$P_n : \mathbb{R}^d \rightarrow \mathbb{R}, \quad P_n(x) = x_1^{-1/3} \cdot \psi_n(x_1)$$

with the understanding that $P_n(x) = 0$ if $x_1 = 0$. We get

$$\begin{aligned} \|P_n\|_{L^2(B^d)}^2 &\leq \int_{[-1,1]^d} |P_n(x)|^2 \, dx \stackrel{\text{Tonelli}}{\leq} \int_{[\frac{1}{n},1]} x_1^{-2/3} \, dx_1 \cdot 2^{d-1} \\ &= (3 - 3 \cdot n^{-1/3}) \cdot 2^{d-1} \leq 3 \cdot 2^{d-1}. \end{aligned}$$

Moreover, taking $n \geq 4\sqrt{d}$, we get

$$\|P_n\|_{L^1(B^d)} \geq \int_{[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]^d} x_1^{-1/3} \cdot \mathbb{1}_{[\frac{2}{n},1]}(x_1) \, d(x_1, \dots, x_d)$$

$$\begin{aligned} & \stackrel{\text{Tonelli}}{=} \int_{[\frac{2}{n}, \frac{1}{\sqrt{d}}]} x_1^{-1/3} dx_1 \cdot \int_{[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]^{d-1}} d(x_2, \dots, x_d) \\ & \geq \left(\frac{2}{\sqrt{d}}\right)^{d-1} \cdot \int_{[\frac{1}{2\sqrt{d}}, \frac{1}{\sqrt{d}}]} x_1^{-1/3} dx_1 =: \theta = \theta(d) > 0. \end{aligned}$$

Moreover, since P_n is continuous,

$$\|P_n\|_{L^\infty(B^d)} \geq P_n(2/n, 0, \dots, 0) = \left(\frac{n}{2}\right)^{1/3} \cdot \psi_n\left(\frac{2}{n}\right) = \left(\frac{n}{2}\right)^{1/3}.$$

Therefore,

$$\frac{2 \cdot \|P_n\|_{L^2(B^d)}^2}{\|P_n\|_{L^\infty(B^d)} \cdot \|P_n\|_{L^1(B^d)}} \leq \frac{6 \cdot 2^{d-1}}{(n/2)^{1/3} \cdot \theta} \rightarrow 0 \quad (n \rightarrow \infty).$$

We can therefore pick $N \in \mathbb{N}$ with

$$\frac{2 \cdot \|P_N\|_{L^2(B^d)}^2}{\|P_N\|_{L^\infty(B^d)}} < \kappa \cdot \|P_N\|_{L^1(B^d)}.$$

Thus, $P := P_N$ is continuous with $P \neq 0$ and

$$\frac{2 \cdot \|P\|_{L^2(B^d)}^2}{\|P\|_{L^\infty(B^d)}} < \kappa \cdot \|P\|_{L^1(B^d)}. \quad (\text{A.2})$$

Let $x^\natural \in B^d$ with $|P(x^\natural)| = \|P\|_{L^\infty}$, which exists since continuous functions on compact sets attain their maximum. Without loss of generality, we may assume $P(x^\natural) > 0$, otherwise replace P by $-P$. Then, by continuity and since B^d is the closure of $(B^d)^\circ$, it is easy to see that there exists a point $x^* \in (B^d)^\circ$ and $\delta > 0$ with the property that $Q := x^* + (-\delta, \delta)^d \subseteq (B^d)^\circ$ and

$$P(x) \geq \frac{P(x^\natural)}{2} > 0 \quad \text{for all } x \in Q. \quad (\text{A.3})$$

We define

$$f := P - c \cdot \mathbb{1}_Q \quad \text{with} \quad c := \frac{\|P\|_{L^2(B^d)}^2}{\int_Q P(x) dx},$$

noting that $\int_Q P(x) dx > 0$ by (A.3). We then compute that

$$\langle P, f \rangle = \|P\|_{L^2}^2 - c \cdot \int_Q P(x) dx = 0,$$

meaning $P \perp f$. Furthermore, since

$$\frac{\|P\|_{L^\infty(B^d)}}{2} \cdot \mathbb{A}^d(Q) = \frac{P(x^{\sharp})}{2} \cdot \mathbb{A}^d(Q) \stackrel{(A.3)}{\leq} \int_Q P(x) \, dx,$$

we get

$$c \leq \frac{2 \cdot \|P\|_{L^2(B^d)}^2}{\|P\|_{L^\infty(B^d)}} \cdot \frac{1}{\mathbb{A}^d(Q)}.$$

This yields that

$$\|P - f\|_{L^1(B^d)} = c \cdot \|\mathbb{1}_Q\|_{L^1(B^d)} = c \cdot \mathbb{A}^d(Q) \leq \frac{2 \cdot \|P\|_{L^2(B^d)}^2}{\|P\|_{L^\infty(B^d)}}.$$

For $V := \text{span } f$ and $U := \text{span } P$, we have $P \in U$ and get

$$\inf_{g \in V} \|P - g\|_{L^1(B^d)} \leq \|P - f\|_{L^1(B^d)} \leq \frac{2 \cdot \|P\|_{L^2(B^d)}^2}{\|P\|_{L^\infty(B^d)}},$$

but $\pi_U(V) = \{0\}$ (since $f \perp P$) and hence

$$\inf_{\tilde{g} \in \pi_U(V)} \|P - \tilde{g}\|_{L^1(B^d)} = \|P\|_{L^1(B^d)}.$$

The claim then follows from (A.2). $\square$

Specifically, the gap in the proof in [18] lies in its Equation (14), where it is claimed that for a fixed polynomial $P \in \mathcal{P}_s(B^d)$ we have

$$\sup_{v \in (\pi_s(U_i))^\perp \cap \mathcal{P}_s(B^d), \|v\|_{L^\infty(B^d)} \leq 1} \langle P, v \rangle = \inf_{h \in \pi_s(U_i)} \|P - h\|_{L^1(B^d)}. \quad (\text{A.4})$$

Here, by definition we have

$$(\pi_s(U_i))^\perp := \left\{ v \in L^\infty(B^d) : \langle v, w \rangle = 0 \text{ for all } w \in \pi_s(U_i) \right\}.$$

For an arbitrary normed space $(\mathcal{X}, \|\cdot\|)$ and a subspace $M \subseteq \mathcal{X}$, it is well-known that the distance between an arbitrary element $x \in \mathcal{X}$ and the space M can be expressed using the *dual* space of $\mathcal{X}$. More precisely, let $\mathcal{X}^*$ denote the normed dual of $\mathcal{X}$ with dual norm $\|\cdot\|_*$. One can then define the *annihilator* of M as

$$M^\perp := \{ \varphi \in \mathcal{X}^* : \varphi|_M \equiv 0 \}.$$

For an arbitrary element $x \in \mathcal{X}$, we then have

$$\inf_{m \in M} \|x - m\| = \sup_{\substack{\varphi \in M^\perp \\ \|\varphi\|_* \leq 1}} \varphi(x),$$

see for instance [16, p.119, Thm.1]. Applying this fact to the case $\mathcal{X} = (\mathcal{P}_s(B^d), \|\cdot\|_{L^1(B^d)})$ and $M = \pi_s(U_i)$ implies

$$\inf_{h \in \pi_s(U_i)} \|P - h\|_{L^1(B^d)} = \sup_{\varphi} \varphi(P),$$

where the supremum is taken over all continuous linear functionals $\varphi : \mathcal{P}_s(B^d) \rightarrow \mathbb{R}$, for which

$$\sup_{\substack{P' \in \mathcal{P}_s(B^d) \\ \|P'\|_{L^1(B^d)} \leq 1}} \varphi(P') \leq 1 \quad \text{and} \quad \varphi(w) = 0 \quad \text{for every } w \in \pi_s(U_i). \quad (\text{A.5})$$

Note that in order for (A.4) (at least with “ $\geq$ ” instead of “ $=$ ”) to hold, it would be sufficient that for each continuous linear functional $\varphi : \mathcal{P}_s(B^d) \rightarrow \mathbb{R}$ satisfying (A.5), there exists a function $v \in \mathcal{P}_s(B^d)$ with $\|v\|_{L^\infty(B^d)} \leq 1$, $\langle v, w \rangle = 0$ for all $w \in \pi_s(U_i)$ and

$$\varphi(P) = \langle v, P \rangle.$$

If one is willing to drop the condition $v \in \mathcal{P}_s(B^d)$, this is in fact true: According to the Hahn-Banach extension theorem, for each such functional $\varphi : \mathcal{P}_s(B^d) \rightarrow \mathbb{R}$ we can pick a (with respect to $\|\cdot\|_{L^1(B^d)}$) continuous linear extension $\varphi' : L^1(B^d) \rightarrow \mathbb{R}$ with $\varphi'|_{\mathcal{P}_s(B^d)} = \varphi$ and

$$\sup_{\substack{f \in L^1(B^d) \\ \|f\|_{L^1(B^d)} \leq 1}} \varphi'(f) = \sup_{\substack{P' \in \mathcal{P}_s(B^d) \\ \|P'\|_{L^1(B^d)} \leq 1}} \varphi(P') \leq 1.$$

Moreover, according to [6, Theorem 6.15] there exists $v \in L^\infty(B^d)$ with $\|v\|_{L^\infty(B^d)} \leq 1$ and

$$\langle v, f \rangle = \varphi'(f) \quad \text{for all } f \in L^1(B^d).$$

In particular, since $\varphi'|_{\mathcal{P}_s(B^d)} = \varphi$, we have $\langle v, w \rangle = \varphi(w) = 0$ for every $w \in \pi_s(U_i)$. However, it is *not* clear (to the authors of the present paper), whether one can pick v to be contained in $\mathcal{P}_s(B^d)$, which would imply (A.4) (with “ $\geq$ ”).

Alternatively, for each continuous linear functional $\varphi : \mathcal{P}_s(B^d) \rightarrow \mathbb{R}$ satisfying (A.5), since $\mathcal{P}_s(B^d)$ is finite-dimensional it is well-known that there exists $v \in \mathcal{P}_s(B^d)$ with

$$\varphi(P') = \langle v, P' \rangle \quad \text{for all } P' \text{ in } \mathcal{P}_s(B^d).$$

By (A.5), we have $\langle v, w \rangle = 0$ for all $w \in \pi_s(U_i)$ and

$$\sup_{\substack{Q \in \mathcal{P}_s(B^d) \\ \|Q\|_{L^1(B^d)} \leq 1}} \langle v, Q \rangle = \sup_{\substack{Q \in \mathcal{P}_s(B^d) \\ \|Q\|_{L^1(B^d)} \leq 1}} \varphi(Q) \leq 1.$$

However, it is *not* clear at all (to the authors of the present paper) if this implies that

$$\|v\|_{L^\infty(B^d)} = \sup_{\substack{f \in L^1(B^d) \\ \|f\|_{L^1(B^d)} \leq 1}} \langle v, f \rangle \leq 1.$$

We emphasize that the preceding discussion including Proposition A.1 does not necessarily imply that the statement [18, Lemma 5] is false. It might well be the case that the stated inequality is true, at least up to a fixed multiplicative constant. However, since the statement does not hold for arbitrary subspaces of $L^2(B^d)$ (see Proposition A.1), this then necessarily relies on specific properties of ridge functions and polynomials that the authors of the present work are not aware of (and that are not mentioned in [18]).

Appendix B: Postponed proofs

B.1 Proof of Lemma 2.6

Proof of Lemma 2.6 We will use throughout the proof without comment that $\binom{k+\sigma}{\sigma} = \binom{k+\sigma}{k}$, so that we need to show

$$\Pr_s f = \sum_{k=0}^{\infty} (\Delta^{\sigma+1} \eta^*)(k) \cdot \binom{k+\sigma}{k} \cdot S_k^\sigma(f).$$

The proof is by induction over $\sigma \in \mathbb{N}_0$. In the case $\sigma = 0$, we get

$$\begin{aligned} \sum_{k=0}^{\infty} \left((\Delta \eta^*)(k) \cdot \binom{k}{k} \cdot S_k^0(f) \right) &= \sum_{k=0}^{\infty} \left[(\Delta \eta^*)(k) \cdot \frac{1}{\binom{k}{0}} \cdot \sum_{j=0}^k \underbrace{\binom{k-j}{0}}_{=1} \cdot \text{proj}_j(f) \right] \\ &= \sum_{k=0}^{\infty} \left[(\eta(k/s) - \eta((k+1)/s)) \cdot \sum_{j=0}^k \text{proj}_j(f) \right] \\ &= \sum_{k=0}^{\infty} \left[\eta(k/s) \cdot \sum_{j=0}^k \text{proj}_j(f) \right] - \sum_{k=0}^{\infty} \left[\eta((k+1)/s) \cdot \sum_{j=0}^k \text{proj}_j(f) \right] \\ &= \sum_{k=0}^{\infty} \left[\eta(k/s) \cdot \sum_{j=0}^k \text{proj}_j(f) \right] - \sum_{k=1}^{\infty} \left[\eta(k/s) \cdot \sum_{j=0}^{k-1} \text{proj}_j(f) \right] \\ &= \sum_{k=0}^{\infty} \eta(k/s) \cdot \text{proj}_k(f) \\ &= \Pr_s f. \end{aligned}$$

We now assume the claim to be true for an arbitrary but fixed $\sigma \in \mathbb{N}_0$. We then get

$$\begin{aligned}
 & \sum_{k=0}^{\infty} \left(\left(\Delta^{\sigma+2} \eta^* \right) (k) \cdot \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right) \\
 &= \sum_{k=0}^{\infty} \left[\left(\left(\Delta^{\sigma+1} \eta^* \right) (k) - \left(\Delta^{\sigma+1} \eta^* \right) (k+1) \right) \cdot \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right] \\
 &= \sum_{k=0}^{\infty} \left[\left(\Delta^{\sigma+1} \eta^* \right) (k) \cdot \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right] \\
 &\quad - \sum_{k=0}^{\infty} \left[\left(\Delta^{\sigma+1} \eta^* \right) (k+1) \cdot \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right] \\
 &= \sum_{k=0}^{\infty} \left[\left(\Delta^{\sigma+1} \eta^* \right) (k) \cdot \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right] \\
 &\quad - \sum_{k=1}^{\infty} \left[\left(\Delta^{\sigma+1} \eta^* \right) (k) \cdot \binom{k+\sigma}{k-1} \cdot S_{k-1}^{\sigma+1}(f) \right] \\
 &= \left(\Delta^{\sigma+1} \eta^* \right) (0) \cdot S_0^{\sigma+1}(f) + \sum_{k=1}^{\infty} \left(\left(\Delta^{\sigma+1} \eta^* \right) (k) \left[\binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) \right. \right. \\
 &\quad \left. \left. - \binom{k+\sigma}{k-1} \cdot S_{k-1}^{\sigma+1}(f) \right] \right) \\
 &= \left(\Delta^{\sigma+1} \eta^* \right) (0) \cdot \binom{0+\sigma}{\sigma} \cdot S_0^{\sigma}(f) \\
 &\quad + \sum_{k=1}^{\infty} \left(\left(\Delta^{\sigma+1} \eta^* \right) (k) \left[\binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) - \binom{k+\sigma}{k-1} \cdot S_{k-1}^{\sigma+1}(f) \right] \right),
 \end{aligned}$$

where the last step used that $S_0^{\gamma}(f) = \text{proj}_0(f)$ for every $\gamma \in \mathbb{N}_0$ as follows from the definition of $S_0^{\gamma}(f)$.

Keeping the induction hypothesis in mind, it therefore suffices to show

$$\binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) - \binom{k+\sigma}{k-1} \cdot S_{k-1}^{\sigma+1}(f) = \binom{k+\sigma}{\sigma} \cdot S_k^{\sigma}(f) \quad \text{for every } k \geq 1.$$

Plugging in the definition of the Cesaro means and using $\binom{k+\sigma+1}{\sigma+1} = \binom{k+\sigma+1}{k}$ and $\binom{k+\sigma}{\sigma+1} = \binom{k+\sigma}{k-1}$, we get

$$\begin{aligned}
 & \binom{k+\sigma+1}{k} \cdot S_k^{\sigma+1}(f) - \binom{k+\sigma}{k-1} \cdot S_{k-1}^{\sigma+1}(f) \\
 &= \frac{\binom{k+\sigma+1}{k}}{\binom{k+\sigma+1}{\sigma+1}} \cdot \sum_{j=0}^k \left[\binom{k-j+\sigma+1}{\sigma+1} \text{proj}_j(f) \right] - \frac{\binom{k+\sigma}{k-1}}{\binom{k+\sigma}{\sigma+1}} \cdot \sum_{j=0}^{k-1} \left[\binom{k-j+\sigma}{\sigma+1} \text{proj}_j(f) \right]
 \end{aligned}$$

$$\begin{aligned}
&= \text{proj}_k(f) + \sum_{j=0}^{k-1} \left[\binom{k-j+\sigma+1}{\sigma+1} - \binom{k-j+\sigma}{\sigma+1} \right] \cdot \text{proj}_j(f) \\
&\stackrel{(*)}{=} \binom{k-k+\sigma}{\sigma} \cdot \text{proj}_k(f) + \sum_{j=0}^{k-1} \binom{k-j+\sigma}{\sigma} \cdot \text{proj}_j(f) \\
&= \sum_{j=0}^k \binom{k-j+\sigma}{\sigma} \cdot \text{proj}_j(f) \\
&= \binom{k+\sigma}{\sigma} \cdot S_k^\sigma(f),
\end{aligned}$$

using Pascal's rule

$$\binom{n}{m} + \binom{n}{m-1} = \binom{n+1}{m} \quad \text{for every } n, m \in \mathbb{N} \text{ with } m \leq n$$

at the step marked with (*). This proves the claim. $\square$

B.2 Proof of Lemma 3.2

We mention again that the proof is essentially taken from [18, Lemma 6], with a few more details added.

Proof (Proof of Lemma 3.2) Let $g(x) := 1 - \frac{1}{2}(1 - 2x)^2 \cdot \log_2(e)$. Then $g(1/2) = 1$ and $g(0) = 1 - \frac{1}{2} \log_2(e) \leq \frac{1}{2}$. By the intermediate value theorem, we can thus choose $a \in (0, \frac{1}{2})$ such that

$$1 - \frac{1}{2}(1 - 2a)^2 \cdot \log_2(e) = \frac{47}{64}.$$

Let $\pi \in \text{sgn}(\Gamma) \subseteq E^m$ be arbitrary and consider the set

$$E_\pi := \left\{ \varepsilon \in E^m : \sum_{i=1}^m |\varepsilon_i - \pi_i| \leq 2am \right\}.$$

Consider the bijection

$$\varphi_\pi : E^m \rightarrow E^m, \quad (\varphi_\pi(\varepsilon))_i = -\pi_i \cdot \varepsilon_i = \begin{cases} \varepsilon_i, & \text{if } \pi_i = -1, \\ -\varepsilon_i, & \text{if } \pi_i = 1. \end{cases}$$

Fix $i \in \{1, \dots, m\}$ for the moment. If $\pi_i = 1$, we observe

$$|(\varphi_\pi(\varepsilon))_i - \pi_i| = -(\varphi_\pi(\varepsilon))_i + 1 = \varepsilon_i + 1$$

and if $\pi_i = -1$, we get

$$|(\varphi_\pi(\varepsilon))_i - \pi_i| = (\varphi_\pi(\varepsilon))_i - (-1) = \varepsilon_i + 1.$$

Since φ_π is a bijection (for instance because $\varphi_\pi \circ \varphi_\pi = \text{id}$), we get

$$\begin{aligned} |E_\pi| &= |\varphi_\pi^{-1}(E_\pi)| = |\{\varepsilon \in E^m : \varphi_\pi(\varepsilon) \in E_\pi\}| = \left| \left\{ \varepsilon \in E^m : \sum_{i=1}^m |(\varphi_\pi(\varepsilon))_i - \pi_i| \leq 2am \right\} \right| \\ &= \left| \left\{ \varepsilon \in E^m : \sum_{i=1}^m (\varepsilon_i + 1) \leq 2am \right\} \right| = \left| \left\{ \varepsilon \in E^m : \sum_{\substack{i \in \{1, \dots, m\} \\ \varepsilon_i = 1}} 2 \leq 2am \right\} \right| \\ &= \left| \left\{ \varepsilon \in E^m : \sum_{\substack{i \in \{1, \dots, m\} \\ \varepsilon_i = 1}} 1 \leq \lfloor am \rfloor \right\} \right| = \sum_{j=0}^{\lfloor am \rfloor} \left| \left\{ \varepsilon \in E^m : \sum_{\substack{i \in \{1, \dots, m\} \\ \varepsilon_i = 1}} 1 = j \right\} \right|. \end{aligned}$$

Each summand of the final sum can be viewed as the number of subsets of $\{1, \dots, m\}$ with cardinality j . We hence get

$$|E_\pi| = \sum_{j=0}^{\lfloor am \rfloor} \binom{m}{j}.$$

This sum can be bounded using Hoeffding's inequality for bounded random variables, see for instance [32, Theorem 2.2.6]. To do so, we denote the binomial distribution with parameters m and $p = 1/2$ by $\text{bin}(m, 1/2)$, i.e., m “tries”, each with success probability $\frac{1}{2}$. We then get

$$\begin{aligned} \sum_{j=0}^{\lfloor am \rfloor} \binom{m}{j} &= 2^m \cdot \mathbb{P}_{X \sim \text{bin}(m, 1/2)} (X \leq \lfloor am \rfloor) \\ &= 2^m \cdot \mathbb{P}_{\sigma_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{0, 1\})} \left(\sum_{i=1}^m \sigma_i \leq \lfloor am \rfloor \right) \\ &= 2^m \cdot \mathbb{P}_{\sigma_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{0, 1\})} \left(\sum_{i=1}^m (-\sigma_i + 1/2) \geq m/2 - \lfloor am \rfloor \right), \end{aligned}$$

where we note that $m/2 - \lfloor am \rfloor > 0$ since $a < 1/2$. Since $-1 \leq -\sigma_i \leq 0$ and $\mathbb{E}[-\sigma_i] = -\frac{1}{2}$, Hoeffding's inequality implies

$$\begin{aligned}
\mathbb{P}_{\sigma_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{0,1\})} \left(\sum_{i=1}^m (-\sigma_i + 1/2) \geq m/2 - \lfloor am \rfloor \right) &\leq \exp \left(-\frac{2(m/2 - \lfloor am \rfloor)^2}{m} \right) \\
&= \exp \left(-2m \cdot (1/2 - \beta)^2 \right) \\
&= \exp \left(-\frac{1}{2} \cdot m \cdot (1 - 2\beta)^2 \right)
\end{aligned}$$

with $\beta := \frac{\lfloor am \rfloor}{m}$. Applying the inequality gives us

$$\begin{aligned}
|E_\pi| = \sum_{j=0}^{\lfloor am \rfloor} \binom{m}{j} &\leq 2^m \cdot \exp \left(-\frac{1}{2} \cdot m \cdot (1 - 2\beta)^2 \right) = 2^{m - \log_2(e) \cdot \frac{1}{2} \cdot m \cdot (1 - 2\beta)^2} \\
&\leq 2^{m(1 - \log_2(e) \cdot \frac{1}{2} \cdot (1 - 2a)^2)} = 2^{m \cdot 47/64},
\end{aligned}$$

where we used that $\beta \leq a \leq \frac{1}{2}$ implies $(1 - 2\beta)^2 \geq (1 - 2a)^2$. Moreover, note that

$$\begin{aligned}
\left| E^m \setminus \bigcup_{\pi \in \text{sgn}(\Gamma)} E_\pi \right| &= |E^m| - \left| \bigcup_{\pi \in \text{sgn}(\Gamma)} E_\pi \right| \\
&\geq |E^m| - |\text{sgn}(\Gamma)| \cdot 2^{m \cdot 47/64} \geq 2^m - 2^{m \cdot 63/64} > 0,
\end{aligned}$$

where we used $|\text{sgn}(\Gamma)| \leq 2^{m/4}$ by assumption. We can thus choose $\varepsilon^* \in E^m \setminus \bigcup_{\pi \in \text{sgn}(\Gamma)} E_\pi$. By definition of the sets E_π , we then have

$$\inf_{\pi \in \text{sgn}(\Gamma)} \|\varepsilon^* - \pi\|_{\ell^1} \geq 2am.$$

It remains to translate this estimate to the set Γ . To this end, let $x \in \mathbb{R}$ be arbitrary and $\varepsilon \in \{-1, 1\}$. If $\text{sgn}(x) = \varepsilon$, we get $|x - \varepsilon| \geq 0 = \frac{1}{2} \cdot |\text{sgn}(x) - \varepsilon|$. If $\text{sgn}(x) \neq \varepsilon$, we get

$$|x - \varepsilon| \geq 1 = \frac{1}{2} \cdot 2 = \frac{1}{2} \cdot |\text{sgn}(x) - \varepsilon|.$$

In any case, we conclude

$$|x - \varepsilon| \geq \frac{1}{2} \cdot |\text{sgn}(x) - \varepsilon|.$$

This yields

$$\inf_{x \in \Gamma} \|\varepsilon^* - x\|_{\ell^1} \geq \frac{1}{2} \cdot \inf_{x \in \Gamma} \|\varepsilon^* - \text{sgn}(x)\|_{\ell^1} = \frac{1}{2} \cdot \inf_{\pi \in \text{sgn}(\Gamma)} \|\varepsilon^* - \pi\|_{\ell^1} \geq am,$$

as desired. $\square$

B.3 Proof of Lemma 3.4

Proof of Lemma 3.4 From [27, Appendix I.1.9], we get the existence of $\gamma_0 = \gamma_0(a), \dots, \gamma_a = \gamma_a(a) \in \mathbb{R}$ as well as $\delta_0 = \delta_0(b), \dots, \delta_b = \delta_b(b) \in \mathbb{R}$ with

$$\cos(\varphi)^a = \sum_{h=0}^a \gamma_h \cos(h\varphi) \quad \text{for all } \varphi \in \mathbb{R}$$

and

$$\sin(\varphi)^b = \begin{cases} \sum_{h=0}^b \delta_h \cos(h\varphi), & \text{if } b \text{ even,} \\ \sum_{h=0}^b \delta_h \sin(h\varphi), & \text{if } b \text{ odd} \end{cases} \quad \text{for all } \varphi \in \mathbb{R}.$$

If b is even, we hence get

$$\cos(\varphi)^a \sin(\varphi)^b = \sum_{h_1=0}^a \sum_{h_2=0}^b \gamma_{h_1} \delta_{h_2} \cdot \cos(h_1\varphi) \cos(h_2\varphi). \quad (\text{B.1})$$

We can use the well-known product-to-sum formula for cosines (see [27, Appendix I.1.8]) and get

$$\cos(h_1\varphi) \cos(h_2\varphi) = \frac{1}{2} \cdot (\cos((h_1 - h_2)\varphi) + \cos((h_1 + h_2)\varphi)).$$

Since $|h_1 \pm h_2| \leq h_1 + h_2 \leq a + b$ for $h_1 \in \{0, \dots, a\}$ and $h_2 \in \{0, \dots, b\}$ and by the symmetry of the cosine, we can thus rearrange (B.1) and get

$$\cos(\varphi)^a \sin(\varphi)^b = \sum_{h=0}^{a+b} \alpha_h \cos(h\varphi)$$

for certain coefficients $\alpha_h = \alpha_h(a, b) \in \mathbb{R}$. Similarly, in the case that b is odd, one obtains

$$\cos(\varphi)^a \sin(\varphi)^b = \sum_{h=0}^{a+b} \beta_h \sin(h\varphi)$$

for coefficients $\beta_h = \beta_h(a, b) \in \mathbb{R}$. This proves the claim. $\square$

B.4 Postponed proofs for identities involving Wirtinger derivatives

In this subsection, we provide rigorous proofs for the identities in Equations (5.3) and (5.5). To this end, we need the following well-known properties of Wirtinger derivatives that can be found for example in [13, E.1a]. Here, we assume that $U \subseteq \mathbb{C}$ is open and $f \in C^1(U; \mathbb{C})$.

- (1) ∂_{wirt} and $\bar{\partial}_{\text{wirt}}$ are both $\mathbb{C}$ -linear operators on the set $C^1(U; \mathbb{C})$.
- (2) f is complex-differentiable at $z \in U$ iff $\bar{\partial}_{\text{wirt}} f(z) = 0$ and in this case the equality

$$\partial_{\text{wirt}} f(z) = f'(z)$$

holds true, with $f'(z)$ denoting the complex derivative of f at z .

- (3) We have the conjugation rules

$$\overline{\partial_{\text{wirt}} f} = \bar{\partial}_{\text{wirt}} \bar{f} \quad \text{and} \quad \overline{\bar{\partial}_{\text{wirt}} f} = \partial_{\text{wirt}} \bar{f}.$$

- (4) If $g \in C^1(U; \mathbb{C})$, the following product rules for Wirtinger derivatives hold for every $z \in U$:

$$\begin{aligned} \partial_{\text{wirt}}(f \cdot g)(z) &= \partial_{\text{wirt}} f(z) \cdot g(z) + f(z) \cdot \partial_{\text{wirt}} g(z), \\ \bar{\partial}_{\text{wirt}}(f \cdot g)(z) &= \bar{\partial}_{\text{wirt}} f(z) \cdot g(z) + f(z) \cdot \bar{\partial}_{\text{wirt}} g(z). \end{aligned}$$

This product rule is not explicitly stated in [13] but follows easily from the product rule for $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$.

- (5) If $V \subseteq \mathbb{C}$ is an open set and $g \in C^1(V; \mathbb{C})$ with $g(V) \subseteq U$, then the following chain rules for Wirtinger derivatives hold true:

$$\begin{aligned} \partial_{\text{wirt}}(f \circ g) &= [(\partial_{\text{wirt}} f) \circ g] \cdot \partial_{\text{wirt}} g + [(\bar{\partial}_{\text{wirt}} f) \circ g] \cdot \partial_{\text{wirt}} \bar{g}, \\ \bar{\partial}_{\text{wirt}}(f \circ g) &= [(\partial_{\text{wirt}} f) \circ g] \cdot \bar{\partial}_{\text{wirt}} g + [(\bar{\partial}_{\text{wirt}} f) \circ g] \cdot \bar{\partial}_{\text{wirt}} \bar{g}. \end{aligned}$$

Using these properties, we can now prove (5.3).

Lemma B.1 *Let $\mathbf{k}, \ell, \mathbf{k}', \ell' \in \mathbb{N}_0^d$ be multiindices with $|\mathbf{k}| = |\mathbf{k}'|$ and $|\ell| = |\ell'|$. Then it holds*

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} (z^{\mathbf{k}'} \bar{z}^{\ell'}) = \mathbb{1}_{(\mathbf{k}, \ell) = (\mathbf{k}', \ell')} \cdot \mathbf{k}! \cdot \ell!.$$

Proof We start by showing that

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} (z^{\mathbf{k}} \bar{z}^{\ell}) = \mathbf{k}! \cdot \ell!. \quad (\text{B.2})$$

Firstly, assume that $\ell = 0$. We show via induction over $|\mathbf{k}|$ that the identity

$$\partial_{\text{wirt}}^{\mathbf{k}} z^{\mathbf{k}} = \mathbf{k}!$$

holds. There is nothing to show in the case $\mathbf{k} = 0$. Therefore, assume that the claim holds for a fixed $\mathbf{k} \in \mathbb{N}_0^d$ and let $j \in \{1, \dots, d\}$ be arbitrary. Then we get

$$\partial_{\text{wirt}}^{\mathbf{k}+e_j} z^{\mathbf{k}+e_j} = \partial_{\text{wirt}}^{\mathbf{k}} \partial_{\text{wirt}}^{e_j} [z^{\mathbf{k}} \cdot z_j].$$

Hence, for fixed variables $z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_d$, we consider $\partial_{\text{wirt}}[z_j \mapsto z^{\mathbf{k}} \cdot z_j]$. From the linearity of ∂_{wirt} we deduce

$$\begin{aligned} \partial_{\text{wirt}}[z_j \mapsto z^{\mathbf{k}} \cdot z_j] &= z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \cdot \partial_{\text{wirt}}[z_j \mapsto z_j^{\mathbf{k}_j+1}] \\ &= (\mathbf{k}_j + 1) \cdot z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \cdot z_j^{\mathbf{k}_j} = (\mathbf{k}_j + 1) \cdot z^{\mathbf{k}}. \end{aligned}$$

Here, we used the fact that $z_j \mapsto z_j^{\mathbf{k}_j+1}$ is holomorphic and the fact that ∂_{wirt} coincides with the regular complex derivative (see (2)) in this case. This gives us

$$\partial_{\text{wirt}}^{\mathbf{k}+e_j} z^{\mathbf{k}+e_j} = \partial_{\text{wirt}}^{\mathbf{k}} \partial_{\text{wirt}}^{e_j} [z^{\mathbf{k}} \cdot z_j] = (\mathbf{k}_j + 1) \cdot \partial_{\text{wirt}}^{\mathbf{k}} z^{\mathbf{k}} = (\mathbf{k} + e_j)!$$

according to the induction hypothesis. By induction, we have thus shown

$$\partial_{\text{wirt}}^{\mathbf{k}} z^{\mathbf{k}} = \mathbf{k}! \quad \text{for every } \mathbf{k} \in \mathbb{N}_0^d.$$

In order to get the full claim, we now perform an additional induction over $|\ell|$ and show that for every $\ell \in \mathbb{N}_0^d$ the statement

$$\forall \mathbf{k} \in \mathbb{N}_0^d : \quad \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} (z^{\mathbf{k}} \bar{z}^{\ell}) = \mathbf{k}! \cdot \ell!$$

holds. Again for $\ell = 0$, the statement is already proven above. Therefore, we pick $j \in \{1, \dots, d\}$ arbitrary and assume that $\ell \in \mathbb{N}_0^d$ satisfies the claim. We then let $\mathbf{k} \in \mathbb{N}_0^d$ be arbitrary and note

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell+e_j} (z^{\mathbf{k}} \bar{z}^{\ell+e_j}) = \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \bar{\partial}_{\text{wirt}}^{e_j} (z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \bar{z}^{\ell-\ell_j \cdot e_j} \cdot z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}).$$

We next note that

$$\bar{\partial}_{\text{wirt}}^{e_j} (z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \bar{z}^{\ell-\ell_j \cdot e_j} \cdot z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}) = \bar{\partial}_{\text{wirt}} [z_j \mapsto z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \bar{z}^{\ell-\ell_j \cdot e_j} \cdot z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}].$$

Again, for fixed variables $z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_d$, due to linearity we get

$$\bar{\partial}_{\text{wirt}} [z_j \mapsto z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \bar{z}^{\ell-\ell_j \cdot e_j} \cdot z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}] = z^{\mathbf{k}-\mathbf{k}_j \cdot e_j} \bar{z}^{\ell-\ell_j \cdot e_j} \cdot \bar{\partial}_{\text{wirt}} [z_j \mapsto z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}].$$

Using the product rule (4), we get

$$\bar{\partial}_{\text{wirt}} [z_j \mapsto z_j^{\mathbf{k}_j} \cdot \bar{z}_j^{\ell_j+1}] = \bar{\partial}_{\text{wirt}} [z_j \mapsto z_j^{\mathbf{k}_j}] \cdot \bar{z}_j^{\ell_j+1} + z_j^{\mathbf{k}_j} \cdot \bar{\partial}_{\text{wirt}} [z_j \mapsto \bar{z}_j^{\ell_j+1}].$$

Note that the first summand vanishes since $z_j \mapsto z_j^{\mathbf{k}_j}$ is holomorphic (see (2)). Moreover, we get

$$\bar{\partial}_{\text{wirt}}[z_j \mapsto \bar{z}_j^{\ell_j+1}] \stackrel{(3)}{=} \overline{\partial_{\text{wirt}}[z_j \mapsto z_j^{\ell_j+1}]} \stackrel{(2)}{=} (\ell_j + 1) \cdot \bar{z}_j^{\ell_j}.$$

All in all, this yields that

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell+e_j} (z^{\mathbf{k}} \bar{z}^{\ell+e_j}) = (\ell_j + 1) \cdot \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} [z^{\mathbf{k}} \bar{z}^{\ell}] = \mathbf{k}! \cdot (\ell + e_j)!,$$

where the induction hypothesis was used in the last equality. By induction, this proves (B.2).

It remains to show that

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} (z^{\mathbf{k}'} \bar{z}^{\ell'}) = 0 \quad \text{if } (\mathbf{k}, \ell) \neq (\mathbf{k}', \ell').$$

Without loss of generality we assume $\mathbf{k} \neq \mathbf{k}'$ (the case $\ell \neq \ell'$ follows analogously). Since $|\mathbf{k}| = |\mathbf{k}'|$, we conclude the existence of an index $j \in \{1, \dots, d\}$ with $\mathbf{k}_j > \mathbf{k}'_j$. Using the commutativity of partial derivatives, it suffices to show that

$$\partial_{\text{wirt}}^{\mathbf{k}_j \cdot e_j} [z^{\mathbf{k}'} \bar{z}^{\ell'}] = 0.$$

Since

$$\partial_{\text{wirt}}^{\mathbf{k}_j \cdot e_j} [z^{\mathbf{k}'} \bar{z}^{\ell'}] = z^{\mathbf{k}' - \mathbf{k}'_j \cdot e_j} \cdot \bar{z}^{\ell' - \ell'_j \cdot e_j} \cdot \partial_{\text{wirt}}^{\mathbf{k}_j} [z_j \mapsto z_j^{\mathbf{k}'_j} \bar{z}_j^{\ell'_j}],$$

it suffices to show that

$$\partial_{\text{wirt}}^{\mathbf{k}_j} [z_j \mapsto z_j^{\mathbf{k}'_j} \bar{z}_j^{\ell'_j}] = 0.$$

From an iterated application of the product rule (4) we get

$$\partial_{\text{wirt}}^{\mathbf{k}'_j} [z_j \mapsto z_j^{\mathbf{k}'_j} \bar{z}_j^{\ell'_j}] = (\mathbf{k}'_j)! \cdot \bar{z}_j^{\ell'_j}.$$

Then we get the claim by noting that

$$\partial_{\text{wirt}} [\bar{z}_j^{\ell'_j}] \stackrel{(3)}{=} \overline{\partial_{\text{wirt}} [z_j^{\ell'_j}]} \stackrel{(2)}{=} 0.$$

□

We continue by proving Equation (5.5).

Lemma B.2 For a fixed vector $a \in \mathbb{C}^d$ and multiindices $\mathbf{k}, \ell \in \mathbb{N}_0^d$ with $|\mathbf{k}| = s$ and $|\ell| = t$ for $s, t \in \mathbb{N}_0$, it holds

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \left((a^T z)^s (\overline{a^T z})^t \right) = s! \cdot t! \cdot a^{\mathbf{k}} \bar{a}^{\ell}.$$

Proof Similar to the proof of Equation (5.3), we first perform induction over s and then over t to obtain the claim. To begin, let $t = 0$ (i.e., $\ell = 0$) and assume that for all $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| = s$ we have

$$\partial_{\text{wirt}}^{\mathbf{k}} \left((a^T z)^s \right) = s! \cdot a^{\mathbf{k}}.$$

Let $j \in \{1, \dots, d\}$ be arbitrary and consider

$$\partial_{\text{wirt}}^{\mathbf{k}+e_j} \left((a^T z)^{s+1} \right) = \partial_{\text{wirt}}^{\mathbf{k}} \partial_{\text{wirt}} \left[z_j \mapsto (a^T z)^{s+1} \right].$$

Using the chain rule (5), we get

$$\partial_{\text{wirt}} \left[z_j \mapsto (a^T z)^{s+1} \right] = (s+1) \cdot (a^T z)^s \cdot \partial_{\text{wirt}} [z_j \mapsto a^T z].$$

Here, we also used (2) and the fact that $w \mapsto w^{s+1}$ is holomorphic, which implies the two properties $\partial_{\text{wirt}} [w \mapsto w^{s+1}] = (s+1) \cdot w^s$ and $\bar{\partial}_{\text{wirt}} [w \mapsto w^{s+1}] = 0$. Moreover, since $z_j \mapsto a^T z$ is holomorphic too, we see that $\partial_{\text{wirt}} [z_j \mapsto a^T z] = a_j$. Overall we get

$$\partial_{\text{wirt}}^{\mathbf{k}+e_j} \left((a^T z)^{s+1} \right) = (s+1) \cdot a_j \cdot \partial_{\text{wirt}}^{\mathbf{k}} \left[(a^T z)^s \right] = (s+1)! \cdot a^{\mathbf{k}+e_j},$$

where the induction hypothesis was used for the last equality. Since the case $\mathbf{k} = 0$ (and hence $s = 0$) is trivial, we conclude by induction that

$$\partial_{\text{wirt}}^{\mathbf{k}} \left((a^T z)^s \right) = s! \cdot a^{\mathbf{k}} \quad \text{for all } \mathbf{k} \in \mathbb{N}_0^d \text{ with } |\mathbf{k}| = s.$$

We now perform an additional induction over $|\ell| = t$ to show that

$$\forall \mathbf{k} \in \mathbb{N}_0^d \text{ with } |\mathbf{k}| = s : \quad \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \left((a^T z)^s (\overline{a^T z})^t \right) = s! \cdot t! \cdot a^{\mathbf{k}} \bar{a}^{\ell}.$$

Note that the case $\ell = 0$ (and hence $t = 0$) is shown above. Therefore, we may assume that the claim holds for all $\ell \in \mathbb{N}_0^d$ with $|\ell| = t$, where $t \in \mathbb{N}_0$ is arbitrary but fixed. We let $j \in \{1, \dots, d\}$ and $\mathbf{k} \in \mathbb{N}_0^d$ with $|\mathbf{k}| = s$ be arbitrary and consider

$$\partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell+e_j} \left((a^T z)^s (\overline{a^T z})^{t+1} \right) = \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \bar{\partial}_{\text{wirt}}^{e_j} \left((a^T z)^s (\overline{a^T z})^{t+1} \right).$$

Note that

$$\bar{\partial}_{\text{wirt}}^{e_j} \left((a^T z)^s (\overline{a^T z})^{t+1} \right) = \bar{\partial}_{\text{wirt}} \left[z_j \mapsto (a^T z)^s (\overline{a^T z})^{t+1} \right].$$

The product rule (4) yields

$$\begin{aligned} & \bar{\partial}_{\text{wirt}} \left[z_j \mapsto (a^T z)^s (\overline{a^T z})^{t+1} \right] \\ &= \bar{\partial}_{\text{wirt}} \left[z_j \mapsto (a^T z)^s \right] \cdot (\overline{a^T z})^{t+1} + (a^T z)^s \cdot \bar{\partial}_{\text{wirt}} \left[z_j \mapsto (\overline{a^T z})^{t+1} \right]. \end{aligned}$$

Note that $\bar{\partial}_{\text{wirt}} [z_j \mapsto (a^T z)^s] = 0$ according to (2). Moreover, (3) and an application of the chain rule (5) similar to above gives us

$$\bar{\partial}_{\text{wirt}} \left[z_j \mapsto (\overline{a^T z})^{t+1} \right] = \overline{\partial_{\text{wirt}} [z_j \mapsto (a^T z)^{t+1}]} = (t+1) \cdot \overline{a_j} \cdot (\overline{a^T z})^t.$$

Putting everything together and using the induction hypothesis, we get

$$\begin{aligned} \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell+e_j} \left((a^T z)^s (\overline{a^T z})^{t+1} \right) &= (t+1) \cdot \overline{a_j} \cdot \partial_{\text{wirt}}^{\mathbf{k}} \bar{\partial}_{\text{wirt}}^{\ell} \left((a^T z)^s (\overline{a^T z})^t \right) \\ &= (t+1)! \cdot s! \cdot a^{\mathbf{k}} \overline{a}^{\ell+e_j}. \end{aligned}$$

The principle of induction thus yields the claim. $\square$

Acknowledgements The authors are grateful to Frank Filbir and Hrushikesh Mhaskar for several helpful and interesting discussions and for pointing us to the reference [4]. FV and PG are grateful for the hospitality of Utrecht University, where this project was initiated during their visit. FV and PG acknowledge support by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1. FV acknowledges support by the Hightech Agenda Bavaria. PS is supported by NWO Talent programme Veni ENW grant, file number VI.Veni.212.176.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

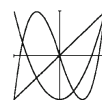
References

1. Bartlett, P.L., Harvey, N., Liaw, C., Mehrabian, A.: Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.* **20**(63), 1–17 (2019)
2. Bassey, J., Qian, L., Li, X.: A Survey of Complex-Valued Neural Networks. 2021. [arXiv: 2101.12249](https://arxiv.org/abs/2101.12249)
3. Candès, E.J.: Ridgelets: estimating with ridge functions. *Ann. Stat.* **31**(5), 1561–1599 (2003)
4. Dai, F., Xu, Y.: *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer, New York (2013)
5. Evans, L. C., Gariepy, R. F.: *Measure theory and fine properties of functions*. Studies in advanced mathematics. Boca Raton: CRC Press, 1992. 268 pp. isbn: 978-0-8493-7157-8

6. Folland, G.: Real Analysis: Modern Techniques and Their Applications. Wiley, 1984. ISBN: 9780471809586
7. Fornasier, M., Vybíral, J., Daubechies, I.: Robust and resource efficient identification of shallow neural networks by fewest samples. In: Information and Inference: A Journal of the IMA 10.2 (Jan. 2021), pp. 625–695. <https://doi.org/10.1093/imaiai/iaaa036>
8. Geuchen, P., Jahn, T., Matt, H.: Universal approximation with complex-valued deep narrow neural networks. 2024. [arXiv: 2305.16910](https://arxiv.org/abs/2305.16910). Accepted for publication in Constructive Approximation
9. Geuchen, P., Voigtlaender, F.: Optimal approximation using complex-valued neural networks. Adv. Neural Inf. Process. Syst. **36**, 15 (2024)
10. Gordon, Y., Maiorov, V., Meyer, M., Reisner, S.: On the best approximation by ridge functions in the uniform norm. Constr. Approx. **18**, 61–85 (2001)
11. Ismailov, V.: Ridge Functions and Applications in Neural Networks. Mathematical Surveys and Monographs. American Mathematical Society, Providence (2021)
12. Ismailov, V.E.: A review of some results on ridge function approximation. Azerb. J. Math. **3**(1), 3–51 (2013)
13. Kaup, L., Kaup, B.: Holomorphic Functions of Several Variables: An Introduction to the Fundamental Theory. De Gruyter (1983). <https://doi.org/10.1515/9783110838350>
14. Lee, C., Hasegawa, H., Gao, S.: Complex-valued neural networks: A comprehensive survey. IEEE/CAA J. Autom. Sinica **9**(8), 1406–1426 (2022)
15. Lin, V.Y., Pinkus, A.: Fundamentality of ridge functions. J. Approx. Theory **75**(3), 295–311 (1993)
16. Luenberger, D.: Optimization by Vector Space Methods. Professional Series. Wiley, 1997. ISBN: 978-0-4711-8117-0
17. Maiorov, V.: On Best Approximation by Ridge Functions. J. Approx. Theory **99**(1), 68–94 (1999). <https://doi.org/10.1006/jath.1998.3304>
18. Maiorov, V.: Best approximation by ridge functions in L_p -spaces. Ukr. Math. J. **62**, 452–466 (2010)
19. Maiorov, V., Pinkus, A.: Lower bounds for approximation by MLP neural networks. Neurocomputing **25**(1–3), 81–91 (1999)
20. Mhaskar, H., Prestin, J.: “On a choice of sampling nodes for optimal approximation of smooth functions by generalized translation networks.” In: Fifth International Conference on Artificial Neural Networks (Conf. Publ. No. 440). 1997, pp. 210–215. <https://doi.org/10.1049/cp:19970728>
21. Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. Neural Comput. **8**(1), 164–177 (1996)
22. Mhaskar, H.N., Micchelli, C.A.: Degree of approximation by neural and translation networks with a single hidden layer. Adv. Appl. Math. **16**(2), 151–183 (1995)
23. Muleshkov, A., Nguyen, T.: Easy proof of the Jacobian for the n -dimensional polar coordinates. Pi Mu Epsilon J. **14**(4), 269–273 (2016)
24. Petersen, P., Voigtlaender, F.: Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Netw. **108**, 296–330 (2018)
25. Pinkus, A.: “Approximating by ridge functions.” Surface fitting and multiresolution methods (1997), pp. 279–292
26. Pinkus, A.: Ridge functions. Cambridge: Cambridge University Press, 2016. ISBN: 978-1-107-12439-4
27. Prudnikov, A., Brychkov, Y., Marichev, O.: Integrals and Series: Volume 1: Elementary Functions. Taylor & Francis, 1986. ISBN: 9782881240973
28. Safran, I., Shamir, O.: “Depth-width tradeoffs in approximating natural functions with neural networks.” In: International conference on machine learning. PMLR. 2017, pp. 2979–2987
29. Shen, Z., Yang, H., Zhang, S.: Deep Network Approximation Characterized by Number of Neurons. Commun. Comput. Phys. **285**, 1768–1811 (2020). <https://doi.org/10.4208/cicp.OA-2020-0149>
30. Shen, Z., Yang, H., Zhang, S.: Optimal approximation rate of ReLU networks in terms of width and depth. Journal de Mathématiques Pures et Appliquées **157**, 101–135 (2022). <https://doi.org/10.1016/j.matpur.2021.07.009>
31. Telgarsky, M.: “Benefits of depth in neural networks.” Conference on learning theory. PMLR. 2016, pp. 1517–1539
32. Vershynin, R.: High-dimensional probability: an introduction with applications in data science. Cambridge University Press, Cambridge (2018)
33. Voigtlaender, F.: The universal approximation theorem for complex-valued neural networks. Appl. Comput. Harmon. Anal. **64**, 33–61 (2023)

34. Yarotsky, D., Zhevnerchuk, A.: The phase diagram of approximation rates for deep neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 13005–13015 (2020)
35. Zhang, S., Shen, Z., Yang, H.: Deep network approximation: achieving arbitrary accuracy with fixed number of neurons. *J. Mach. Learning Res.* **23**(276), 1–60 (2022)

B.5. Universal approximation using complex-valued deep narrow neural networks



Universal approximation with complex-valued deep narrow neural networks

Paul Geuchen¹ · Thomas Jahn² · Hannes Matt¹

Received: 17 July 2023 / Revised: 26 November 2024 / Accepted: 15 March 2025
© The Author(s) 2025

Abstract

We study the universality of complex-valued neural networks with bounded widths and arbitrary depths. Under mild assumptions, we give a full description of those activation functions $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ that have the property that their associated networks are universal, i.e., are capable of approximating continuous functions to arbitrary accuracy on compact domains. Precisely, we show that deep narrow complex-valued networks are universal if and only if their activation function is neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine. This is a much larger class of functions than in the dual setting of arbitrary width and fixed depth. Unlike in the real case, the sufficient width differs significantly depending on the considered activation function. We show that a width of $2n + 2m + 5$ is always sufficient and that in general a width of $\max \{2n, 2m\}$ is necessary. We prove, however, that a width of $n + m + 3$ suffices for a rich subclass of the admissible activation functions. Here, n and m denote the input and output dimensions of the considered networks. Moreover, for the case of smooth and non-polyharmonic activation functions, we provide a quantitative approximation bound in terms of the depth of the considered networks.

Keywords Complex-valued neural networks · Holomorphic function · Polyharmonic function · Uniform approximation · Universality

Communicated by Wolfgang Dahmen.

All authors contributed equally to this work.

✉ Hannes Matt
hannes.matt@ku.de

Paul Geuchen
paul.geuchen@ku.de

Thomas Jahn
jahn.thomas@unijena.de

¹ Mathematical Institute for Machine Learning and Data Science (MIDS), Catholic University of Eichstätt–Ingolstadt (KU), Auf der Schanz 49, 85049 Ingolstadt, Germany

² Institute for Mathematics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 1-2, 07737 Jena, Germany

Mathematics Subject Classification 68T07 · 41A30 · 41A63 · 31A30 · 30E10

1 Introduction

This paper addresses the universality of deep narrow complex-valued neural networks (CVNNs), i.e., the density of neural networks with arbitrarily large depths but bounded widths, in spaces of continuous functions over compact domains with respect to the uniform norm. Our main theorem is as follows.

Theorem 1.1 *Let $n, m \in \mathbb{N}$, and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ be a continuous function which at some point is real differentiable with non-vanishing derivative. Then $\mathcal{NN}_{n,m,2n+2m+5}^{\varrho}$ is universal if and only if ϱ is neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine.*

Here $\mathcal{NN}_{n,m,W}^{\varrho}$ denotes the set of complex-valued neural networks with input dimension n , output dimension m , activation function ϱ , and W neurons per hidden layer. These neural networks are alternating compositions

$$V_L \circ \varrho^{\times W} \circ \dots \circ \varrho^{\times W} \circ V_0 : \mathbb{C}^n \rightarrow \mathbb{C}^m$$

of affine maps $V_0 : \mathbb{C}^n \rightarrow \mathbb{C}^W$, $V_1, \dots, V_{L-1} : \mathbb{C}^W \rightarrow \mathbb{C}^W$, $V_L : \mathbb{C}^W \rightarrow \mathbb{C}^m$, and componentwise applications of the activation function ϱ , see Section 2 for a detailed definition.

Studying the expressivity of neural networks is an important part of the mathematical analysis of deep learning. Theorem 1.1 is a *qualitative* result in that direction. Such qualitative results naturally precede the investigation of approximation rates, i.e., the decay of approximation errors as the class of approximants increases. Our focus is on qualitative results, but to show how our methods can also be used to derive quantitative bounds, we prove such a result for the case of smooth and non-polyharmonic activation functions, i.e., we derive an upper bound on the depth necessary to achieve an approximation accuracy less than ε , for a prescribed approximation accuracy $\varepsilon > 0$; see Theorem 5.3.

1.1 Complex-valued neural networks

Although mostly real-valued neural networks (RVNNs) are used in the field of Deep Learning, recent years have shown a growing interest in the use of complex-valued neural networks in various application areas [2, 20, 21, 32, 40], for instance Magnetic Resonance Imaging (MRI) [9, 19, 36] and Polarimetric Synthetic Aperture Radar (PolSAR) Imaging [3, 33, 41]. These application areas are usually characterized by the fact that complex numbers naturally occur as inputs for machine learning models. In such areas, CVNNs are, in contrast to real-valued neural networks, able to handle the complex-valued nature of the inputs in a faithful way, for instance by using a phase-preserving activation function. Note that this behavior cannot be achieved if one applies a non-trivial real-valued activation function to real and imaginary part of an input separately.

The identification $\mathbb{R}^2 \cong \mathbb{C}$ might suggest that most theoretical properties of CVNNs can directly be derived by those of RVNNs. However, the two network classes differ in the following two deciding aspects:

- The activation function in a CVNN is a function $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ (i.e., $\mathbb{R}^2 \rightarrow \mathbb{R}^2$), whereas the activation function in an RVNN is a function $\mathbb{R} \rightarrow \mathbb{R}$. Hence, regarding the activation function, CVNNs are more *versatile* than RVNNs.
- The affine maps in a CVNN are required to be $\mathbb{C}$ -affine, whereas the affine maps in an RVNN only need to be $\mathbb{R}$ -affine. Therefore, regarding the affine maps, CVNNs are more *restrictive* than RVNNs.

The observation that CVNNs are on the one hand more versatile and on the other hand more restrictive than RVNNs shows that it is *not* possible to obtain theoretical properties of CVNNs as a special case of those of RVNNs or vice versa. In fact, studying the universality of neural networks of fixed depth [22, 37] has already uncovered significant differences between RVNNs and CVNNs; see also Section 1.3.

1.2 Related work

In the neural network context, universal approximation theorems date back to the 1980s and 1990s [10, 22], where it was shown that real-valued shallow neural networks with output dimension 1 and a fixed continuous activation function are universal if and only if the activation function is not a polynomial. Modifications of the setting in which universal approximation is studied appear in the neural network literature over the past decades. These variants of the problem refer to, e.g., the input and the output dimension, the target space (typically L_p for $1 \leq p \leq \infty$, continuous functions, also modulo the action of a group), the choice of activation functions (only ReLU vs. any continuous non-polynomial function), constraints on either the width or depth of the neural network (narrow vs. wide, shallow vs. deep networks), or constraints on the norm or the sparsity of the weights, see for instance [7, 15, 17, 24, 27, 29, 39] and the references therein. Furthermore, changes in the network architecture [31, 42], the incorporation of randomness [25], and changes in the nature of the inputs [6, 23] are also subjects of investigation in the literature on universal approximation.

Moreover, the literature contains numerous quantitative statements about the approximation properties of neural networks (see for instance [4, 26, 30, 38] and the references therein). We explicitly mention the paper [18], which provides quantitative approximation bounds for deep narrow RVNNs in terms of the depth of the considered networks (see [18, Proposition 53]). In the present work, we prove a statement similar to [18, Proposition 53(i)] for CVNNs with a smooth and non-polyharmonic activation function (see Theorem 5.3).

Remarkably, the theory mostly covers real-valued neural networks. Yet, the fact that CVNNs are applied successfully in various application areas (see Section 1.1) motivates the theoretical study of their approximative capabilities. To the best of our knowledge, the only qualitative and quantitative results in that direction are [28, 37], and [8, 13], respectively. The article [37] provides a characterization of activation functions, for which shallow CVNNs are universal. This characterization is crucial for the purposes of the paper at hand and is therefore given in Theorem 4.1 below. For both

the shallow and the deep narrow (real-valued and complex-valued) neural networks with analytic activation functions, it is shown in [28] that the closures of these classes in $C(K)$ coincides with the closure of polynomials, where K is a compact subset of $\mathbb{R}^n$ or $\mathbb{C}^n$. An application of the Stone–Weierstrass theorem or Mergelyan’s theorem then yields universality of neural networks in the set of continuous functions in the real-valued case or holomorphic functions in the complex-valued case. The authors of [8] prove quantitative bounds for the approximation of C^k -functions on $\mathbb{C}^n$ using complex-valued neural networks with the modReLU activation function. Those results have recently been generalized in [13], where the same approximation bounds have been proven for the rich class of complex-valued activation functions that are smooth and non-polyharmonic on some non-empty open set. This class in particular includes the modReLU.

1.3 Contribution

In the real-valued case, under mild assumptions on their regularity, activation functions that yield universal neural networks have been characterized in the literature. In the complex-valued case, however, such a characterization is only known in the case of neural networks with fixed depths and arbitrary widths. To complete the picture, we give in Theorem 1.1 a characterization of activation functions for which CVNNs with bounded widths and arbitrary depths are universal.

Recall that polynomial activation functions are precisely the ones for which real-valued neural networks with arbitrary widths and fixed depth are not universal, as shown by Kidger and Lyons in [17, Section 1]. Yet, in the dual situation, where depth is arbitrary and widths are bounded, polynomial activation functions (of minimum degree 2) do give rise to universal real-valued neural networks, see again [17, Theorem 3.2]. The situation is different in the complex-valued case. For continuous activation functions, Voigtlaender shows in [37, Theorem 1.3] that shallow CVNNs are universal if and only if the activation function is non-polyharmonic. CVNNs of arbitrary widths and fixed depth > 1 are universal if and only if the activation function does not coincide with a polynomial in z and $\bar{z}$, and is neither holomorphic nor anti-holomorphic, cf. [37, Theorem 1.4]. In the deep narrow regime studied in Theorem 1.1, the requirements on the activation function for universality are again weaker in the sense that not being a polynomial in z and $\bar{z}$ is replaced by not being $\mathbb{R}$ -affine.

In this work, we consider continuous complex-valued activation functions which have non-vanishing derivative (in the sense of real variables) at some point. Our analysis roughly splits into two parts: polyharmonic and non-polyharmonic activation functions. While this distinction mainly impacts the proof techniques, the bounds on the widths of the CVNNs are actually governed by the properties of the Wirtinger derivatives of the activation function at the point of differentiability. Fig. 1 is a graphical guide through our results.

For non-polyharmonic activation functions, we show universality of CVNNs with input dimension n and output dimension m , where the number of neurons per hidden layer is $2n + 2m + 1$ or even $n + m + 1$, see Theorem 4.2. The proofs in that case are based on the fact that for non-polyharmonic activation functions shallow CVNNs

Is $\varrho \in C(\mathbb{C}; \mathbb{C})$ neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine?

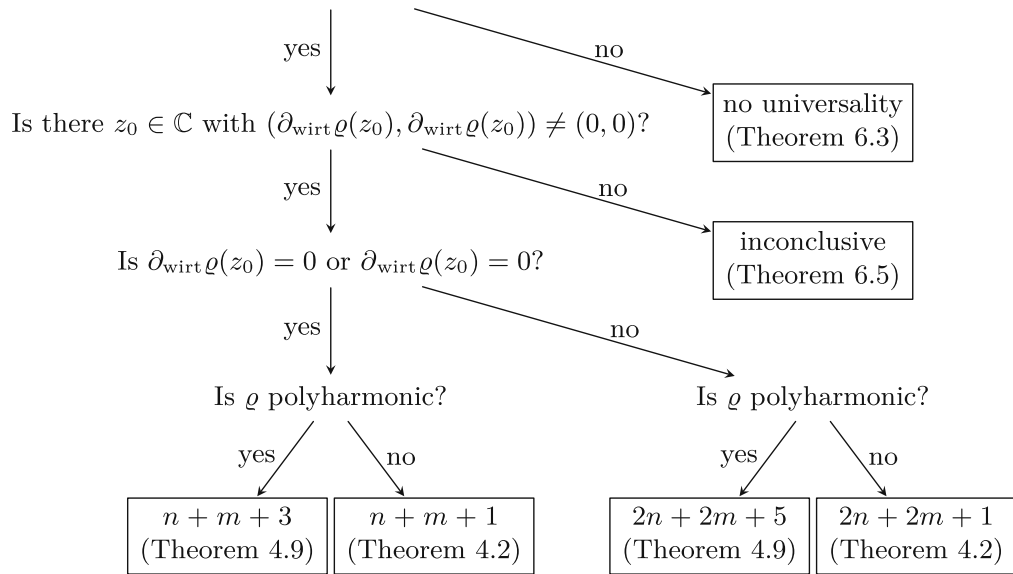


Fig. 1 Our results in a nutshell

are universal, as shown by Voigtlaender in [37, Theorem 1.3]. We combine this with an adaptation of the register model technique developed by Kidger and Lyons [17, Theorem 3.2]. There, they used this technique to deduce universality of deep narrow *real-valued* neural networks from the classical result on universality of shallow *real-valued* neural networks [10, 22]. For polyharmonic activation functions, we show that CVNNs with input dimension n , output dimension m , and $2n + 2m + 5$ or even $n + m + 3$ neurons per hidden layer are universal, see Theorem 4.9. This is done by approximating polynomials in z and $\bar{z}$ uniformly on compact sets and invoking the Stone–Weierstrass theorem.

Moreover, based on the ideas from [18, Proposition 59], we provide a quantitative approximation bound in terms of the depth of the considered networks for the case of a smooth and non-polyharmonic activation function $\varrho \in C(\mathbb{C}; \mathbb{C})$. Precisely, given a function $f \in C([-1, 1]^n + i \cdot [-1, 1]^n; \mathbb{C}^m)$ and $\varepsilon > 0$, we show that one can approximate f up to precision ε with deep narrow networks using activation function ϱ and a depth of at most

$$\left(32 \cdot \left[\omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} \cdot (1 + \frac{n}{2})} \right) \right]^{-2} + 9 \right)^{2n};$$

see Theorem 5.3. Here, $\omega^{-1}(f, \cdot)$ denotes the inverse modulus of continuity of the function f , see (5.1). We believe that our techniques combined with those from [18, Appendix B.2] can also be used to derive depth estimates for deep narrow CVNNs for the case of more general activation functions $\varrho \in C(\mathbb{C}; \mathbb{C})$. However, since the present paper focuses on the aspect of universality, this is left as future work.

1.4 Organization of our paper

In Section 2 we fix our notation and recall some basics from complex and functional analysis and the theory of neural networks. Section 3 introduces the register model and shows how the identity map on $\mathbb{C}$ and complex conjugation can be approximated using CVNNs. For non-polyharmonic functions, the proof of Theorem 1.1 can be found in Section 4.1. In Section 4.2, we present the proof of universality claimed in Theorem 1.1 in the case of polyharmonic activation functions which are neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine. In Section 5, we prove the quantitative approximation bound in terms of the depth of the considered networks. In Section 6, we show that CVNNs whose activation function is holomorphic or antiholomorphic or $\mathbb{R}$ -affine are never universal, regardless of the number of neurons per hidden layer. Moreover, we show that there exist activation functions satisfying the assumptions from Theorem 1.1 for which a width of $\max\{2n, 2m\}$ is necessary in order to provide universal CVNNs. In the appendix, we provide basics on the relationship between local uniform convergence and universal approximation, and on Taylor approximations in terms of Wirtinger derivatives.

2 Preliminaries

In this section, we recall facts from complex analysis, functional analysis, and the theory of neural networks behind the phrases in Theorem 1.1. The presentation is loosely based on [34, Chapter 7], [35, Chapter 11], and [14, Section 1].

2.1 Complex and functional analysis

We use the symbols $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$ to denote the natural, real, and complex numbers, respectively. By $\operatorname{Re}(z)$, $\operatorname{Im}(z)$, and $\bar{z}$, we denote the componentwise real part, imaginary part, and complex conjugate of a vector $z \in \mathbb{C}^n$, respectively. We call the function $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ *partially differentiable* at z_0 , if the *partial derivatives*

$$\frac{\partial \varrho}{\partial x}(z_0) := \lim_{\mathbb{R} \setminus \{0\} \ni h \rightarrow 0} \frac{\varrho(z_0 + h) - \varrho(z_0)}{h}$$

and

$$\frac{\partial \varrho}{\partial y}(z_0) := \lim_{\mathbb{R} \setminus \{0\} \ni h \rightarrow 0} \frac{\varrho(z_0 + ih) - \varrho(z_0)}{h}$$

exist. Higher-order partial derivatives are defined in the standard manner. We write $\varrho \in C^k(\mathbb{C}; \mathbb{C})$ if ϱ admits partial derivatives up to order k at each point of $\mathbb{C}$ and the k th-order partial derivatives are continuous functions $\mathbb{C} \rightarrow \mathbb{C}$. Likewise, we write $\varrho \in C^\infty(\mathbb{C}; \mathbb{C})$ if $\varrho \in C^k(\mathbb{C}; \mathbb{C})$ for all $k \in \mathbb{N}$. If $\frac{\partial \varrho}{\partial x}(z_0)$ and $\frac{\partial \varrho}{\partial y}(z_0)$ exist and the identity

$$\lim_{\mathbb{C} \setminus \{0\} \ni h \rightarrow 0} \frac{\varrho(z_0 + h) - \varrho(z_0) - \frac{\partial \varrho}{\partial x}(z_0) \operatorname{Re}(h) - \frac{\partial \varrho}{\partial y}(z_0) \operatorname{Im}(h)}{h} = 0 \quad (2.1)$$

holds true, then ϱ is called *real differentiable at z_0* with derivative $(\frac{\partial \varrho}{\partial x}(z_0), \frac{\partial \varrho}{\partial y}(z_0))$. Similarly, ϱ is called *complex differentiable at z_0* if

$$\lim_{\mathbb{C} \setminus \{0\} \ni h \rightarrow 0} \frac{\varrho(z_0 + h) - \varrho(z_0) - ch}{h} = 0$$

for some number $c \in \mathbb{C}$, which in that case is given by

$$\partial_{\text{wirt}} \varrho(z_0) := \frac{1}{2} \left(\frac{\partial \varrho}{\partial x}(z_0) - i \frac{\partial \varrho}{\partial y}(z_0) \right).$$

Complex differentiability of ϱ at z_0 can be equivalently stated as

$$\bar{\partial}_{\text{wirt}} \varrho(z_0) := \frac{1}{2} \left(\frac{\partial \varrho}{\partial x}(z_0) + i \frac{\partial \varrho}{\partial y}(z_0) \right) = 0.$$

The differential operators ∂_{wirt} and $\bar{\partial}_{\text{wirt}}$ are called *Wirtinger derivatives*. If $\bar{\partial}_{\text{wirt}} \varrho(z) = 0$ for all $z \in \mathbb{C}$, then ϱ is a *holomorphic* function. The function ϱ is called *antiholomorphic* if the function $\bar{\varrho} : \mathbb{C} \rightarrow \mathbb{C}$, $\bar{\varrho}(z) := \text{Re}(\varrho(z)) - i \text{Im}(\varrho(z))$ is holomorphic or, equivalently, $\partial_{\text{wirt}} \varrho(z) = 0$ for all $z \in \mathbb{C}$. As the linear operator that maps the partial derivatives onto the Wirtinger derivatives is invertible, it follows that $(\frac{\partial \varrho}{\partial x}(z_0), \frac{\partial \varrho}{\partial y}(z_0)) = (0, 0)$ if and only if $(\partial_{\text{wirt}} \varrho(z_0), \bar{\partial}_{\text{wirt}} \varrho(z_0)) = (0, 0)$. Furthermore, the symmetry of mixed partial derivatives implies for $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ that

$$4\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho = 4\bar{\partial}_{\text{wirt}} \partial_{\text{wirt}} \varrho = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \varrho =: \Delta \varrho. \quad (2.2)$$

If $\varrho \in C^\infty(\mathbb{C}; \mathbb{C})$ and $\Delta^m \varrho = 0$ for some $m \in \mathbb{N}$, then ϱ is called *polyharmonic of order m* . Because of (2.2), holomorphic and antiholomorphic functions are *harmonic*, i.e., polyharmonic of order 1.

The following well-known lemma generalizes the classical real-valued Taylor expansion to the complex-valued setting; see Lemma B.1 for a proof.

Lemma 2.1 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$ and $z, z_0 \in \mathbb{C}$. If ϱ is real differentiable at z_0 , then*

$$\varrho(z + z_0) = \varrho(z_0) + \partial_{\text{wirt}} \varrho(z_0)z + \bar{\partial}_{\text{wirt}} \varrho(z_0)\bar{z} + \Theta_1(z) \quad (2.3)$$

for a function $\Theta_1 : \mathbb{C} \rightarrow \mathbb{C}$ with $\lim_{\mathbb{C} \setminus \{0\} \ni z \rightarrow 0} \frac{\Theta_1(z)}{z} = 0$. If $\varrho \in C^2(\mathbb{C}; \mathbb{C})$, then

$$\begin{aligned} \varrho(z + z_0) &= \varrho(z_0) + \partial_{\text{wirt}} \varrho(z_0)z + \bar{\partial}_{\text{wirt}} \varrho(z_0)\bar{z} + \frac{1}{2} \partial_{\text{wirt}}^2 \varrho(z_0)z^2 + \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0)z\bar{z} \\ &\quad + \frac{1}{2} \bar{\partial}_{\text{wirt}}^2 \varrho(z_0)\bar{z}^2 + \Theta_2(z) \end{aligned} \quad (2.4)$$

for a function $\Theta_2 : \mathbb{C} \rightarrow \mathbb{C}$ with $\lim_{\mathbb{C} \setminus \{0\} \ni z \rightarrow 0} \frac{\Theta_2(z)}{z^2} = 0$.

On $\mathbb{C}^m$, we consider the topology induced by the Euclidean norm

$$\|(z_1, \dots, z_m)\|_{\mathbb{C}^m} = \left(\sum_{j=1}^m |z_j|^2 \right)^{\frac{1}{2}}.$$

For $K \subseteq \mathbb{C}^n$, we denote the vector space of continuous functions $K \rightarrow \mathbb{C}^m$ by $C(K; \mathbb{C}^m)$. When $K \subseteq \mathbb{C}^n$ is compact, the expression

$$\|f\|_{C(K; \mathbb{C}^m)} := \sup_{z \in K} \|f(z)\|_{\mathbb{C}^m}$$

defines a norm on $C(K; \mathbb{C}^m)$, called the *uniform norm*, which renders $C(K; \mathbb{C}^m)$ a Banach space. The convergence of a sequence $(f_j)_{j \in \mathbb{N}}$ of elements $f_j \in C(K; \mathbb{C}^m)$ to a limit $f \in C(K; \mathbb{C}^m)$ with respect to $\|\cdot\|_{C(K; \mathbb{C}^m)}$ is written as $f_j \rightarrow f$ as $j \rightarrow \infty$, *uniformly on K* . Similarly, a sequence $(f_j)_{j \in \mathbb{N}}$ of functions $f_j \in C(\mathbb{C}^n; \mathbb{C}^m)$ is said to converge *locally uniformly* to a function $f \in C(\mathbb{C}^n; \mathbb{C}^m)$ if it converges uniformly to f on every compact subset $K \subseteq \mathbb{C}^n$. Since compositions of continuous functions are continuous, we have $\mathcal{NN}_{n,m,W}^{\varrho} \subseteq C(\mathbb{C}^n; \mathbb{C}^m)$ when $\varrho \in C(\mathbb{C}; \mathbb{C})$. The main objective of the paper at hand is to show that under certain additional assumptions on ϱ and W , the elements of $\mathcal{NN}_{n,m,W}^{\varrho}$ are arbitrarily close to the elements of $C(\mathbb{C}^n; \mathbb{C}^m)$ in the following sense.

Definition 2.2 We say that a function class $\mathcal{F} \subseteq C(\mathbb{C}^n; \mathbb{C}^m)$ has the *universal approximation property* (or is *universal*) if for every function $g \in C(\mathbb{C}^n; \mathbb{C}^m)$, every compact subset $K \subseteq \mathbb{C}^n$ and every $\varepsilon > 0$ there exists a function $f \in \mathcal{F}$ such that

$$\sup_{z \in K} \|f(z) - g(z)\|_{\mathbb{C}^m} < \varepsilon.$$

The universal approximation property of $\mathcal{F}$ is equivalent to saying that for every given function $g \in C(\mathbb{C}^n; \mathbb{C}^m)$ there exists a sequence $(f_j)_{j \in \mathbb{N}}$ with $f_j \in \mathcal{F}$ for every $j \in \mathbb{N}$ that converges locally uniformly to g . Likewise, this is equivalent to saying that the class $\mathcal{F}$ is dense in $C(\mathbb{C}^n; \mathbb{C}^m)$ with respect to the *compact-open topology*. We elaborate this equivalence in Appendix A. By $\overline{\mathcal{F}}$ we denote the closure of $\mathcal{F}$ with respect to the compact-open topology. Notice that we also use the notation $\bar{z}$ to denote the complex conjugate of $z \in \mathbb{C}$. The precise meaning will be clear from the context.

2.2 Neural networks

A (fully connected feed-forward) *complex-valued neural network* (CVNN) is a function

$$V_L \circ \varrho^{\times N_L} \dots \circ \varrho^{\times N_1} \circ V_0 : \mathbb{C}^{N_0} \rightarrow \mathbb{C}^{N_{L+1}}$$

where

- $L \in \mathbb{N}$ is called the *depth* of the CVNN,
- $N_j \in \mathbb{N}$ is the *width* of the j th layer,

- $\max \{N_0, \dots, N_{L+1}\}$ is the *width* of the CVNN,
- $V_j : \mathbb{C}^{N_j} \rightarrow \mathbb{C}^{N_{j+1}}$ is a $\mathbb{C}$ -affine map, abbreviated as $V_j \in \text{Aff}(\mathbb{C}^{N_j}; \mathbb{C}^{N_{j+1}})$, i.e., there exist $A_j \in \mathbb{C}^{N_{j+1} \times N_j}$ and $b_j \in \mathbb{C}^{N_{j+1}}$ such that $V_j(z) = A_j z + b_j$ for all $z \in \mathbb{C}^{N_j}$,
- $\varrho^{\times N_j}(z_1, \dots, z_{N_j}) = (\varrho(z_1), \dots, \varrho(z_{N_j}))$ is the componentwise application of a (potentially non-affine) map $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ called the *activation function*.

We refer to the numbers N_0 and N_{L+1} as the *input dimension* and *output dimension*, respectively. The numbers $N_1, \dots, N_L$ are the widths of the hidden layers of the CVNN. Since it is always possible to pad matrices and vectors by additional zero rows and columns, we may and will assume without loss of generality that $N_1 = N_2 = \dots = N_L$.

We introduce a short-hand notation for the CVNNs that arise this way.

Definition 2.3 Let $n, m, W, L \in \mathbb{N}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$. We denote by $\mathcal{NN}_{n,m,W,L}^\varrho$ the set of CVNNs with depth L , input dimension n , output dimension m , and $N_j = W$ for $j \in \{1, \dots, L\}$. In view of cases where the depth is not relevant, we let

$$\mathcal{NN}_{n,m,W}^\varrho := \bigcup_{L \in \mathbb{N}} \mathcal{NN}_{n,m,W,L}^\varrho.$$

The elements of $\mathcal{NN}_{n,m,W}^\varrho$ are thus alternating compositions

$$V_L \circ \varrho^{\times W} \circ \dots \circ \varrho^{\times W} \circ V_0 : \mathbb{C}^n \rightarrow \mathbb{C}^m \quad (2.5)$$

of $\mathbb{C}$ -affine maps V_j and the componentwise applications of the activation function ϱ . In the subsequent Sections 4.2 and 4.1, we have $W \geq \max \{n, m\}$, such that W turns out to be the width of the neural networks under consideration.

A typical way of thinking about neural networks is viewing the component functions of $\varrho^{\times N_{j+1}} \circ V_j$ as building blocks called *neurons*. Each neuron performs a computation of the form

$$z \mapsto \varrho(w^\top z + b)$$

where z is the output of the previous layer, w a vector of *weights*, and b a number called *bias*.

Since the composition of affine maps is affine, it is also possible to think about neural networks as maps

$$(\Psi_L \circ \varrho^{\times W} \circ \Phi_L) \circ (\Psi_{L-1} \circ \varrho^{\times W} \circ \Phi_{L-1}) \circ \dots \circ (\Psi_1 \circ \varrho^{\times W} \circ \Phi_1),$$

where each of the maps Φ_k, Ψ_k is affine. This allows to perceive shallow networks, see Definition 2.4, as building blocks for neural networks. This point of view is similar to the notion of *enhanced neurons* in [17].

For a fixed activation function ϱ , different choices of the $\mathbb{C}$ -affine functions V_j may lead to the same composite function (2.5). In view of this, both depth and width of a CVNN are not properties of the function (2.5) but of the tuple $(V_1, \dots, V_L)$. For this

reason, a different terminology is sometimes used in the literature where $(V_1, \dots, V_L)$ is called the *neural network* and (2.5) is its *realization*, cf. [14, Section 1].

Apart from the choice of the activation function ϱ , restrictions on the depth or the width are common ingredients in the analysis of (fully connected feed-forward) neural networks. A CVNN is called *shallow* if its depth equals 1, and *deep* otherwise. Since shallow networks play a special role in our analysis, we introduce an own notation for them.

Definition 2.4 Let $n, m, W \in \mathbb{N}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$. We denote by

$$\mathcal{SN}_{n,m,W}^{\varrho} := \mathcal{NN}_{n,m,W,1}^{\varrho}$$

the set of *shallow CVNNs* with W hidden neurons. We write

$$\mathcal{SN}_{n,m}^{\varrho} := \bigcup_{W \in \mathbb{N}} \mathcal{SN}_{n,m,W}^{\varrho}.$$

In contrast to shallowness, *narrowness* is not an individual property of CVNNs but a class property. A set $\mathcal{F}$ of CVNNs is said to be *narrow* if it does not contain CVNNs of arbitrarily large widths, i.e., if $\mathcal{F} \subseteq \mathcal{NN}_{n,m,W}^{\varrho}$ for suitable $n, m, W \in \mathbb{N}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$.

3 Building blocks and register model

In this section, we prove several preliminary statements that are crucial for the results derived in Section 5 and 4. Specifically, in Section 3.1, we construct building blocks to approximate elementary functions locally uniformly by shallow CVNNs. In Section 3.2, we consider the fundamental concept of the register model to transform shallow networks into deep narrow networks.

3.1 Building blocks

In this section we introduce various *building blocks* for complex-valued networks, i.e., small neural network blocks that are able to represent elementary functions (e.g., the complex identity $\text{id}_{\mathbb{C}}$ or complex conjugation $\overline{\text{id}_{\mathbb{C}}}$) up to an arbitrarily small approximation error. These building blocks are used in Sections 4.2 and 4.1 to construct the deep narrow networks that we use to approximate a given continuous function. Throughout the chapter, we assume that the used activation function $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ is differentiable (in the real sense) at one point with non-vanishing derivative at that point. In fact, the strategy for constructing these building blocks is always going to be similar: By using the first- and second-order Taylor expansion of the activation function ϱ as introduced in Lemma 2.1 one can localize the activation function around its point of differentiability where it behaves like a complex polynomial in z and $\bar{z}$ of degree 1 and 2, respectively. This enables us to extract elementary functions from that Taylor expansion.

Proposition 3.1 is fundamental for the universality results introduced in Sections 4.2 and 4.1. It shows that it is possible to uniformly approximate the complex identity or complex conjugation on compact sets using neural networks with a single hidden layer and width at most 2. If the activation function (or its complex conjugate) is not just real but even complex differentiable, the width can be reduced to 1. See Fig. 2 for an illustration of the building blocks.

Proposition 3.1 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$. Assume that there exists $z_0 \in \mathbb{C}$ such that ϱ is real differentiable at z_0 with $(\partial_{\text{wirt}}\varrho(z_0), \bar{\partial}_{\text{wirt}}\varrho(z_0)) \neq (0, 0)$. Furthermore, let $K \subseteq \mathbb{C}$ be compact and $\varepsilon > 0$.*

(i) *If $\partial_{\text{wirt}}\varrho(z_0) \neq 0$ and $\bar{\partial}_{\text{wirt}}\varrho(z_0) = 0$, there exist $\phi, \psi \in \text{Aff}(\mathbb{C}; \mathbb{C})$ such that*

$$\sup_{z \in K} |(\psi \circ \varrho \circ \phi)(z) - z| < \varepsilon.$$

(ii) *If $\partial_{\text{wirt}}\varrho(z_0) = 0$ and $\bar{\partial}_{\text{wirt}}\varrho(z_0) \neq 0$, there exist $\phi, \psi \in \text{Aff}(\mathbb{C}; \mathbb{C})$ such that*

$$\sup_{z \in K} |(\psi \circ \varrho \circ \phi)(z) - \bar{z}| < \varepsilon.$$

(iii) *If $\partial_{\text{wirt}}\varrho(z_0) \neq 0 \neq \bar{\partial}_{\text{wirt}}\varrho(z_0)$, there exist $\phi \in \text{Aff}(\mathbb{C}; \mathbb{C}^2)$ and $\psi \in \text{Aff}(\mathbb{C}^2; \mathbb{C}^2)$ such that*

$$\sup_{z \in K} \left\| (\psi \circ \varrho^{\times 2} \circ \phi)(z) - (z, \bar{z}) \right\|_{\mathbb{C}^2} < \varepsilon$$

Proof Recall that Lemma 2.1 yields the existence of a function $\Theta : \mathbb{C} \rightarrow \mathbb{C}$ satisfying $\lim_{z \rightarrow 0} \frac{\Theta(z)}{z} = 0$ and

$$\varrho(z + z_0) = \varrho(z_0) + \partial_{\text{wirt}}\varrho(z_0)z + \bar{\partial}_{\text{wirt}}\varrho(z_0)\bar{z} + \Theta(z)$$

for every $z \in \mathbb{C}$.

If $\bar{\partial}_{\text{wirt}}\varrho(z_0) = 0$ and $\partial_{\text{wirt}}\varrho(z_0) \neq 0$ we see for all $h > 0$ and $z \in K$ that

$$\frac{\varrho(z_0 + hz) - \varrho(z_0)}{\partial_{\text{wirt}}\varrho(z_0)h} = z + \frac{\Theta(hz)}{\partial_{\text{wirt}}\varrho(z_0)h}. \quad (3.1)$$

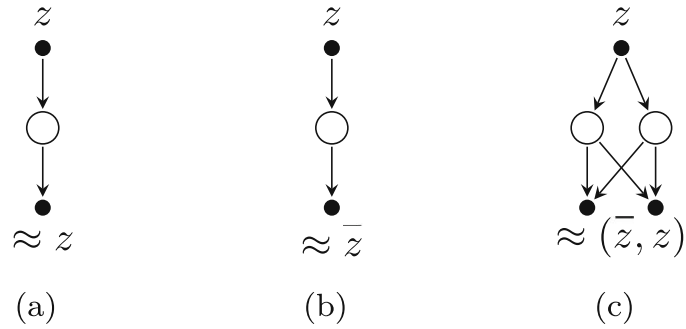
Similarly, if $\bar{\partial}_{\text{wirt}}\varrho(z_0) \neq 0$ and $\partial_{\text{wirt}}\varrho(z_0) = 0$, we get for all $h > 0$ and $z \in K$ that

$$\frac{\varrho(z_0 + hz) - \varrho(z_0)}{\bar{\partial}_{\text{wirt}}\varrho(z_0)h} = \bar{z} + \frac{\Theta(hz)}{\bar{\partial}_{\text{wirt}}\varrho(z_0)h}. \quad (3.2)$$

If $\partial_{\text{wirt}}\varrho(z_0) \neq 0 \neq \bar{\partial}_{\text{wirt}}\varrho(z_0)$, consider

$$\frac{i\varrho(z_0 + hz) + \varrho(z_0 + ihz) - (1 + i)\varrho(z_0)}{2ih\partial_{\text{wirt}}\varrho(z_0)} = z + \frac{i\Theta(hz) + \Theta(ihz)}{2ih\partial_{\text{wirt}}\varrho(z_0)} \quad (3.3)$$

Fig. 2 Illustration of the neural network building blocks from Proposition 3.1. Neurons in the input and output layers are depicted in filled dots at the top and bottom, respectively. Applications of the activation function ϱ are shown as circles



as well as

$$\frac{-i\varrho(z_0 + hz) + \varrho(z_0 + ihz) - (1 - i)\varrho(z_0)}{-2ih\bar{\partial}_{\text{wirt}}\varrho(z_0)} = \bar{z} + \frac{-i\Theta(hz) + \Theta(ihz)}{-2ih\bar{\partial}_{\text{wirt}}\varrho(z_0)}. \quad (3.4)$$

It remains to show that the second summands on the right-hand sides of (3.1), (3.2), (3.3), and (3.4) tend to 0 as $h \downarrow 0$. Since K is compact there exists $L > 0$ satisfying $|z| \leq L$ for all $z \in K$. Let $\varepsilon' > 0$ be arbitrary and take $\delta > 0$ such that

$$\left| \frac{\Theta(w)}{w} \right| < \frac{\varepsilon'}{L}$$

for every $w \in \mathbb{C} \setminus \{0\}$ with $|w| < \delta$. Let $h \in (0, \delta/L)$ and $z \in K$. Since $|hz| < \delta$ we see for every $z \in K \setminus \{0\}$ that

$$\left| \frac{\Theta(hz)}{h} \right| \leq L \cdot \left| \frac{\Theta(hz)}{hz} \right| < \varepsilon'$$

and since ε' has been taken arbitrarily

$$\limsup_{h \downarrow 0} \sup_{z \in K} \frac{\Theta(hz)}{h} = 0.$$

Here, we concluded $\Theta(0) = 0$ from (2.3) to also cover the case $z = 0$. $\square$

Proposition 3.2 is important for the case of polyharmonic activation functions which is considered in Section 4.2. It essentially states that, given an activation function which is not $\mathbb{R}$ -affine, one can approximate one of the functions $z \mapsto z\bar{z}$, $z \mapsto z^2$ or $z \mapsto \bar{z}^2$ by using a shallow neural network of width 4, see Fig. 3 for an illustration.

Proposition 3.2 *Let*

$$\begin{aligned} f_1 : \mathbb{C} &\rightarrow \mathbb{C}, & f_1(z) &= z\bar{z}, & f_2 : \mathbb{C} &\rightarrow \mathbb{C}, & f_2(z) &= z^2, & \text{and} \\ f_3 : \mathbb{C} &\rightarrow \mathbb{C}, & f_3(z) &= \bar{z}^2. \end{aligned}$$

Moreover, let $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ be not $\mathbb{R}$ -affine. Then there exists a function $f \in \{f_1, f_2, f_3\}$ with the following property: For every compact subset $K \subseteq \mathbb{C}$ and every $\varepsilon > 0$ there exist affine maps $\phi \in \text{Aff}(\mathbb{C}; \mathbb{C}^4)$ and $\psi \in \text{Aff}(\mathbb{C}^4; \mathbb{C})$ such that

$$\sup_{z \in K} |(\psi \circ \varrho^{\times 4} \circ \phi)(z) - f(z)| < \varepsilon$$

holds true.

Proof Since ϱ is not $\mathbb{R}$ -affine, there exists $z_0 \in \mathbb{C}$ such that $\partial_{\text{wirt}}^2 \varrho(z_0) \neq 0$, $\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) \neq 0$, or $\bar{\partial}_{\text{wirt}}^2 \varrho(z_0) \neq 0$, see, e.g., Proposition B.2. Using the second-order Taylor expansion stated in Lemma 2.1, there exists a function $\Theta : \mathbb{C} \rightarrow \mathbb{C}$ satisfying $\lim_{w \rightarrow 0} \frac{\Theta(w)}{w^2} = 0$ and

$$\begin{aligned} \varrho(z_0 + w) &= \varrho(z_0) + \partial_{\text{wirt}} \varrho(z_0) w + \bar{\partial}_{\text{wirt}} \varrho(z_0) \bar{w} + \frac{1}{2} \partial_{\text{wirt}}^2 \varrho(z_0) w^2 \\ &\quad + \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) w \bar{w} + \frac{1}{2} \bar{\partial}_{\text{wirt}}^2 \varrho(z_0) \bar{w}^2 + \Theta(w) \end{aligned}$$

for every $w \in \mathbb{C}$. Applying this identity to $-w$ in place of w and adding up, we infer for any $w \in \mathbb{C}$ that

$$\begin{aligned} \varrho(z_0 + w) + \varrho(z_0 - w) &= 2\varrho(z_0) + \partial_{\text{wirt}}^2 \varrho(z_0) w^2 + 2\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) w \bar{w} \\ &\quad + \bar{\partial}_{\text{wirt}}^2 \varrho(z_0) \bar{w}^2 + \Theta(w) + \Theta(-w). \end{aligned}$$

Let $h > 0$ and $z \in K$. If $\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) \neq 0$, we see with $w = hz$ and $w = ihz$ that

$$\begin{aligned} &\frac{\varrho(z_0 + hz) + \varrho(z_0 - hz) + \varrho(z_0 + ihz) + \varrho(z_0 - ihz) - 4\varrho(z_0)}{4h^2 \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0)} \\ &= z\bar{z} + \frac{\Theta(hz) + \Theta(-hz) + \Theta(ihz) + \Theta(-ihz)}{4h^2 \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0)}. \end{aligned} \quad (3.5)$$

If $\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) = 0$ and $\partial_{\text{wirt}}^2 \varrho(z_0) \neq 0$, consider $w = hz$ and $w = \sqrt{i}hz$, where $\sqrt{i}$ is a fixed square root of i :

$$\begin{aligned} &\frac{\varrho(z_0 + hz) + \varrho(z_0 - hz) - i\varrho(z_0 + \sqrt{i}hz) - i\varrho(z_0 - \sqrt{i}hz) + 2(-1 + i)\varrho(z_0)}{2h^2 \partial_{\text{wirt}}^2 \varrho(z_0)} \\ &= z^2 + \frac{\Theta(hz) + \Theta(-hz) - i\Theta(\sqrt{i}hz) - i\Theta(-\sqrt{i}hz)}{2h^2 \partial_{\text{wirt}}^2 \varrho(z_0)}. \end{aligned} \quad (3.6)$$

Last, if $\partial_{\text{wirt}}^2 \varrho(z_0) = \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) = 0$, consider $w = hz$:

$$\frac{\varrho(z_0 + hz) + \varrho(z_0 - hz) - 2\varrho(z_0)}{h^2 \bar{\partial}_{\text{wirt}}^2 \varrho(z_0)} = \bar{z}^2 + \frac{\Theta(hz) + \Theta(-hz)}{h^2 \bar{\partial}_{\text{wirt}}^2 \varrho(z_0)}. \quad (3.7)$$

It remains to show that the second summands on the right-hand sides of (3.5), (3.6), and (3.7) tend to 0 as $h \downarrow 0$. Since K is bounded, there exists $L > 0$ with $|z| \leq L$ for every $z \in K$. For given $\varepsilon' > 0$ there exists $\delta > 0$ such that

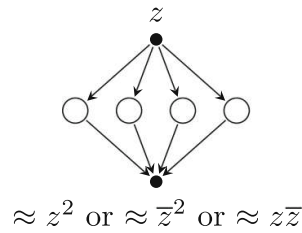


Fig. 3 Illustration of the neural network building block from Proposition 3.2. Neurons in the input and output layers are depicted in filled dots at the top and bottom, respectively. Applications of the activation function ϱ are shown as circles

$$\left| \frac{\Theta(w)}{w^2} \right| < \frac{\varepsilon'}{L^2}$$

for every $w \in \mathbb{C} \setminus \{0\}$ with $|w| < \delta$. Hence, we see for every $h \in (0, \delta/L)$ and all $z \in K \setminus \{0\}$ that

$$\left| \frac{\Theta(hz)}{h^2} \right| \leq L^2 \cdot \left| \frac{\Theta(hz)}{(hz)^2} \right| < \varepsilon'$$

where we used that $|hz| < \delta$. Therefore, we conclude

$$\limsup_{h \downarrow 0} \sup_{z \in K} \frac{\Theta(hz)}{h^2} = 0,$$

using $\Theta(0) = 0$ from (2.4) to also cover the case $z = 0$. $\square$

In order to approximate arbitrary polynomials in the variables $z_1, \dots, z_n$ and $\bar{z}_1, \dots, \bar{z}_n$, we will compute iterative products of two complex numbers in Theorem 4.9. The following result enables the approximation of such products. An illustration of the CVNN blocks appearing in the proof are given in Fig. 4.

Proposition 3.3 *Let*

$$\begin{aligned} \text{mul}_1 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_1(z, w) &= zw, \\ \text{mul}_2 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_2(z, w) &= z\bar{w}, \\ \text{mul}_3 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_3(z, w) &= \bar{z}w. \end{aligned}$$

Moreover, let $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ be not $\mathbb{R}$ -affine. Then there exists $\text{mul} \in \{\text{mul}_1, \text{mul}_2, \text{mul}_3\}$ with the following property: For every compact subset $K \subseteq \mathbb{C}^2$ and $\varepsilon > 0$ there exist $\phi \in \text{Aff}(\mathbb{C}^2; \mathbb{C}^{12})$ and $\psi \in \text{Aff}(\mathbb{C}^{12}; \mathbb{C})$ such that

$$\sup_{(z, w) \in K} \left| (\psi \circ \varrho^{\times 12} \circ \phi)(z, w) - \text{mul}(z, w) \right| < \varepsilon$$

holds true.

Proof The main steps of the proof are to use a variant of the polarization identity to reconstruct the three multiplication operators from their values on the diagonal, and then to apply Proposition 3.2 to approximate the latter.

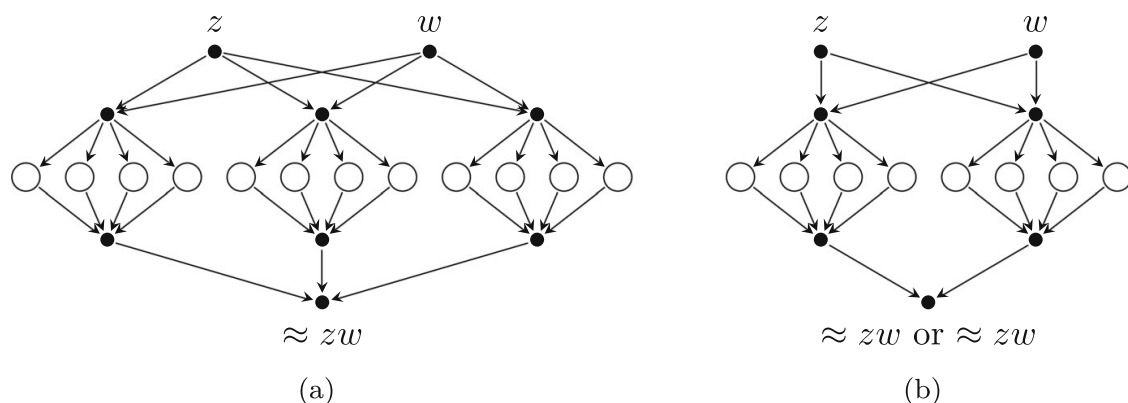


Fig. 4 Illustration of the neural network building block from Proposition 3.3. Neurons in the input and output layers are depicted in filled dots at the top and bottom, respectively. Applications of the activation function ϱ are shown as circles. From the input values z and w , three or two linear combinations are computed. Then the building block from Fig. 3 is inserted to approximate $z \mapsto z^2$, $z \mapsto \bar{z}^2$, or $z \mapsto z\bar{z}$. The results are again combined linearly

Precisely, the construction is as follows: If $\zeta \mapsto \zeta\bar{\zeta} = |\zeta|^2$ can be approximated according to the first case of Proposition 3.2, use the identity

$$\left(\frac{1}{4} + \frac{i}{4}\right) |z + w|^2 + \left(-\frac{1}{4} + \frac{i}{4}\right) |z - w|^2 - \frac{i}{2} |z - iw|^2 = z\bar{w}.$$

Thus, in order to approximate $(z, w) \mapsto z\bar{w}$, one needs 4 hidden neurons to approximate $\zeta \mapsto |\zeta|^2$ for each of the 3 linear combinations of z and w , resulting in a total amount of 12 hidden neurons.

If we have the second case of Proposition 3.2, we can approximate $\zeta \mapsto \zeta^2$ using 4 hidden neurons. In this case, consider

$$\frac{1}{4} \left[(z + w)^2 - (z - w)^2 \right] = zw,$$

so that in total one needs 8 neurons in order to approximate $(z, w) \mapsto zw$. In the last case of Proposition 3.2 we can approximate $\zeta \mapsto \bar{\zeta}^2$ using 4 hidden neurons. Considering

$$\frac{1}{4} \left[\overline{(z + w)^2} - \overline{(z - w)^2} \right] = \bar{z}\bar{w},$$

we infer that $(z, w) \mapsto \bar{z}\bar{w}$ can be approximated using 8 hidden neurons. $\square$

It remains open whether the number 12 in Proposition 3.3 can be reduced. However, it should be noted that the exact number is not crucial for the final result Theorem 4.9, since in its proof the shallow networks obtained in Proposition 3.3 are transformed to deep narrow networks according to Propositions 3.6 and 3.8 and the width of these networks is in fact independent of the number of hidden neurons in the original shallow networks.

3.2 Register model

In this section, we introduce the fundamental concept of the *register model*. This construction has been heavily used in [17] to prove the real-valued counterpart of the theorem established in the present paper.

Definition 3.4 Let $n, m, W, L \in \mathbb{N}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$. Denote by $\mathcal{I}_{n,m,W,L}^\varrho$ the set of *register models*

$$T_L \circ \tilde{\varrho} \circ \dots \circ \tilde{\varrho} \circ T_0,$$

where $T_0 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^W)$, $T_L \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^m)$, $T_\ell \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^W)$ and

$$\tilde{\varrho} : \mathbb{C}^W \rightarrow \mathbb{C}^W, \quad (\tilde{\varrho}(z_1, \dots, z_W))_j = \begin{cases} \varrho(z_W), & j = W, \\ z_j, & j \neq W. \end{cases}$$

In view of cases where the depth L does not matter, we set

$$\mathcal{I}_{n,m,W}^\varrho := \bigcup_{L \in \mathbb{N}} \mathcal{I}_{n,m,W,L}^\varrho.$$

Remark 3.5 The set $\mathcal{I}_{n,m,W,L}^\varrho$ may be viewed as the set of CVNNs with n input neurons, m output neurons, a width of W and a depth of L , where in every hidden layer the first $W - 1$ neurons use the identity as activation function and in the last neuron, ϱ is used as activation function. In fact, since we can apply permutations to the entries of a vector before and after each layer, it is irrelevant in which neuron the application of ϱ takes place. We choose the last neuron for convenience.

One can transform a shallow network into a deep narrow register model by “flipping” the shallow network and only performing one computation per layer. This is formalized in Proposition 3.6 and illustrated in Fig. 5.

Proposition 3.6 Let $n, m, L \in \mathbb{N}$. Let $f \in \mathcal{SN}_{n,m,L}^\varrho$ and let $\tilde{f} : \mathbb{C}^n \rightarrow \mathbb{C}^n \times \mathbb{C}^m$, $\tilde{f}(z) := (z, f(z))$. Then we have $\tilde{f} \in \mathcal{I}_{n,n+m,n+m+1,L}^\varrho$. In particular, we have $\mathcal{SN}_{n,m,L}^\varrho \subseteq \mathcal{I}_{n,m,n+m+1,L}^\varrho$.

Proof Let $f \in \mathcal{SN}_{n,m}^\varrho$. Then there exist $V_1 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^L)$, and $V_2 \in \text{Aff}(\mathbb{C}^L; \mathbb{C}^m)$ such that $f = V_2 \circ \varrho^{\times L} \circ V_1$. For $j \in \{1, \dots, m\}$, the j th component function f_j of f can be written as

$$f_j(z) = \left(\sum_{k=1}^L c_{k,j} \varrho(a_k^\top z + b_k) \right) + d_j$$

with suitably chosen $a_k \in \mathbb{C}^n$ and $c_{k,j}, b_k, d_j \in \mathbb{C}$. We define

$$T_0 : \mathbb{C}^n \rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, \quad T_1(z) = (z, a_1^\top z + b_1, 0).$$

For $\ell \in \{1, \dots, L-1\}$, we define

$$T_\ell : \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m \rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, \quad T_\ell(z, u, (w_1, \dots, w_m)) = \begin{pmatrix} z \\ a_{\ell+1}^\top z + b_{\ell+1} \\ w_1 + c_{\ell,1}u \\ w_2 + c_{\ell,2}u \\ \vdots \\ w_m + c_{\ell,m}u \end{pmatrix}.$$

Last, we set

$$T_L : \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m \rightarrow \mathbb{C}^n \times \mathbb{C}^m, \quad T_L(z, u, (w_1, \dots, w_m)) = \begin{pmatrix} z \\ w_1 + c_{L,1}u + d_1 \\ w_2 + c_{L,2}u + d_2 \\ \vdots \\ w_m + c_{L,m}u + d_m \end{pmatrix}.$$

Then we have

$$\tilde{f} = T_L \circ \tilde{q} \circ \dots \circ \tilde{q} \circ T_0 \quad (3.8)$$

with

$$\tilde{q} : \mathbb{C}^{n+m+1} \rightarrow \mathbb{C}^{n+m+1}, \quad (\tilde{q}(z_1, \dots, z_{n+m+1}))_j = \begin{cases} \varrho(z_{n+1}) & \text{if } j = n+1, \\ z_j & \text{if } j \neq n+1, \end{cases}$$

which clearly yields $\tilde{f} \in \mathcal{I}_{n,n+m,n+m+1,L}^\varrho$ (see Remark 3.5). The second part of the statement follows by applying a projection onto the last m coordinates after the last layer. $\square$

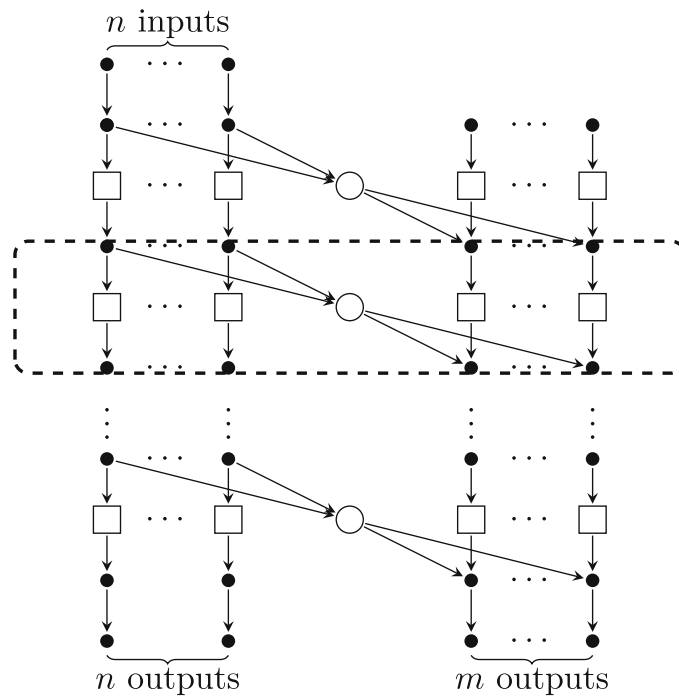
We illustrate the proof of Proposition 3.6 in the following. To this end, let us adopt some terminology from [17, Proof of Proposition 4.6]. The neurons that use ϱ as the activation function (the one with index $n+1$ in each layer) shall be referred to as *computation neurons*. We call the neurons with indices $< n+1$ in each layer the *in-register neurons*. Here the inputs are just passed through the different layers unaltered. In other words, the restriction of $\tilde{q} \circ \dots \circ \tilde{q} \circ T_0$ in (3.8) to the components with indices $< n+1$ is just the identity on $\mathbb{C}^n$. The remaining neurons (the ones with indices $> n+1$ in each layer) are called *out-register neurons*. Here the outputs of the computation neurons are assembled to form the final outputs.

The following proposition enables us to approximate CVNNs that use $\bar{\varrho}$ as activation function locally uniformly by CVNNs that use ϱ as activation function, if there exists a point $z_0 \in \mathbb{C}$ with $\partial_{\text{wirt}}\varrho(z_0) = 0 \neq \bar{\partial}_{\text{wirt}}\varrho(z_0)$.

Proposition 3.7 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$ and $n, m, W, L \in \mathbb{N}$. Assume that there exists $z_0 \in \mathbb{C}$ such that ϱ is real differentiable at z_0 with*

$$\partial_{\text{wirt}}\varrho(z_0) = 0 \quad \text{and} \quad \bar{\partial}_{\text{wirt}}\varrho(z_0) \neq 0.$$

Fig. 5 Illustration of the register model from Proposition 3.6. Neurons where the complex identity is used as activation function are visualized as squares, whereas neurons using ϱ as activation function are visualized using circles. The in-register neurons (squares on the left) store the input values and pass them to the computation neurons (middle circles). The result of the computations are added up and stored in the out-register neurons (squares on the right). The dashed box highlights one of the blocks that are later replaced using approximations of the complex identity



Then

$$\mathcal{NN}_{n,m,W,L}^{\overline{\varrho}} \subseteq \overline{\mathcal{NN}_{n,m,W,2L}^{\varrho}},$$

where the closure is taken with respect to the compact-open topology.

Proof Let $f \in \mathcal{NN}_{n,m,W}^{\overline{\varrho}}$ be arbitrary and consider the decomposition

$$f = V_L \circ \overline{\varrho}^{\times W} \circ V_{L-1} \circ \dots \circ \overline{\varrho}^{\times W} \circ V_0$$

with $V_0 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^W)$, $V_\ell \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^W)$, and $V_L \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^m)$ for every $\ell \in \{1, \dots, L-1\}$. From Proposition 3.1(ii) and Proposition A.2 we infer the existence of sequences $(\Phi_j)_{j \in \mathbb{N}}$ and $(\Psi_j)_{j \in \mathbb{N}}$ of affine maps $\Phi_j, \Psi_j \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^W)$ for $j \in \mathbb{N}$ such that

$$\Psi_j \circ \varrho^{\times W} \circ \Phi_j \xrightarrow{j \rightarrow \infty} \overline{\text{id}_{\mathbb{C}}}^{\times W}$$

locally uniformly. But then we see

$$V_L \circ (\Psi_j \circ \varrho^{\times W} \circ \Phi_j) \circ \varrho^{\times W} \circ V_{L-1} \circ \dots \circ (\Psi_j \circ \varrho^{\times W} \circ \Phi_j) \circ \varrho^{\times W} \circ V_0 \xrightarrow{j \rightarrow \infty} f$$

locally uniformly, where the left-hand side is an element of $\mathcal{NN}_{n,m,W,2L}^{\varrho}$ for every $j \in \mathbb{N}$. Here we applied Proposition A.7. The claim then follows from Proposition A.4. $\square$

Using the previous proposition, we can now show that every register model of width W can be approximated by CVNNs, where in *every* hidden neuron the function ϱ is used as activation function. To do so, it is necessary to approximate the identity connections that appear in the register model. By assumption, the activation function

ϱ is real differentiable at some point $z_0 \in \mathbb{C}$ with non-zero derivative. We consider three different cases. First, if ϱ is even *complex* differentiable at z_0 , Proposition 3.1 yields an approximation of $\text{id}_{\mathbb{C}}$ using shallow CVNNs with activation function ϱ and a width of 1. In that case, replacing the identity connections in the register model by these approximations results in a sufficient width of W . Second, if $\bar{\varrho}$ is complex differentiable at z_0 , we use Proposition 3.7 to get the desired result. Last, if neither of the two former cases happens, we show that a width of $2W - 1$ is sufficient for CVNNs with activation function ϱ in order to be universal, by using the third building block described in Proposition 3.1. Note that in each layer, there are $W - 1$ identity connections to be replaced (resulting in a width of $2(W - 1) = 2W - 2$ for the identity connections) and one additional application of the activation function ϱ , which in total gives us a width of $2W - 1$.

Proposition 3.8 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$. Assume that there exists a point $z_0 \in \mathbb{C}$ such that ϱ is real differentiable at z_0 with non-zero derivative. Let $n, m, W, L \in \mathbb{N}$.*

- (i) *If $\partial_{\text{wirt}}\varrho(z_0) \neq 0$ and $\bar{\partial}_{\text{wirt}}\varrho(z_0) = 0$, then $\mathcal{I}_{n,m,W,L}^{\varrho} \subseteq \overline{\mathcal{NN}_{n,m,W,L}^{\varrho}}$.*
- (ii) *If $\partial_{\text{wirt}}\varrho(z_0) = 0$ and $\bar{\partial}_{\text{wirt}}\varrho(z_0) \neq 0$, then $\mathcal{I}_{n,m,W,L}^{\varrho} \subseteq \overline{\mathcal{NN}_{n,m,W,2L}^{\varrho}}$.*
- (iii) *If $\partial_{\text{wirt}}\varrho(z_0) \neq 0$ and $\bar{\partial}_{\text{wirt}}\varrho(z_0) \neq 0$, then $\mathcal{I}_{n,m,W,L}^{\varrho} \subseteq \overline{\mathcal{NN}_{n,m,2W-1,L}^{\varrho}}$.*

Here, the closure is taken with respect to the compact-open topology.

Proof Let $g \in \mathcal{I}_{n,m,W,L}^{\varrho}$. This means that there exist maps $T_0 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^W)$, $T_L \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^m)$, and $T_{\ell} \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^W)$ for $\ell \in \{1, \dots, L-1\}$ such that

$$g = T_L \circ \tilde{\varrho} \circ T_{L-1} \circ \dots \circ \tilde{\varrho} \circ T_0,$$

where $\tilde{\varrho} : \mathbb{C}^W \rightarrow \mathbb{C}^W$ is given by

$$(\tilde{\varrho}(z_1, \dots, z_W))_j = \begin{cases} z_j & \text{if } j \in \{1, \dots, W-1\}, \\ \varrho(z_W) & \text{if } j = W. \end{cases}$$

We first prove (i). Combining Proposition 3.1(i) and Proposition A.2, we deduce the existence of sequences $(\Psi_j)_{j \in \mathbb{N}}$ and $(\Phi_j)_{j \in \mathbb{N}}$ with $\Psi_j, \Phi_j \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^W)$ for $j \in \mathbb{N}$, satisfying

$$\Psi_j \circ \varrho^{\times W} \circ \Phi_j \xrightarrow{j \rightarrow \infty} \tilde{\varrho}$$

locally uniformly. Proposition A.7 now implies that the sequence $(g_j)_{j \in \mathbb{N}}$ given by

$$g_j = T_L \circ (\Psi_j \circ \varrho^{\times W} \circ \Phi_j) \circ T_{L-1} \circ \dots \circ (\Psi_j \circ \varrho^{\times W} \circ \Phi_j) \circ T_0$$

converges locally uniformly to g as $j \rightarrow \infty$. Moreover, since the composition of affine maps is affine, we have $g_j \in \mathcal{NN}_{n,m,W,L}^{\varrho}$ for all $j \in \mathbb{N}$. Claim (i) now follows from Proposition A.4.

We now deal with (ii). By the fundamental properties of Wirtinger derivatives (see for instance [16, E. 1a]) we compute

$$\partial_{\text{wirt}} \bar{\varrho}(z_0) = \overline{\partial_{\text{wirt}} \varrho(z_0)} \neq 0 \quad \text{and} \quad \bar{\partial}_{\text{wirt}} \bar{\varrho}(z_0) = \overline{\bar{\partial}_{\text{wirt}} \varrho(z_0)} = 0.$$

From (i) we infer $\mathcal{I}_{n,m,W,L}^{\varrho} \subseteq \overline{\mathcal{NN}_{n,m,W,L}^{\varrho}}$. A direct application of Proposition 3.7 yields (ii).

The proof of (iii) is analogous to the proof of (i). Gluing together one copy of the complex identity $\text{id}_{\mathbb{C}}$ and $W - 1$ copies of the function ϕ or (the projection onto the first component of) ψ of Proposition 3.1(iii), respectively, we construct sequences $(\Psi_j)_{j \in \mathbb{N}}, (\Phi_j)_{j \in \mathbb{N}}$ satisfying $\Psi_j \in \text{Aff}(\mathbb{C}^{2W-1}; \mathbb{C}^W)$ and $\Phi_j \in \text{Aff}(\mathbb{C}^W; \mathbb{C}^{2W-1})$ for $j \in \mathbb{N}$ such that

$$\Psi_j \circ \varrho^{\times(2W-1)} \circ \Phi_j \xrightarrow{j \rightarrow \infty} \tilde{\varrho}$$

locally uniformly. Because each of the building blocks consists of 2 neurons in this case, the resulting approximating neural network has width $2W - 1$ instead of W . $\square$

4 Proof of the main result

In this section, we prove Theorem 1.1. The analysis is split into the case of non-polyharmonic activation functions (see Section 4.1) and polyharmonic activation functions (see Section 4.2).

4.1 The non-polyharmonic case

When the activation function ϱ is not polyharmonic, the universal approximation theorem for shallow CVNNs from [37, Theorem 1.3] is applicable. For convenience, we state the following special case relevant for our investigations.

Theorem 4.1 *Let $n \in \mathbb{N}$ and $\varrho \in C(\mathbb{C}; \mathbb{C})$. Then $\mathcal{SN}_{n,1}^{\varrho}$ is universal if and only if ϱ is not polyharmonic.*

Since local uniform convergence on $\mathbb{C}^m$ is equivalent to componentwise local uniform convergence, we conclude that, if ϱ is non-polyharmonic, the set $\mathcal{SN}_{n,m}^{\varrho}$ is universal for every $m \in \mathbb{N}$.

Since the set of shallow networks with non-polyharmonic activation function is universal and each shallow network can be represented by a suitable register model (see Proposition 3.6), which in turn can be locally uniformly approximated by deep narrow CVNNs (see Proposition 3.8), we get the following result.

Theorem 4.2 *Let $n, m \in \mathbb{N}$. Assume that $\varrho \in C(\mathbb{C}; \mathbb{C})$ is not polyharmonic and that there exists $z_0 \in \mathbb{C}$ such that ϱ is real differentiable at z_0 with non-zero derivative.*

- (i) *If $\partial_{\text{wirt}} \varrho(z_0) \neq 0$ and $\bar{\partial}_{\text{wirt}} \varrho(z_0) = 0$, then the set $\mathcal{NN}_{n,m,n+m+1}^{\varrho}$ is universal.*
- (ii) *If $\partial_{\text{wirt}} \varrho(z_0) = 0$ and $\bar{\partial}_{\text{wirt}} \varrho(z_0) \neq 0$, then the set $\mathcal{NN}_{n,m,n+m+1}^{\varrho}$ is universal.*
- (iii) *If $\partial_{\text{wirt}} \varrho(z_0) \neq 0$ and $\bar{\partial}_{\text{wirt}} \varrho(z_0) \neq 0$, then the set $\mathcal{NN}_{n,m,2n+2m+1}^{\varrho}$ is universal.*

Proof Note that

$$\mathcal{SN}_{n,m}^{\varrho} \subseteq \mathcal{I}_{n,m,n+m+1}^{\varrho}$$

according to the second part of Proposition 3.6. Since ϱ is non-polyharmonic, $\mathcal{SN}_{n,m}^{\varrho}$ is universal according to Theorem 4.1 and we hence have

$$C(\mathbb{C}^n; \mathbb{C}^m) = \overline{\mathcal{SN}_{n,m}^{\varrho}} \subseteq \overline{\mathcal{I}_{n,m,n+m+1}^{\varrho}}.$$

The claim then follows from Propositions 3.8 and A.6. $\square$

Next we provide two examples of activation functions which are used in practice and to which Theorem 4.2 applies.

Example 4.3 The modReLU function has for example been proposed in [2] as a generalization of the classical ReLU to the complex plane. For a parameter $b < 0$ it is defined as

$$\text{modReLU}_b : \mathbb{C} \rightarrow \mathbb{C}, \quad \text{modReLU}_b(z) := \begin{cases} (|z| + b) \frac{z}{|z|} & \text{if } |z| + b \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

An application of Theorem 4.2(iii) shows that for $n, m \in \mathbb{N}$, and $b < 0$, the set $\mathcal{NN}_{n,m,2n+2m+1}^{\text{modReLU}_b}$ is universal.

To this end, let us verify the assumptions of Theorem 4.2 in detail. Since the continuity of modReLU_b is immediate for $z \in \mathbb{C}$ with $|z| \neq -b$, it remains to check the case $|z| = -b$. Take any sequence $(z_j)_{j \in \mathbb{N}}$ with $z_j \rightarrow z$ as $j \rightarrow \infty$, where we assume without loss of generality $|z_j| \geq -b$ for every $j \in \mathbb{N}$. Then $|\text{modReLU}_b(z_j)| = |z_j| + b \rightarrow |z| + b = 0$ as $j \rightarrow \infty$. This shows the continuity of modReLU_b .

In [13, Corollary 5.4] it is shown that for all $z \in \mathbb{C}$ with $|z| > -b$ and all $k, \ell \in \mathbb{N}_0$ one has

$$\partial_{\text{wirt}}^k \bar{\partial}_{\text{wirt}}^{\ell} \text{modReLU}_b(z) \neq 0.$$

This implies that modReLU_b is non-polyharmonic and $\partial_{\text{wirt}} \text{modReLU}_b(z) \neq 0 \neq \bar{\partial}_{\text{wirt}} \text{modReLU}_b(z)$ for all $z \in \mathbb{C}$ with $|z| > -b$.

Further note that the result from Theorem 4.2 cannot be used to reduce the sufficient width to $n + m + 1$ since it holds $\partial_{\text{wirt}} \text{modReLU}_b(z) \neq 0 \neq \bar{\partial}_{\text{wirt}} \text{modReLU}_b(z)$ for all $z \in \mathbb{C}$ with $|z| > -b$, $\partial_{\text{wirt}} \text{modReLU}_b(z) = \bar{\partial}_{\text{wirt}} \text{modReLU}_b(z) = 0$ for every $z \in \mathbb{C}$ with $|z| < -b$ and modReLU_b is not real differentiable at any $z \in \mathbb{C}$ with $|z| = -b$.

Example 4.4 The complex cardioid function has been used in [36] in the context of MRI fingerprinting, where complex-valued neural networks significantly outperformed their real-valued counterparts. It is defined as

$$\text{card} : \mathbb{C} \rightarrow \mathbb{C}, \quad \text{card}(z) := \begin{cases} \frac{1}{2} \left(1 + \frac{\text{Re}(z)}{|z|} \right) z & \text{if } z \in \mathbb{C} \setminus \{0\}, \\ 0 & \text{if } z = 0. \end{cases}$$

An application of Theorem 4.2(i) shows that for $n, m \in \mathbb{N}$ the set $\mathcal{NN}_{n,m,n+m+1}^{\text{card}}$ is universal.

To this end, let us verify the assumptions of Theorem 4.2 in detail. The continuity of card on $\mathbb{C} \setminus \{0\}$ is immediate. Further note that

$$|\text{card}(z)| = \left| \frac{1}{2} \left(1 + \frac{\text{Re}(z)}{|z|} \right) z \right| \leq |z| \rightarrow 0$$

as $z \rightarrow 0$, showing the continuity of card on the entire complex plane $\mathbb{C}$. Now [13, Corollaries 5.6 and 5.7] show that card is non-polyharmonic and

$$\partial_{\text{wirt}} \text{card}(z) = \frac{1}{2} + \frac{1}{8} \cdot \frac{\bar{z}}{|z|} + \frac{3}{8} \cdot \frac{z}{|z|}, \quad \bar{\partial}_{\text{wirt}} \text{card}(z) = -\frac{1}{8} \cdot \frac{z^3}{|z|^3} + \frac{1}{8} \cdot \frac{z}{|z|}$$

for every $z \in \mathbb{C} \setminus \{0\}$. Hence, we see $\partial_{\text{wirt}} \text{card}(1) = 1$ and $\bar{\partial}_{\text{wirt}} \text{card}(1) = 0$.

4.2 The polyharmonic case

In this section, we deal with activation functions $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ that are polyharmonic. However, it turns out that this property can be relaxed to only requiring that $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ in order for the proofs to work. Note that we still assume that the activation function ϱ is neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine. That these assumptions cannot be neglected is shown in Theorem 6.3. The main assumptions of this subsection can therefore be stated as follows.

Assumption 4.5 Let $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ be a function satisfying the following conditions:

- (i) $\varrho \in C^2(\mathbb{C}; \mathbb{C})$,
- (ii) ϱ is not holomorphic,
- (iii) ϱ is not antiholomorphic,
- (iv) ϱ is not $\mathbb{R}$ -affine.

Using Proposition 3.1, we derive the following Proposition 4.6. It states that the function $z \mapsto (z, \bar{z})$ can be uniformly approximated on compact sets by a shallow network of width 2.

Proposition 4.6 Let ϱ satisfy Assumption 4.5, let $K \subseteq \mathbb{C}$ be compact, and $\varepsilon > 0$. Then there exist maps $\phi \in \text{Aff}(\mathbb{C}; \mathbb{C}^2)$ and $\psi \in \text{Aff}(\mathbb{C}^2; \mathbb{C}^2)$ such that

$$\sup_{z \in K} \left\| (\psi \circ \varrho^{\times 2} \circ \phi)(z) - (z, \bar{z}) \right\|_{\mathbb{C}^2} < \varepsilon.$$

Proof If there exists a point $z_0 \in \mathbb{C}$ with $\partial_{\text{wirt}} \varrho(z_0) \neq 0 \neq \bar{\partial}_{\text{wirt}} \varrho(z_0)$, we can directly apply Proposition 3.1(iii). If there does not exist such a point z_0 , we can still find $z_1 \in \mathbb{C}$ with $\partial_{\text{wirt}} \varrho(z_1) \neq 0$ and $z_2 \in \mathbb{C}$ with $\bar{\partial}_{\text{wirt}} \varrho(z_2) \neq 0$, since ϱ is neither holomorphic nor antiholomorphic. By assumption of the nonexistence of $z_0 \in \mathbb{C}$ with $\partial_{\text{wirt}} \varrho(z_0) \neq 0 \neq \bar{\partial}_{\text{wirt}} \varrho(z_0)$, it follows that $\bar{\partial}_{\text{wirt}} \varrho(z_1) = 0$ and $\partial_{\text{wirt}} \varrho(z_2) = 0$, and we can thus apply Proposition 3.1(i) and (ii). $\square$

Proposition 4.6 is a central finding, since it implies that we can build from the activation function ϱ constructions that are similar to the register model construction from Proposition 3.6.

The Stone–Weierstrass theorem states that any continuous function can be arbitrarily well approximated by complex polynomials in $z_1, \dots, z_n$ and $\overline{z_1}, \dots, \overline{z_n}$ in the uniform norm on compact subsets of $\mathbb{C}^n$. In order to show universality of CVNNs, it therefore suffices to show that such polynomials can be approximated by deep narrow CVNNs to arbitrary precision in the uniform norm on compact sets. Motivated by this observation, for $N, n \in \mathbb{N}$, we define

$$\mathcal{P}_N^n := \left\{ p : \mathbb{C}^n \rightarrow \mathbb{C}, \quad p(z) = \sum_{\substack{\mathbf{m} \in \mathbb{N}_0^n \\ \mathbf{m} \leq N}} \sum_{\substack{\ell \in \mathbb{N}_0^n \\ \ell \leq N}} a_{\mathbf{m}, \ell} z^{\mathbf{m}} \overline{z}^{\ell} : a_{\mathbf{m}, \ell} \in \mathbb{C} \forall \mathbf{m}, \ell \right\} \quad (4.1)$$

as the set of complex polynomials on $\mathbb{C}^n$ of degree less than N in each variable. Here, the notation $\mathbf{m} \leq N$ means $\mathbf{m}_j \leq N$ for every $j \in \{1, \dots, n\}$ and

$$z^{\mathbf{m}} \overline{z}^{\ell} := \prod_{j=1}^n z_j^{\mathbf{m}_j} \overline{z_j}^{\ell_j}$$

for $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_n), \ell = (\ell_1, \dots, \ell_n) \in \mathbb{N}_0^n$. We follow an approach similar to that of the register model by preserving the inputs and outputs from layer to layer while gradually performing multiplications to approximate the individual monomials. This is described in the following lemma. In its statement and proof, we use $\prod_{k=1}^n f_k^{\beta_k}$ as an abbreviation for the composition

$$\underbrace{(f_n \circ \dots \circ f_n)}_{\beta_n \text{ many}} \circ \dots \circ \underbrace{(f_2 \circ \dots \circ f_2)}_{\beta_2 \text{ many}} \circ \underbrace{(f_1 \circ \dots \circ f_1)}_{\beta_1 \text{ many}}.$$

Lemma 4.7 *Let $p : \mathbb{C}^n \rightarrow \mathbb{C}^m$, $p(z) = (p_1(z), \dots, p_m(z))$ such that $p_j \in \mathcal{P}_N^n$ for every $j \in \{1, \dots, m\}$ for a suitable choice of $N \in \mathbb{N}$. Let mul be one of the three maps*

$$\begin{aligned} \text{mul}_1 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_1(z, w) &= zw, \\ \text{mul}_2 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_2(z, w) &= z\overline{w}, \\ \text{mul}_3 : \mathbb{C} \times \mathbb{C} &\rightarrow \mathbb{C}, & \text{mul}_3(z, w) &= \overline{zw}, \end{aligned}$$

as in Proposition 3.3. Further, let

$$f_k^{\text{mul}} = f_k : \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m \rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, \quad f_k(z, w, u) = (z, \text{mul}(w, z_k), u)$$

for $k \in \{1, \dots, n\}$, and

$$h_k : \mathbb{C}^{n+1+m} \rightarrow \mathbb{C}^{n+1+m}, \quad h_k = \text{id}_{\mathbb{C}^{k-1}} \times \overline{\text{id}_{\mathbb{C}}} \times \text{id}_{\mathbb{C}^{n+m-k}},$$

for $k \in \{1, \dots, n+1\}$. Then there exist $L \in \mathbb{N}$, maps $g_j : \mathbb{C}^{n+m+1} \rightarrow \mathbb{C}^{n+m+1}$ for $j \in \{1, \dots, L\}$ each of which is a finite composition of the maps $f_1, \dots, f_n, h_1, \dots, h_{n+1}$ defined above, and affine mappings $T_0 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^{n+m+1})$, $T_\ell \in \text{Aff}(\mathbb{C}^{n+m+1}; \mathbb{C}^{n+m+1})$ for $\ell \in \{1, \dots, L-1\}$ and further $T_L \in \text{Aff}(\mathbb{C}^{n+m+1}; \mathbb{C}^m)$ such that

$$p = T_L \circ g_L \circ \dots \circ T_1 \circ g_1 \circ T_0.$$

Proof By definition of $\mathcal{P}_N^n$, there exist $a_{j,\mathbf{m},\ell} \in \mathbb{C}$ for $\mathbf{m}, \ell \in \mathbb{N}_0^n$ with $\mathbf{m}, \ell \leq N$ and $j \in \{1, \dots, m\}$ such that

$$p_j(z) = \sum_{\substack{\mathbf{m}, \ell \in \mathbb{N}_0^n \\ \mathbf{m}, \ell \leq N}} a_{j,\mathbf{m},\ell} z^{\mathbf{m}} \bar{z}^\ell \quad \text{for every } j \in \{1, \dots, m\}.$$

By enumerating all the occurring monomials from 1 to L , we can rewrite

$$p_j(z) = \sum_{\ell=1}^L c_{j,\ell} \prod_{k=1}^n z_k^{\alpha_{\ell,k}} \bar{z}_k^{\beta_{\ell,k}}$$

with $\alpha_{\ell,k}, \beta_{\ell,k} \in \{0, \dots, N\}$ and $c_{j,\ell} \in \mathbb{C}$ for every $\ell \in \{1, \dots, L\}$, $k \in \{1, \dots, n\}$, and $j \in \{1, \dots, m\}$.

We then set

$$\begin{aligned} T_0: \mathbb{C}^n &\rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, & T_0(z) &= (z, 1, 0), \\ T_\ell: \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m &\rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, & T_\ell(z, w, u) &= \left(z, 1, w \sum_{j=1}^m c_{j,\ell} e_j + u \right), \end{aligned}$$

and

$$T_L: \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m \rightarrow \mathbb{C}^m, \quad T_L(z, w, u) = w \sum_{j=1}^m c_{j,L} e_j + u,$$

where $\ell \in \{1, \dots, L-1\}$. Here, e_j denotes the j th standard basis vector in $\mathbb{C}^m$. Clearly, the maps $T_0, \dots, T_L$ are $\mathbb{C}$ -affine. Moreover, for $\ell \in \{1, \dots, L\}$, let

$$g_\ell = \begin{cases} \prod_{k=1}^n f_k^{\alpha_{\ell,k}} \circ h_{n+1} \circ \prod_{k=1}^n f_k^{\beta_{\ell,k}} & \text{if } \text{mul} = \text{mul}_1, \\ \prod_{k=1}^n f_k^{\beta_{\ell,k}} \circ h_{n+1} \circ \prod_{k=1}^n f_k^{\alpha_{\ell,k}} & \text{if } \text{mul} = \text{mul}_2, \\ \prod_{k=1}^n (h_{n+1} \circ f_k)^{\alpha_{\ell,k}} \circ \prod_{k=1}^n (h_{n+1} \circ h_k \circ f_k \circ h_k)^{\beta_{\ell,k}} & \text{if } \text{mul} = \text{mul}_3. \end{cases}$$

Then

$$g_\ell(z, 1, u) = \left(z, \prod_{k=1}^n z_k^{\alpha_{\ell,k}} \bar{z}_k^{\beta_{\ell,k}}, u \right) \quad (4.2)$$

for all $\ell \in \{1, \dots, L\}$, $z \in \mathbb{C}^n$, and $u \in \mathbb{C}$, i.e., the $(n+1)$ -entry of $g_\ell(z, 1, u)$ is the ℓ th monomial. The application of T_ℓ adds this monomial with the correct coefficients to the cumulative sums in the entries indexed $n+2, \dots, n+1+m$, and resets the $(n+1)$ -entry to 1 for the assembly of the next monomial. This proves

$$p = T_L \circ g_L \circ \dots \circ T_1 \circ g_1 \circ T_0.$$

To see (4.2), note that for example

$$\begin{aligned} & (h_{n+1} \circ h_k \circ f_k \circ h_k)(z_1, \dots, z_k, \dots, z_n, w, u) \\ &= (h_{n+1} \circ h_k \circ f_k)(z_1, \dots, \overline{z_k}, \dots, z_n, w, u) \\ &= (h_{n+1} \circ h_k)(z_1, \dots, \overline{z_k}, \dots, z_n, z_k \overline{w}, u) \\ &= h_{n+1}(z_1, \dots, z_k, \dots, z_n, z_k \overline{w}, u) \\ &= (z_1, \dots, z_k, \dots, z_n, \overline{z_k} w, u) \end{aligned}$$

when $\text{mul} = \text{mul}_3$. □

Lemma 4.7 states that every function from $\mathbb{C}^n$ to $\mathbb{C}^m$ whose components are polynomials can be written as the composition of affine maps and the maps f_k and h_k that are defined in Lemma 4.7. We thus aim to show that each of the f_k and h_k can be locally uniformly approximated by narrow CVNNs. For h_k , this is a direct consequence of Proposition 4.6. For f_k , since each function mul can be approximated by a shallow CVNN of width 12 (see Proposition 3.3), together with Proposition 4.6 it is easy to see that $f_k \in \overline{\mathcal{NN}_{n+m+1, n+m+1, 2n+2m+12}^Q}$. However, a careful analysis shows that $f_k \in \mathcal{NN}_{n+m+1, n+m+1, 2n+2m+5}^Q$; see the proof of Theorem 4.9. The next proposition is a crucial ingredient for this and shows that each f_k can be approximated by register models with a width of $n+m+3$.

Proposition 4.8 *Let $Q \in C^2(\mathbb{C}; \mathbb{C})$ be not $\mathbb{R}$ -affine, $n, m \in \mathbb{N}$, and $k \in \{1, \dots, n\}$. Let further mul be chosen according to Proposition 3.3 and $f_k^{\text{mul}} = f_k : \mathbb{C}^{n+m+1} \rightarrow \mathbb{C}^{n+m+1}$ be defined as in Lemma 4.7. Then we have*

$$f_k \in \overline{\mathcal{I}_{n+m+1, n+m+1, n+m+3}^Q},$$

where the closure is taken with respect to the compact-open topology.

Proof According to Proposition 3.3 and Proposition A.2, we may pick a sequence $(\varphi_j)_{j \in \mathbb{N}}$ with $\varphi_j \in \mathcal{NN}_{2,1}^Q$ for $j \in \mathbb{N}$ that satisfies

$$\varphi_j \rightarrow \text{mul} \quad \text{locally uniformly.}$$

We define

$$\eta_j : \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m \rightarrow \mathbb{C}^n \times \mathbb{C} \times \mathbb{C}^m, \quad \eta_j(z, w, u) := (z, \varphi_j(w, z_k), u)$$

and note that

$$\eta_j \rightarrow f_k \text{ locally uniformly.}$$

Hence, if we can show that $\eta_j \in \mathcal{I}_{n+m+1, n+m+1, n+m+3}^{\mathcal{O}}$ for every $j \in \mathbb{N}$, the proof is completed according to Proposition A.4.

To this end, we take $j \in \{1, \dots, n\}$ and define

$$\begin{aligned} \gamma_j: \mathbb{C}^2 &\rightarrow \mathbb{C}^3, & \gamma_j(w, z) &= (w, z, \varphi_j(w, z)) \quad \text{and} \\ \theta_j: \mathbb{C}^2 &\rightarrow \mathbb{C}^2, & \theta_j(w, z) &= (z, \varphi_j(w, z)). \end{aligned}$$

The first part of Proposition 3.6 then shows $\gamma_j \in \mathcal{I}_{2,3,4}^{\mathcal{O}}$, and an application of a projection onto the last two coordinates after the last layer shows $\theta_j \in \mathcal{I}_{2,2,4}^{\mathcal{O}}$. Let $\widehat{z}_k \in \mathbb{C}^{n-1}$ denote the vector that arises from $z \in \mathbb{C}^n$ by omitting the k th entry. Then, we can write

$$\eta_j(z, w, u) = \pi_k(\widehat{z}_k, \theta_j(w, z_k), u),$$

where $\pi_k: \mathbb{C}^{n+m+1} \rightarrow \mathbb{C}^{n+m+1}$ is the permutation of the entries of a vector that satisfies

$$\pi_k(\widehat{z}_k, z_k, w, u) = (z, w, u)$$

for every $z \in \mathbb{C}^n$, $w \in \mathbb{C}$ and $u \in \mathbb{C}^m$. From this representation, we clearly see

$$\eta_j \in \mathcal{I}_{n+m+1, n+m+1, n+m+3}^{\mathcal{O}},$$

as desired. $\square$

We can now prove the final bound.

Theorem 4.9 *Let $n, m \in \mathbb{N}$. Assume that $\varrho \in C(\mathbb{C}; \mathbb{C})$ satisfies Assumption 4.5. Then the set $\mathcal{NN}_{n,m,2n+2m+5}^{\mathcal{O}}$ is universal. Moreover, if there exists a point $z_0 \in \mathbb{C}$ with either*

$$\partial_{\text{wirt}} \varrho(z_0) \neq 0 = \overline{\partial}_{\text{wirt}} \varrho(z_0) \quad \text{or} \quad \partial_{\text{wirt}} \varrho(z_0) = 0 \neq \overline{\partial}_{\text{wirt}} \varrho(z_0),$$

then $\mathcal{NN}_{n,m,n+m+3}^{\mathcal{O}}$ is universal.

Proof From the Stone–Weierstrass theorem [12, Theorem 4.51], we know that the set of complex polynomials in $z_1, \dots, z_n$ and $\overline{z}_1, \dots, \overline{z}_n$ is dense in $C(K; \mathbb{C})$ with respect to the uniform norm on any compact set $K \subset \mathbb{C}^n$. Hence, it suffices to show that each function $p: \mathbb{C}^n \rightarrow \mathbb{C}^m$, whose components are polynomials in $z_1, \dots, z_n$ and $\overline{z}_1, \dots, \overline{z}_n$ can be uniformly approximated on K by CVNNs of appropriate width. Equivalently, it suffices to show that $p \in \overline{\mathcal{NN}_{n,m,W}^{\mathcal{O}}}$, where W is the desired width and the closure is taken with respect to the compact-open topology (see Propositions A.2, A.4, and A.5).

Let $\text{mul}: \mathbb{C}^2 \rightarrow \mathbb{C}$ be chosen according to Proposition 3.3 (depending on ϱ). From Lemma 4.7, we infer that there exists a natural number $L \in \mathbb{N}$ such that

$$p = T_L \circ g_L \circ \dots \circ T_1 \circ g_1 \circ T_0,$$

where $T_0 \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^{n+m+1})$, $T_L \in \text{Aff}(\mathbb{C}^{n+m+1}; \mathbb{C}^m)$ and $T_\ell \in \text{Aff}(\mathbb{C}^{n+m+1}; \mathbb{C}^{n+m+1})$ for every $\ell \in \{1, \dots, L-1\}$. Moreover, each function $g_\ell : \mathbb{C}^{n+m+1} \rightarrow \mathbb{C}^{n+m+1}$ is a finite composition of the $f_k = f_k^{\text{mul}}$ and h_k as defined in Lemma 4.7. In view of Proposition A.7, it hence suffices to show that for every $k \in \{1, \dots, n\}$, we have

$$f_k, h_k \in \overline{\mathcal{NN}_{n+m+1, n+m+1, 2n+2m+5}^\varrho},$$

where the closure is taken with respect to the compact-open topology.

To this end, let $k \in \{1, \dots, n\}$. Since $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ and ϱ is by assumption clearly not constant, there exists a point $z_1 \in \mathbb{C}$ at which ϱ is real differentiable with non-zero derivative. Propositions 3.8 and 4.8 then show

$$f_k \in \overline{\mathcal{I}_{n+m+1, n+m+1, n+m+3}^\varrho} \subseteq \overline{\mathcal{NN}_{n+m+1, n+m+1, 2n+2m+5}^\varrho}.$$

Moreover, since both the identity and the conjugation can be locally uniformly approximated on $\mathbb{C}$ by shallow networks of width 2 according to Proposition 4.6, we get

$$h_k \in \overline{\mathcal{SN}_{n+m+1, n+m+1, 2n+2m+2}^\varrho} \subseteq \overline{\mathcal{NN}_{n+m+1, n+m+1, 2n+2m+5}^\varrho}.$$

This proves that $\mathcal{NN}_{n+m+1, n+m+1, 2n+2m+5}^\varrho$ is universal (see Proposition A.6).

Let us now assume that there exists a point $z_0 \in \mathbb{C}$ with

$$\partial_{\text{wirt}}\varrho(z_0) \neq 0 = \bar{\partial}_{\text{wirt}}\varrho(z_0).$$

In that case, Propositions 3.8 and 4.8 yield

$$f_k \in \overline{\mathcal{I}_{n+m+1, n+m+1, n+m+3}^\varrho} \subseteq \overline{\mathcal{NN}_{n+m+1, n+m+1, n+m+3}^\varrho}.$$

Moreover, since identities can be locally uniformly approximated by shallow CVNNs of width 1 according to Proposition 3.1, we get

$$h_k \in \overline{\mathcal{SN}_{n+m+1, n+m+1, n+m+2}^\varrho} \subseteq \overline{\mathcal{NN}_{n+m+1, n+m+1, n+m+3}^\varrho},$$

where we again used Proposition 4.6 to approximate the conjugation (which might possibly require a width of 2).

It remains to deal with the case

$$\partial_{\text{wirt}}\varrho(z_0) = 0 \neq \bar{\partial}_{\text{wirt}}\varrho(z_0).$$

Note that $\bar{\varrho}$ satisfies Assumption 4.5 and by the fundamental properties of Wirtinger derivatives (see for instance [16, E. 1a]) we compute

$$\partial_{\text{wirt}}\bar{\varrho}(z_0) = \overline{\partial_{\text{wirt}}\varrho(z_0)} \neq 0 \quad \text{and} \quad \bar{\partial}_{\text{wirt}}\bar{\varrho}(z_0) = \overline{\bar{\partial}_{\text{wirt}}\varrho(z_0)} = 0.$$

Hence, by what we have previously shown, we infer that $\mathcal{NN}_{n, m, n+m+3}^{\bar{\varrho}}$ is universal. Proposition 3.7 then yields that $\mathcal{NN}_{n, m, n+m+3}^\varrho$ is universal. $\square$

5 A quantitative bound in terms of the depth

In this section, we provide a quantitative approximation bound in terms of the depth of the considered networks. For $n \in \mathbb{N}$, we let

$$\Omega_n := [-1, 1]^n + i \cdot [-1, 1]^n$$

denote the $2n$ -dimensional unit cube embedded into $\mathbb{C}^n$. Let $f \in C(\Omega_n; \mathbb{C}^m)$ be a given continuous function defined on that cube and let $\varepsilon > 0$ be a prescribed approximation accuracy. According to the results established in Section 4.2 and 4.1, we can approximate the function f up to arbitrary precision on Ω_n with deep narrow networks. However, these statements are of a qualitative nature and do not address the question of how deep the networks have to be in order to guarantee an approximation accuracy less than ε . In this section, we prove such a quantitative statement for the case of activation functions that are smooth and non-polyharmonic on some non-empty open subset of $\mathbb{C}$. Our bound heavily relies on the *modulus of continuity* of the given function f : For a compact set $K \subseteq \mathbb{C}^n$, a function $f \in C(K; \mathbb{C}^m)$ and $h > 0$, we let

$$\omega(f, h) := \sup \{ \|f(z_1) - f(z_2)\|_{\mathbb{C}^m} : z_1, z_2 \in K, \|z_1 - z_2\|_{\mathbb{C}^n} < h \}.$$

We get the following result, which is a generalization of [18, Proposition 48] to the complex-valued setting.

Proposition 5.1 *Let $f \in C(\Omega_n; \mathbb{C})$. Then for every $k \in \mathbb{N}$ there exists $p \in \mathcal{P}_{2k}^n$ with*

$$\|f - p\|_{C(\Omega_n; \mathbb{C})} \leq \left(\sqrt{2} + \frac{n}{\sqrt{2}} \right) \omega\left(f, \frac{2}{\sqrt{k}}\right).$$

Here, $\mathcal{P}_{2k}^n$ is as defined in (4.1).

Proof We define

$$\tilde{f} : [0, 1]^n + i \cdot [0, 1]^n \rightarrow \mathbb{C}, \quad \tilde{f}(z) = f(2z - \mathbb{1}),$$

where $\mathbb{1} \in \mathbb{C}^n$ is the vector with every entry equal to $1 + i$. According to [18, Proposition 48], we then observe the existence of a function $\tilde{p}_1 : \mathbb{C}^n \rightarrow \mathbb{R}$ with

$$\begin{aligned} \|\operatorname{Re}(\tilde{f}) - \tilde{p}_1\|_{C([0, 1]^n + i \cdot [0, 1]^n; \mathbb{R})} &\leq \left(1 + \frac{2n}{4}\right) \omega\left(\operatorname{Re}(\tilde{f}), \frac{1}{\sqrt{k}}\right) \\ &= \left(1 + \frac{n}{2}\right) \omega\left(\operatorname{Re}(\tilde{f}), \frac{1}{\sqrt{k}}\right) \end{aligned}$$

and $\tilde{p}_1$ is of the form

$$\tilde{p}_1(z) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq k}} a_{\mathbf{k}} \prod_{j=1}^n \operatorname{Re}(z_j)^{\mathbf{k}_j} \operatorname{Im}(z_j)^{\mathbf{k}_{n+j}},$$

with $a_{\mathbf{k}} \in \mathbb{R}$ for every $\mathbf{k} \in \mathbb{N}_0^{2n}$ with $\mathbf{k} \leq k$. Similarly, we obtain a function $\tilde{p}_2 : \mathbb{C}^n \rightarrow \mathbb{R}$ of the form

$$\tilde{p}_2(z) = \sum_{\substack{\mathbf{k} \in \mathbb{N}_0^{2n} \\ \mathbf{k} \leq k}} b_{\mathbf{k}} \prod_{j=1}^n \operatorname{Re}(z_j)^{\mathbf{k}_j} \operatorname{Im}(z_j)^{\mathbf{k}_{n+j}}$$

with $b_{\mathbf{k}} \in \mathbb{R}$ satisfying

$$\|\operatorname{Im}(\tilde{f}) - \tilde{p}_2\|_{C([0,1]^n + i \cdot [0,1]^n; \mathbb{R})} \leq \left(1 + \frac{n}{2}\right) \omega\left(\operatorname{Im}(\tilde{f}), \frac{1}{\sqrt{k}}\right).$$

We set $\tilde{p} := \tilde{p}_1 + i \cdot \tilde{p}_2$ and get

$$\begin{aligned} & \|\tilde{f} - \tilde{p}\|_{C([0,1]^n + i \cdot [0,1]^n; \mathbb{C})} \\ &= \sup_{z \in [0,1]^n + i \cdot [0,1]^n} \sqrt{(\operatorname{Re}(\tilde{f}(z)) - \operatorname{Re}(\tilde{p}(z)))^2 + (\operatorname{Im}(\tilde{f}(z)) - \operatorname{Im}(\tilde{p}(z)))^2} \\ &\leq \sqrt{\|\operatorname{Re}(\tilde{f}) - \tilde{p}_1\|_{C([0,1]^n + i \cdot [0,1]^n; \mathbb{R})}^2 + \|\operatorname{Im}(\tilde{f}) - \tilde{p}_2\|_{C([0,1]^n + i \cdot [0,1]^n; \mathbb{R})}^2} \\ &\leq \left(\sqrt{2} + \frac{n}{\sqrt{2}}\right) \omega\left(\tilde{f}, \frac{1}{\sqrt{k}}\right). \end{aligned}$$

Using the substitutions $\operatorname{Re}(z) = \frac{1}{2}(z + \bar{z})$ and $\operatorname{Im}(z) = \frac{1}{2i}(z - \bar{z})$, we obtain $\tilde{p} \in \mathcal{P}_{2k}^n$. In the end, we set $p(z) := \tilde{p}(\frac{1}{2}z + \frac{1}{2}\mathbb{1})$ and note $p \in \mathcal{P}_{2k}^n$. Moreover, we have

$$\|f - p\|_{C(\Omega_n; \mathbb{C})} \leq \left(\sqrt{2} + \frac{n}{\sqrt{2}}\right) \omega\left(\tilde{f}, \frac{1}{\sqrt{k}}\right).$$

It remains to show that $\omega(\tilde{f}, \frac{1}{\sqrt{k}}) = \omega(f, \frac{2}{\sqrt{k}})$. To this end, we observe

$$\begin{aligned} & \omega\left(\tilde{f}, \frac{1}{\sqrt{k}}\right) \\ &= \sup \left\{ |\tilde{f}(\tilde{z}_1) - \tilde{f}(\tilde{z}_2)| : \tilde{z}_1, \tilde{z}_2 \in [0,1]^n + i \cdot [0,1]^n, \|\tilde{z}_1 - \tilde{z}_2\|_{\mathbb{C}^n} \leq \frac{1}{\sqrt{k}} \right\} \\ &= \sup \{ |f(2\tilde{z}_1 - 1) - f(2\tilde{z}_2 - 1)| : \tilde{z}_1, \tilde{z}_2 \in [0,1]^n + i \cdot [0,1]^n, \\ & \quad \|\tilde{z}_1 - \tilde{z}_2\|_{\mathbb{C}^n} \leq \frac{1}{\sqrt{k}} \} \\ &= \sup \left\{ |f(z_1) - f(z_2)| : z_1, z_2 \in \Omega_n, \left\| \frac{1}{2}z_1 + \frac{1}{2} - \left(\frac{1}{2}z_2 + \frac{1}{2} \right) \right\|_{\mathbb{C}^n} \leq \frac{1}{\sqrt{k}} \right\} \\ &= \omega\left(f, \frac{2}{\sqrt{k}}\right). \end{aligned}$$

□

The following result shows how polynomials can be approximated by shallow CVNNs. Remarkably, the size of the networks needed does only depend on the degree of the polynomial and not on the approximation accuracy. Moreover, if one aims to approximate polynomials from a bounded subset of $\mathcal{P}_k^n$ (with respect to any norm on $\mathcal{P}_k^n$), one can choose the weights connecting the input and the hidden layer of the shallow network independent of the particular polynomial p which is to be approximated; only the weights connecting hidden and output layer have to be adjusted to p .

Proposition 5.2 (cf. [13, Theorem 3.1]) *Let $k, n \in \mathbb{N}$, $\varepsilon > 0$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ be smooth and non-polyharmonic on a non-empty open set $U \subseteq \mathbb{C}$. Let $\mathcal{P}' \subseteq \mathcal{P}_k^n$ be bounded (with respect to some norm on $\mathcal{P}_k^n$) and set $N := (4k + 1)^{2n}$. Then there exists a map $\varphi \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^N)$ with the following property: For every $p \in \mathcal{P}'$ there exists a map $\psi \in \text{Aff}(\mathbb{C}^N; \mathbb{C})$ with*

$$\left\| p - \left(\psi \circ \varrho^{\times N} \circ \varphi \right) \right\|_{C(\Omega_n; \mathbb{C})} \leq \varepsilon.$$

We can now prove the desired quantitative approximation statement. It is based on Propositions 5.1 and 5.2 and uses the fact that each shallow network can be approximated up to arbitrary precision by deep narrow CVNNs according to Propositions 3.8 and 3.6. The depth of this CVNN is determined by the number of hidden neurons in the shallow network, which we can quantify according to Propositions 5.1 and 5.2. To formulate the final result, we further introduce the notation

$$\omega^{-1}(f, \varepsilon) := \sup \{ \delta > 0 : \omega(f, \delta) \leq \varepsilon \} \quad (5.1)$$

for $\varepsilon > 0$ and $f \in C(\Omega_n; \mathbb{C}^m)$.

Theorem 5.3 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$ be smooth and non-polyharmonic on some non-empty open set $\emptyset \neq U \subseteq \mathbb{C}$. Moreover, let $f = (f_1, \dots, f_m) \in C(\Omega_n; \mathbb{C}^m)$ and $\varepsilon > 0$ be given. Then there exists a number $N \in \mathbb{N}$ with*

$$N \leq \left(32 \cdot \left[\omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} \cdot (1 + \frac{n}{2})} \right) \right]^{-2} + 9 \right)^{2n} \quad (5.2)$$

and a network $\Phi \in \mathcal{NN}_{n,m,2n+2m+1,N}^{\varrho}$ with $\|f - \Phi\|_{C(\Omega_n; \mathbb{C}^m)} \leq \varepsilon$.

Moreover, if there exists a point $z_0 \in \mathbb{C}$ where ϱ is real differentiable with

$$\partial_{\text{wirt}} \varrho(z_0) \neq 0 = \bar{\partial}_{\text{wirt}} \varrho(z_0),$$

we may pick $\Phi \in \mathcal{NN}_{n,m,n+m+1,N}^{\varrho}$. If, on the other hand, we have

$$\partial_{\text{wirt}} \varrho(z_0) = 0 \neq \bar{\partial}_{\text{wirt}} \varrho(z_0),$$

for some $z_0 \in \mathbb{C}^n$, we may pick $\Phi \in \mathcal{NN}_{n,m,n+m+1,2N}^{\varrho}$.

Proof We set

$$k := \left\lceil \left(\omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} (1 + \frac{n}{2})} \right) \right)^{-2} \cdot 4 \right\rceil.$$

Note that we have

$$k \geq \left(\omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} (1 + \frac{n}{2})} \right) \right)^{-2} \cdot 4,$$

which implies

$$\frac{2}{\sqrt{k}} \leq \omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} (1 + \frac{n}{2})} \right).$$

By (5.1), we get

$$\omega \left(f, \frac{2}{\sqrt{k}} \right) \leq \frac{\varepsilon}{3 \cdot \sqrt{2m} \cdot (1 + \frac{n}{2})}.$$

According to Theorem 5.1, there exist polynomials $p_1, \dots, p_m \in \mathcal{P}_{2k}^n$ with

$$\|f_j - p_j\|_{C(\Omega_n; \mathbb{C})} \leq \left(\sqrt{2} + \frac{n}{\sqrt{2}} \right) \omega \left(f_j, \frac{2}{\sqrt{k}} \right)$$

for every $j \in \{1, \dots, m\}$. Letting $p := (p_1, \dots, p_m)$, we then have

$$\begin{aligned} \|f - p\|_{C(\Omega_n; \mathbb{C}^m)} &\leq \left(\sum_{j=1}^m \|f_j - p_j\|_{C(\Omega_n; \mathbb{C})}^2 \right)^{1/2} \\ &\leq \left(\sqrt{2} + \frac{n}{\sqrt{2}} \right) \cdot \left(\sum_{j=1}^m \left[\omega \left(f_j, \frac{2}{\sqrt{k}} \right) \right]^2 \right)^{1/2} \\ &\leq \left(\sqrt{2} + \frac{n}{\sqrt{2}} \right) \cdot \left(\sum_{j=1}^m \left[\omega \left(f, \frac{2}{\sqrt{k}} \right) \right]^2 \right)^{1/2} \\ &= \left(\sqrt{2} + \frac{n}{\sqrt{2}} \right) \cdot \sqrt{m} \cdot \omega \left(f, \frac{2}{\sqrt{k}} \right) \leq \frac{\varepsilon}{3}. \end{aligned}$$

We set $N := (8k+1)^{2n}$. Applying Proposition 5.2 to the finite (and therefore bounded) set given by $\mathcal{P}' := \{p_1, \dots, p_m\}$ yields the existence of functions $\varphi \in \text{Aff}(\mathbb{C}^n; \mathbb{C}^N)$ and $\psi \in \text{Aff}(\mathbb{C}^N; \mathbb{C}^m)$ with

$$\left\| p - (\psi \circ \varrho^{\times N} \circ \varphi) \right\|_{C(\Omega_n; \mathbb{C}^m)} \leq \frac{\varepsilon}{3}.$$

Note that N is independent of m since the weights connecting the input and the hidden layer can be chosen independent of the polynomial p_j (see Proposition 5.2). Since ϱ

is smooth and non-polyharmonic on a non-empty open set (in particular not constant) there exists $z_0 \in \mathbb{C}$ such that ϱ is real differentiable at z_0 with non-zero derivative. According to Proposition 3.6 and Proposition 3.8(iii) we have

$$\psi \circ \varrho^{\times N} \circ \varphi \in \mathcal{I}_{n,m,n+m+1,N}^{\varrho} \subseteq \overline{\mathcal{NN}_{n,m,2n+2m+1,N}^{\varrho}}. \quad (5.3)$$

Hence, there exists $\Phi \in \mathcal{NN}_{n,m,2n+2m+1,N}^{\varrho}$ with

$$\left\| \Phi - (\psi \circ \varrho^{\times N} \circ \varphi) \right\|_{C(\Omega_n; \mathbb{C}^m)} \leq \frac{\varepsilon}{3}.$$

By the triangle inequality, we get

$$\begin{aligned} \|f - \Phi\|_{C(\Omega_n; \mathbb{C}^m)} &\leq \|f - p\|_{C(\Omega_n; \mathbb{C}^m)} + \left\| p - (\psi \circ \varrho^{\times N} \circ \varphi) \right\|_{C(\Omega_n; \mathbb{C}^m)} \\ &\quad + \left\| \Phi - (\psi \circ \varrho^{\times N} \circ \varphi) \right\|_{C(\Omega_n; \mathbb{C}^m)} \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \leq \varepsilon, \end{aligned}$$

as desired.

It remains to estimate the depth N of Φ . Note that we have

$$k \leq \left(\omega^{-1} \left(f, \frac{\varepsilon}{3 \cdot \sqrt{2m} (1 + \frac{n}{2})} \right) \right)^{-2} \cdot 4 + 1$$

by definition of k . Combining this with $N = (8k + 1)2^n$, we obtain the upper bound (5.2) for N .

The case that there exists $z_0 \in \mathbb{C}$ such that ϱ is differentiable at z_0 with

$$(\partial_{\text{wirt}} \varrho(z_0), \bar{\partial}_{\text{wirt}} \varrho(z_0)) \neq (0, 0)$$

follows analogously, by using Proposition 3.8(i) and (ii) in (5.3). $\square$

6 Necessity of our assumptions

The proof of Theorem 1.1 is not yet complete. So far, we have proven that activation functions $\varrho \in C(\mathbb{C}; \mathbb{C})$ which are neither holomorphic, nor antiholomorphic, nor $\mathbb{R}$ -affine yield universality of CVNNs of width $2n + 2m + 5$, with indicated improvements under additional assumptions. The necessity part is done in this section: If ϱ is holomorphic, antiholomorphic, or $\mathbb{R}$ -affine, then even the set of CVNNs using activation function ϱ with *arbitrary widths and depths* is not universal, cf. Theorem 6.3. Furthermore, Theorem 1.1 states that under the mentioned constraints on the activation function, a width of $2n + 2m + 5$ is sufficient for universality of CVNNs with input dimension n and output dimension m . But could we have done better? In Theorem 6.4

below, we show that for a family of real-valued activation functions, the set $\mathcal{NN}_{n,m,W}^q$ is not universal when $W < \max\{2n, 2m\}$. In our final result Theorem 6.5, we show that real differentiability of the activation function at one point with non-vanishing derivative is not necessary for the universal approximation property of deep narrow CVNNs.

We prepare the proof of Theorem 6.3 by two lemmas, the first of which is about uniform convergence of $\mathbb{R}$ -affine functions.

Lemma 6.1 *Let $n, m \in \mathbb{N}$, $(f_k)_{k \in \mathbb{N}}$ be a sequence of $\mathbb{R}$ -affine functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Let $(f_k)_{k \in \mathbb{N}}$ converge locally uniformly to f . Then f is $\mathbb{R}$ -affine too.*

Proof Let $A_k \in \mathbb{R}^{m \times n}$ and $b_k \in \mathbb{R}^m$ such that $f_k(x) = A_k x + b_k$. Let $b := f(0)$. Then we have $b_k = f_k(0) \rightarrow f(0) = b$.

Furthermore we see for every $j \in \{1, \dots, n\}$ that

$$\|A_k e_j - A_\ell e_j\|_{\mathbb{R}^m} \leq \|A_k e_j + b_k - A_\ell e_j - b_\ell\|_{\mathbb{R}^m} + \|b_k - b_\ell\|_{\mathbb{R}^m} \rightarrow 0$$

as $k, \ell \rightarrow \infty$, uniformly over j , meaning $\max_{j \in \{1, \dots, n\}} \|A_k e_j - A_\ell e_j\|_{\mathbb{R}^m} \rightarrow 0$ as $k, \ell \rightarrow \infty$. Here e_j denotes the element of $\mathbb{R}^n$ whose entries are 0 except for the j th which is 1.

Consequently, $(A_k)_{k \in \mathbb{N}}$ is a Cauchy sequence and thus converges to some $A \in \mathbb{R}^{m \times n}$. We claim $f(x) = Ax + b$ for every $x \in \mathbb{R}^n$. Indeed, this follows from

$$\begin{aligned} \|A_k x + b_k - Ax + b\|_{\mathbb{R}^m} &\leq \|A_k x - Ax\|_{\mathbb{R}^m} + \|b_k - b\|_{\mathbb{R}^m} \\ &\leq \|A_k - A\|_{\mathbb{R}^{m \times n}} \|x\|_{\mathbb{R}^n} + \|b_k - b\|_{\mathbb{R}^m} \rightarrow 0 \end{aligned}$$

as $k \rightarrow \infty$. □

Our second lemma in preparation of the proof of Theorem 6.3 concerns locally uniform limits of sequences of functions that are either holomorphic or antiholomorphic.

Lemma 6.2 *Let $\mathcal{F} := \{F : \mathbb{C} \rightarrow \mathbb{C} : F \text{ holomorphic or antiholomorphic}\}$ and $(f_k)_{k \in \mathbb{N}}$ be a sequence of functions with $f_k \in \mathcal{F}$ for every $k \in \mathbb{N}$. Let $f : \mathbb{C} \rightarrow \mathbb{C}$ be such that $f_k \rightarrow f$ locally uniformly. Then it holds $f \in \mathcal{F}$.*

Proof We distinguish two cases:

- (i) If there exists a subsequence of $(f_k)_{k \in \mathbb{N}}$ consisting of holomorphic functions, the limit f of this subsequence also has to be holomorphic (see for instance [35, Theorem 10.28]).
- (ii) If there exists a subsequence of $(f_k)_{k \in \mathbb{N}}$ consisting of antiholomorphic functions, the limit f of this subsequence also has to be antiholomorphic, where we again apply [35, Theorem 10.28] to the complex conjugates of the functions in this subsequence. □

The necessity part of Theorem 1.1 is covered by the following theorem.

Theorem 6.3 *Let $n, m \in \mathbb{N}$ and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ be holomorphic or antiholomorphic or $\mathbb{R}$ -affine. Then $\mathcal{NN}_{n,m}^{\varrho} := \bigcup_{W \in \mathbb{N}} \mathcal{NN}_{n,m,W}^{\varrho}$ is not universal.*

Proof It suffices to show the claim for $m = 1$, since local uniform approximation in $\mathbb{C}^m$ means componentwise local uniform approximation.

We start with the case $n = 1$. If ϱ is holomorphic or antiholomorphic it follows that all the elements of $\mathcal{NN}_{1,1}^{\varrho}$ are holomorphic or antiholomorphic (see, e.g., [37, Proof of Eq. (4.15), p. 28]. But then it follows from Lemma 6.2 and Proposition A.2 that $\mathcal{NN}_{1,1}^{\varrho}$ is not universal. If ϱ is $\mathbb{R}$ -affine, each element of $\mathcal{NN}_{1,1}^{\varrho}$ is $\mathbb{R}$ -affine (as the composition of $\mathbb{R}$ -affine functions). By Lemma 6.1 and Proposition A.2 it follows that $\mathcal{NN}_{1,1}^{\varrho}$ is not universal.

The case $n > 1$ can be reduced to the case $n = 1$ in the following way: Assume that $\mathcal{NN}_{n,1}^{\varrho}$ is universal and pick any arbitrary function $f \in C(\mathbb{C}; \mathbb{C})$. Let $\pi : \mathbb{C}^n \rightarrow \mathbb{C}$, $\pi(z_1, \dots, z_n) = z_1$ and $\tilde{\pi} : \mathbb{C} \rightarrow \mathbb{C}^n$, $\tilde{\pi}(z) = (z, 0, \dots, 0)$. Note that it holds $\pi \circ \tilde{\pi} = \text{id}_{\mathbb{C}}$. By assumption, there exists a sequence $(g_k)_{k \in \mathbb{N}}$ with $g_k \in \mathcal{NN}_{n,1}^{\varrho}$ for $k \in \mathbb{N}$ and $g_k \rightarrow f \circ \pi$ locally uniformly. From Proposition A.7 it follows $g_k \circ \tilde{\pi} \rightarrow f$ locally uniformly. Since $g_k \circ \tilde{\pi} \in \mathcal{NN}_{1,1}^{\varrho}$ for every $k \in \mathbb{N}$ it follows that $\mathcal{NN}_{1,1}^{\varrho}$ is universal, in contradiction to what has just been shown. $\square$

In the previous sections, we showed that for a large class of activation functions a width of $2n + 2m + 5$ is sufficient for universality of CVNNs with input dimension n and output dimension m . Following the lines of [7, Lemma 1], we show next that for some activation functions, a width of at least $\max\{2n, 2m\}$ is necessary to guarantee universality.

Theorem 6.4 *Let $n, m \in \mathbb{N}$.*

- (i) *Let $\phi \in C(\mathbb{R}; \mathbb{C})$, and $\varrho : \mathbb{C} \rightarrow \mathbb{C}$ be given by $\varrho(z) := \phi(\text{Re}(z))$. Then $\mathcal{NN}_{n,m,2n-1}^{\varrho}$ is not universal.*
- (ii) *Let $\varrho : \mathbb{C} \rightarrow \mathbb{R}$. Then $\mathcal{NN}_{n,m,2m-1}^{\varrho}$ is not universal.*

Proof We start with (i). Let $K := [-2, 2]^n + i[-2, 2]^n \subseteq \mathbb{C}^n$ and $f(z) := (\|z\|_{\mathbb{C}^n}, 0, \dots, 0)$ for $z \in \mathbb{C}^n$. Let $g \in \mathcal{NN}_{n,m,2n-1}^{\varrho}$ be arbitrary. From the definition of ϱ , it follows that we may write g as

$$g(z) = \psi(\text{Re}(Vz) + b),$$

where $\psi : \mathbb{C}^{2n-1} \rightarrow \mathbb{C}^m$ is some function, $V \in \mathbb{C}^{(2n-1) \times n}$, $b \in \mathbb{R}^{2n-1}$, and the real part Re is taken componentwise. Interpreting $\text{Re} \circ V$ as an $\mathbb{R}$ -linear function from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2n-1}$, we conclude from $2n - 1 < 2n$ that there exists $v \in \mathbb{C}^n$ with $\|v\|_{\mathbb{C}^n} = 1$ satisfying $\text{Re}(Vv) = 0$ and hence

$$g(z + v) = g(z) \text{ for any } z \in \mathbb{C}^n. \quad (6.1)$$

We set

$$A := \left\{ z \in \mathbb{C}^n : \|z\|_{\mathbb{C}^n} < \frac{1}{10} \right\} \quad \text{and} \quad B := \left\{ z \in \mathbb{C}^n : \|z - v\|_{\mathbb{C}^n} < \frac{1}{10} \right\}$$

and compute

$$\begin{aligned} \int_K \|f(z) - g(z)\|_{\mathbb{C}^m} dz &\geq \int_A \|f(z) - g(z)\|_{\mathbb{C}^m} dz + \int_B \|f(z) - g(z)\|_{\mathbb{C}^m} dz \\ &\stackrel{B=A+v}{=} \int_A (\|f(z) - g(z)\|_{\mathbb{C}^m} + \|f(z+v) - g(z+v)\|_{\mathbb{C}^m}) dz \\ &\stackrel{(6.1)}{\geq} \int_A \|f(z) - f(z+v)\|_{\mathbb{C}^m} dz \geq 0.8 \cdot \lambda^{2n}(A) \end{aligned}$$

with λ^{2n} denoting the $2n$ -dimensional Lebesgue measure. In the last inequality we used

$$|\|z\|_{\mathbb{C}^n} - \|z+v\|_{\mathbb{C}^n}| \geq \|z+v\|_{\mathbb{C}^n} - \|z\|_{\mathbb{C}^n} \geq \|v\|_{\mathbb{C}^n} - 2\|z\|_{\mathbb{C}^n} \geq 0.8.$$

Hence it follows that $\mathcal{NN}_{n,m,2n-1}^{\mathcal{Q}}$ is not dense in $C(K; \mathbb{C}^m)$ with respect to the L^1 -norm and thus, using Hölder's inequality, it follows that $\mathcal{NN}_{n,m,2n-1}^{\mathcal{Q}}$ is not dense in $C(K; \mathbb{C}^m)$ with respect to the L^p -norm for any $p \in [1, \infty]$, so in particular for $p = \infty$ which shows that $\mathcal{NN}_{n,m,2n-1}^{\mathcal{Q}}$ is not universal.

Next, we prove (ii). To this end, we construct a function $f \in C(\mathbb{C}^n; \mathbb{C}^m)$, a compact set $K \subset \mathbb{C}^n$, and a number $\varepsilon > 0$ such that

$$\sup_{z \in K} \|f(z) - g(z)\|_{\mathbb{C}^m} \geq \varepsilon$$

for all $g \in \mathcal{NN}_{n,m,2m-1}^{\mathcal{Q}}$. For a moment, fix $g \in \mathcal{NN}_{n,m,2m-1}^{\mathcal{Q}}$. Since the activation function $\mathcal{Q}$ is real-valued, the output of the last but one layer of g is a function $\psi : \mathbb{C}^n \rightarrow \mathbb{R}^{2m-1}$. Also, there exist a matrix $V \in \mathbb{C}^{m \times (2m-1)}$ and a vector $b \in \mathbb{C}^m$ such that $g(z) = V\psi(z) + b$. With $\mathbb{C}^m \cong \mathbb{R}^{2m}$, we may view the restriction of V to $\mathbb{R}^{2m-1}$ as an $\mathbb{R}$ -linear map $\mathbb{R}^{2m-1} \rightarrow \mathbb{R}^{2m}$, and the range $\{g(z) : z \in \mathbb{C}\}$ of g is thus contained in a $(2m-1)$ -dimensional affine subspace $U = U(g)$ of $\mathbb{R}^{2m}$. As

$$\sup_{z \in K} \|f(z) - g(z)\|_{\mathbb{R}^{2m}} \geq \sup_{z \in K} \inf_{u \in U(g)} \|f(z) - u\|_{\mathbb{R}^{2m}},$$

it is sufficient for our purposes to find a function $f \in C(\mathbb{C}^n; \mathbb{C}^m)$, a compact set $K \subset \mathbb{C}^n$, and a number $\varepsilon > 0$ such that

$$\inf_U \sup_{z \in K} \inf_{u \in U} \|f(z) - u\|_{\mathbb{R}^{2m}} \geq \varepsilon$$

where the outermost infimum traverses the $(2m-1)$ -dimensional affine subspaces U of $\mathbb{R}^{2m}$. This is achieved by a function f whose range $\{f(z) : z \in \mathbb{C}^n\}$ is not contained in any $(2m-1)$ -dimensional affine subspace U of $\mathbb{R}^{2m}$. A semi-explicit construction is like this:

Let $K := \{(\lambda, 0, \dots, 0) : \lambda \in \mathbb{R}, 0 \leq \lambda \leq 1\} \subseteq \mathbb{C}^n$ and

$$f_1 : \mathbb{C}^n \rightarrow [0, 1], \quad f_1(z_1, \dots, z_n) = \max\{0, \min\{1, \operatorname{Re}(z_1)\}\}.$$

Further let $f_2 : [0, 1] \rightarrow \mathbb{C}^m$ be a parameterization of a curve that along the edges of the cube $Q := [0, 1]^m + i[0, 1]^m \subseteq \mathbb{C}^m \cong \mathbb{R}^{2m}$ passes through all of its vertices $\{0, 1\}^m + i\{0, 1\}^m \subseteq \mathbb{C}^m \cong \mathbb{R}^{2m}$, and $f = f_2 \circ f_1$. From [5, Table 1], we deduce

$$\inf_U \sup_{z \in K} \inf_{u \in U} \|f(z) - u\|_{\mathbb{R}^{2m}} \geq \frac{1}{2}$$

and this finishes the proof. $\square$

Note that there are non-polyharmonic functions like $\varrho(z) = e^{\operatorname{Re}(z)}$ but also polyharmonic functions that are neither holomorphic, nor anti-holomorphic, nor $\mathbb{R}$ -affine like $\varrho(z) = \operatorname{Re}(z)^2$ that meet the assumptions made in Theorem 6.4.

Following the ideas from [17, Proposition 4.15] we also want to add a short note on the necessity of the differentiability of the activation function. It turns out that the differentiability of the activation function is *not* a necessary condition for the fact that narrow networks with width $n + m + 1$ using this activation function have the universal approximation property. The proof is in fact identical to the proof presented in [17]. However, we include a detailed proof to clarify that the reasoning also works in the case of activation functions $\mathbb{C} \rightarrow \mathbb{C}$.

Theorem 6.5 *Take any function $w \in C(\mathbb{C}; \mathbb{C})$ which is bounded and nowhere real differentiable. Then $\varrho(z) := \sin(z) + w(z) \exp(-z)$ is also nowhere differentiable and $\mathcal{NN}_{n,m,n+m+1}^{\varrho}$ is universal.*

Proof Since ϱ is non-polyharmonic (since it is nowhere differentiable) it suffices to show that the identity function can be uniformly approximated on compact sets using compositions that have the form $\psi \circ \varrho \circ \phi$ with $\phi, \psi \in \operatorname{Aff}(\mathbb{C}; \mathbb{C})$. Then the statement can be derived similarly to the proof of Theorem 4.2(i).

Therefore, take any compact set $K \subseteq \mathbb{C}$ and $\varepsilon > 0$. Choose $M_1 > 0$ with $|z| \leq M_1$ for every $z \in K$. Take $h > 0$ arbitrary and consider

$$\sup_{z \in K \setminus \{0\}} \left| \frac{\sin(hz) - hz}{h} \right| \leq \sup_{z \in K \setminus \{0\}} M_1 \left| \frac{\sin(hz)}{hz} - 1 \right| \rightarrow 0$$

as $h \rightarrow 0$. Therefore, we may take $h > 0$ with

$$\left| \frac{\sin(hz) - hz}{h} \right| < \frac{\varepsilon}{2}$$

for every $z \in K$. Furthermore, choose $M_2 > 0$ with $|w(z)| \leq M_2$ for every $z \in \mathbb{C}$ and pick $k \in \mathbb{N}$ large enough such that

$$\frac{|\exp(-hz)| |\exp(-2\pi k)|}{h} < \frac{\varepsilon}{2M_2}$$

for all $z \in K$. Hence we derive

$$\left| \frac{\sin(hz + 2\pi k) + w(hz + 2\pi k) \exp(-hz - 2\pi k)}{h} - z \right|$$

$$\begin{aligned}
 &= \left| \frac{\sin(hz) + w(hz + 2\pi k) \exp(-hz - 2\pi k)}{h} - z \right| \\
 &\leq \left| \frac{\sin(hz) - hz}{h} \right| + M_2 \frac{|\exp(-hz)| |\exp(-2\pi k)|}{h} < \varepsilon.
 \end{aligned}$$

Thus, we get the claim by defining $\phi(z) := hz + 2\pi k$ and $\psi(z) := \frac{1}{h}z$. $\square$

A Topological notes on locally uniform convergence

In this appendix, we discuss the relationship between locally uniform convergence, the compact open topology, and the universal approximation property introduced in Definition 2.2. Note that [28, Appendix B] is another account on the same topic.

Although locally uniform convergence can be studied more generally for functions defined on a topological space and taking values in a metric space, we restrict ourselves to functions $\mathbb{C}^n \rightarrow \mathbb{C}^m$.

Definition A.1 Let $(f_k)_{k \in \mathbb{N}}$ be a sequence of functions $f_k : \mathbb{C}^n \rightarrow \mathbb{C}^m$ and $f : \mathbb{C}^n \rightarrow \mathbb{C}^m$. The sequence $(f_k)_{k \in \mathbb{N}}$ converges locally uniformly to f , if for every compact set $K \subseteq \mathbb{C}^n$ we have

$$\sup_{z \in K} \|f_k(z) - f(z)\|_{\mathbb{C}^m} \xrightarrow{k \rightarrow \infty} 0.$$

There is a certain equivalence between locally uniform convergence and the universal approximation property introduced in Definition 2.2.

Proposition A.2 Let $\mathcal{F} \subseteq C(\mathbb{C}^n; \mathbb{C}^m)$ and $f \in C(\mathbb{C}^n; \mathbb{C}^m)$. Then the following are equivalent:

(i) For every compact set $K \subseteq \mathbb{C}^n$ and $\varepsilon > 0$, there exists a function $g \in \mathcal{F}$ satisfying

$$\|f - g\|_{C(K; \mathbb{C}^m)} < \varepsilon.$$

(ii) There exists a sequence $(f_k)_{k \in \mathbb{N}}$ with $f_k \in \mathcal{F}$ for $k \in \mathbb{N}$ such that $(f_k)_{k \in \mathbb{N}}$ converges locally uniformly to f .

Proof We start with the implication (i) $\Rightarrow$ (ii). Let $f \in C(\mathbb{C}^n; \mathbb{C}^m)$. For every $k \in \mathbb{N}$, choose $f_k \in \mathcal{F}$ with

$$\|f_k - f\|_{C(\overline{B}_k(0); \mathbb{C}^m)} \leq \frac{1}{k}.$$

Then $(f_k)_{k \in \mathbb{N}}$ converges locally uniformly to f , since every compact set $K \subseteq \mathbb{C}^n$ is contained in the closed ball $\overline{B}_k(0)$ with radius k and center 0 for all $k \geq J$ for some $J \in \mathbb{N}$.

Now we show the implication (ii) $\Rightarrow$ (i). For any compact set $K \subseteq \mathbb{C}^n$ and $\varepsilon > 0$ we know by definition of locally uniform convergence that there exists $k \in \mathbb{N}$ satisfying

$$\|f_k - f\|_{C(K; \mathbb{C}^m)} < \varepsilon.$$

Since $f_k \in \mathcal{F}$, this shows (i). $\square$

Next, we show that locally uniform convergence of sequences $(f_k)_{k \in \mathbb{N}}$ of elements $f_k \in C(\mathbb{C}^n; \mathbb{C}^m)$ coincides with convergence with respect to the *compact-open topology*, cf. [11, Definition XII.1.1]. Hence, the compact-open topology is the topology in charge when we speak about universality of a set of continuous functions.

Definition A.3 For each pair of sets $A \subseteq \mathbb{C}^n$, $B \subseteq \mathbb{C}^m$, we denote

$$(A, B) := \{f \in C(\mathbb{C}^n; \mathbb{C}^m) : f(A) \subseteq B\}.$$

The *compact-open topology* on $C(\mathbb{C}^n; \mathbb{C}^m)$ is then the smallest topology containing the sets (K, V) , where $K \subseteq \mathbb{C}^n$ is compact and $V \subseteq \mathbb{C}^m$ is open.

That this topology indeed induces locally uniform convergence is a direct consequence of [11, Theorem XII.7.2] and $\mathbb{C}^m$ being a metric space.

Proposition A.4 Let $(f_k)_{k \in \mathbb{N}}$ be a sequence of functions with $f_k \in C(\mathbb{C}^n; \mathbb{C}^m)$ and $f \in C(\mathbb{C}^n; \mathbb{C}^m)$. Then the following statements are equivalent.

- (i) The sequence $(f_k)_{k \in \mathbb{N}}$ converges to f in the compact-open topology.
- (ii) The sequence $(f_k)_{k \in \mathbb{N}}$ converges to f locally uniformly.

The next result says that in $C(\mathbb{C}^n; \mathbb{C}^m)$ with the compact-open topology, closures of subsets can be characterized by limits of sequences.

Proposition A.5 Let $f \in C(\mathbb{C}^n; \mathbb{C}^m)$ and $\mathcal{F} \subseteq C(\mathbb{C}^n; \mathbb{C}^m)$. Denote by $\overline{\mathcal{F}}$ the closure of $\mathcal{F}$ with respect to the compact-open topology. Then $f \in \overline{\mathcal{F}}$ if and only if there exists a sequence $(f_k)_{k \in \mathbb{N}}$ with $f_k \in \mathcal{F}$ for $k \in \mathbb{N}$ that converges to f in the compact-open topology.

Proof Since both $\mathbb{C}^n$ and $\mathbb{C}^m$ are second countable, we infer by [11, Theorem XII.5.2] that $C(\mathbb{C}^n; \mathbb{C}^m)$ equipped with the compact-open topology is second countable, too. For second countable topological spaces it is well-known that closures of subsets can be characterized by limits of sequences; see [11, Ex. 3 on p. 186 and Theorem X.6.2]. $\square$

In particular, Propositions A.2, A.4, and A.5 yield the following equivalence, which we state in the following proposition.

Proposition A.6 Let $\mathcal{F} \subseteq C(\mathbb{C}^n; \mathbb{C}^m)$. Then the following are equivalent:

- (i) The set $\mathcal{F}$ has the universal approximation property.
- (ii) For every $f \in C(\mathbb{C}^n; \mathbb{C}^m)$, there exists a sequence $(f_k)_{k \in \mathbb{N}}$ of elements $f_k \in \mathcal{F}$ such that $(f_k)_{k \in \mathbb{N}}$ converges locally uniformly to f .
- (iii) The set $\mathcal{F}$ is dense in $C(\mathbb{C}^n; \mathbb{C}^m)$ with respect to the compact-open topology.

In the present paper, it is of particular importance that the composition of functions is compatible with locally uniform convergence.

Proposition A.7 Let $(f_k)_{k \in \mathbb{N}}$ and $(g_k)_{k \in \mathbb{N}}$ be two sequences of functions with $f_k \in C(\mathbb{C}^{n_1}; \mathbb{C}^{n_2})$ and $g_k \in C(\mathbb{C}^{n_2}; \mathbb{C}^{n_3})$ for $k \in \mathbb{N}$. Let $f \in C(\mathbb{C}^{n_1}; \mathbb{C}^{n_2})$ and $g \in C(\mathbb{C}^{n_2}; \mathbb{C}^{n_3})$ such that $f_k \rightarrow f$ and $g_k \rightarrow g$ locally uniformly. Then we have

$$g_k \circ f_k \xrightarrow{k \rightarrow \infty} g \circ f$$

locally uniformly.

Proof Using [11, Theorem XII.2.2], we know that the map

$$C(\mathbb{C}^{n_1}; \mathbb{C}^{n_2}) \times C(\mathbb{C}^{n_2}; \mathbb{C}^{n_3}) \rightarrow C(\mathbb{C}^{n_1}; \mathbb{C}^{n_3}), \quad (h_1, h_2) \mapsto h_2 \circ h_1$$

is continuous, where each space $C(\mathbb{C}^{n_j}; \mathbb{C}^{n_k})$ is equipped with the compact-open topology and Cartesian products of spaces are equipped with the product topology. Note here that we use the fact that $\mathbb{C}^{n_1}$ and $\mathbb{C}^{n_3}$ are Hausdorff spaces and $\mathbb{C}^{n_2}$ is locally compact. Then the claim follows from Proposition A.4. $\square$

Note that the statement of Proposition A.7 can inductively be extended to the composition of L functions, where L is any natural number.

B Taylor expansion using Wirtinger derivatives

In this appendix we give some details about the Taylor expansion introduced in Lemma 2.1. Furthermore we show that an activation function which is not $\mathbb{R}$ -affine necessarily admits a point where one of the second-order Wirtinger derivatives does not vanish. We begin by restating and proving Lemma 2.1.

Lemma B.1 *Let $\varrho \in C(\mathbb{C}; \mathbb{C})$ and $z, z_0 \in \mathbb{C}$. If ϱ is real differentiable at z_0 , then*

$$\varrho(z + z_0) = \varrho(z_0) + \partial_{\text{wirt}}\varrho(z_0)z + \bar{\partial}_{\text{wirt}}\varrho(z_0)\bar{z} + \Theta_1(z) \quad (\text{B.1})$$

for a function $\Theta_1 : \mathbb{C} \rightarrow \mathbb{C}$ with $\lim_{\mathbb{C} \setminus \{0\} \ni z \rightarrow 0} \frac{\Theta_1(z)}{z} = 0$. If $\varrho \in C^2(\mathbb{C}; \mathbb{C})$, then

$$\begin{aligned} \varrho(z + z_0) &= \varrho(z_0) + \partial_{\text{wirt}}\varrho(z_0)z + \bar{\partial}_{\text{wirt}}\varrho(z_0)\bar{z} + \frac{1}{2}\partial_{\text{wirt}}^2\varrho(z_0)z^2 + \partial_{\text{wirt}}\bar{\partial}_{\text{wirt}}\varrho(z_0)z\bar{z} \\ &\quad + \frac{1}{2}\bar{\partial}_{\text{wirt}}^2\varrho(z_0)\bar{z}^2 + \Theta_2(z) \end{aligned} \quad (\text{B.2})$$

for a function $\Theta_2 : \mathbb{C} \rightarrow \mathbb{C}$ with $\lim_{\mathbb{C} \setminus \{0\} \ni z \rightarrow 0} \frac{\Theta_2(z)}{z^2} = 0$.

Proof Eq. (B.1) follows from the definition of real differentiability (2.1) by using

$$\begin{aligned} \frac{\partial \varrho}{\partial x}(z_0) \operatorname{Re}(z) + \frac{\partial \varrho}{\partial y}(z_0) \operatorname{Im}(z) &= \frac{\partial \varrho}{\partial x}(z_0) \cdot \frac{1}{2}(z + \bar{z}) + \frac{\partial \varrho}{\partial y} \cdot \frac{1}{2i}(z - \bar{z}) \\ &= \partial_{\text{wirt}}\varrho(z_0)z + \bar{\partial}_{\text{wirt}}\varrho(z_0)\bar{z}. \end{aligned}$$

In order to prove (B.2) we use the second-order Taylor expansion of ϱ around z_0 which can be found for instance in [1, Theorem VII.5.11] and obtain

$$\varrho(z + z_0) = \varrho(z_0) + \frac{\partial \varrho}{\partial x}(z_0)x + \frac{\partial \varrho}{\partial y}(z_0)y + \frac{1}{2}\frac{\partial^2 \varrho}{\partial x^2}(z_0)x^2 + \frac{\partial^2 \varrho}{\partial x \partial y}(z_0)xy$$

$$+ \frac{1}{2} \frac{\partial^2 \varrho}{\partial y^2}(z_0) y^2 + \Theta_2(z)$$

where $\Theta_2 : \mathbb{C} \rightarrow \mathbb{C}$ satisfies $\lim_{\mathbb{C} \setminus \{0\} \ni z \rightarrow 0} \frac{\Theta_2(z)}{z^2} = 0$. Furthermore, we use the notation $x = \operatorname{Re}(z)$ and $y = \operatorname{Im}(z)$. Letting $x = \frac{1}{2}(z + \bar{z})$, $y = \frac{1}{2i}(z - \bar{z})$ and using

$$\frac{1}{4} \begin{pmatrix} 1 & -2i & -1 \\ 1 & 0 & 1 \\ 1 & 2i & -1 \end{pmatrix} \begin{pmatrix} \frac{\partial^2}{\partial x^2} \\ \frac{\partial^2}{\partial x \partial y} \\ \frac{\partial^2}{\partial y^2} \end{pmatrix} = \begin{pmatrix} \partial_{\text{wirt}}^2 \\ \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \\ \bar{\partial}_{\text{wirt}}^2 \end{pmatrix} \quad (\text{B.3})$$

yields the claim. $\square$

The following Proposition is required in the proof of Proposition 3.2.

Proposition B.2 *Let $\varrho \in C^2(\mathbb{C}; \mathbb{C})$ be not $\mathbb{R}$ -affine. Then there exists a point $z_0 \in \mathbb{C}$ such that either $\partial_{\text{wirt}}^2 \varrho(z_0) \neq 0$, $\partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho(z_0) \neq 0$ or $\bar{\partial}_{\text{wirt}}^2 \varrho(z_0) \neq 0$.*

Proof Assume $\partial_{\text{wirt}}^2 \varrho \equiv \partial_{\text{wirt}} \bar{\partial}_{\text{wirt}} \varrho \equiv \bar{\partial}_{\text{wirt}}^2 \varrho \equiv 0$. From the fact that the matrix on the left-hand side in (B.3) is invertible it follows $\frac{\partial^2 \varrho}{\partial x^2} \equiv \frac{\partial^2 \varrho}{\partial x \partial y} \equiv \frac{\partial^2 \varrho}{\partial y^2} \equiv 0$. Since ϱ is $\mathbb{R}$ -affine if and only if $\operatorname{Re}(\varrho)$ and $\operatorname{Im}(\varrho)$ are both $\mathbb{R}$ -affine, we may assume that ϱ is real-valued. It is a well-known fact that a C^1 -function with vanishing gradient is necessarily constant. Applying this fact to $\frac{\partial \varrho}{\partial x}$ and $\frac{\partial \varrho}{\partial y}$ separately shows

$$\nabla \varrho \equiv a$$

for a constant $a \in \mathbb{R}^2$. Let $f(z) := z^\top a$ where $z \in \mathbb{C}$ is treated as an element of $\mathbb{R}^2$. Then the gradient of $\varrho - f$ vanishes identically and hence it holds $\varrho - f \equiv b$ for a constant $b \in \mathbb{R}$. This yields

$$\varrho(z) = z^\top a + b \quad \text{for all } z \in \mathbb{C}.$$

But then ϱ is $\mathbb{R}$ -affine. $\square$

Acknowledgements The authors thank Felix Voigtlaender for his helpful comments and the anonymous reviewers for their feedback which helped improving the presentation of the material. PG acknowledges support by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1-1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Amann, H., Escher, J.: Analysis II. Birkhäuser Verlag, Basel (2008)
2. Arjovsky, M., Shah, A., Bengio, Y.: Unitary evolution recurrent neural networks, pp. 1120–1128. PMLR, International conference on machine learning (2016)
3. Barrachina, J.A., Ren, C., Morisseau, C., Vieillard, G., Ovarlez, J.-P.: Comparison between equivalent architectures of complex-valued and real-valued neural networks-application on polarimetric SAR image segmentation. *Journal of Signal Processing Systems* **95**(1), 57–66 (2023). <https://doi.org/10.1007/s11265-022-01793-0>
4. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**(3), 930–945 (1993)
5. Brandenberg, R.: Radii of regular polytopes. *Discrete Comput. Geom.* **33**(1), 43–55 (2005). <https://doi.org/10.1007/s00454-004-1127-1>
6. Bueno, C.: *Universal approximation for neural nets on sets*, PhD thesis, University of California, Santa Barbara, (2021)
7. Cai, Y.: *Achieve the minimum width of neural networks for universal approximation*, arXiv preprint, (2022), [arXiv:2209.11395](https://arxiv.org/abs/2209.11395)
8. Caragea, A., Lee, D.G., Maly, J., Pfander, G., Voigtlaender, F.: Quantitative approximation results for complex-valued neural networks. *SIAM J. Math. Data Sci.* **4**(2), 553–580 (2022). <https://doi.org/10.1137/21M1429540>
9. Cole, E., Cheng, J., Pauly, J., Vasanawala, S.: Analysis of deep complex-valued convolutional neural networks for mri reconstruction and phase-focused applications. *Magnetic Resonance in Medicine* **86**(2), 1093–1109 (2021). <https://doi.org/10.1002/mrm.28733>
10. Cybenko, G.: Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2**(4), 303–314 (1989). <https://doi.org/10.1007/BF02551274>
11. Dugundji, J.: Topology. Allyn and Bacon Inc, Boston, MA (1966)
12. Folland, G.B.: Real analysis: Modern techniques and their applications, 2nd, ed Pure and applied mathematics, Wiley, New York (1999)
13. Geuchen, P., Voigtlaender, F.: *Optimal approximation using complex-valued neural networks*, Advances in Neural Information Processing Systems (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, Curran Associates, Inc., 2023, pp. 1681–1737, https://proceedings.neurips.cc/paper_files/paper/2023/file/05b69cc4c8ff6e24c5de1ecd27223d37-Paper-Conference.pdf
14. Gribonval, R., Kutyniok, G., Nielsen, M., Voigtlaender, F.: Approximation spaces of deep neural networks. *Constr. Approx.* **55**(1), 259–367 (2022). <https://doi.org/10.1007/s00365-021-09543-4>
15. Ismailov, V.E., Savas, E.: Measure theoretic results for approximation by neural networks with limited weights. *Numer. Funct. Anal. Optim.* **38**(7), 819–830 (2017). <https://doi.org/10.1080/01630563.2016.1254654>
16. Kaup, L., Kaup, B.: *Holomorphic functions of several variables: An introduction to the fundamental theory*, De Gruyter Studies in Mathematics, vol. 3, Walter de Gruyter & Co., Berlin, 1983, <https://doi.org/10.1515/9783110838350>
17. Kidger, P., Lyons, T.: *Universal approximation with deep narrow networks*, Proceedings of Thirty Third Conference on Learning Theory (J. Abernethy and S. Agarwal, eds.), vol. 125, 2020, pp. 2306–2327, <https://proceedings.mlr.press/v125/kidger20a.html>
18. Kratsios, A., Papon, L.: Universal approximation theorems for differentiable geometric deep learning. *Journal of Machine Learning Research* **23**(196), 1–73 (2022)
19. Küstner, T., Fuin, N., Hammernik, K., Bustin, A., Qi, H., Hajhosseiny, R., Masci, P. G., Neji, R., Rueckert, D., Botnar, R. M., Prieto, C.: *CINENet: deep learning-based 3d cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions*, Scientific Reports **10** (2020), pp. article no. 13710, 13 pp., <https://doi.org/10.1038/s41598-020-70551-8>
20. Lee, C.Y., Hasegawa, H., Gao, S.C.: Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica* **9**(8), 1406–1426 (2022). <https://doi.org/10.1109/JAS.2022.105743>
21. Lei, Z., Gao, S., Hasegawa, H., Zhang, Z., Zhou, M., Sedraoui, K.: Fully complex-valued gated recurrent neural network for ultrasound imaging. *IEEE Transactions on Neural Networks and Learning Systems* **35**(10), 14918–14931 (2024). <https://doi.org/10.1109/TNNLS.2023.3282231>

22. Leshno, M., Lin, V. Ya., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6**(6), 861–867 (1993). [https://doi.org/10.1016/S0893-6080\(05\)80131-5](https://doi.org/10.1016/S0893-6080(05)80131-5)
23. Li, X., Wang, G.: Universal approximation of polygonal fuzzy neural networks in sense of K -integral norms. *Sci. China Inf. Sci.* **54**(11), 2307–2323 (2011). <https://doi.org/10.1007/s11432-011-4364-y>
24. Lu, Zh., Pu, H., Wang, F., Hu, Zh., Wang, L.: *The expressive power of neural networks: A view from the width*, Advances in Neural Information Processing Systems (I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., (2017), pp. 1–9, <https://proceedings.neurips.cc/paper/2017/file/32cbf687880eb1674a07bf717761dd3a-Paper.pdf>
25. Merkh, T., Montúfar, G.: Stochastic feedforward neural networks: universal approximation, *Mathematical aspects of deep learning*, pp. 267–314. Cambridge Univ. Press, Cambridge (2023)
26. Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. *Neural computation* **8**(1), 164–177 (1996)
27. Murata, N., Sonoda, S.: Neural network with unbounded activation functions is universal approximator. *Appl. Comput. Harmon. Anal.* **43**(2), 233–268 (2017). <https://doi.org/10.1016/j.acha.2015.12.005>
28. Park, J., Wojtowysch, S.: *Qualitative neural network approximation over $\mathbb{R}$ and $\mathbb{C}$: Elementary proofs for analytic and polynomial activation*, Explorations in the Mathematics of Data Science: The Inaugural Volume of the Center for Approximation and Mathematical Data Analytics, Springer, (2024), pp. 31–64
29. Park, S., Yun, C., Lee, J., Shin, J.: *Minimum width for universal approximation*, arXiv preprint, (2020), [arXiv:2006.08859](https://arxiv.org/abs/2006.08859)
30. Petersen, P., Voigtlaender, F.: Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks* **108**, 296–330 (2018)
31. Petersen, P., Voigtlaender, F.: Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proc. Amer. Math. Soc.* **148**(4), 1567–1581 (2020). <https://doi.org/10.1090/proc/14789>
32. Qu, Y.-D., Zhang, R.-Q., Shen, S.-Q., Yu, J., Li, M.: *Entanglement detection with complex-valued neural networks*, International Journal of Theoretical Physics **62** (2023), no. 9, pp. article no. 206, 15 pp., <https://doi.org/10.1007/s10773-023-05460-3>
33. Ren, Y., Jiang, W., Liu, Y.: *A new architecture of a complex-valued convolutional neural network for PolSAR image classification*, Remote Sensing **15** (2023), no. 19, pp. article no. 4801, 27 pp., <https://doi.org/10.3390/rs15194801>
34. Rudin, W.: *Principles of mathematical analysis*, third ed., International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., New York-Auckland-Düsseldorf, (1976)
35. Virtue, P., Yu, S.X., Lustig, M.: Real and complex analysis, 3rd edn. McGraw-Hill Book Co., New York (1987)
36. Virtue, P., Yu, S. X., Lustig, M.: *Better than real: Complex-valued neural nets for MRI fingerprinting*, 2017 IEEE International Conference on Image Processing (ICIP), IEEE, (2017), pp. 3953–3957
37. Voigtlaender, F.: The universal approximation theorem for complex-valued neural networks. *Appl. Comput. Harmon. Anal.* **64**, 33–61 (2023). <https://doi.org/10.1016/j.acha.2022.12.002>
38. Yarotsky, D.: *Optimal approximation of continuous functions by very deep ReLU networks*, Conference on learning theory, PMLR, (2018), pp. 639–649
39. Yarotsky, D.: Universal approximations of invariant maps by neural networks. *Constr. Approx.* **55**(1), 407–474 (2022). <https://doi.org/10.1007/s00365-021-09546-1>
40. Zhang, H., Gu, M., Jiang, X. D., Thompson, J., Cai, H., Paesani, S., Santagati, R., Laing, A., Zhang, Y., Yung, M. H., Shi, Y. Z., Muhammad, F. K., Lo, G. Q., Luo, X. S., Dong, B., Kwong, D. L., Kwek, L. C., Liu, A. Q.: *An optical neural chip for implementing complex-valued neural network*, Nature Communications **12** (2021), pp. article no. 457, 11 pp., <https://doi.org/10.1038/s41467-020-20719-7>
41. Zhang, Z., Wang, H., Xu, F., Jin, Y.-Q.: Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55**(12), 7177–7188 (2017). <https://doi.org/10.1109/TGRS.2017.2743222>
42. Zhou, D.-X.: Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.* **48**(2), 787–794 (2020). <https://doi.org/10.1016/j.acha.2019.06.004>