DEEP NEURAL NETWORK APPROXIMATION THEORY

DISS. ETH NO. 27547

DEEP NEURAL NETWORK APPROXIMATION THEORY

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

DMYTRO PEREKRESTENKO

M.Sc. in Electrical and Electronic Engineering, École polytechnique fédérale de Lausanne

born on 04.04.1993
citizen of Ukraine

accepted on the recommendation of

Prof. Dr. Helmut Bölcskei, examiner
Prof. Dr. Dmitry Yarotsky, coexaminer

2021

To those hungry for more in life and daring
to believe that anything is possible

Abstract

The first part of this thesis develops fundamental limits of deep neural network learning by characterizing what is possible if no constraints are imposed on the learning algorithm and on the amount of training data. Concretely, we consider Kolmogorov-optimal approximation through deep neural networks with the guiding theme being a relation between the complexity of the function (class) to be approximated and the complexity of the approximating network in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The theory we develop establishes that deep networks are Kolmogorov-optimal approximants for markedly different function classes, such as unit balls in Besov spaces and modulation spaces. In addition, deep networks provide exponential approximation accuracy—i.e., the approximation error decays exponentially in the number of nonzero weights in the network—of the multiplication operation, polynomials, sinusoidal functions, and certain smooth functions. Moreover, this holds true even for one-dimensional oscillatory textures and the Weierstrass function—a fractal function, neither of which has previously known methods achieving exponential approximation accuracy. We also show that in the approximation of sufficiently smooth functions finite-width deep networks require strictly smaller connectivity than finite-depth wide networks.

The second part of this thesis shows that every d-dimensional probability distribution with bounded support can be generated through deep ReLU networks out of a one-dimensional uniform input distribu-

tion. What is more, this is possible without incurring a cost—in terms of approximation error measured in Wasserstein-distance—relative to generating the d-dimensional target distribution from d independent random variables. This is enabled by a vast generalization of the space-filling approach discovered recently in (Bailey and Telgarsky, 2018). Moreover, our construction elicits the importance of network depth in driving the Wasserstein distance between the target distribution and its neural network approximation to zero. Finally, we demonstrate that, for histogram target distributions, the number of bits needed to uniquely encode the corresponding generative network is close to the fundamental limit as dictated by quantization theory (Graf and Luschgy, 2000).

Kurzfassung

Im ersten Teil dieser Arbeit werden fundamentale Grenzen des maschinellen Lernens mit tiefen neuronalen Netzen entwickelt. Konkret wird charakterisiert was prinzipiell möglich ist, wenn keine Beschränkungen an den Lernalgorithmus und an die Menge der verfügbaren Trainingsdaten bestehen. Dies führt zum neuartigen Konzept der Kolmogorov-optimalen Approximation durch tiefe neuronale Netze. Das zugehörige Leitthema ist eine Beziehung zwischen der Komplexität der zu approximierenden Funktion (bzw. Funktionenklasse) und der Komplexität des zugehörigen approximierenden neuronalen Netzes in Bezug auf dessen Konnektivität (bzw. die Anzahl der zur Darstellung der Netzwerktopologie und der zugehörigen quantisierten Gewichte) nötigen Bits. Die resultierende Theorie besagt, dass tiefe neuronale Netze Kolmogorov-optimale Approximation für grundlegend verschiedene Funktionenklassen, wie z. B. Einheitskugeln in Besov-Räumen und Modulationsräumen, erlauben. Darüber hinaus ermöglichen tiefe neuronale Netze exponentielle Approximationsgenauigkeit—d. h., der Approximationsfehler klingt exponentiell in der Anzahl der von Null verschiedenen Gewichte im Netz ab—für die Multiplikationsoperation, Polynome, Sinusfunktionen und bestimmte glatte Funktionen. Dies ist sogar für eindimensionale stark oszillierende Funktionen sowie die Weierstraß-Funktion—eine fraktale Funktion, möglich; für diese beiden Funktionenklassen ist bisher keine Methode bekannt, die exponentielle Approximationsgenauigkeit erzielt. Wir zeigen auch, dass in der Approximation hinre-

ichend glatter Funktionen tiefe Netze endlicher Breite strikt kleinere Konnektivität benötigen als breite Netze endlicher Tiefe.

Im zweiten Teil der Arbeit wird gezeigt, dass jede d-dimensionale Zufallsvariable mit kompakt getragener Wahrscheinlichkeitsverteilung durch tiefe ReLU-Netze aus einer eindimensionalen gleichverteilten Zufallsvariable erzeugt werden kann. Dies ist darüber hinaus möglich ohne Abstriche im erreichbaren Approximationsfehler—gemessen in der Wasserstein-Distanz—relativ zur Erzeugung der d-dimensionalen Zieldichte aus d unabhängigen Zufallsvariablen in Kauf nehmen zu müssen. Die zugrundeliegende Idee basiert auf einer weitreichenden Verallgemeinerung des Ansatzes über raumfüllende Kurven, der kürzlich in (Bailey and Telgarsky, 2018) entdeckt wurde. Die resultierende neuartige Konstruktion zeigt auch die Bedeutung von Netzwerktiefe in der Approximation der Zielverteilung mit beliebiger Genauigkeit auf. Schließlich zeigen wir, dass die zur eindeutigen Darstellung von generativen Netzen, die Histogramm-Zielverteilungen approximieren, benötigte Anzahl an Bits nahe an der durch die Quantisierungstheorie vorgegebenen fundamentalen Grenze liegt (Graf and Luschgy, 2000).

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Helmut Bölcskei, for his excellent guidance, patience and support.

I would like to thank Prof. Helmut Bölcskei and Prof. Dmitry Yarotsky for acting as examiners for this thesis.

I would also like to thank Michael Lerjen for his constant help and kindness, Recep Gül for being the closest soul to talk to, Michael Tschannen for being a strong work culture example, Verner Vlačić for his mathematical rigorousness, Erwin Riegler for his infinite positivity, Weigutian Ou for his wonderful jokes, Thomas Allard for his openness and sincerity, and thanks to Diyora Salimova, Thomas Wiatowski, Céline Aubel, the great people around me.

Finally, I would like to thank my family, wife and friends for sharing with me all these years.

Contents

1 Deep neural network approximation theory 1
1.1 Introduction 1
1.2 Setup and basic ReLU calculus 5
1.3 Approximation of multiplication, polynomials, smooth functions, and sinusoidals 10
1.4 Approximation of Function Classes and Metric Entropy 27
A Kolmogorov-Donoho Rate Distortion Theory . 29
B Metric entropy 31
1.5 Approximation with Dictionaries 38
1.6 Approximation with Deep Neural Networks 51
1.7 The Transference Principle 65
1.8 Affine Dictionaries are Effectively Representable by Neural Networks 71
A Affine Dictionaries with Canonical Ordering . 73
B Invariance to Affine Transformations 74
C Canonically Ordered Affine Dictionaries are Effectively Representable 76
D Spline wavelets 79
1.9 Weyl-Heisenberg dictionaries 84
1.10 Improving Polynomial Approximation Rates to Exponential Rates 99
1.11 Impossibility results for finite-depth networks 105
1.12 Appendices 108

A Auxiliary neural network constructions 108
B Tail compactness for Besov spaces 117
C Tail compactness for modulation spaces 118

2 High-dimensional distribution generation through deep neural networks 121
2.1 Introduction 121
2.2 Definitions and notation 124
2.3 Sawtooth functions 126
2.4 ReLU networks generate histogram distributions . . . 129
2.5 Increasing distribution dimensionality 133
2.6 Realization of transport map through quantized networks 157
2.7 Approximation of arbitrary distributions on $[0,1]^d$ by quantized histogram distributions 163
2.8 Approximation of arbitrary distributions on bounded subsets of $\mathbb{R}^d$ with generative ReLU networks 178
2.9 Complexity of generative networks 180
2.10 Appendix 184
A Proof of Lemma 27 184

3 Publications 191

References 193

CHAPTER 1

Deep neural network approximation theory

1.1. INTRODUCTION

Triggered by the availability of vast amounts of training data and drastic improvements in computing power, deep neural networks have become state-of-the-art technology for a wide range of practical machine learning tasks such as image classification (Krizhevsky et al., 2012), handwritten digit recognition (LeCun et al., 1995), speech recognition (Hinton et al., 2012), or game intelligence (D. Silver et al., 2016). For an in-depth overview, we refer to the survey paper (LeCun et al., 2015) and the recent book (Goodfellow et al., 2016).

A neural network effectively implements a mapping approximating a function that is learned based on a given set of input-output value pairs, typically through the backpropagation algorithm (Rumelhart et al., 1986). Characterizing the fundamental limits of approximation through neural networks shows what is possible if no constraints are imposed on the learning algorithm and on the amount of training data (Anthony and Bartlett, 1999).

The theory of function approximation through neural networks has a long history dating back to the work by McCulloch and Pitts (McCulloch and Pitts, 1943) and the seminal paper by Kolmogorov (Kol-

mogorov, 1957), who showed, when interpreted in neural network parlance, that any continuous function of n variables can be represented exactly through a 2-layer neural network of width $2n+1$. However, the nonlinearities in Kolmogorov's neural network are highly nonsmooth and the outer nonlinearities, i.e., those in the output layer, depend on the function to be represented. In modern neural network theory, one is usually interested in networks with nonlinearities that are independent of the function to be realized and exhibit, in addition, certain smoothness properties. Significant progress in understanding the approximation capabilities of such networks has been made in (Cybenko, 1989; Hornik, 1991), where it was shown that single-hidden-layer neural networks can approximate continuous functions on bounded domains arbitrarily well, provided that the activation function satisfies certain (mild) conditions and the number of nodes is allowed to grow arbitrarily large. In practice one is, however, often interested in approximating functions from a given function class $\mathcal{C}$ determined by the application at hand. It is therefore natural to ask how the complexity of a neural network approximating every function in $\mathcal{C}$ to within a prescribed accuracy depends on the complexity of $\mathcal{C}$ (and on the desired approximation accuracy). The recently developed Kolmogorov-Donoho rate-distortion theory for neural networks (Bölcskei et al., 2019) formalizes this question by relating the complexity of $\mathcal{C}$—in terms of the number of bits needed to describe any element in $\mathcal{C}$ to within prescribed accuracy—to network complexity in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The theory is based on a framework for quantifying the fundamental limits of nonlinear approximation through dictionaries as introduced by Donoho (Donoho, 1993, 1996).

The purpose of this chapter is to provide a comprehensive, principled, and self-contained introduction to Kolmogorov-Donoho rate-distortion optimal approximation through deep neural networks. The idea is to equip the reader with a working knowledge of the mathematical tools underlying the theory at a level that is sufficiently deep to enable further research in the field. Part of this chapter is based on (Bölcskei

et al., 2019), but extends the theory therein to the rectified linear unit (ReLU) activation function and to networks with depth scaling in the approximation error.

The theory we develop educes remarkable universality properties of finite-width deep networks. Specifically, deep networks are Kolmogorov-Donoho optimal approximants for vastly different function classes such as unit balls in Besov spaces (Mallat, 2008) and modulation spaces (Gröchenig, 2013). This universality is afforded by a concurrent invariance property of deep networks to time-shifts, scalings, and frequency-shifts. In addition, deep networks provide exponential approximation accuracy-i.e., the approximation error decays exponentially in the number of parameters employed in the approximant, namely the number of nonzero weights in the network-for vastly different functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures (Demanet and Ying, 2007) and the Weierstrass function - a fractal function, neither of which has known methods achieving exponential approximation accuracy.

While we consider networks based on the ReLU[1] activation function throughout, certain parts of our theory carry over to strongly sigmoidal activation functions of order $k \geq 2$ as defined in (Bölcskei et al., 2019). For the sake of conciseness, we refrain from providing these extensions.

Outline of the chapter. In Section 1.2, we introduce notation, formally define neural networks, and record basic elements needed in the neural network constructions throughout the chapter. Section 1.3 presents an algebra of function approximation by neural networks. In Section 1.4, we develop the Kolmogorov-Donoho rate-distortion framework that will allow us to characterize the fundamental limits of deep neural network learning of function classes. This theory is based on the concept of metric entropy, which is introduced and reviewed starting from first principles. Section 1.5 then puts the Kolmogorov-Donoho framework to work in the context of nonlinear function approximation

[1] ReLU stands for the Rectified Linear Unit nonlinearity defined as $x \mapsto \max\{0, x\}$.

with dictionaries. This discussion serves as a basis for the development of the concept of best M-weight approximation in neural networks presented in Section 1.6. We proceed, in Section 1.7, with the development of a method—termed the transference principle—for transferring results on function approximation through dictionaries to results on approximation by neural networks. The purpose of Section 1.8 is to demonstrate that function classes that are optimally approximated by affine dictionaries (e.g., wavelets), are optimally approximated by neural networks as well. In Section 1.9, we show that this optimality transfer extends to function classes that are optimally approximated by Weyl-Heisenberg dictionaries. Section 1.10 demonstrates that neural networks can improve the best-known approximation rates for two example functions, namely oscillatory textures and the Weierstrass function, from polynomial to exponential. The final Section 1.11 makes a formal case for depth in neural network approximation by establishing a provable benefit of deep networks over shallow networks in the approximation of sufficiently smooth functions. The Appendices collect ancillary technical results.

Notation. For a function $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ and a set $\Omega \subseteq \mathbb{R}^d$, we define $\|f\|_{L^\infty(\Omega)} := \sup\{|f(x)| : x \in \Omega\}$. $L^p(\mathbb{R}^d)$ and $L^p(\mathbb{R}^d, \mathbb{C})$ denote the space of real-valued, respectively complex-valued, L^p-functions. When dealing with the approximation error for simple functions such as, e.g., $(x, y) \mapsto xy$, we will for brevity of exposition and with slight abuse of notation, make the arguments inside the norm explicit according to $\|f(x, y) - xy\|_{L^p(\Omega)}$. For a vector $b \in \mathbb{R}^d$, we let $\|b\|_\infty := \max_{i=1,\ldots,d} |b_i|$, similarly we write $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ for the matrix $A \in \mathbb{R}^{m\times n}$. We denote the identity matrix of size $n \times n$ by $\mathbb{I}_n$. $\log$ stands for the logarithm to base 2. For a set $X \in \mathbb{R}^d$, we write $|X|$ for its Lebesgue measure. Constants like C are understood to be allowed to take on different values in different uses.

1.2. SETUP AND BASIC RELU CALCULUS

This section defines neural networks, introduces the basic setup as well as further notation, and lists basic elements needed in the neural network constructions considered throughout, namely compositions and linear combinations of neural networks. There is a plethora of neural network architectures and activation functions in the literature. Here, we restrict ourselves to the ReLU activation function and consider the following general network architecture.

Definition 1. *Let $L \in \mathbb{N}$ and $N_0, N_1, \ldots, N_L \in \mathbb{N}$. A* ReLU neural network Φ *is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3 \end{cases}, \quad (1.1)$$

where, for $\ell \in \{1, 2, \ldots, L\}$, $W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(x) := A_\ell x + b_\ell$ are the associated affine transformations with matrices $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $b_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho \colon \mathbb{R} \to \mathbb{R}$, $\rho(x) := \max(0, x)$ acts component-wise, i.e., $\rho(x_1, \ldots, x_N) := (\rho(x_1), \ldots, \rho(x_N))$. We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. Moreover, we define the following quantities related to the notion of size of the ReLU network Φ:

- *the* connectivity $\mathcal{M}(\Phi)$ *is the total number of nonzero entries in the matrices A_ℓ, $\ell \in \{1, 2, \ldots, L\}$, and the vectors b_ℓ, $\ell \in \{1, 2, \ldots, L\}$,*
- depth $\mathcal{L}(\Phi) := L$,
- width $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$,
- weight magnitude $\mathcal{B}(\Phi) := \max_{\ell=1,\ldots,L} \max\{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$.

Remark 1. *Note that for a given function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$, which can be expressed according to* (1.1)*, the underlying affine transformations W_ℓ are highly nonunique in general (Fefferman, 1994; Elbrächter et al., 2019). The question of uniqueness in this context is of independent interest and was addressed recently in (Vlačić and Bölcskei, 2021b,a). Whenever we talk about a given ReLU network Φ, we will either explicitly or implicitly associate Φ with a given set of affine transformations W_ℓ.*

N_0 is the dimension of the input layer *indexed as the 0-th layer, $N_1, \dots, N_{L-1}$ are the* dimensions of the $L-1$ hidden layers*, and N_L is the* dimension of the output layer. *Our definition of depth $\mathcal{L}(\Phi)$ counts the number of affine transformations involved in the representation (1.1). Single-hidden-layer neural networks hence have depth 2 in this terminology. Finally, we consider standard affine transformations as neural networks of depth 1 for technical purposes.*

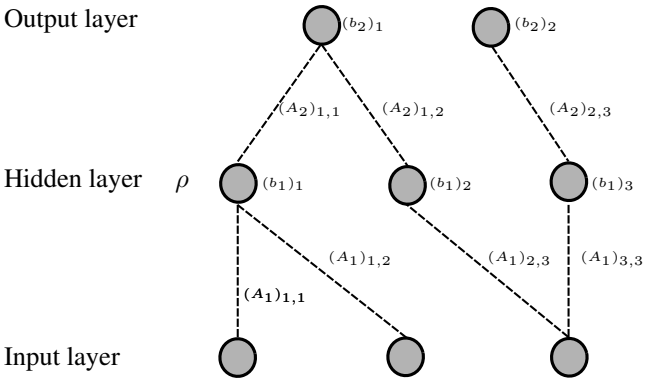
The matrix entry $(A_\ell)_{i,j}$ represents the weight associated with the edge between the j-th node in the $(\ell-1)$-th layer and the i-th node in the ℓ-th layer*, $(b_\ell)_i$ is* the weight associated with the i-th node in the ℓ-th layer. *These assignments are schematized in Figure 1.1. The real numbers $(A_\ell)_{i,j}$ and $(b_\ell)_i$ are referred to as the network's edge weights and node weights, respectively.*

Throughout the chapter, we assume that every node in the input layer and in layers $1, \dots, L-1$ has at least one outgoing edge and every node in the output layer L has at least one incoming edge. These nondegeneracy assumptions are basic as nodes that do not satisfy them can be removed without changing the functional relationship realized by the network.

Finally, we note that the connectivity satisfies

$$\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1).$$

The term "network" stems from the interpretation of the mapping Φ as a weighted acyclic directed graph with nodes arranged in hierarchical layers and edges only between adjacent layers.



$$A_2 = \begin{pmatrix} (A_2)_{1,1} & (A_2)_{1,2} & 0 \\ 0 & 0 & (A_2)_{2,3} \end{pmatrix}$$

$$A_1 = \begin{pmatrix} (A_1)_{1,1} & (A_1)_{1,2} & 0 \\ 0 & 0 & (A_1)_{2,3} \\ 0 & 0 & (A_1)_{3,3} \end{pmatrix}$$

Fig. 1.1: Assignment of the weights $(A_\ell)_{i,j}$ and $(b_\ell)_i$ of a two-layer network to the edges and nodes, respectively.

We mostly consider the case $\Phi : \mathbb{R}^d \to \mathbb{R}$, i.e., $N_L = 1$, but emphasize that our results readily generalize to $N_L > 1$.

The neural network constructions provided in the chapter frequently make use of basic elements introduced next, namely compositions and linear combinations of networks (Petersen and Voigtlaender, 2018).

Lemma 1. *Let $d_1, d_2, d_3 \in \mathbb{N}$, $\Phi_1 \in \mathcal{N}_{d_1,d_2}$, and $\Phi_2 \in \mathcal{N}_{d_2,d_3}$. Then, there exists a network $\Psi \in \mathcal{N}_{d_1,d_3}$ with $\mathcal{L}(\Psi) = \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_2)$, $\mathcal{M}(\Psi) \leq 2\mathcal{M}(\Phi_1) + 2\mathcal{M}(\Phi_2)$, $\mathcal{W}(\Psi) \leq \max\{2d_2, \mathcal{W}(\Phi_1), \mathcal{W}(\Phi_2)\}$, $\mathcal{B}(\Psi) = \max\{\mathcal{B}(\Phi_1), \mathcal{B}(\Phi_2)\}$, and satisfying*

$$\Psi(x) = (\Phi_2 \circ \Phi_1)(x) = \Phi_2(\Phi_1(x)), \quad \textit{for all } x \in \mathbb{R}^{d_1}.$$

Proof. The proof is based on the identity $x = \rho(x) - \rho(-x)$. First, note that by Definition 1, we can write

$$\Phi_1 = W^1_{L_1} \circ \rho \circ W^1_{L_1-1} \circ \cdots \circ \rho \circ W^1_1$$

and

$$\Phi_2 = W^2_{L_2} \circ \rho \circ \cdots \circ W^2_2 \circ \rho \circ W^2_1.$$

Next, let $N^1_{L_1-1}$ denote the width of layer $L_1 - 1$ in Φ_1 and let N^2_1 denote the width of layer 1 in Φ_2. We define the affine transformations $\widetilde{W}^1_{L_1} : \mathbb{R}^{N^1_{L_1-1}} \mapsto \mathbb{R}^{2d_2}$ and $\widetilde{W}^2_1 : \mathbb{R}^{2d_2} \mapsto \mathbb{R}^{N^2_1}$ according to

$$\widetilde{W}^1_{L_1}(x) := \begin{pmatrix} \mathbb{I}_{d_2} \\ -\mathbb{I}_{d_2} \end{pmatrix} W^1_{L_1}(x) \text{ and } \widetilde{W}^2_1(y) := W^2_1 \left(\begin{pmatrix} \mathbb{I}_{d_2} & -\mathbb{I}_{d_2} \end{pmatrix} y \right).$$

The proof is finalized by noting that the network

$$\Psi := W^2_{L_2} \circ \rho \circ \cdots \circ W^2_2 \circ \rho \circ \widetilde{W}^2_1 \circ \rho \circ \widetilde{W}^1_{L_1} \circ \rho \circ W^1_{L_1-1} \circ \ldots \circ \rho \circ W^1_1$$

satisfies the claimed properties. □

Unless explicitly stated otherwise, the composition of two neural networks will be understood in the sense of Lemma 1.

In order to formalize the concept of a linear combination of networks with possibly different depths, we need the following two technical lemmas which show how to augment network depth while retaining the network's input-output relation and how to parallelize networks.

Lemma 2. *Let $d_1, d_2, K \in \mathbb{N}$, and $\Phi \in \mathcal{N}_{d_1,d_2}$ with $\mathcal{L}(\Phi) < K$. Then, there exists a network $\Psi \in \mathcal{N}_{d_1,d_2}$ with $\mathcal{L}(\Psi) = K$, $\mathcal{M}(\Psi) \leq \mathcal{M}(\Phi) + d_2\mathcal{W}(\Phi) + 2d_2(K - \mathcal{L}(\Phi))$, $\mathcal{W}(\Psi) = \max\{2d_2, \mathcal{W}(\Phi)\}$, $\mathcal{B}(\Psi) = \max\{1, \mathcal{B}(\Phi)\}$, and satisfying $\Psi(x) = \Phi(x)$ for all $x \in \mathbb{R}^{d_1}$.*

Proof. Let $\widetilde{W}_j(x) := \operatorname{diag}\left(\mathbb{I}_{d_2}, \mathbb{I}_{d_2}\right) x$, for $j \in \{\mathcal{L}(\Phi)+1, \ldots, K-1\}$, $\widetilde{W}_K(x) := \begin{pmatrix} \mathbb{I}_{d_2} & -\mathbb{I}_{d_2} \end{pmatrix} x$, and note that with

$$\Phi = W_{\mathcal{L}(\Phi)} \circ \rho \circ W_{\mathcal{L}(\Phi)-1} \circ \rho \circ \cdots \circ \rho \circ W_1,$$

the network

$$\Psi := \widetilde{W}_K \circ \rho \circ \widetilde{W}_{K-1} \circ \rho \circ \cdots \circ \rho \circ \widetilde{W}_{\mathcal{L}(\Phi)+1} \circ \rho \circ \begin{pmatrix} W_{\mathcal{L}(\Phi)} \\ -W_{\mathcal{L}(\Phi)} \end{pmatrix} \circ \rho \circ W_{\mathcal{L}(\Phi)-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

satisfies the claimed properties. □

For the sake of simplicity of exposition, we state the following two lemmas only for networks of the same depth, the extension to the general case follows by straightforward application of Lemma 2. The first of these two lemmas formalizes the notion of neural network parallelization, concretely of combining neural networks implementing the functions f and g into a neural network realizing the mapping $x \mapsto (f(x), g(x))$.

Lemma 3. *Let $n, L \in \mathbb{N}$ and, for $i \in \{1, 2, \ldots, n\}$, let $d_i, d_i' \in \mathbb{N}$ and $\Phi_i \in \mathcal{N}_{d_i, d_i'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{\sum_{i=1}^n d_i, \sum_{i=1}^n d_i'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) = \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) = \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i \mathcal{B}(\Phi_i)$, and satisfying*

$$\Psi(x) = (\Phi_1(x_1), \Phi_2(x_2), \ldots, \Phi_n(x_n)) \in \mathbb{R}^{\sum_{i=1}^n d_i'},$$

for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{\sum_{i=1}^n d_i}$ with $x_i \in \mathbb{R}^{d_i}$, $i \in \mathbb{N}$.

Proof. We write the networks Φ_i as

$$\Phi_i = W_L^i \circ \rho \circ W_{L-1}^i \circ \rho \circ \cdots \circ \rho \circ W_1^i,$$

with $W_\ell^i(x) = A_\ell^i x + b_\ell^i$. Furthermore, we denote the layer dimensions of Φ_i by $N_0^i, \ldots, N_L^i$ and set $N_\ell := \sum_{i=1}^n N_\ell^i$, for $\ell \in \{0, 1, \ldots, L\}$. Next, define, for $\ell \in \{1, 2, \ldots, L\}$, the block-diagonal matrices $A_\ell := \operatorname{diag}(A_\ell^1, A_\ell^2, \ldots, A_\ell^n)$, the vectors $b_\ell = (b_\ell^1, b_\ell^2, \ldots, b_\ell^n)$, and the affine transformations $W_\ell(x) := A_\ell x + b_\ell$. The proof is concluded by noting that

$$\Psi := W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

satisfies the claimed properties. □

We are now ready to formalize the concept of a linear combination of neural networks.

Lemma 4. *Let $n, L, d' \in \mathbb{N}$ and, for $i \in \{1, 2, \ldots, n\}$, let $d_i \in \mathbb{N}$, $a_i \in \mathbb{R}$, and $\Phi_i \in \mathcal{N}_{d_i,d'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{\sum_{i=1}^n d_i,d'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) \leq \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i\{|a_i|\mathcal{B}(\Phi_i)\}$, and satisfying*

$$\Psi(x) = \sum_{i=1}^{n} a_i \Phi_i(x_i) \in \mathbb{R}^{d'},$$

for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{\sum_{i=1}^n d_i}$ with $x_i \in \mathbb{R}^{d_i}$, $i \in \{1, 2, \ldots, n\}$.

Proof. The proof follows by taking the construction in Lemma 3, replacing A_L by $(a_1 A_L^1, a_2 A_L^2, \ldots, a_n A_L^n)$, b_L by $\sum_{i=1}^n a_i b_L^i$, and noting that the resulting network satisfies the claimed properties. □

1.3. APPROXIMATION OF MULTIPLICATION, POLYNOMIALS, SMOOTH FUNCTIONS, AND SINUSOIDALS

This section constitutes the first part of the chapter dealing with the approximation of basic function "templates" through neural networks. Specifically, we shall develop an algebra of neural network approximation by starting with the squaring function, building thereon to approximate the multiplication function, proceeding to polynomials and general smooth functions, and ending with sinusoidal functions.

The basic element of the neural network algebra we develop is based on an approach by Yarotsky (Yarotsky, 2017) and by Schmidt-Hieber (Schmidt-Hieber, 2020), both of whom, in turn, employed the "sawtooth" construction from (Telgarsky, 2015).

We start by reviewing the sawtooth construction underlying our program. Consider the hat function $g : \mathbb{R} \to [0, 1]$,

$$g(x) = 2\rho(x) - 4\rho(x - \tfrac{1}{2}) + 2\rho(x - 1) = \begin{cases} 2x, & \text{if } 0 \le x < \frac{1}{2} \\ 2(1 - x), & \text{if } \frac{1}{2} \le x \le 1 \\ 0, & \text{else} \end{cases},$$

let $g_0(x) = x, g_1(x) = g(x)$, and define the s-th order sawtooth function g_s as the s-fold composition of g with itself, i.e.,

$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \quad s \ge 2. \tag{1.2}$$

We note that g can be realized by a 2-layer network $\Phi_g \in \mathcal{N}_{1,1}$ according to $\Phi_g := W_2 \circ \rho \circ W_1 = g$ with

$$W_1(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x - \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}, \qquad W_2(x) = \begin{pmatrix} 2 & -4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The s-th order sawtooth function g_s can hence be realized by a network $\Phi_g^s \in \mathcal{N}_{1,1}$ according to

$$\Phi_g^s := W_2 \circ \rho \circ \underbrace{W_g \circ \rho \circ \cdots \circ W_g \circ \rho}_{s-1} \circ W_1 = g_s \tag{1.3}$$

with

$$W_g(x) = \begin{pmatrix} 2 & -4 & 2 \\ 2 & -4 & 2 \\ 2 & -4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}.$$

The following restatement of (Telgarsky, 2015, Lemma 2.4) summarizes the self-similarity and symmetry properties of $g_s(x)$ we will frequently make use of.

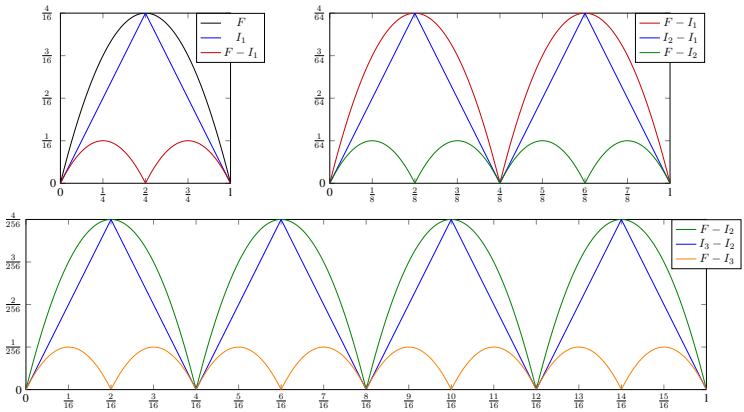


Fig. 1.2: First three steps of approximating $F(x) = x - x^2$ by an equispaced linear interpolation I_m at $2^m + 1$ points.

Lemma 5. *For $s \in \mathbb{N}$, $k \in \{0, 1, \ldots, 2^{s-1} - 1\}$, it holds that $g(2^{s-1} \cdot - k)$ is supported in $\left[\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\right]$,*

$$g_s(x) = \sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k), \quad \text{for } x \in [0, 1],$$

and

$$g_s\left(\tfrac{k}{2^{s-1}} + x\right) = g_s\left(\tfrac{k+1}{2^{s-1}} - x\right), \quad \text{for } x \in \left[0, \tfrac{1}{2^{s-1}}\right].$$

We are now ready to proceed with the statement of the basic building block of our neural network algebra, namely the approximation of the squaring function through deep ReLU networks.

Proposition 1. *There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \leq C \log(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon) = 3$, $\mathcal{B}(\Phi_\varepsilon) = 1$, $\Phi_\varepsilon(0) = 0$, satisfying*

$$\|\Phi_\varepsilon(x) - x^2\|_{L^\infty([0,1])} \leq \varepsilon.$$

Proof. The proof builds on two rather elementary observations. The first one concerns the linear interpolation $I_m\colon [0,1] \to \mathbb{R}$, $m \in \mathbb{N}$, of the function $F(x) := x - x^2$ at the points $\frac{j}{2^m}$, $j \in \{0, 1, \ldots, 2^m\}$, and in particular the self-similarity of the refinement step $I_m \to I_{m+1}$. For every $m \in \mathbb{N}$, the residual $F - I_m$ is identical on each interval between two points of interpolation (see Figure 1.2). Concretely, let $f_m\colon [0, 2^{-m}] \to [0, 2^{-2m-2}]$ be defined as $f_m(x) = 2^{-m}x - x^2$ and consider its linear interpolation $h_m\colon [0, 2^{-m}] \to [0, 2^{-2m-2}]$ at the midpoint and the endpoints of the interval $[0, 2^{-m}]$ given by

$$h_m(x) := \begin{cases} 2^{-m-1}x, & x \in [0, 2^{-m-1}] \\ -2^{-m-1}x + 2^{-2m-1}, & x \in [2^{-m-1}, 2^{-m}] \end{cases}.$$

Direct calculation shows that

$$f_m(x) - h_m(x) = \begin{cases} f_{m+1}(x), & x \in [0, 2^{-m-1}] \\ f_{m+1}(x - 2^{-m-1}), & x \in [2^{-m-1}, 2^{-m}] \end{cases}.$$

As $F = f_0$ and $I_1 = h_0$ this implies that, for all $m \in \mathbb{N}$,

$$F(x) - I_m(x) = f_m(x - \tfrac{j}{2^m}), \text{ for } x \in [\tfrac{j}{2^m}, \tfrac{j+1}{2^m}], \quad j \in \{0, 1, \ldots, 2^m - 1\}$$

and $I_m = \sum_{k=0}^{m-1} H_k$, where $H_k\colon [0,1] \to \mathbb{R}$ is given by

$$H_k(x) = h_k(x - \tfrac{j}{2^k}), \text{ for } x \in [\tfrac{j}{2^k}, \tfrac{j+1}{2^k}], \ j \in \{0, 1, \ldots, 2^k - 1\}.$$

Thus, we have

$$\begin{aligned} \sup_{x\in[0,1]} |x^2 - (x - I_m(x))| &= \sup_{x\in[0,1]} |F(x) - I_m(x)| \\ &= \sup_{x\in[0,2^{-m}]} |f_m(x)| = 2^{-2m-2}. \end{aligned} \tag{1.4}$$

The second observation we build on is a manifestation of the sawtooth construction described above and leads to economic realizations of the H_k through k-layer networks with two neurons in each layer;

a third neuron is used to realize the approximation $x - I_m(x)$ to x^2. Concretely, let $s_k(x) := 2^{-1}\rho(x) - \rho(x - 2^{-2k-1})$, and note that, for $x \in [0,1]$, $H_0 = s_0$, we get $H_k = s_k \circ H_{k-1}$. We can thus construct a network realizing $x - I_m(x)$, for $x \in [0,1]$, as follows. Let $A_1 := (1,1,1)^T \in \mathbb{R}^{3\times 1}$, $b_1 := (0, -2^{-1}, 0)^T \in \mathbb{R}^3$,

$$A_\ell := \begin{pmatrix} 2^{-1} & -1 & 0 \\ 2^{-1} & -1 & 0 \\ -2^{-1} & 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad b_\ell := \begin{pmatrix} 0 \\ -2^{-2\ell+1} \\ 0 \end{pmatrix} \in \mathbb{R}^3,$$

for $\ell \in \{2, \ldots, m\}$,

and $A_{m+1} := (-2^{-1}, 1, 1) \in \mathbb{R}^{1\times 3}$, $b_{m+1} = 0$. Setting $W_\ell(x) := A_\ell x + b_\ell$, $\ell \in \{1, 2, \ldots, m+1\}$, and

$$\widetilde{\Phi}_m := W_{m+1} \circ \rho \circ W_m \circ \rho \circ \cdots \circ \rho \circ W_1,$$

a direct calculation yields $\widetilde{\Phi}_m(x) = x - \sum_{k=0}^{m-1} H_k(x)$, for $x \in [0,1]$. The proof is completed upon noting that the networks $\Phi_\varepsilon := \widetilde{\Phi}_{\lceil \log(\varepsilon^{-1})/2 \rceil}$ satisfy the claimed properties. □

The symmetry properties of $g_s(x)$ according to Lemma 5 lead to the interpolation error in the proof of Proposition 1 being identical in each interval, with the maximum error taken on at the centers of the respective intervals. More importantly, however, the approximating neural networks realize linear interpolation at a number of points that grows exponentially in network depth. This is a manifestation of the fact that the number of linear regions in the sawtooth construction (1.3) grows exponentially with depth, which, owing to Lemma 18, is optimal. We emphasize that the theory developed in this chapter hinges critically on this optimality property, which, however, is brittle in the sense that networks with weights obtained through training will, as observed in (Hanin and Rolnick, 2019), in general, not exhibit exponential growth of the number of linear regions with network depth. An interesting approach to neural network training which manages to partially circumvent this problem was proposed recently in (Fokina

and Oseledets, 2019). Understanding how the number of linear regions grows in general trained networks and quantifying the impact of this—possibly subexponential—growth behavior on the approximation-theoretic fundamental limits of neural networks constitutes a major open problem.

We proceed to the construction of networks that approximate the multiplication function over the interval $[-D, D]$. This will be effected by using the result on the approximation of x^2 just established combined with the polarization identity $xy = \frac{1}{4}((x+y)^2 - (x-y)^2)$, the fact that $\rho(x) + \rho(-x) = |x|$, and a scaling argument exploiting that the ReLU function is positive homogeneous, i.e., $\rho(\lambda x) = \lambda\rho(x)$, for all $\lambda \geq 0$, $x \in \mathbb{R}$.

Proposition 2. *There exists a constant $C > 0$ such that, for all $D \in \mathbb{R}_+$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{N}_{2,1}$ with $\mathcal{L}(\Phi_{D,\varepsilon}) \leq C(\log(\lceil D \rceil) + \log(\varepsilon^{-1}))$, $\mathcal{W}(\Phi_{D,\varepsilon}) \leq 5$, $\mathcal{B}(\Phi_{D,\varepsilon}) = 1$, satisfying $\Phi_{D,\varepsilon}(0, x) = \Phi_{D,\varepsilon}(x, 0) = 0$, for all $x \in \mathbb{R}$, and*

$$\|\Phi_{D,\varepsilon}(x, y) - xy\|_{L^\infty([-D,D]^2)} \leq \varepsilon. \tag{1.5}$$

Proof. We first note that, w.l.o.g., we can assume $D \geq 1$ in the following, as for $D < 1$, we can simply employ the network constructed for $D = 1$ to guarantee the claimed properties. The proof builds on the polarization identity and essentially constructs two squaring networks according to Proposition 1 which share the neuron responsible for summing up the H_k, preceded by a layer mapping (x, y) to $(|x+y|/(2D), |x-y|/(2D))$ and followed by layers realizing the multiplication by D^2 through weights bounded by 1. Specifically, consider the network $\widetilde{\Psi}_m$ with associated matrices A_ℓ and vectors b_ℓ given by

$$A_1 := \frac{1}{2D}\begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 2}, \quad b_1 := 0 \in \mathbb{R}^4,$$

$$A_2 := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{5\times 4}, \quad b_2 := \begin{pmatrix} 0 \\ -2^{-1} \\ 0 \\ 0 \\ -2^{-1} \end{pmatrix},$$

$$A_\ell := \begin{pmatrix} 2^{-1} & -1 & 0 & 0 & 0 \\ 2^{-1} & -1 & 0 & 0 & 0 \\ -2^{-1} & 1 & 1 & 2^{-1} & -1 \\ 0 & 0 & 0 & 2^{-1} & -1 \\ 0 & 0 & 0 & 2^{-1} & -1 \end{pmatrix} \in \mathbb{R}^{5\times 5},$$

$$b_\ell := \begin{pmatrix} 0 \\ -2^{-2\ell+3} \\ 0 \\ 0 \\ -2^{-2\ell+3} \end{pmatrix}, \quad \text{for } \ell \in \{3, \dots, m+1\},$$

and $A_{m+2} := (-2^{-1}, 1, 1, 2^{-1}, -1) \in \mathbb{R}^{1\times 5}$, $b_{m+2} := 0$. A direct calculation yields

$$\begin{aligned} \widetilde{\Psi}_m(x,y) &= \left(\tfrac{|x+y|}{2D} - \sum_{k=0}^{m-1} H_k\big(\tfrac{|x+y|}{2D}\big) \right) \\ &\quad - \left(\tfrac{|x-y|}{2D} - \sum_{k=0}^{m-1} H_k\big(\tfrac{|x-y|}{2D}\big) \right) \\ &= \widetilde{\Phi}_m\left(\tfrac{|x+y|}{2D}\right) - \widetilde{\Phi}_m\left(\tfrac{|x-y|}{2D}\right), \end{aligned} \tag{1.6}$$

with H_k and $\widetilde{\Phi}_m$ as defined in the proof of Proposition 1. With (1.4)

this implies

$$\begin{aligned}
&\sup_{(x,y)\in[-D,D]^2} \left|\widetilde{\Psi}_m(x,y) - \tfrac{xy}{D^2}\right| \\
&= \sup_{(x,y)\in[-D,D]^2} \Big| \left(\widetilde{\Phi}_m\left(\tfrac{|x+y|}{2D}\right) - \widetilde{\Phi}_m\left(\tfrac{|x-y|}{2D}\right)\right) \\
&- \left(\left(\tfrac{|x+y|}{2D}\right)^2 - \left(\tfrac{|x-y|}{2D}\right)^2\right)\Big| \\
&\leq 2 \sup_{z\in[0,1]} |\widetilde{\Phi}_m(z) - z^2| \leq 2^{-2m-1}.
\end{aligned} \tag{1.7}$$

Next, let $\Psi_D(x) = D^2 x$ be the scalar multiplication network according to Lemma 14 and take $\Phi_{D,\varepsilon} := \Psi_D \circ \widetilde{\Psi}_{m(D,\varepsilon)}$, where $m(D,\varepsilon) := \lceil 2^{-1}(1+\log(D^2\varepsilon^{-1}))\rceil$. Then, the error estimate (1.5) follows directly from (1.7) and Lemma 1 establishes the desired bounds on depth, width, and weight magnitude. Finally, $\Phi_{D,\varepsilon}(0,x) = \Phi_{D,\varepsilon}(x,0) = 0$, for all $x \in \mathbb{R}$, follows directly from (1.6). □

Remark 2. *Note that the multiplication network just constructed has weights bounded by* 1 *irrespectively of the size* D *of the domain. This is accomplished by trading network depth for weight magnitude according to Lemma 14.*

We proceed to the approximation of polynomials, effected by networks that realize linear combinations of monomials, which, in turn, are built by composing multiplication networks. Before presenting the specifics of this construction, we hasten to add that a similar approach was considered previously in (Yarotsky, 2017) and (Schmidt-Hieber, 2020). While there are slight differences in formulation, the main distinction between our construction and those in (Yarotsky, 2017) and (Schmidt-Hieber, 2020) resides in their purpose. Specifically, the goal in (Yarotsky, 2017) and (Schmidt-Hieber, 2020) is to establish, by way of local Taylor-series approximation, that d-variate, k-times (weakly) differentiable functions can be approximated in L^∞-norm to within error ε with networks of connectivity scaling according to

$\varepsilon^{-d/k}\log(\varepsilon^{-1})$. Here, on the other hand, we will be interested in functions that allow approximation with networks of connectivity scaling polylogarithmically in ε^{-1} (i.e., as a polynomial in $\log(\varepsilon^{-1})$). Moreover, for ease of exposition, we will employ finite-width networks. Polylogarithmic connectivity scaling will turn out to be crucial (see Sections 1.6-1.9) in establishing Kolmogorov-Donoho rate-distortion optimality of neural networks in the approximation of a variety of prominent function classes. Finally, we would like to mention related recent work (Schwab and Zech, 2019; Opschoor et al., 2020), (Gühring et al., 2020) on the approximation of Sobolev-class functions in certain Sobolev norms enabled by neural network approximations of the multiplication operation and of polynomials.

Proposition 3. *There exists a constant $C > 0$ such that for all $m \in \mathbb{N}$, $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$, $D \in \mathbb{R}_+$, and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_{a,D,\varepsilon}) \leq Cm(\log(\varepsilon^{-1}) + m\log(\lceil D\rceil) + \log(m) + \log(\lceil\|a\|_\infty\rceil))$, $\mathcal{W}(\Phi_{a,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Phi_{a,D,\varepsilon}) \leq 1$, and satisfying*

$$\|\Phi_{a,D,\varepsilon}(x) - \sum_{i=0}^m a_i x^i\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Proof. As in the proof of Proposition 2 and for the same reason, it suffices to consider the case $D \geq 1$. For $m = 1$, we simply have an affine transformation and the statement follows directly from Corollary 2. The proof for $m \geq 2$ will be effected by realizing the monomials $x^k, k \geq 2$, through iterative composition of multiplication networks and combining this with a construction that uses the network realizing x^k not only as a building block in the network implementing x^{k+1} but also to approximate the partial sum $\sum_{i=0}^k a_i x^i$ in parallel.

We start by setting $B_k = B_k(D,\eta) := \lceil D\rceil^k + \eta\sum_{s=0}^{k-2}\lceil D\rceil^s$, $k \in \mathbb{N}, \eta \in \mathbb{R}_+$ and take $\Phi_{B_k,\eta}$ to be the multiplication network from Proposition 2. Next, we recursively define the functions

$$f_{k,D,\eta}(x) = \Phi_{B_{k-1},\eta}(x, f_{k-1,D,\eta}(x)), \quad k \geq 2,$$

with $f_{0,D,\eta}(x) = 1$ and $f_{1,D,\eta}(x) = x$. For notational simplicity, we use the abbreviation $f_k = f_{k,D,\eta}$ in the following. First, we verify that the $f_{k,D,\eta}$ approximate monomials sufficiently well. Specifically, we prove by induction that

$$\|f_k(x) - x^k\|_{L^\infty([-D,D])} \leq \eta \sum_{s=0}^{k-2} \lceil D \rceil^s, \tag{1.8}$$

for all $k \geq 2$. The base case $k = 2$, i.e.,

$$\|f_2(x) - x^2\|_{L^\infty([-D,D])} = \|\Phi_{B_1,\eta}(x, x) - x^2\|_{L^\infty([-D,D])} \leq \eta,$$

follows directly from Proposition 2 upon noting that $D \leq B_1 = \lceil D \rceil$ (we take the sum in the definition of B_k to equal zero when the upper limit of summation is negative). We proceed to establish the induction step $(k-1) \to k$ with the induction assumption given by

$$\|f_{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \leq \eta \sum_{s=0}^{k-3} \lceil D \rceil^s.$$

As

$$\begin{aligned}
&\|f_{k-1}\|_{L^\infty([-D,D])} \\
&\leq \|x^{k-1}\|_{L^\infty([-D,D])} + \|f_{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \\
&\leq B_{k-1},
\end{aligned}$$

application of Proposition 2 yields

$$\begin{aligned}
&\|f_k(x) - x^k\|_{L^\infty([-D,D])} \\
&\leq \|f_k(x) - x f_{k-1}(x)\|_{L^\infty([-D,D])} + \|x f_{k-1}(x) - x^k\|_{L^\infty([-D,D])} \\
&\leq \|\Phi_{B_{k-1},\eta}(x, f_{k-1}(x)) - x f_{k-1}(x)\|_{L^\infty([-D,D])} \\
&+ D\|f_{k-1}(x) - x^{k-1}\|_{L^\infty([-D,D])} \\
&\leq \eta + \lceil D \rceil \eta \sum_{s=0}^{k-3} \lceil D \rceil^s \\
&= \eta \sum_{s=0}^{k-2} \lceil D \rceil^s,
\end{aligned}$$

which completes the induction.

We now construct the network $\Phi_{a,D,\varepsilon}$ approximating the polynomial $\sum_{i=0}^{m} a_i x^i$. To this end, note that there exists a constant C' such that for all $m \geq 2$, $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$, and $i \in \{1, \ldots, m-1\}$, there is a network $\Psi_{a,D,\eta}^i \in \mathcal{N}_{3,3}$ with $\mathcal{L}(\Psi_{a,D,\eta}^i) \leq C'(\log(\eta^{-1}) + \log(\lceil B_i \rceil) + \log(\|a\|_\infty))$, $\mathcal{W}(\Psi_{a,D,\eta}^i) \leq 9$, $\mathcal{B}(\Psi_{a,D,\eta}^i) \leq 1$, and satisfying

$$\Psi_{a,D,\eta}^i(x, s, y) = (x, s + a_i y, \Phi_{B_i,\eta}(x, y)).$$

To see that this is, indeed, the case, consider the following chain of mappings

$$(x, s, y) \xrightarrow{(I)} (x, s, y, y) \xrightarrow{(II)} (x, s + a_i y, y) \xrightarrow{(III)} (x, s + a_i y, x, y)$$
$$\xrightarrow{(IV)} (x, s + a_i y, \Phi_{B_i,\eta}(x, y)).$$

Observe that the mapping (I) is an affine transformation with coefficients in $\{0, 1\}$, which we can simply consider to be a depth-1 network. The mapping (II) is obtained by using Corollary 2 in order to implement the affine transformation $(s, y) \mapsto s + a_i y$ with weights bounded by 1, followed by application of Lemmas 2 and 3 to put this network in parallel with two networks realizing the identity mapping according to $x = \rho(x) - \rho(-x)$. Mapping (III) is obtained along the same lines by putting the result of mapping (II) in parallel with another network realizing the identity mapping. Finally, mapping (IV) is realized by putting the network $\Phi_{B_i,\eta}$ in parallel with two identity networks. Composing these four networks according to Lemma 1 yields, for $i \in \{1, \ldots, m-1\}$, a network $\Psi_{a,D,\eta}^i$ with the claimed properties. Next, we employ Corollary 2 to get networks $\Psi_{a,D,\eta}^0$ which implement $x \mapsto (x, a_0, x)$ as well as networks $\Psi_{a,D,\eta}^m$ realizing $(x, s, y) \mapsto s + a_m y$. Let now $\eta = \eta(a, D, \varepsilon) := (\|a\|_\infty (m-1)^2 \lceil D \rceil^{m-2})^{-1} \varepsilon$ and define

$$\Phi_{a,D,\varepsilon} := \Psi_{a,D,\eta}^m \circ \Psi_{a,D,\eta}^{m-1} \circ \cdots \circ \Psi_{a,D,\eta}^1 \circ \Psi_{a,D,\eta}^0.$$

A direct calculation yields

$$\Phi_{a,D,\varepsilon} = \sum_{i=0}^{m} a_i f_{i,D,\eta}.$$

Hence (1.8) implies

$$\begin{aligned}
&\Big\|\Phi_{a,D,\varepsilon}(x) - \sum_{i=0}^{m} a_i x^i\Big\|_{L^\infty([-D,D])} \\
&\leq \sum_{i=0}^{m} |a_i| \|f_{i,D,\eta}(x) - x^i\|_{L^\infty([-D,D])} \\
&\leq \sum_{i=2}^{m} |a_i| \Big(\eta \sum_{s=0}^{i-2} \lceil D \rceil^s\Big) \leq \|a\|_\infty \eta \sum_{k=0}^{m-2} (m-1-k) \lceil D \rceil^k \\
&\leq \|a\|_\infty (m-1)^2 \lceil D \rceil^{m-2} \eta = \varepsilon.
\end{aligned}$$

Lemma 1 now establishes that $\mathcal{W}(\Phi_{a,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Phi_{a,D,\varepsilon}) \leq 1$, and

$$\begin{aligned}
\mathcal{L}(\Phi_{a,D,\varepsilon}) &\leq \sum_{i=0}^{m} \mathcal{L}(\Psi^i_{a,D,\eta}) \\
&\leq 2(\log(\lceil \|a\|_\infty \rceil) + 5) \\
&\quad + \sum_{i=1}^{m-1} C'(\log(\eta^{-1}) + \log(\lceil B_{i-1} \rceil) + \log(\lceil \|a\|_\infty \rceil)) \\
&\leq Cm(\log(\varepsilon^{-1}) + m \log(\lceil D \rceil) + \log(m) + \log(\lceil \|a\|_\infty \rceil))
\end{aligned}$$

for a suitably chosen absolute constant C. This completes the proof. □

Next, we recall that the Weierstrass approximation theorem states that every continuous function on a closed interval can be approximated to within arbitrary accuracy by a polynomial.

Theorem 1 ((Stone, 1948))**.** *Let $[a, b] \subseteq \mathbb{R}$ and $f \in C([a, b])$. Then, for every $\varepsilon > 0$, there exists a polynomial π such that*

$$\|f - \pi\|_{L^\infty([a,b])} \leq \varepsilon.$$

Proposition 3 hence allows us to conclude that every continuous function on a closed interval can be approximated to within arbitrary accuracy by a deep ReLU network of width no more than 9. This amounts to a variant of the universal approximation theorem (Cybenko, 1989; Hornik, 1991) for finite-width deep ReLU networks. A quantitative statement in terms of making the approximating network's width, depth, and weight bounds explicit can be obtained for (very) smooth functions by applying Proposition 3 to Lagrangian interpolation with Chebyshev points.

Lemma 6. *Consider the set*

$$\mathcal{S}_{[-1,1]} := \big\{ f \in C^\infty([-1,1],\mathbb{R})\colon \|f^{(n)}(x)\|_{L^\infty([-1,1])} \leq n!, \text{ for all } n \in \mathbb{N}_0 \big\}.$$

There exists a constant $C > 0$ such that for all $f \in \mathcal{S}_{[-1,1]}$ and $\varepsilon \in (0,1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq 1$, and satisfying

$$\|\Psi_{f,\varepsilon} - f\|_{L^\infty([-1,1])} \leq \varepsilon.$$

Proof. A fundamental result on Lagrangian interpolation with Chebyshev points (see e.g. (Liang and Srikant, 2017, Lemma 3)) guarantees, for all $f \in \mathcal{S}_{[-1,1]}$, $m \in \mathbb{N}$, the existence of a polynomial $P_{f,m}$ of degree m such that

$$\|f - P_{f,m}\|_{L^\infty([-1,1])} \leq \tfrac{1}{(m+1)!2^m} \|f^{(m+1)}\|_{L^\infty([-1,1])} \leq \tfrac{1}{2^m}.$$

Note that $P_{f,m}$ can be expressed in the Chebyshev basis (see e.g. (Gil et al., 2007, Section 3.4.1)) according to $P_{f,m} = \sum_{j=0}^m c_{f,m,j} T_j(x)$ with $|c_{f,m,j}| \leq 2$ and the Chebyshev polynomials defined through the two-term recursion $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, $k \geq 2$, with $T_0(x) = 1$ and $T_1(x) = x$. We can moreover use this recursion to conclude that the coefficients of the T_k in the monomial basis are upper-bounded by 3^k. Consequently, we can express $P_{f,m}$ according

to $P_{f,m} = \sum_{j=0}^{m} a_{f,m,j} x^j$ with

$$A_{f,m} := \max_{j=0,\ldots,m} |a_{f,m,j}| \leq 2(m+1)3^m.$$

Application of Proposition 3 to $P_{f,m}$ in the monomial basis, with $m = \lceil \log(2/\varepsilon) \rceil$ and approximation error $\varepsilon/2$, completes the proof upon noting that

$$C' m(\log(2/\varepsilon) + \log(m) + \log(|A_{f,m}|)) \leq C(\log(\varepsilon^{-1}))^2$$

for some absolute constant C. □

An extension of Lemma 6 to approximation over general intervals is provided in Lemma 17. While Lemma 6 shows that a specific class of C^∞-functions, namely those whose derivatives are suitably bounded, can be approximated by neural networks with connectivity growing polylogarithmically in ε^{-1}, it turns out that this is not possible for general (Sobolev-class) k-times differentiable functions (Yarotsky, 2017, Thm. 4).

We are now ready to proceed to the approximation of sinusoidal functions. Before stating the corresponding result, we comment on the basic idea enabling the approximation of oscillatory functions through deep neural networks. In essence, we exploit the optimality of the sawtooth construction (1.3) in terms of achieving exponential—in network depth—growth in the number of linear regions. As indicated in Figure 1.3, the composition of the cosine function (realized according to Lemma 6) with the sawtooth function, combined with the symmetry properties of the cosine function and the sawtooth function, yields oscillatory behavior that increases exponentially with network depth.

Theorem 2. *There exists a constant $C > 0$ such that for every $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{a,D,\varepsilon}) \leq 1$, and satisfying*

$$\|\Psi_{a,D,\varepsilon}(x) - \cos(ax)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Proof. Note that $f(x) := (6/\pi^3)\cos(\pi x)$ is in $\mathcal{S}_{[-1,1]}$. Thus, by Lemma 6, there exists a constant $C > 0$ such that for every $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Phi_\varepsilon) \leq 9$, $\mathcal{B}(\Phi_\varepsilon) \leq 1$, and satisfying

$$\|\Phi_\varepsilon - f\|_{L^\infty([-1,1])} \leq \tfrac{6}{\pi^3}\varepsilon. \tag{1.9}$$

We now extend this result to the approximation of $x \mapsto \cos(ax)$ on the interval $[-1, 1]$ for arbitrary $a \in \mathbb{R}_+$. This will be accomplished by exploiting that $x \mapsto \cos(\pi x)$ is 2-periodic and even. Let $g_s \colon [0, 1] \to [0, 1]$, $s \in \mathbb{N}$, be the s-th order sawtooth functions as defined in (1.2) and note that, due to the periodicity and the symmetry of the cosine function (see Figure 1.3 for illustration), we have for all $s \in \mathbb{N}_0$, $x \in [-1, 1]$,

$$\cos(\pi 2^s x) = \cos(\pi g_s(|x|)).$$

For $a > \pi$, we define $s = s(a) := \lceil \log(a) - \log(\pi) \rceil$ and $\alpha = \alpha(a) := (\pi 2^s)^{-1} a \in (1/2, 1]$, and note that

$$\cos(ax) = \cos(\pi 2^s \alpha x) = \cos(\pi g_s(\alpha|x|)), \quad x \in [-1, 1].$$

As $g_s(\alpha|x|) \in [0, 1]$, it follows from (1.9) that

$$\begin{aligned} &\|\tfrac{\pi^3}{6}\Phi_\varepsilon(g_s(\alpha|x|)) - \cos(ax)\|_{L^\infty([-1,1])} \\ &= \tfrac{\pi^3}{6}\|\Phi_\varepsilon(g_s(\alpha|x|)) - f(g_s(\alpha|x|))\|_{L^\infty([-1,1])} \leq \varepsilon. \end{aligned} \tag{1.10}$$

In order to realize $\Phi_\varepsilon(g_s(\alpha|x|))$ as a neural network, we start from the networks Φ_g^s defined in (1.3) and apply Proposition 9 to convert them into networks $\Psi_g^s(x) = g_s(x)$, for $x \in [0, 1]$, with $\mathcal{B}(\Psi_g^s) \leq 1$, $\mathcal{L}(\Psi_g^s) = 7(s + 1)$, and $\mathcal{W}(\Psi_g^s) = 3$. Furthermore, let $\Psi(x) := \alpha\rho(x) - \alpha\rho(-x) = \alpha|x|$ and take $\Phi^{\text{mult}}_{\pi^3/6}$ to be the scalar multiplication network from Lemma 14. Noting that $\Psi_{a,\varepsilon} := \Phi^{\text{mult}}_{\pi^3/6} \circ \Phi_\varepsilon \circ \Psi_g^s \circ \Psi = \Phi_\varepsilon(g_s(\alpha|x|))$ and concluding from Lemma 1 that $\mathcal{L}(\Psi_{a,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil a \rceil))$, $\mathcal{W}(\Psi_{a,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{a,\varepsilon}) \leq 1$, together with (1.10), establishes the desired result for

$a > \pi$ and for approximation over the interval $[-1, 1]$. For $a \in (0, \pi)$, we can simply take $\Psi_{a,\varepsilon} := \Phi^{\text{mult}}_{\pi^3/6} \circ \Phi_\varepsilon$ as $x \mapsto (6/\pi^3)\cos(ax)$ is in $\mathcal{S}_{[-1,1]}$ in this case.

Finally, we consider the approximation of $x \mapsto \cos(ax)$ on intervals $[-D, D]$, for arbitrary $D \geq 1$. To this end, we define the networks $\Psi_{a,D,\varepsilon}(x) := \Psi_{aD,\varepsilon}(\frac{x}{D})$ and observe that

$$\begin{aligned} &\sup_{x\in[-D,D]} |\Psi_{a,D,\varepsilon}(x) - \cos(ax)| \\ &= \sup_{y\in[-1,1]} |\Psi_{a,D,\varepsilon}(Dy) - \cos(aDy)| \\ &= \sup_{y\in[-1,1]} |\Psi_{aD,\varepsilon}(y) - \cos(aDy)| \\ &\leq \varepsilon. \end{aligned} \tag{1.11}$$

This concludes the proof. □

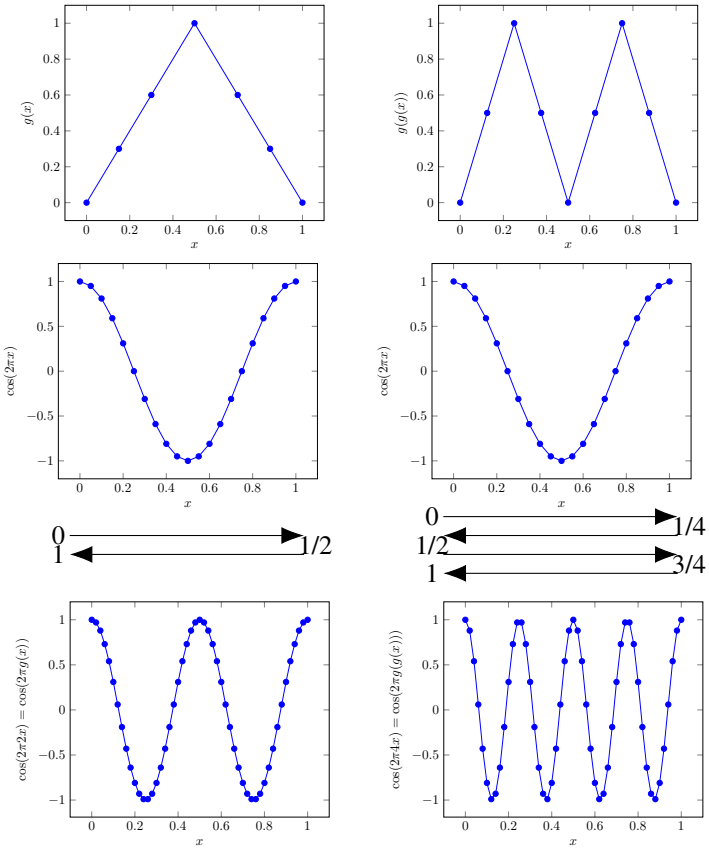


Fig. 1.3: Approximation of the function $\cos(2\pi a x)$ according to Theorem 2 using "sawtooth" functions $g_s(x)$ as per (1.2), left $a = 2$, right $a = 4$.

The result just obtained extends to the approximation of $x \mapsto \sin(ax)$, formalized next, simply by noting that $\sin(x) = \cos(x - \pi/2)$.

Corollary 1. *There exists a constant $C > 0$ such that for every $a, D \in \mathbb{R}_+$, $b \in \mathbb{R}$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{a,b,D,\varepsilon} \in \mathcal{N}_{1,1}$ with*

$\mathcal{L}(\Psi_{a,b,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil aD + |b| \rceil))$, $\mathcal{W}(\Psi_{a,b,D,\varepsilon}) \leq 9$, $\mathcal{B}(\Psi_{a,b,D,\varepsilon}) \leq 1$, *and satisfying*

$$\|\Psi_{a,b,D,\varepsilon}(x) - \cos(ax - b)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Proof. For given $a, D \in \mathbb{R}_+$, $b \in \mathbb{R}$, $\varepsilon \in (0, 1/2)$, consider the network $\Psi_{a,b,D,\varepsilon}(x) := \Psi_{a,D+\frac{|b|}{a},\varepsilon}\left(x - \frac{b}{a}\right)$ with $\Psi_{a,D,\varepsilon}$ as defined in the proof of Theorem 2, and observe that, owing to (1.11),

$$\begin{aligned} &\sup_{x\in[-D,D]} |\Psi_{a,b,D,\varepsilon}(x) - \cos(ax - b)| \\ &\leq \sup_{y\in\left[-(D+\frac{|b|}{a}),D+\frac{|b|}{a}\right]} |\Psi_{a,D+\frac{|b|}{a},\varepsilon}(y) - \cos(ay)| \leq \varepsilon. \end{aligned}$$

□

Remark 3. *The results in this section all have approximating networks of finite width and depth scaling polylogarithmically in ε^{-1}. Owing to*

$$\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi) + 1)$$

this implies that the connectivity scales no faster than polylogarithmic in ε^{-1}. It therefore follows that the approximation error ε decays (at least) exponentially fast in the connectivity or equivalently in the number of parameters the approximant (i.e., the neural network) employs. We say that the network provides exponential approximation accuracy.

1.4. APPROXIMATION OF FUNCTION CLASSES AND METRIC ENTROPY

So far we considered the explicit construction of deep neural networks for the approximation of a wide range of functions, namely polynomials, smooth functions, and sinusoidal functions, in all cases with exponential accuracy, i.e., with an approximation error that decays

exponentially in network connectivity. We now proceed to lay the foundation for the development of a framework that allows us to characterize the fundamental limits of deep neural network approximation of entire function classes. But first, we provide a review of relevant literature.

The best-known results on approximation by neural networks are the universal approximation theorems of Hornik (Hornik, 1991) and Cybenko (Cybenko, 1989), stating that continuous functions on bounded domains can be approximated arbitrarily well by a single-hidden-layer ($L = 2$ in our terminology) neural network with sigmoidal activation function. The literature on approximation-theoretic properties of networks with a single hidden layer continuing this line of work is abundant. Without any claim to completeness, we mention work on approximation error bounds in terms of the number of neurons for functions with Fourier transforms of bounded first moments (Barron, 1993), (Barron, 1994), the nonexistence of localized approximations (Chui et al., 1994), a fundamental lower bound on approximation rates (DeVore et al., 1996; Candès, 1998), and the approximation of smooth or analytic functions (Mhaskar, 1996; Mhaskar and Micchelli, 1995).

Approximation-theoretic results for networks with multiple hidden layers were obtained in (Hornik et al., 1989; Mhaskar, 1993) for general functions, in (Funahashi, 1989) for continuous functions, and for functions together with their derivatives in (Nguyen-Thien and Tran-Cong, 1999). In (Chui et al., 1994) it was shown that for certain approximation tasks deep networks can perform fundamentally better than single-hidden-layer networks. We also highlight two recent papers, which investigate the benefit—from an approximation-theoretic perspective—of multiple hidden layers. Specifically, in (Eldan and Shamir, 2016) it was shown that there exists a function which, although expressible through a small three-layer network, can only be represented through a very large two-layer network; here size is measured in terms of the total number of neurons in the network.

In the setting of deep convolutional neural networks first results of a nature similar to those in (Eldan and Shamir, 2016) were reported

in (Mhaskar and Poggio, 2016). Linking the expressivity properties of neural networks to tensor decompositions, (Cohen et al., 2016; Cohen and Shashua, 2016) established the existence of functions that can be realized by relatively small deep convolutional networks but require exponentially larger shallow convolutional networks.

We conclude by mentioning recent results bearing witness to the approximation power of deep ReLU networks in the context of PDEs. Specifically, it was shown in (Schwab and Zech, 2019) that deep ReLU networks can approximate very effectively certain solution families of parametric PDEs depending on a large (possibly infinite) number of parameters. The series of papers (Grohs et al., 2018; Berner et al., 2020; Beck et al., 2018; Elbrächter et al., 2018) constructs and analyzes a deep-learning-based numerical solver for Black-Scholes PDEs.

For survey articles on approximation-theoretic aspects of neural networks, we refer the interested reader to (Ellacott, 1994) and (Pinkus, 1999) as well as the very recent (DeVore et al., 2020). Most closely related to the framework we develop here is the paper by Shaham, Cloninger, and Coifman (Shaham et al., 2018), which shows that for functions that are sparse in specific wavelet frames, the best M-weight approximation rate (see Definition 8 below) of three-layer neural networks is at least as large as the best M-term approximation rate in piecewise linear wavelet frames.

We begin the development of our framework with a review of a widely used theoretical foundation for deterministic lossy data compression (DeVore and Lorentz, 1993; DeVore, 1998). Our presentation essentially follows (Donoho, 1993; Grohs, 2015).

A. Kolmogorov-Donoho Rate Distortion Theory

Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and consider a set of functions $\mathcal{C} \subseteq L^2(\Omega)$, which we will frequently refer to as *function class*. Then, for each $\ell \in \mathbb{N}$, we denote by

$$\mathfrak{E}^\ell := \left\{ E : \mathcal{C} \to \{0,1\}^\ell \right\}$$

the set of *binary encoders of $\mathcal{C}$ of length ℓ*, and we let

$$\mathfrak{D}^\ell := \left\{D : \{0,1\}^\ell \to L^2(\Omega)\right\}$$

be the set of *binary decoders of length ℓ*. An encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ is said to *achieve uniform error ε over the function class $\mathcal{C}$*, if

$$\sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon.$$

Note that here we quantified the approximation error in $L^2(\Omega)$-norm, whereas in the previous section we used the $L^\infty(\Omega)$-norm. While results in terms of $L^\infty(\Omega)$-norm are stronger, we shall employ the $L^2(\Omega)$-norm in order to parallel the Kolmogorov-Donoho framework for nonlinear approximation through dictionaries (Donoho, 1993, 1996). We furthermore note that for sets Ω of finite Lebesgue measure $|\Omega|$, the two norms are related through $\|f\|_{L^2(\Omega)} \leq |\Omega|^{1/2}\|f\|_{L^\infty(\Omega)}$. Finally, whenever we talk about compactness and related topological notions, we shall always mean w.r.t. the topology induced by the $L^2(\Omega)$-norm.

A quantity of central interest is the minimal length $\ell \in \mathbb{N}$ for which there exists an encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ that achieves uniform error ε over the function class $\mathcal{C}$, along with its asymptotic behavior as made precise in the following definition.

Definition 2. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. Then, for $\varepsilon > 0$, the* minimax code length $L(\varepsilon, \mathcal{C})$ *is*

$$L(\varepsilon,\mathcal{C}) := \min\Big\{\ell \in \mathbb{N} : \exists (E,D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon\Big\}. \tag{1.12}$$

Moreover, the optimal exponent $\gamma^*(\mathcal{C})$ *is defined as*

$$\gamma^*(\mathcal{C}) := \sup\left\{\gamma \in \mathbb{R} : L(\varepsilon,\mathcal{C}) \in \mathcal{O}\left(\varepsilon^{-1/\gamma}\right),\ \varepsilon \to 0\right\}.$$

The optimal exponent $\gamma^*(\mathcal{C})$ determines the minimum growth rate of $L(\varepsilon, \mathcal{C})$ as the error ε tends to zero and can hence be seen as quantifying the "description complexity" of the function class $\mathcal{C}$. Larger $\gamma^*(\mathcal{C})$ results in smaller growth rate and hence smaller memory requirements for storing functions $f \in \mathcal{C}$ such that reconstruction with uniformly bounded error is possible.

Remark 4. *The optimal exponent $\gamma^*(\mathcal{C})$ can equivalently be thought of as quantifying the asymptotic behavior of the minimal achievable error for the function class $\mathcal{C}$ with a given code length. Specifically, we have*

$$\begin{aligned}\gamma^*(\mathcal{C}) &= \sup\left\{\gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\big(\varepsilon^{-1/\gamma}\big),\, \varepsilon \to 0\right\} \\ &= \sup\left\{\gamma \in \mathbb{R} : \varepsilon(L) \in \mathcal{O}\big(L^{-\gamma}\big),\, L \to \infty\right\},\end{aligned} \tag{1.13}$$

where

$$\varepsilon(L) := \inf_{(E,D)\in \mathfrak{E}^L \times \mathfrak{D}^L} \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)}.$$

The quantity $\gamma^*(\mathcal{C})$ is closely related to the concept of Kolmogorov-Tikhomirov epsilon entropy a.k.a. metric entropy (Ott, 2002). We next make this connection explicit.

B. Metric entropy

Most of the discussion in this subsection, which is almost exclusively of review nature, follows very closely (Wainwright, 2019, Chapter 5). Consider the metric space $(\mathcal{X}, \rho)$ with $\mathcal{X}$ a nonempty set and $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a distance function. A natural measure for the size of a compact subset $\mathcal{C}$ of $\mathcal{X}$ is given by the number of balls of a fixed radius ε required to cover $\mathcal{C}$, a quantity known as the covering number (for covering radius ε).

Definition 3. *(Wainwright, 2019) Let $(\mathcal{X}, \rho)$ be a metric space. An ε-covering of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \ldots, x_N\} \subseteq \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in$*

$\{1, \ldots, N\}$ *so that* $\rho(x, x_i) \leq \varepsilon$. *The* ε*-covering number* $N(\varepsilon; \mathcal{C}, \rho)$ *is the cardinality of the smallest* ε*-covering.*

An ε-covering is a collection of balls of radius ε that cover the set $\mathcal{C}$, i.e.,

$$\mathcal{C} \subseteq \bigcup_{i=1}^{N(\varepsilon; \mathcal{C}, \rho)} B(x_i, \varepsilon),$$

where $B(x_i, \varepsilon)$ is a ball—in the metric ρ—of radius ε centered at x_i. The covering number is nonincreasing in ε, i.e., $N(\varepsilon; \mathcal{C}, \rho) \geq N(\varepsilon'; \mathcal{C}, \rho)$, for all $\varepsilon \leq \varepsilon'$. When the set $\mathcal{C}$ is not finite, the covering number goes to infinity as ε goes to zero. We shall be interested in the corresponding rate of growth, more specifically in the quantity $\log N(\varepsilon; \mathcal{C}, \rho)$ known as the metric entropy of $\mathcal{C}$ with respect to ρ. Recall that $\log$ is to the base 2, hence the unit of metric entropy is "bits". The operational significance of metric entropy follows from the question: What is the minimum number of bits needed to represent any element $x \in \mathcal{C}$ with error—quantified in terms of the distance measure ρ—of at most ε? By what was just developed, the answer to this question is $\lceil \log N(\varepsilon; \mathcal{C}, \rho) \rceil$. Specifically, for a given $x \in \mathcal{X}$, the corresponding encoder $E(x)$ simply identifies the closest ball center x_i and encodes the index i using $\lceil \log N(\varepsilon; \mathcal{C}, \rho) \rceil$ bits. The corresponding decoder D delivers the ball center x_i, which guarantees that the resulting error satisfies $\|D(E(x)) - x\| \leq \varepsilon$.

We proceed with a simple example ((Wainwright, 2019, Example 5.2)) computing an upper bound on the metric entropy of the interval $\mathcal{C} = [-1, 1]$ in $\mathbb{R}$ with respect to the metric $\rho(x, x') = |x - x'|$. To this end, we divide $\mathcal{C}$ into intervals of length 2ε by setting $x_i = -1 + 2(i-1)\varepsilon$, for $i \in [1, L]$, where $L = \lfloor \frac{1}{\varepsilon} \rfloor + 1$. This guarantees that, for every point $x \in [-1, 1]$, there is an $i \in [1, L]$ such that $|x - x_i| \leq \varepsilon$, which, in turn, establishes

$$N(\varepsilon; \mathcal{C}, \rho) \leq \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 \leq \frac{1}{\varepsilon} + 1$$

and hence yields an upper bound on metric entropy according to[2]

$$\log N(\varepsilon; \mathcal{C}, \rho) \leq \log\left(\frac{1}{\varepsilon} + 1\right) \asymp \log(\varepsilon^{-1}), \quad \text{as } \varepsilon \to 0. \tag{1.14}$$

This result can be generalized to the d-dimensional unit cube to yield $\log(N(\varepsilon; \mathcal{C}, \rho)) \leq d \log(1/\varepsilon + 1) \asymp d \log(\varepsilon^{-1})$. In order to show that the upper bound (1.14) correctly reflects metric entropy scaling for $\mathcal{C} = [-1, 1]$ with respect to $\rho(x, x') = |x - x'|$, we would need a lower bound on $N(\varepsilon; \mathcal{C}, \rho)$ that exhibits the same scaling (in ε) behavior. A systematic approach to establishing lower bounds on metric entropy is through the concept of packing, which will be introduced next.

We start with the definition of the packing number of a compact set $\mathcal{C}$ in a metric space $(\mathcal{X}, \rho)$.

Definition 4. *(Wainwright, 2019, Definition 5.4) Let $(\mathcal{X}, \rho)$ be a metric space. An ε-packing of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, \ldots, x_N\} \subseteq \mathcal{C}$ such that $\rho(x_i, x_j) > \varepsilon$, for all distinct i, j. The ε-packing number $M(\varepsilon; \mathcal{X}, \rho)$ is the cardinality of the largest ε-packing.*

An ε-packing is a collection of nonintersecting balls of radius $\varepsilon/2$ and centered at elements in $\mathcal{X}$. Although different, the covering number and the packing number provide essentially the same measure of size of a set as formalized next.

Lemma 7. *(Wainwright, 2019, Lemma 5.5) Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C}$ a compact set in $\mathcal{X}$. For all $\varepsilon > 0$, the packing and the covering number are related according to*

$$M(2\varepsilon; \mathcal{C}, \rho) \leq N(\varepsilon; \mathcal{C}, \rho) \leq M(\varepsilon; \mathcal{C}, \rho).$$

Proof. (Wainwright, 2019; Prosser, 1966) First, choose a minimal ε-covering and a maximal 2ε-packing of $\mathcal{C}$. Since no two centers of the

[2]The notation $f(\varepsilon) \asymp g(\varepsilon)$, as $\varepsilon \to 0$, means that there are constants $c, C, \varepsilon_0 > 0$ such that $cf(\varepsilon) \leq g(\varepsilon) \leq Cf(\varepsilon)$, for all $\varepsilon \leq \varepsilon_0$. For ease of exposition, we shall usually omit the qualifier $\varepsilon \to 0$.

2ε-packing can lie in the same ball of the ε-covering, it follows that $M(2\varepsilon; \mathcal{C}, \rho) \leq N(\varepsilon; \mathcal{C}, \rho)$. To establish $N(\varepsilon; \mathcal{C}, \rho) \leq M(\varepsilon; \mathcal{C}, \rho)$, we note that, given a maximal packing $M(\varepsilon; \mathcal{C}, \rho)$, for any $x \in \mathcal{C}$, we have the center of at least one of the balls in the packing within distance less than ε. If this were not the case, we could add another ball to the packing thereby violating its maximality. This maximal packing hence also provides an ε-covering and since $N(\varepsilon; \mathcal{C}, \rho)$ is a minimal covering, we must have $N(\varepsilon; \mathcal{C}, \rho) \leq M(\varepsilon; \mathcal{C}, \rho)$. □

We now return to the example in which we computed an upper bound on the metric entropy of $\mathcal{C} = [-1, 1]$ with respect to $\rho(x, x') = |x - x'|$ and show how Lemma 7 can be employed to establish the scaling behavior of metric entropy. To this end, we simply note that the points $x_i = -1 + 2(i-1)\varepsilon$, $i \in [1, L]$, are separated according to $|x_i - x_j| = 2\varepsilon > \varepsilon$, for all $i \neq j$, which implies that $M(\varepsilon; \mathcal{C}, |\cdot|) \geq L = \lfloor 1/\varepsilon \rfloor + 1 \geq \frac{1}{\varepsilon}$. Combining this with the upper bound (1.14) and Lemma 7, we obtain $\log N(\varepsilon; \mathcal{C}, |\cdot|) \asymp \log(\varepsilon^{-1})$. Likewise, it can be established that $\log N(\varepsilon; \mathcal{C}, \|\cdot\|) \asymp d \log(\varepsilon^{-1})$ for the d-dimensional unit cube. This illustrates how an explicit construction of a packing set can be used to determine the scaling behavior of metric entropy.

We next formalize the notion that metric entropy is determined by the volume of the corresponding covering balls. Specifically, the following result establishes a relationship between a certain volume ratio and metric entropy.

Lemma 8. *(Wainwright, 2019, Lemma 5.7) Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^d$, and let $\mathcal{B}$ and $\mathcal{B}'$ be their corresponding unit balls, i.e., $\mathcal{B} = \{x \in \mathbb{R}^d | \|x\| \leq 1\}$ and $\mathcal{B}' = \{x \in \mathbb{R}^d | \|x\|' \leq 1\}$. Then, the ε-covering number of $\mathcal{B}$ in the $\|\cdot\|'$-norm satisfies*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{vol(\mathcal{B})}{vol(\mathcal{B}')} \leq N(\varepsilon; \mathcal{B}, \|\cdot\|') \leq \frac{vol(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}')}{vol(\mathcal{B}')}. \tag{1.15}$$

Proof. (Wainwright, 2019) Let $\{x_1, \ldots, x_{N(\varepsilon; \mathcal{B}, \|\cdot\|')}\}$ be an ε-covering

of $\mathcal{B}$ in $\|\cdot\|'$-norm. Then, we have

$$\mathcal{B} \subseteq \bigcup_{j=1}^{N(\varepsilon;\mathcal{B},\|\cdot\|')} \{x_j + \varepsilon\mathcal{B}'\},$$

which implies $vol(\mathcal{B}) \leq N(\varepsilon;\mathcal{B},\|\cdot\|')\,\varepsilon^d\,vol(\mathcal{B}')$, thus establishing the lower bound in (1.15). The upper bound is obtained by starting with a maximal ε-packing $\{x_1,\ldots,x_{M(\varepsilon;\mathcal{B},\|\cdot\|')}\}$ of $\mathcal{B}$ in the $\|\cdot\|'$-norm. The balls $\{x_j + \frac{\varepsilon}{2}\mathcal{B}', j = 1,\ldots,M(\varepsilon;\mathcal{B},\|\cdot\|')\}$ are all disjoint and contained within $\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'$. We can therefore conclude that

$$\sum_{j=1}^{M(\varepsilon;\mathcal{B},\|\cdot\|')} vol\Big(x_j + \frac{\varepsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'\Big),$$

and hence

$$M(\varepsilon;\mathcal{B},\|\cdot\|')\,vol\Big(\frac{\varepsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}'\Big).$$

Finally, we have $vol(\frac{\varepsilon}{2}\mathcal{B}') = (\frac{\varepsilon}{2})^d vol(\mathcal{B}')$ and

$$vol(\mathcal{B} + \frac{\varepsilon}{2}\mathcal{B}') = (\frac{\varepsilon}{2})^d vol(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}'),$$

which, together with $M(\varepsilon;\mathcal{B},\|\cdot\|') \geq N(\varepsilon;\mathcal{B},\|\cdot\|')$ due to Lemma 7, yields the upper bound in (1.15). □

This result now allows us to establish the scaling of the metric entropy of unit balls in terms of their own norm, thus yielding a measure of the massiveness of unit balls in d-dimensional spaces. Specifically, we set $\mathcal{B}' = \mathcal{B}$ in Lemma 8 and get

$$vol\left(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}'\right) = vol\left(\left(\frac{2}{\varepsilon} + 1\right)\mathcal{B}\right) = \left(\frac{2}{\varepsilon} + 1\right)^d vol(\mathcal{B}),$$

which when used in (1.15) yields $N(\varepsilon;\mathcal{B},\|\cdot\|) \asymp \varepsilon^{-d}$ and hence results in metric entropy scaling according to $\log(N(\varepsilon;\mathcal{B},\|\cdot\|)) \asymp d\log(\varepsilon^{-1})$.

Particularizing this result to the unit ball $\mathcal{B}_\infty^d = [-1,1]^d$ and the metric $\|\cdot\|_\infty$, we recover the result of our direct analysis in the example above.

So far we have been concerned with the metric entropy of subsets of $\mathbb{R}^d$. We now proceed to analyzing the metric entropy of function classes, which will eventually allow us to establish the desired connection between the optimal exponent $\gamma^*(\mathcal{C})$ and metric entropy. We begin with the simple one-parameter function class considered in (Wainwright, 2019, Example 5.9) and follow closely the exposition in (Wainwright, 2019). For a fixed θ, define the real-valued function $f_\theta(x) = 1 - e^{-\theta x}$, and consider the class

$$\mathcal{P} = \{f_\theta : [0,1] \to \mathbb{R} \,|\, \theta \in [0,1]\}.$$

The set $\mathcal{P}$ constitutes a metric space under the sup-norm given by $\|f-g\|_{L^\infty([0,1])} = \sup_{x\in[0,1]} |f(x)-g(x)|$. We show that the covering number of $\mathcal{P}$ satisfies

$$1 + \left\lfloor \frac{1 - 1/e}{2\varepsilon} \right\rfloor \leq N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \leq \frac{1}{2\varepsilon} + 2,$$

which leads to the scaling behavior

$$N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \asymp \varepsilon^{-1}$$

and hence to metric entropy scaling according to

$$\log(N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])})) \asymp \log(\varepsilon^{-1}).$$

We start by establishing the upper bound. For given $\varepsilon \in [0,1]$, set $T = \lfloor \frac{1}{2\varepsilon} \rfloor$, and define the points $\theta_i = 2\varepsilon i$, for $i = 0, 1, \ldots, T$. By also adding the point $\theta_{T+1} = 1$, we obtain a collection of $T+2$ points $\{\theta_0, \theta_1, \ldots, \theta_{T+1}\}$ in $[0,1]$. We show that the associated functions $\{f_{\theta_0}, f_{\theta_1}, \ldots, f_{\theta_{T+1}}\}$ form an ε-covering for $\mathcal{P}$. Indeed, for any $f_\theta \in \mathcal{P}$, we can find some θ_i in the covering such that $|\theta - \theta_i| \leq \varepsilon$. We then have

$$\|f_\theta - f_{\theta_i}\|_{L^\infty([0,1])} = \max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| \leq |\theta - \theta_i|,$$

where we used, for $\theta < \theta_i$,

$$\begin{aligned}\max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| &= \max_{x\in[0,1]} (e^{-\theta x} - e^{-\theta_i x}) \\ &= \max_{x\in[0,1]} e^{-\theta x}(1 - e^{-(\theta_i-\theta)x}) \\ &\leq \max_{x\in[0,1]} (1 - e^{-(\theta_i-\theta)x}) \\ &\leq \max_{x\in[0,1]} (\theta_i - \theta)x \leq \theta_i - \theta \\ &= |\theta - \theta_i|,\end{aligned}$$

as a consequence of $1-e^{-x} \leq x$, for $x \in [0,1]$, which is easily verified by noting that the function $g(x) = 1 - e^{-x} - x$ satisfies $g(0) = 0$ and $g'(x) \leq 0$, for $x \in [0,1]$. The case $\theta > \theta_i$ follows similarly. In summary, we have shown that $N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \leq T+2 \leq \frac{1}{2\varepsilon}+2$.

In order to derive the lower bound, we first bound the packing number from below and then use Lemma 7. We start by constructing an explicit packing as follows. Set $\theta_0 = 0$ and define $\theta_i = -\log(1 - \varepsilon i)$, for all i such that $\theta_i \leq 1$. The largest index T such that this holds is given by $T = \lfloor \frac{1-1/e}{\varepsilon} \rfloor$. Moreover, note that for all i, j with $i \neq j$, we have $\|f_{\theta_i} - f_{\theta_j}\|_{L^\infty([0,1])} \geq |f_{\theta_i}(1) - f_{\theta_j}(1)| = |\varepsilon(i-j)| \geq \varepsilon$. We can therefore conclude that $M(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq \lfloor \frac{1-1/e}{\varepsilon} \rfloor + 1$, and hence, due to the lower bound in Lemma 7,

$$N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq M(2\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])}) \geq \left\lfloor \frac{1-1/e}{2\varepsilon} \right\rfloor + 1,$$

as claimed. We have thus established that the function class $\mathcal{P}$ has metric entropy scaling according to

$$\log(N(\varepsilon; \mathcal{P}, \|\cdot\|_{L^\infty([0,1])})) \asymp \log(1/\varepsilon), \text{ as } \varepsilon \to 0.$$

This rate is typical for one-parameter function classes.

We now turn our attention to richer function classes and start by considering Lipschitz functions on the d-dimensional unit cube, meaning real-valued functions on $[0,1]^d$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_\infty, \qquad \text{for all} \quad x, y \in [0,1]^d.$$

This class, denoted as $\mathcal{F}_L([0,1]^d)$, has metric entropy scaling (Kolmogorov and Tikhomirov, 1959; Wainwright, 2019)

$$\log N(\varepsilon; \mathcal{F}_L, \|\cdot\|_{L^\infty([0,1]^d)}) \asymp (L/\varepsilon)^d. \tag{1.16}$$

Contrasting the exponential dependence of metric entropy in (1.16) on the ambient dimension d to the linear dependence we identified earlier for simpler sets such as unit balls in $\mathbb{R}^d$, where we had

$$\log N(\varepsilon; \mathcal{B}, \|\cdot\|_\infty) \asymp d \log(\varepsilon^{-1}),$$

shows that $\mathcal{F}_L([0,1]^d)$ is significantly more massive.

We are now ready to relate the optimal exponent $\gamma^*(\mathcal{C})$ in Definition 2 to metric entropy scaling. All the examples of metric entropy scaling we have seen exhibit a behavior that fits the law $\log(N(\varepsilon; \mathcal{C}, \|\cdot\|)) \asymp \varepsilon^{-1/\gamma}$ or $\log(N(\varepsilon; \mathcal{C}, \|\cdot\|)) \asymp \varepsilon^{-1/\gamma} \log(\varepsilon^{-1})^\beta$. The optimal exponent is hence a crude measure of growth insensitive to $\log$-factors or similar factors that are dominated by the growth of $\varepsilon^{-1/\gamma}$.

While we restrict ourselves to the approximation of functions on Euclidean domains, the framework described in this section can be extended to functions on manifolds (see e.g. (Ehler and Filbir, 2018)). As such, an interesting direction for future research would be the extension of the deep neural network approximation theory developed in this chapter to functions on manifolds. First results on the neural network approximation of functions on manifolds have been reported in (Shaham et al., 2018; Bölcskei et al., 2019; Schmidt-Hieber, 2019). For further reading on the general subject of function approximation on manifolds, we recommend (Mhaskar, 2020) and references therein.

1.5. APPROXIMATION WITH DICTIONARIES

We now show how Kolmogorov-Donoho rate-distortion theory can be put to work in the context of optimal approximation with dictionaries. Again, this subsection is of review nature. We start with a brief

discussion of basics on optimal approximation in Hilbert spaces. Specifically, we shall consider two types of approximation, namely linear and nonlinear.

Let $\mathcal{H}$ be a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|_{\mathcal{H}}$ and let e_k, $k = 1, 2, \ldots$, be an orthonormal basis for $\mathcal{H}$. For linear approximation, we use the linear space $\mathcal{H}_M := \text{span}\{e_k : 1 \leq k \leq M\}$ to approximate a given element $f \in \mathcal{H}$. We measure the approximation error by

$$E_M(f) := \inf_{g \in \mathcal{H}_M} \|f - g\|_{\mathcal{H}}.$$

In nonlinear approximation, we consider best M-term approximation, which replaces $\mathcal{H}_M$ by the set Σ_M consisting of all elements $g \in \mathcal{H}$ that can be expressed as

$$g = \sum_{k \in \Lambda} c_k e_k,$$

where $\Lambda \subseteq \mathbb{N}$ is a set of indices with $|\Lambda| \leq M$. Note that, in contrast to $\mathcal{H}_M$, the set Σ_M is not a linear space as a linear combination of two elements in Σ_M will, in general, need $2M$ terms in its representation by the e_k. Analogous to E_M, we define the error of best M-term approximation

$$\Gamma_M(f) := \inf_{g \in \Sigma_M} \|f - g\|_{\mathcal{H}}.$$

The key difference between linear and nonlinear approximation resides in the fact that in nonlinear approximation, we can choose the M elements e_k participating in the approximation of f freely from the entire orthonormal basis whereas in linear approximation we are constrained to the first M elements. A classical example for linear approximation is the approximation of periodic functions by the Fourier series elements corresponding to the M lowest frequencies (assuming natural ordering of the dictionary). This approach clearly leads to poor approximation if the function under consideration consists of high-frequency components. In contrast, in nonlinear approximation we would seek the M

frequencies that yield the smallest approximation error. In summary, it is clear that (nonlinear) best M-term approximation can achieve smaller approximation error than linear M-term approximation.

We shall consider nonlinear approximation in arbitrary, possibly redundant, dictionaries, i.e., in frames (Morgenshtern and Bölcskei, 2012), and will exclusively be interested in the case $\mathcal{H} = L^2(\Omega)$, in particular the approximation error will be measured in terms of $L^2(\Omega)$-norm. Specifically, let $\mathcal{C}$ be a set of functions in $L^2(\Omega)$ and consider a countable family of functions $\mathcal{D} := (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$, termed *dictionary*.

We consider the *best M-term approximation error* of $f \in \mathcal{C}$ in $\mathcal{D}$ defined as follows.

Definition 5. *(DeVore and Lorentz, 1993) Given $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, a function class $\mathcal{C} \subseteq L^2(\Omega)$, and a dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{D}}(f) := \inf_{\substack{I_{f,M} \subseteq \mathbb{N}, \\ |I_{f,M}|=M, (c_i)_{i\in I_{f,M}}}} \left\| f - \sum_{i\in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} . \qquad (1.17)$$

We call $\Gamma_M^{\mathcal{D}}(f)$ the best M-term approximation error of f in $\mathcal{D}$. *Every $f_M = \sum_{i\in I_{f,M}} c_i\varphi_i$ attaining the infimum in* (1.17) *is referred to as a* best M-term approximation *of f in $\mathcal{D}$. The supremal $\gamma > 0$ such that*

$$\sup_{f\in\mathcal{C}} \Gamma_M^{\mathcal{D}}(f) \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

will be denoted by $\gamma^*(\mathcal{C},\mathcal{D})$. We say that the best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$ *is $\gamma^*(\mathcal{C},\mathcal{D})$.*

Function classes $\mathcal{C}$ widely studied in the approximation theory literature include unit balls in Lebesgue, Sobolev, or Besov spaces (DeVore, 1998), as well as α-cartoon-like functions (Grohs et al., 2016a). A wealth of structured dictionaries $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets (Daubechies, 1992), followed

by ridgelets (Candès, 1998), curvelets (Candès and Donoho, 2002), shearlets (Guo et al., 2006), parabolic molecules (Grohs and Kutyniok, 2014), and most generally α-molecules (Grohs et al., 2016a), which include all previously named dictionaries as special cases. Further examples are Gabor frames (Gröchenig, 2013), Wilson bases (Gröchenig and Samarah, 2000), and wave atoms (Demanet and Ying, 2007).

The best M-term approximation rate $\gamma^*(\mathcal{C}, \mathcal{D})$ according to Definition 5 quantifies how difficult it is to approximate a given function class $\mathcal{C}$ in a fixed dictionary $\mathcal{D}$. It is sensible to ask whether for given $\mathcal{C}$, there is a fundamental limit on $\gamma^*(\mathcal{C}, \mathcal{D})$ when one is allowed to vary over $\mathcal{D}$. To answer this question, we first note that for every dense (and countable) $\mathcal{D}$, for any given $f \in \mathcal{C}$, by density of $\mathcal{D}$, there exists a single dictionary element that approximates f to within arbitrary accuracy thereby effectively realizing a 1-term approximation for arbitrary approximation error ε. Formally, this can be expressed through $\gamma^*(\mathcal{C}, \mathcal{D}) = \infty$. Identifying this single dictionary element or, more generally, the M elements participating in the best M-term approximation is in general, however, practically infeasible as it entails searching through the infinite set $\mathcal{D}$ and requires an infinite number of bits to describe the indices of the participating elements. This insight leads to the concept of "best M-term approximation subject to polynomial-depth search" as introduced by Donoho in (Donoho, 1996). Here, the basic idea is to restrict the search for the elements in $\mathcal{D}$ participating in the best M-term approximation to the first $\pi(M)$ elements of $\mathcal{D}$, with π a polynomial. We formalize this under the name of effective best M-term approximation as follows.

Definition 6. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\mathcal{C} \subseteq L^2(\Omega)$ be compact, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$. We define for $M \in \mathbb{N}$ and π a polynomial*

$$\varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) := \sup_{f\in\mathcal{C}} \inf_{\substack{I_{f,M}\subseteq\{1,2,\dots,\pi(M)\},\\ |I_{f,M}|=M,\,|c_i|\leq\pi(M)}} \left\| f - \sum_{i\in I_{f,M}} c_i\varphi_i \right\|_{L^2(\Omega)} \tag{1.18}$$

and

$$\gamma^{*,\textit{eff}}(\mathcal{C},\mathcal{D}) := \sup\{\gamma \geq 0 \colon \exists \,\text{polynomial}\, \pi \\ \text{s.t.}\, \varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) \in \mathcal{O}(M^{-\gamma}),\ M \to \infty\}. \tag{1.19}$$

We refer to $\gamma^{*,\textit{eff}}(\mathcal{C},\mathcal{D})$ *as the* effective best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$.

Note that we required the coefficients c_i in the approximant in Definition 6 to be polynomially bounded in M. This condition, not present in (Donoho, 1993; Grohs, 2015) and easily met for generic $\mathcal{C}$ and $\mathcal{D}$, is imposed for technical reasons underlying the transference results in Section 1.7. Strictly speaking—relative to (Donoho, 1993; Grohs, 2015)—we hence get a subtly different notion of approximation rate. Exploring the implications of this difference is certainly worthwhile, but deemed beyond the scope of this chapter.

We next present a central result in best M-term approximation theory stating that for compact $\mathcal{C} \subseteq L^2(\Omega)$, the effective best M-term approximation rate in any dictionary $\mathcal{D}$ is upper-bounded by $\gamma^*(\mathcal{C})$ and hence limited by the "description complexity" of $\mathcal{C}$. This endows $\gamma^*(\mathcal{C})$ with operational meaning.

Theorem 3. *(Donoho, 1993; Grohs, 2015) Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. The effective best M-term approximation rate of the function class $\mathcal{C} \subseteq L^2(\Omega)$ in the dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ satisfies*

$$\gamma^{*,\textit{eff}}(\mathcal{C},\mathcal{D}) \leq \gamma^*(\mathcal{C}).$$

In light of this result the following definition is natural (see also (Grohs, 2015)).

Definition 7. *(Kolmogorov-Donoho optimality) Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. If the effective best M-term approximation rate of the function class $\mathcal{C} \subseteq L^2(\Omega)$ in the dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ satisfies*

$$\gamma^{*,\textit{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C}),$$

we say that the function class $\mathcal{C}$ is optimally representable *by $\mathcal{D}$.*

As the ideas underlying the proof of Theorem 3 are essential ingredients in the development of a kindred theory of best M-weight approximation rates for neural networks, we present a detailed proof, which is similar to that in (Grohs, 2015). We perform, however, some minor technical modifications with an eye towards rendering the proof a suitable genesis for the new theory of best M-weight approximation with neural networks, developed in the next section. The spirit of the proof is to construct, for every given $M \in \mathbb{N}$ an encoder that, for each $f \in \mathcal{C}$, maps the indices of the dictionary elements participating in the effective best M-term approximation[3] of f, along with the corresponding coefficients c_i, to a bitstring. This bitstring needs to be of sufficient length for the decoder to be able to reconstruct an approximation to f with an error which is of the same order as that of the best M-term approximation we started from. As elucidated in the proof, this can be accomplished while ensuring that the length of the bitstring is proportional to $M \log(M)$, which upon noting that $\varepsilon = M^{-\gamma}$ implies $M = \varepsilon^{-1/\gamma}$, establishes optimality.

Proof of Theorem 3. The proof will be based on showing that for every $\gamma \in \mathbb{R}_+$ the following Implication (I) holds: Assume that there exist a constant $C > 0$ and a polynomial π such that for every $M \in \mathbb{N}$, the following holds: For every $f \in \mathcal{C}$, there are an index set $I_{f,M} \subseteq \{1, 2, \dots, \pi(M)\}$ and coefficients $(c_i)_{i \in I_{f,M}} \subseteq \mathbb{R}$ with $|c_i| \leq \pi(M)$ so that

$$\Big\| f - \sum_{i \in I_{f,M}} c_i \varphi_i \Big\|_{L^2(\Omega)} \leq C M^{-\gamma}. \tag{1.20}$$

This implies the existence of a constant $C' > 0$ such that for every $M \in \mathbb{N}$, there is an encoder-decoder pair $(E_M, D_M) \in \mathfrak{E}^{\ell(M)} \times \mathfrak{D}^{\ell(M)}$

[3] Note that as we have an infimum in (1.18) an effective best M-term approximation need not exist, but we can pick an M-term approximation that yields an error arbitrarily close to the infimum.

with $\ell(M) \leq C' M \log(M)$ and

$$\|f - D_M(E_M(f))\|_{L^2(\Omega)} \leq C' M^{-\gamma}. \tag{1.21}$$

The implication will be proven by explicit construction. For a given $f \in \mathcal{C}$, we pick an M-term approximation according to (1.20) and encode the associated index set $I_{f,M}$ and weights c_i as follows. First, note that owing to $|I_{f,M}| \leq \pi(M)$, each index in $I_{f,M}$ can be represented by at most $C_\pi \log(M)$ bits; this results in a total of $C_\pi M \log(M)$ bits needed to encode the indices of all dictionary elements participating in the M-term approximation. The encoder and the decoder are assumed to know C_π, which allows stacking of the binary representations of the indices such that the decoder can read them off uniquely from the sequence of their binary representations.

We proceed to the encoding of the coefficients c_i. First, note that even though the c_i are bounded (namely, polynomially in M) by assumption, we did not impose bounds on the norms of the dictionary elements $\{\varphi_i\}_{i \in I_{f,M}}$ participating in the M-term approximation under consideration. Hence, we can not, in general, expect to be able to control the approximation error incurred by reconstructing f from quantized c_i. We can get around this by performing a Gram-Schmidt orthogonalization on the dictionary elements $\{\varphi_i\}_{i \in I_{f,M}}$ and, as will be seen later, using the fact that the function class $\mathcal{C}$ was assumed to be compact. Specifically, this Gram-Schmidt orthogonalization yields a set of functions $\{\widetilde{\varphi}_i\}_{i \in \widetilde{I}_{f,M}}$, with $\widetilde{M} \leq M$, that has the same span as $\{\varphi_i\}_{i \in I_{f,M}}$. Next, we define (implicitly) the coefficients $\widetilde{c}_i$ according to

$$\sum_{i \in \widetilde{I}_{f,M}} \widetilde{c}_i \widetilde{\varphi}_i = \sum_{i \in I_{f,M}} c_i \varphi_i. \tag{1.22}$$

Now, note that

$$\left\| \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)}^2 = \left\| f - (f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i) \right\|_{L^2(\Omega)}^2$$
$$\leq \|f\|_{L^2(\Omega)}^2 + \left\| f - \sum_{i \in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)}^2 .$$

Making use of the orthonormality of the $\widetilde{\varphi}_i$, we can conclude that

$$\sum_{i \in \widetilde{I}_{f,\widetilde{M}}} |\widetilde{c}_i|^2 \leq \sup_{f \in \mathcal{C}} \|f\|_{L^2(\Omega)}^2 + C^2 M^{-2\gamma}.$$

As $\mathcal{C}$ is compact by assumption, we have $\sup_{f \in \mathcal{C}} \|f\|_{L^2(\Omega)}^2 < \infty$, which establishes that the coefficients $\widetilde{c}_i$ are uniformly bounded. This, in turn, allows us to quantize them, specifically, we shall round the $\widetilde{c}_i$ to integer multiples of $M^{-(\gamma+1/2)}$, and denote the resulting rounded coefficients by $\hat{c}_i$. As the $\widetilde{c}_i$ are uniformly bounded, this results in a number of quantization levels that is proportional to $M^{(\gamma+1/2)}$. The number of bits needed to store the binary representations of the quantized coefficients is therefore proportional to $M \log(M)$. Again, the proportionality constant is assumed known to encoder and decoder, which allows us to stack the binary representations of the quantized coefficients in a uniquely decodable manner. The resulting bitstring is then appended to the bitstring encoding the indices of the participating dictionary elements. We finally note that the specific choice of the exponent $\gamma + 1/2$ is informed by the upper bound on the reconstruction error we are allowed, this will be made explicit below in the description of the decoder.

In summary, we have mapped the function f to a bitstring of length $\mathcal{O}(M \log(M))$. The decoder is presented with this bitstring and reconstructs an approximation to f as follows. It first reads out the indices of the set $I_{f,M}$ and the quantized coefficients $\hat{c}_i$. Recall that this is

uniquely possible. Next, the decoder performs a Gram-Schmidt orthonormalization on the set of dictionary elements indexed by $I_{f,M}$. The error resulting from reconstructing the function f from the quantized coefficients $\hat{c}_i$ rather than the exact coefficients $\widetilde{c}_i$ can be bounded according to

$$\begin{aligned}
&\left\| f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \hat{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)} \\
&= \left\| f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i + \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \hat{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)} \\
&\leq \left\| f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)} + \left\| \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} (\widetilde{c}_i - \hat{c}_i) \widetilde{\varphi}_i \right\|_{L^2(\Omega)} \\
&= \left\| f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \widetilde{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)} + \left(\sum_{i \in \widetilde{I}_{f,\widetilde{M}}} |\widetilde{c}_i - \hat{c}_i|^2 \right)^{1/2},
\end{aligned} \tag{1.23}$$

where in the last step we again exploited the orthonormality of the $\widetilde{\varphi}_i$. Next, note that due to the choice of the quantizer resolution, we have $|\widetilde{c}_i - \hat{c}_i|^2 \leq C'' M^{-2\gamma-1}$ for some constant C''. With $\widetilde{M} \leq M$ this yields

$$\sum_{i \in \widetilde{I}_{f,\widetilde{M}}} |\widetilde{c}_i - \hat{c}_i|^2 \leq C'' M^{-2\gamma}.$$

Combining (1.20), (1.22), and (1.23), we obtain

$$\left\| f - \sum_{i \in \widetilde{I}_{f,\widetilde{M}}} \hat{c}_i \widetilde{\varphi}_i \right\|_{L^2(\Omega)} \leq C' M^{-\gamma},$$

for some constant C'. As the length of the bitstring used in this construction is proportional to $M \log(M)$, the claim (1.21) is established.

Now, we note that the antecedent of Implication (I) holds for all $\gamma < \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$. Assume next, towards a contradiction, that the antecedent holds for a $\gamma > \gamma^*(\mathcal{C})$. This would imply that for any $\gamma' < \gamma$,

$$\inf_{(E,D)\in\mathfrak{E}^L\times\mathfrak{D}^L} \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \in \mathcal{O}\big(L^{-\gamma'}\big),\ L\to\infty. \tag{1.24}$$

In particular, (1.24) would hold for some $\gamma' > \gamma^*(\mathcal{C})$ which, owing to (1.13) stands in contradiction to the definition of $\gamma^*(\mathcal{C})$. This completes the proof. □

Space		$\mathcal{C}$	Optimal dictionary	$\gamma^*(\mathcal{C})$	
L^2-Sobolev	$W_2^m([0,1])$	$\mathcal{U}(W_2^m([0,1]))$	Fourier/Wavelet basis	m	(Donoho et al., 1998, Sec. 14.2)
Hölder	$C^\alpha([0,1])$	$\mathcal{U}(C^\alpha([0,1]))$	Wavelet basis	α	(Donoho et al., 1998, Sec. 14.2)
Bump Algebra	$B^1_{1,1}([0,1])$	$\mathcal{U}(B^1_{1,1}([0,1]))$	Wavelet basis	1	(Donoho et al., 1998, Sec. 14.2)
Bounded Variation	$BV([0,1])$	$\mathcal{U}(BV([0,1]))$	Haar basis	1	(Donoho et al., 1998, Sec. 14.2)
L^p-Sobolev[4]	$W_p^m(\Omega)$	$\mathcal{U}(W_p^m(\Omega))$	Wavelet frame	$\frac{m}{d}$	(Grohs et al., 2020, Thm. 1.3)
Besov[5]	$B^m_{p,q}(\Omega)$	$\mathcal{U}(B^m_{p,q}(\Omega))$	Wavelet frame	$\frac{m}{d}$	(Grohs et al., 2020, Thm. 1.3)
Modulation[6]	$M^s_{p,p}(\mathbb{R}^d)$	$\mathcal{U}(M^s_{p,p}(\mathbb{R}^d))$	Wilson basis	$(\frac{1}{p}-\frac{1}{2}+\frac{2s}{d})^{-1}$	(Hinrichs et al., 2008, Thm. 4.4)
Cartoon functions[7]		$\mathcal{E}^\beta([-\frac{1}{2},\frac{1}{2}]^d)$	α-Curvelet frame[8]	$\frac{\beta(d-1)}{2}$	(Petersen and Voigtlaender, 2018)

Table 1: Optimal exponents and corresponding optimal dictionaries. $\mathcal{U}(X) = \{f \in X : \|f\|_X \leq 1\}$ denotes the unit ball in the space X and $\Omega \subseteq \mathbb{R}^d$ is a Lipschitz domain. Recall that compactness of these unit balls is w.r.t. L^2-norm.

The optimal exponent $\gamma^*(\mathcal{C})$ is known for various function classes such as unit balls in Besov spaces $B^m_{p,q}(\mathbb{R}^d)$ with $p, q \in (0,\infty]$ and $m > d(1/p - 1/2)_+$, where $\gamma^*(\mathcal{C}) = m/d$ (see (Grohs et al., 2020)), and unit balls in (polynomially) weighted modulation spaces $M^s_{p,p}(\mathbb{R}^d)$ with $p \in (1,2)$ and $s \in \mathbb{R}_+$, where $\gamma^*(\mathcal{C}) = (\frac{1}{p} - \frac{1}{2} + \frac{2s}{d})^{-1}$ (see (Hinrichs et al., 2008)). A further example is the set of β-cartoon-like functions, which are β-smooth on some bounded d-dimensional domain with sufficiently smooth boundary and zero otherwise. Here, we have $\gamma^*(\mathcal{C}) = \beta(d-1)/2$ (see (Donoho, 2001; Grohs et al., 2016b; Petersen and Voigtlaender, 2018)). These examples along with additional ones

are summarized in Table 1. For an extensive summary of metric entropy results and techniques for their derivation, we also refer to (Kolmogorov and Tikhomirov, 1959).

We conclude this section with general remarks on certain formal aspects of the Kolmogorov-Donoho rate-distortion framework. First, we note that for the set $\mathcal{C} \subseteq L^2(\Omega)$ to have a well-defined optimal exponent it must be relatively compact[9]. This follows from the fact that the set over which the minimum in the definition (1.12) of $L(\varepsilon, \mathcal{C})$ is taken must be nonempty for every $\varepsilon \in (0, \infty)$. To see this, note that every length-$L(\varepsilon, \mathcal{C})$ encoder-decoder pair induces an ε-covering of $\mathcal{C}$ with at most $2^{L(\varepsilon,\mathcal{C})}$ balls (and ball centers $\{D(E(f))\}_{f\in\mathcal{C}}$). It hence follows that $\mathcal{C}$ must be totally bounded and thus relatively compact as a consequence of $L^2(\Omega)$ being a complete metric space (Munkres, 2000, Thm. 45.1).

As shown in the proof of Theorem 3, effective best M-term approximations construct encoder-decoder pairs and thereby induce ε-coverings. By the arguments just made, this implies that also $\gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$ is well-defined only for compact function classes $\mathcal{C}$.

A consequence of the compactness requirement on $\mathcal{C}$ is that the spaces in Table 1 either consist of functions on bounded domains or, in the case of modulation spaces, are equipped with a weighted norm. In order to provide intuition on why this must be so, let us consider a function space $(X, \|\cdot\|_X)$ with $X \subseteq L^2(\mathbb{R}^d)$ and $\|\cdot\|_X$ translation invariant. Take $\varepsilon > 0$ and $f \in X$ with $\|f\|_X = 1$ and choose $C > 0$ such that $\|f\|_{L^2([-C,C]^d)} > \frac{4}{5}\|f\|_{L^2(\mathbb{R}^d)}$. Now, consider the family of translates of f given by $f_i(x) := f(x - 2Ci)$, $i \in \mathbb{Z}^d$, and note that $\|f_i\|_X = 1$ for all $i \in \mathbb{Z}^d$ by translation invariance of $\|\cdot\|_X$. Furthermore, we have

$$\begin{aligned}\|f_i\|_{L^2([-C,C]^d)} &= \left(\|f_i\|^2_{L^2(\mathbb{R}^d)} - \|f_i\|^2_{L^2(\mathbb{R}^d\setminus[-C,C]^d)}\right)^{\frac{1}{2}} \\ &\leq \left(\|f\|^2_{L^2(\mathbb{R}^d)} - \|f\|^2_{L^2([-C,C]^d)}\right)^{\frac{1}{2}} < \tfrac{3}{5}\|f\|_{L^2(\mathbb{R}^d)}\end{aligned}$$

[9]For the sake of simplicity, we assume, however, compactness throughout even though relative compactness (i.e. having a compact closure) would be sufficient.

for all $i \in \mathbb{Z}^d \backslash \{0\}$ by construction. This, in turn, implies

$$\begin{aligned} \|f_i - f_j\|_{L^2(\mathbb{R}^d)} &= \|f_{i-j} - f\|_{L^2(\mathbb{R}^d)} \\ &\geq \|f_{i-j} - f\|_{L^2([-C,C]^d)} \\ &> \tfrac{1}{5}\|f\|_{L^2(\mathbb{R}^d)} \end{aligned} \tag{1.25}$$

for all $i, j \in \mathbb{Z}^d$, with $i \neq j$, by the reverse triangle inequality. As such no ε-ball (w.r.t. $L^2(\mathbb{R}^d)$-norm) with $\varepsilon \leq \frac{1}{10}\|f\|_{L^2(\mathbb{R}^d)}$ can contain more than one of the infinitely many $(f_i)_{i\in\mathbb{Z}^d}$ which are, however, all contained in the unit ball $\mathcal{U}(X)$ of the space $(X, \|\cdot\|_X)$. This implies that $\mathcal{U}(X)$ cannot be totally bounded and thereby not relatively compact (w.r.t. $L^2(\mathbb{R}^d)$-norm). Somewhat nonchalantly speaking, for spaces equipped with translation-invariant norms this issue can be avoided by considering functions that live on a bounded domain, which ensures that (1.25) pertains only to a finite number of translates. Alternatively, for spaces of functions living on unbounded domains once can consider weighted norms that are not translation invariant. Here, the weighting effectively constrains the functions to a bounded domain.

The less restrictive concept of best M-term approximation rate $\gamma^*(\mathcal{C}, \mathcal{D})$ (see Definition 5) is, in apparent contrast, often studied for noncompact function classes $\mathcal{C}$.

In (Donoho et al., 1998, Sec. 15.2) a condition for $\gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$ and $\gamma^*(\mathcal{C}, \mathcal{D})$ to coincide is presented. Specifically, this condition, referred to as tail compactness, is expressed as follows. Let $\mathcal{C} \subseteq L^2(\Omega)$ be bounded and let $\mathcal{D} = \{\varphi_i\}_{i\in\mathbb{N}}$ be an ordered orthonormal basis for $\mathcal{C}$. We say that tail compactness holds if there exist $C, \beta > 0$ such that for all $N \in \mathbb{N}$,

$$\sup_{f\in\mathcal{C}} \left\| f - \sum_{i=1}^{N} \langle f, \varphi_i\rangle \varphi_i \right\|_{L^2(\Omega)} \leq CN^{-\beta}. \tag{1.26}$$

In order to see that (1.26) implies $\gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C}, \mathcal{D})$, we consider, for fixed $f \in \mathcal{C}$, the (unconstrained) best M-term approximation $f_M = \sum_{i\in I}\langle f, \varphi_i\rangle\varphi_i$ with $I \subseteq \mathbb{N}$, $|I| = M$. We now modify this

M-term approximation by letting $\alpha := \lceil \gamma^*(\mathcal{C},\mathcal{D})/\beta \rceil \in \mathbb{N}$ and removing, in the expansion $f_M = \sum_{i\in I}\langle f,\varphi_i\rangle\varphi_i$, all terms corresponding to indices that are larger than M^α. Recalling that in Definition 6 the same polynomial π bounds the search depth and the size of the coefficients, it follows that the modified approximation we just constructed obeys a polynomial depth search constraint with constraining polynomial $\pi_\alpha(x) = x^\alpha + S$, where $S := \sup_{f\in\mathcal{C}} \|f\|_{L^2(\Omega)}$. Here, owing to orthonormality of $\mathcal{D}$, S accounts for the size of the expansion coefficients $\langle f,\varphi_i\rangle$. In order to complete the argument, we need to show that the additional approximation error incurred by removing terms in $f_M = \sum_{i\in I}\langle f,\varphi_i\rangle\varphi_i$ is in $\mathcal{O}(M^{-\gamma^*(\mathcal{C},\mathcal{D})})$, i.e., it is of the same order as the error corresponding to the original (unconstrained) best M-term approximation. Due to orthonormality of $\mathcal{D}$ this additional error is given by the norm of $\sum_{i\in I, i>\pi_\alpha(M)}\langle f,\varphi_i\rangle\varphi_i$ and can, by virtue of (1.26), be bounded as

$$\begin{aligned}\left\|\sum_{i\in I, i>\pi_\alpha(M)}\langle f,\varphi_i\rangle\varphi_i\right\|_{L^2(\Omega)} &\leq \left\|\sum_{i=\pi_\alpha(M)+1}^{\infty}\langle f,\varphi_i\rangle\varphi_i\right\|_{L^2(\Omega)}\\ &= \left\|f-\sum_{i=1}^{\pi_\alpha(M)}\langle f,\varphi_i\rangle\varphi_i\right\|_{L^2(\Omega)}\\ &\leq C(\pi_\alpha(M))^{-\beta} \in \mathcal{O}(M^{-\gamma^*(\mathcal{C},\mathcal{D})}),\end{aligned}$$

which establishes the claim. We have hence shown that under tail compactness of arbitrary rate $\beta > 0$, $\gamma^*(\mathcal{C},\mathcal{D}) = \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$, and hence there is no cost incurred by imposing a polynomial depth search constraint combined with a polynomial bound on the size of the expansion coefficients. We hasten to add that the assumptions stated at the beginning of this paragraph together with what was just established imply that $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$ is, indeed, well-defined. For the more general case of $\mathcal{D}$ a frame, we refer to (Grohs, 2015, Sec. 5.4.3) for analogous arguments. Finally, we remark that the tail compactness inequality (1.26) can be interpreted as quantifying the rate of linear approximation for $\mathcal{C}$

in $\mathcal{D}$. Two examples of pairs $(\mathcal{C}, \mathcal{D})$ satisfying tail compactness, namely Besov spaces with wavelet bases and modulation spaces with Wilson bases, are provided in Appendices B and C, respectively.

As already mentioned, a larger optimal exponent $\gamma^*(\mathcal{C})$ leads to faster error decay (specifically according to $L^{-\gamma^*(\mathcal{C})}$) and hence corresponds to a function class of smaller complexity. As such, techniques for deriving lower bounds on the optimal exponent are often based on variations of the approach employed in the proof of Theorem 3, namely on the explicit construction of encoder-decoder pairs (in the case of the proof of Theorem 3 by encoding the dictionary elements participating in the M-term approximation). A powerful method for deriving upper bounds on the optimal exponent is the hypercube embedding approach proposed by Donoho in (Donoho, 2001); the basic idea here is to show that the function class $\mathcal{C}$ under consideration contains a sufficiently complex embedded set of orthogonal hypercubes and to then find the exponent corresponding to this set. An interesting alternative technique for deriving optimal exponents was proposed in the context of modulation spaces in (Hinrichs et al., 2008). The essence of this approach is to exploit the isomorphism between weighted modulation spaces and weighted mixed-norm sequence spaces (Gröchenig, 2013) and to then utilize results about entropy numbers of operators between sequence spaces.

1.6. APPROXIMATION WITH DEEP NEURAL NETWORKS

Inspired by the theory of best M-term approximation with dictionaries, we now develop the new concept of best M-weight approximation through neural networks. At the heart of this theory lies the interpretation of the network weights as the counterpart of the coefficients c_i in best M-term approximation. In other words, parsimony in terms of the number of participating elements in a dictionary is replaced by parsimony in terms of network connectivity. Our development will

parallel that for best M-term approximation in the previous section.

Before proceeding to the specifics, we would like to issue a general remark. While the neural network approximation results in Section 1.3 were formulated in terms of L^∞-norm, we shall be concerned with L^2-norm approximation here, on the one hand paralleling the use of L^2-norm in the context of best M-term approximation, and on the other hand allowing for the approximation of discontinuous functions by ReLU neural networks, which, owing to the continuity of the ReLU nonlinearity, necessarily realize continuous functions.

We start by introducing the concept of best M-weight approximation rate.

Definition 8. *Given $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and a function class $\mathcal{C} \subseteq L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,*

$$\Gamma_M^{\mathcal{N}}(f) := \inf_{\substack{\Phi \in \mathcal{N}_{d,1} \\ \mathcal{M}(\Phi) \leq M}} \|f - \Phi\|_{L^2(\Omega)}. \tag{1.27}$$

We call $\Gamma_M^{\mathcal{N}}(f)$ the best M-weight approximation error of f. *The supremal $\gamma > 0$ such that*

$$\sup_{f \in \mathcal{C}} \Gamma_M^{\mathcal{N}}(f) \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

will be denoted by $\gamma_{\mathcal{N}}^*(\mathcal{C})$. We say that the best M-weight approximation rate of $\mathcal{C}$ by neural networks *is $\gamma_{\mathcal{N}}^*(\mathcal{C})$.*

We emphasize that the infimum in (1.27) is taken over all networks with fixed input dimension d, no more than M nonzero (edge and node) weights, and arbitrary depth L. In particular, this means that the infimum is with respect to all possible network topologies and weight choices. The best M-weight approximation rate is fundamental as it benchmarks all algorithms that map a function f and an $\varepsilon > 0$ to a neural network approximating f with error no more than ε.

The two restrictions underlying the concept of effective best M-term approximation through dictionaries, namely polynomial depth

search and polynomially bounded coefficients, are next addressed in the context of approximation through deep neural networks. We start by noting that the need for the former is obviated by the tree-like-structure of neural networks. To see this, first note that $\mathcal{W}(\Phi) \leq \mathcal{M}(\Phi)$ and $\mathcal{L}(\Phi) \leq \mathcal{M}(\Phi)$. As the total number of nonzero weights in the network can not exceed $\mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1)$, this yields at most $\mathcal{O}(\mathcal{M}(\Phi)^3)$ possibilities for the "locations" (in terms of entries in the A_ℓ and the b_ℓ) of the $\mathcal{M}(\Phi)$ nonzero weights. Encoding the locations of the $\mathcal{M}(\Phi)$ nonzero weights hence requires $\log(\binom{C\mathcal{M}(\Phi)^3}{\mathcal{M}(\Phi)}) = \mathcal{O}(\mathcal{M}(\Phi)\log(\mathcal{M}(\Phi)))$ bits. This assumes, however, that the architecture of the network, i.e., the number of layers $\mathcal{L}(\Phi)$ and the N_k are known. Proposition 4 below shows that the architecture can, indeed, also be encoded with $\mathcal{O}(\mathcal{M}(\Phi)\log(\mathcal{M}(\Phi)))$ bits. In summary, we can therefore conclude that the tree-like-structure of neural networks automatically guarantees what we had to enforce through the polynomial depth search constraint in the case of best M-term approximation.

Inspection of the approximation results in Section 1.3 reveals that a sublinear growth restriction on $\mathcal{L}(\Phi)$ as a function of $\mathcal{M}(\Phi)$ is natural. Specifically, the approximation results in Section 1.3 all have $\mathcal{L}(\Phi)$ proportional to a polynomial in $\log(\varepsilon^{-1})$. As we are interested in approximation error decay according to $\mathcal{M}(\Phi)^{-\gamma}$, see Definition 8, this suggests to restrict $\mathcal{L}(\Phi)$ to growth that is polynomial in $\log(\mathcal{M}(\Phi))$.

The second restriction imposed in the definition of effective best M-term approximation, namely polynomially bounded coefficients, will be imposed in monomorphic manner on the magnitude of the weights. This growth condition will turn out natural in the context of the approximation results we are interested in and will, together with polylogarithmic depth growth, be seen below to allow rate-distortion-optimal quantization of the network weights. We remark, however, that networks with weights growing polynomially in $\mathcal{M}(\Phi)$ can be converted into networks with uniformly bounded weights at the expense of increased—albeit still of polylogarithmic scaling in $\mathcal{M}(\Phi)$—depth (see Proposition 9). In summary, we will develop the concept of "best M-

weight approximation subject to polylogarithmic depth and polynomial weight growth".

We start by introducing the following notation for neural networks with depth and weight magnitude bounded polylogarithmically respectively polynomially w.r.t. their connectivity.

Definition 9. *For $M, d, d' \in \mathbb{N}$, and π a polynomial, we define*

$$\mathcal{N}_{M,d,d'}^{\pi} := \Big\{\Phi \in \mathcal{N}_{d,d'} \colon \mathcal{M}(\Phi) \leq M, \mathcal{L}(\Phi) \leq \pi(\log(M)), \\ \mathcal{B}(\Phi) \leq \pi(M)\Big\}.$$

Next, we formalize the notion of effective best M-weight approximation rate subject to polylogarithmic depth and polynomial weight growth.

Definition 10. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. We define for $M \in \mathbb{N}$ and π a polynomial*

$$\varepsilon_{\mathcal{N}}^{\pi}(M) := \sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{N}_{M,d,1}^{\pi}} \|f - \Phi\|_{L^2(\Omega)}$$

and

$$\gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C}) := \sup\{\gamma \geq 0 \colon \exists \text{ polynomial } \pi \text{ s.t. } \varepsilon_{\mathcal{N}}^{\pi}(M) \in \mathcal{O}(M^{-\gamma}), \\ M \to \infty\}.$$

We refer to $\gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C})$ as the effective best M-weight approximation rate of $\mathcal{C}$.

We now state the equivalent of Theorem 3 for approximation by deep neural networks. Specifically, we establish that the optimal exponent $\gamma^*(\mathcal{C})$ constitutes a fundamental bound on the effective best M-weight approximation rate of $\mathcal{C}$ as well.

Theorem 4. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. Then, we have*

$$\gamma_{\mathcal{N}}^{*,\mathit{eff}}(\mathcal{C}) \leq \gamma^*(\mathcal{C}).$$

The key ingredients of the proof of Theorem 4 are developed throughout this section and the formal proof appears at the end of the section. Before getting started, we note that, in analogy to Definition 7, what we just found suggests the following.

Definition 11. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and let $\mathcal{C} \subseteq L^2(\Omega)$ be compact. We say that the function class $\mathcal{C} \subseteq L^2(\Omega)$ is* optimally representable by neural networks *if*

$$\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) = \gamma^*(\mathcal{C}).$$

It is interesting to observe that the fundamental limits of effective best M-term approximation (through dictionaries) and effective best M-weight approximation in neural networks are determined by the same quantity, although the approximants in the two cases are vastly different. We have linear combinations of elements of a dictionary under polynomial weight growth of the coefficients and with the participating functions identified subject to a polynomial-depth search constraint in the former, and concatenations of affine functions followed by nonlinearities under polynomial growth constraints on the coefficients of the affine functions and with a polylogarithmic growth constraint on the number of concatenations in the latter case.

We now commence the program developing the proof of Theorem 4. As in the arguments in the proof sketch of Theorem 3, the main idea is to compare the length of the bitstring needed to encode the approximating network to the minimax code length of the function class $\mathcal{C}$ to be approximated. To this end, we will need to represent the approximating network's nonzero weights, its architecture, i.e., L and the N_k, and the nonzero weights' locations as a bitstring. As the weights are real numbers and hence require, in principle, an infinite number of bits for their binary representations, we will have to suitably quantize them. In particular, the resolution of the corresponding quantizer will have to increase appropriately with decreasing ε. To formalize this idea, we start by defining the quantization employed.

Definition 12. *Let $m \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$. The network Φ is said*

to have (m, ε)*-quantized weights if all its weights are elements of* $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$.

A key ingredient of the proof of Theorem 4 is the following result, which establishes a fundamental lower bound on the connectivity of networks with quantized weights achieving uniform error ε over a given function class $\mathcal{C}$.

Proposition 4. *Let* $d, d' \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\mathcal{C} \subseteq L^2(\Omega)$, *and let* π *be a polynomial. Further, let*

$$\Psi : \left(0, \tfrac{1}{2}\right) \times \mathcal{C} \to \mathcal{N}_{d,d'}$$

be a map such that for every $\varepsilon \in (0, 1/2)$, $f \in \mathcal{C}$, *the network* $\Psi(\varepsilon, f)$ *has* $(\lceil \pi(\log(\varepsilon^{-1}))\rceil, \varepsilon)$*-quantized weights and satisfies*

$$\sup_{f\in\mathcal{C}} \|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon.$$

Then,

$$\sup_{f\in\mathcal{C}} \mathcal{M}(\Psi(\varepsilon, f)) \notin \mathcal{O}\left(\varepsilon^{-1/\gamma}\right),\ \varepsilon \to 0, \quad \textit{for all } \gamma > \gamma^*(\mathcal{C}).$$

Proof. The proof is by contradiction. Let $\gamma > \gamma^*(\mathcal{C})$ and assume that $\sup_{f\in\mathcal{C}} \mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}), \varepsilon \to 0$. The contradiction will be effected by constructing encoder-decoder pairs $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$ achieving uniform error ε over $\mathcal{C}$ with

$$\begin{aligned}
&\ell(\varepsilon)\\
&\leq C_0 \cdot \sup_{f\in\mathcal{C}} \left(\mathcal{M}(\Psi(\varepsilon, f)) \log(\mathcal{M}(\Psi(\varepsilon, f))) + 1\right) (\log(\varepsilon^{-1}))^q\\
&\leq C_0 \left(\varepsilon^{-1/\gamma} \log(\varepsilon^{-1/\gamma}) + 1\right) (\log(\varepsilon^{-1}))^q\\
&\leq C_1 \left(\varepsilon^{-1/\gamma} (\log(\varepsilon^{-1}))^{q+1} + (\log(\varepsilon^{-1}))^q\right)\\
&\in \mathcal{O}\left(\varepsilon^{-1/\nu}\right), \text{ for } \varepsilon \to 0,
\end{aligned}$$

where $C_0, C_1, q > 0$ are constants not depending on f, ε and $\gamma > \nu > \gamma^*(\mathcal{C})$. The specific form of the upper bound (1.28) will become

apparent in the construction of the bitstring representing Ψ detailed below.

We proceed to the construction of the encoder-decoder pairs $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$, which will be accomplished by encoding the network architecture, its topology, and the quantized weights in bitstrings of length $\ell(\varepsilon)$ satisfying (1.28) while guaranteeing unique reconstruction (of the network). For the sake of notational simplicity, we fix $\varepsilon \in (0, 1/2)$ and $f \in \mathcal{C}$ and set $\Psi := \Psi(\varepsilon, f)$, $M := \mathcal{M}(\Psi)$, and $L := \mathcal{L}(\Psi)$. Recall that the number of nodes in layers $0, \dots, L$ is denoted by $N_0, \dots, N_L$ and that $N_0 = d, N_L = d'$ (see Definition 1). Moreover, note that due to our nondegeneracy assumption (see Remark 1) we have $\sum_{\ell=0}^{L} N_\ell \leq 2M$ and $L \leq M$. The bitstring representing Ψ is constructed according to the following steps.

Step 1: If $M = 0$, we encode the network by a single 0. Using the convention $0 \log(0) = 0$, we then note that (1.28) holds trivially and we terminate the encoding procedure. Else, we encode the network connectivity, M, by starting the overall bitstring with M 1's followed by a single 0. The length of this bitstring is therefore given by $M + 1$.

Step 2: We continue by encoding the number of layers which, due to $L \leq M$, requires no more than $\lceil \log(M) \rceil$ bits. We thus reserve the next $\lceil \log(M) \rceil$ bits for the binary representation of L.

Step 3: Next, we store the layer dimensions $N_0, \dots, N_L$. As $L \leq M$ and $N_\ell \leq M$, for all $\ell \in \{0, \dots, L\}$, owing to nondegeneracy, we can encode the layer dimensions using $(M + 1)\lceil \log(M) \rceil$ bits. In combination with Steps 1 and 2 this yields an overall bitstring of length at most

$$M\lceil \log(M) \rceil + M + 2\lceil \log(M) \rceil + 1. \tag{1.28}$$

Step 4: We encode the topology of the graph associated with the network Ψ. To this end, we enumerate all nodes by assigning a unique index i to each one of them, starting from the 0-th layer and increasing from left to right within a given layer. The indices range from 1 to $N := \sum_{\ell=0}^{L} N_\ell \leq 2M$. Each of these indices can be encoded by a bitstring of length $\lceil \log(N) \rceil$. We denote the bitstring corresponding

to index i by $b(i) \in \{0,1\}^{\lceil \log(N) \rceil}$ and let for all nodes, except for those in the last layer, $n(i)$ be the number of children of the node with index i, i.e., the number of nodes in the next layer connected to the node with index i via an edge. For each of these nodes i, we form a bitstring of length $n(i)\lceil \log(N) \rceil$ by concatenating the bitstrings indexing its children. We follow this string with an all-zeros bitstring of length $\lceil \log(N) \rceil$ to signal that all children of the current node have been encoded. Overall, this yields a bitstring of length

$$\sum_{i=1}^{N-d'} (n(i)+1)\lceil \log(N) \rceil \leq 3M \lceil \log(2M) \rceil, \tag{1.29}$$

where we used $\sum_{i=1}^{N-d'} n(i) \leq M$.

Step 5: We encode the weights of Ψ. By assumption, Ψ has $(\lceil \pi(\log(\varepsilon^{-1})) \rceil, \varepsilon)$-quantized weights, which means that each weight of Ψ can be represented by no more than

$$B_\varepsilon := 2(\lceil \pi(\log(\varepsilon^{-1})) \rceil \lceil \log(\varepsilon^{-1}) \rceil + 1)$$

bits. For each node $i = 1, \ldots, N$, we reserve the first B_ε bits to encode its associated node weight and, for each of its children a bitstring of length B_ε to encode the weight corresponding to the edge between the current node and that child. Concatenating the results in ascending order of child node indices, we get a bitstring of length $(n(i)+1)B_\varepsilon$ for node i, and an overall bitstring of length

$$\sum_{i=1}^{N-d'} (n(i)+1)B_\varepsilon + d'B_\varepsilon \leq 3MB_\varepsilon$$

representing the weights. Combining this with (1.28) and (1.29), we find that the overall number of bits needed to encode the network architecture, topology, and weights is no more than

$$3MB_\varepsilon + 3M\lceil \log(2M) \rceil + (M+2)\lceil \log(M) \rceil + M + 1. \tag{1.30}$$

The network can be recovered by sequentially reading out M, L, the N_ℓ, the topology, and the quantized weights from the overall bitstring. It is not difficult to verify that the individual steps in the encoding procedure were crafted such that this yields unique recovery. As (1.30) can be upper-bounded by

$$C_0(M\log(M)+1)(\log\big(\varepsilon^{-1}\big))^q$$

for constants $C_0, q > 0$ depending on π only, we have constructed an encoder-decoder pair $(E_\varepsilon, D_\varepsilon) \in \mathfrak{E}^{\ell(\varepsilon)} \times \mathfrak{D}^{\ell(\varepsilon)}$ with $\ell(\varepsilon)$ satisfying (1.28). This concludes the proof. □

Proposition 4 states that the connectivity growth rate of networks with quantized weights achieving uniform approximation error ε over a function class $\mathcal{C}$ must exceed $\mathcal{O}\big(\varepsilon^{-1/\gamma^*(\mathcal{C})}\big)$, $\varepsilon \to 0$. As Proposition 4 applies to networks that have each weight represented by a finite number of bits scaling polynomially in $\log(\varepsilon^{-1})$, while guaranteeing that the underlying encoder-decoder pair achieves uniform error ε over $\mathcal{C}$, it remains to establish that such a compatibility is, indeed, possible. Specifically, this requires a careful interplay between the network's depth and connectivity scaling, and its weight growth, all as a function of ε. Establishing that this delicate balancing is implied by our technical assumptions is the subject of the remainder of this section. We start with a perturbation result quantifying how the error induced by weight quantization in the network translates to the output function realized by the network.

Lemma 9. *Let $d, d', k \in \mathbb{N}$, $D \in \mathbb{R}_+$, $\Omega \subseteq [-D, D]^d$, $\varepsilon \in (0, 1/2)$, let $\Phi \in \mathcal{N}_{d,d'}$ with $\mathcal{M}(\Phi) \leq \varepsilon^{-k}$, $\mathcal{B}(\Phi) \leq \varepsilon^{-k}$, and let $m \in \mathbb{N}$ satisfy*

$$m \geq 3k\mathcal{L}(\Phi) + \log(\lceil D \rceil). \tag{1.31}$$

Then, there exists a network $\widetilde{\Phi} \in \mathcal{N}_{d,d'}$ with (m, ε)-quantized weights satisfying

$$\sup_{x\in\Omega} \|\Phi(x) - \widetilde{\Phi}(x)\|_\infty \leq \varepsilon.$$

More specifically, the network $\widetilde{\Phi}$ can be obtained simply by replacing every weight in Φ by a closest element in $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$.

Proof of Theorem 9. We first consider the case $\mathcal{L}(\Phi) = 1$. Here, it follows from Definition 1 that the network simply realizes an affine transformation and hence

$$\sup_{x\in\Omega} \|\Phi(x) - \widetilde{\Phi}(x)\|_\infty \leq \mathcal{M}(\Phi)\lceil D\rceil 2^{-m\lceil \log(\varepsilon^{-1})\rceil - 1} \leq \varepsilon.$$

In the remainder of the proof, we can therefore assume that $\mathcal{L}(\Phi) \geq 2$. For simplicity of notation, we set $L := \mathcal{L}(\Phi)$, $M := \mathcal{M}(\Phi)$, and, as usual, write

$$\Phi = W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

with $W_\ell(x) = A_\ell x + b_\ell$, $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{N_\ell}$. We now consider the partial networks $\Phi^\ell \colon \Omega \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L-1\}$, given by

$$\Phi^\ell := \begin{cases} \rho \circ W_1, & \ell = 1 \\ \rho \circ W_2 \circ \rho \circ W_1, & \ell = 2 \\ \rho \circ W_\ell \circ \rho \circ W_{\ell-1} \circ \cdots \circ \rho \circ W_1, & \ell = 3, \ldots, L-1, \end{cases}$$

and set $\Phi^L := \Phi$. We hasten to add that we decided—for ease of exposition—to deviate from the convention used in Definition 1 and to have the partial networks include the application of ρ at the end. Now, for $\ell \in \{1, 2, \ldots, L\}$, let $\widetilde{\Phi}^\ell$ be the (partial) network obtained by replacing all the entries of the A_ℓ and b_ℓ by a closest element in $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$. We denote these replacements by $\widetilde{A}_\ell$ and $\widetilde{b}_\ell$, respectively, and note that

$$\begin{aligned} \max_{i,j} |A_{\ell,i,j} - \widetilde{A}_{\ell,i,j}| &\leq \tfrac{1}{2}\, 2^{-m\lceil \log(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\,\varepsilon^m, \\ \max_{i,j} |b_{\ell,i,j} - \widetilde{b}_{\ell,i,j}| &\leq \tfrac{1}{2}\, 2^{-m\lceil \log(\varepsilon^{-1})\rceil} \leq \tfrac{1}{2}\,\varepsilon^m. \end{aligned} \tag{1.32}$$

The proof will be effected by upper-bounding the error building up across layers as a result of this quantization. To this end, we define, for $\ell \in \{1, 2, \ldots, L\}$, the error in the ℓ-th layer as

$$e_\ell := \sup_{x \in \Omega} \|\Phi^\ell(x) - \widetilde{\Phi}^\ell(x)\|_\infty.$$

We further set $C_0 := \lceil D \rceil$ and $C_\ell := \max\{1, \sup_{x\in\Omega} \|\Phi^\ell(x)\|_\infty\}$. As each entry of the vector $\Phi^\ell(x) \in \mathbb{R}^{N_\ell}$ is obtained by applying[10] the 1-Lipschitz function ρ to the sum of a weighted sum of at most $N_{\ell-1}$ components of the vector $\Phi^{\ell-1}(x) \in \mathbb{R}^{N_{\ell-1}}$ and a bias component $b_{\ell,i}$, and $\mathcal{B}(\Phi) \leq \varepsilon^{-k}$ by assumption, we have for all $\ell \in \{1, 2, \ldots, L\}$,

$$C_\ell \leq N_{\ell-1}\varepsilon^{-k} C_{\ell-1} + \varepsilon^{-k} \leq (N_{\ell-1} + 1)\, \varepsilon^{-k} C_{\ell-1},$$

which implies, for all $\ell \in \{1, 2, \ldots, L\}$, that

$$C_\ell \leq C_0\, \varepsilon^{-k\ell} \prod_{i=0}^{\ell-1} (N_i + 1). \tag{1.33}$$

Next, note that the components $(\widetilde{\Phi}^1(x))_i, i \in \{1, 2, \ldots, N_1\}$, of the vector $\widetilde{\Phi}^1(x) \in \mathbb{R}^{N_1}$ can be written as

$$(\widetilde{\Phi}^1(x))_i = \rho\left(\left(\sum_{j=1}^{N_0} \widetilde{A}_{1,i,j} x_j\right) + \widetilde{b}_{1,i}\right),$$

which, combined with (1.32) and the fact that ρ is 1-Lipschitz implies

$$e_1 \leq C_0 N_0 \tfrac{\varepsilon^m}{2} + \tfrac{\varepsilon^m}{2} \leq C_0 (N_0 + 1) \tfrac{\varepsilon^m}{2}. \tag{1.34}$$

Due to ρ and the identity mapping being 1-Lipschitz, we have, for

[10] Note that going from Φ_{L-1} to Φ_L the activation function is not applied anymore, which nevertheless leads to the same estimate as the identity mapping is 1-Lipschitz.

$\ell = 1, \ldots, L$,

$$
\begin{aligned}
e_\ell &= \sup_{x\in\Omega} \|\Phi^\ell(x) - \widetilde{\Phi}^\ell(x)\|_\infty \\
&= \sup_{x\in\Omega, i\in\{1,\ldots,N_\ell\}} |(\Phi^\ell(x))_i - (\widetilde{\Phi}^\ell(x))_i| \\
&\leq \sup_{x\in\Omega, i\in\{1,\ldots,N_\ell\}} \Bigg| \left[\left(\sum_{j=1}^{N_{\ell-1}} A_{\ell,i,j} (\Phi^{\ell-1}(x))_j \right) + b_{\ell,i} \right] \\
&\quad - \left[\left(\sum_{j=1}^{N_{\ell-1}} \widetilde{A}_{\ell,i,j} (\widetilde{\Phi}^{\ell-1}(x))_j \right) + \widetilde{b}_{\ell,i} \right] \Bigg| \\
&\leq \sup_{x\in\Omega, i\in\{1,\ldots,N_\ell\}} \Bigg[\Bigg(\sum_{j=1}^{N_{\ell-1}} \Big| A_{\ell,i,j} (\Phi^{\ell-1}(x))_j \\
&\quad - \widetilde{A}_{\ell,i,j} (\widetilde{\Phi}^{\ell-1}(x))_j \Big| \Bigg) + \Big| b_{\ell,i} - \widetilde{b}_{\ell,i} \Big| \Bigg].
\end{aligned}
\tag{1.35}
$$

As $|(\Phi^{\ell-1}(x))_j - (\widetilde{\Phi}^{\ell-1}(x))_j| \leq e_{\ell-1}$ and $|(\Phi^{\ell-1}(x))_j| \leq C_{\ell-1}$ for all $x \in \Omega$, $j \in \{1, \ldots, N_{\ell-1}\}$ by definition, and $|A_{\ell,i,j}| \leq \varepsilon^{-k}$ by assumption, upon invoking (1.32), we get

$$
\begin{aligned}
&|A_{\ell,i,j} (\Phi^{\ell-1}(x))_j - \widetilde{A}_{\ell,i,j} (\widetilde{\Phi}^{\ell-1}(x))_j| \\
&\leq e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2} + e_{\ell-1}\tfrac{\varepsilon^m}{2}.
\end{aligned}
$$

Since $\varepsilon \in (0, 1/2)$, it therefore follows from (1.35), that for all $\ell \in \{2, \ldots, L\}$,

$$
\begin{aligned}
e_\ell &\leq N_{\ell-1}(e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2} + e_{\ell-1}\tfrac{\varepsilon^m}{2}) + \tfrac{\varepsilon^m}{2} \\
&\leq (N_{\ell-1}+1)(2e_{\ell-1}\varepsilon^{-k} + C_{\ell-1}\tfrac{\varepsilon^m}{2}).
\end{aligned}
\tag{1.36}
$$

We now claim that, for all $\ell \in \{2, \ldots, L\}$,

$$
e_\ell \leq \tfrac{1}{2}(2^\ell - 1) C_0 \varepsilon^{m-(\ell-1)k} \prod_{i=0}^{\ell-1} (N_i + 1), \tag{1.37}
$$

which we prove by induction. The base case $\ell = 1$ was already established in (1.34). For the induction step we assume that (1.37) holds for a given ℓ which, in combination with (1.33) and (1.36), implies

$$\begin{aligned} e_{\ell+1} &\leq \big(N_\ell + 1\big)\big(2e_\ell\varepsilon^{-k} + C_\ell\tfrac{\varepsilon^m}{2}\big) \\ &\leq (N_\ell+1)\Bigg((2^\ell-1)C_0\varepsilon^{m-(\ell-1)k}\varepsilon^{-k}\prod_{i=0}^{\ell-1}(N_i+1) \\ &\quad + C_0\varepsilon^{-k\ell}\tfrac{\varepsilon^m}{2}\prod_{i=0}^{\ell-1}(N_i+1)\Bigg) \\ &= \frac{1}{2}(2^{\ell+1}-1)C_0\varepsilon^{m-\ell k}\prod_{i=0}^{\ell}(N_i+1). \end{aligned}$$

This completes the induction argument and establishes (1.37). Using $2^{L-1} \leq \varepsilon^{-(L-1)}$, $\prod_{i=0}^{L-1}(N_i+1) \leq M^L \leq \varepsilon^{-kL}$, and $m \geq 3kL + \log(\lceil D\rceil)$ by assumption, we get

$$\begin{aligned} \sup_{x\in\Omega}\|\Phi(x)-\widetilde{\Phi}(x)\|_\infty = e_L &\leq \tfrac{1}{2}(2^L-1)C_0\varepsilon^{m-(L-1)k}\prod_{i=0}^{L-1}(N_i+1) \\ &\leq \varepsilon^{m-(L-1+kL-k+\log(\lceil D\rceil)+kL)} \\ &\leq \varepsilon^{m-(3kL+\log(\lceil D\rceil)-1)} \leq \varepsilon. \end{aligned}$$

This completes the proof. □

We are now ready to finalize the proof of Theorem 4.

Proof of Theorem 4. Suppose towards a contradiction that $\gamma^{*,\mathrm{eff}}_{\mathcal{N}}(\mathcal{C}) > \gamma^*(\mathcal{C})$ and let $\gamma \in \big(\gamma^*(\mathcal{C}), \gamma^{*,\mathrm{eff}}_{\mathcal{N}}(\mathcal{C})\big)$. Then, by Definition 10, there exist a polynomial π and a constant $C > 0$ such that

$$\sup_{f\in\mathcal{C}}\inf_{\Phi\in\mathcal{N}^\pi_{M,d,1}}\|f-\Phi\|_{L^2(\Omega)} \leq CM^{-\gamma}, \text{ for all } M \in \mathbb{N}.$$

Setting $M_\varepsilon := \big\lceil(\varepsilon/(4C))^{-1/\gamma}\big\rceil$, it follows that, for every $f \in \mathcal{C}$ and every $\varepsilon \in (0, 1/2)$, there exists a neural network $\Phi_{\varepsilon,f} \in \mathcal{N}^\pi_{M_\varepsilon,d,1}$ such

that

$$\|f - \Phi_{\varepsilon,f}\|_{L^2(\Omega)} \leq 2 \sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{N}^{\pi}_{M_\varepsilon,d,1}} \|f - \Phi\|_{L^2(\Omega)} \tag{1.38}$$

$$\leq 2CM_\varepsilon^{-\gamma} \leq \frac{\varepsilon}{2}. \tag{1.39}$$

By Lemma 9 there exists a polynomial π^* such that for every $f \in \mathcal{C}, \varepsilon \in (0, 1/2)$, there is a network $\widetilde{\Phi}_{\varepsilon,f}$ with $(\lceil \pi^*(\log(\varepsilon^{-1}))\rceil, \varepsilon)$-quantized weights satisfying

$$\left\|\Phi_{\varepsilon,f} - \widetilde{\Phi}_{\varepsilon,f}\right\|_{L^2(\Omega)} \leq \frac{\varepsilon}{2}. \tag{1.40}$$

The conditions of Lemma 9 are satisfied as M_ε can be upper-bounded by ε^{-k} with a suitably chosen k, the weights in $\Phi_{\varepsilon,f}$ are polynomially bounded in M_ε, and (1.31) follows from the depth of networks in $\Phi \in \mathcal{N}^{\pi}_{M_\varepsilon,d,1}$ being polylogarithmically bounded in M_ε due to Definition 9. Now, defining

$$\Psi\colon \left(0, \tfrac{1}{2}\right) \times \mathcal{C} \to \mathcal{N}_{d,1}, \quad (\varepsilon, f) \mapsto \widetilde{\Phi}_{\varepsilon,f},$$

it follows from (1.38) and (1.40), by application of the triangle inequality, that

$$\sup_{f\in\mathcal{C}} \|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$$

with

$$\sup_{f\in\mathcal{C}} \mathcal{M}(\Psi(\varepsilon, f)) \leq M_\varepsilon \in \mathcal{O}\big(\varepsilon^{-1/\gamma}\big),\ \varepsilon \to 0.$$

The proof is concluded by noting that $\Psi(\varepsilon, f)$ violates Proposition 4. □

We conclude this section with a discussion of the conceptual implications of the results established above. Proposition 4 combined with Lemma 9 establishes that neural networks achieving uniform approximation error ε while having weights that are polynomially bounded in ε^{-1} and depth growing polylogarithmically in ε^{-1} cannot exhibit connectivity growth rate smaller than $\mathcal{O}(\varepsilon^{-1/\gamma^*(\mathcal{C})}), \varepsilon \to 0$; in other words, a decay of the uniform approximation error, as a function of M, faster than $\mathcal{O}(M^{-\gamma^*(\mathcal{C})}), M \to \infty$, is not possible.

1.7. THE TRANSFERENCE PRINCIPLE

We have seen that a wide array of function classes can be approximated in Kolmogorov-Donoho optimal fashion through dictionaries, provided that the dictionary $\mathcal{D}$ is chosen to consort with the function class $\mathcal{C}$ according to $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$. Examples of such pairs are unit balls in Besov spaces with wavelet bases and unit balls in weighted modulation spaces with Wilson bases. A more extensive list of optimal pairs is provided in Table 1. On the other hand, as shown in (Donoho, 1993), Fourier bases are strictly suboptimal—in terms of approximation rate—for balls $\mathcal{C}$ of finite radius in the spaces $BV(\mathbb{R})$ and $W_p^m(\mathbb{R})$.

In light of what was just said, it is hence natural to let neural networks play the role of the dictionary $\mathcal{D}$ and to ask which function classes $\mathcal{C}$ are approximated in Kolmogorov-Donoho-optimal fashion by neural networks. Towards answering this question, we next develop a general framework for transferring results on function approximation through dictionaries to results on approximation by neural networks. This will eventually lead us to a characterization of function classes $\mathcal{C}$ that are optimally representable by neural networks in the sense of Definition 11.

We start by introducing the notion of effective representability of dictionaries through neural networks.

Definition 13. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ be a dictionary. We call $\mathcal{D}$* effectively representable by neural networks*, if there exists a bivariate polynomial π such that for all $i \in \mathbb{N}$, $\varepsilon \in (0,1/2)$, there is a neural network $\Phi_{i,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying $\mathcal{M}(\Phi_{i,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}),\log(i))$, $\mathcal{B}(\Phi_{i,\varepsilon}) \leq \pi(\varepsilon^{-1},i)$, and*

$$\|\varphi_i - \Phi_{i,\varepsilon}\|_{L^2(\Omega)} \leq \varepsilon.$$

The next result will allow us to conclude that optimality—in the sense of Definition 7—of a dictionary $\mathcal{D}$ for a function class $\mathcal{C}$ combined with effective representability of $\mathcal{D}$ by neural networks implies optimal representability of $\mathcal{C}$ by neural networks. The proof is, in

essence, effected by noting that every element of the effectively representable $\mathcal{D}$ participating in a best M-term-rate achieving approximation f_M of $f \in \mathcal{C}$ can itself be approximated by neural networks well enough for an overall network to approximate f_M with connectivity $M\pi(\log(M))$. As this connectivity is only polylogarithmically larger than the number of terms M participating in the best M-term approximation f_M, we will be able to conclude that the optimal approximation rate, indeed, transfers from approximation in $\mathcal{D}$ to approximation in neural networks. The conditions on $\mathcal{M}(\Phi_{i,\varepsilon})$ and $\mathcal{B}(\Phi_{i,\varepsilon})$ in Definition 13 guarantee precisely that the connectivity increase is at most by a polylogarithmic factor. To see this, we first recall that effective best M-term approximation has a polynomial depth search constraint, which implies that the indices i under consideration are upper-bounded by a polynomial in M. In addition, the approximation error behavior we are interested in is $\varepsilon = M^{-\gamma}$. Combining these two insights, it follows that $\mathcal{M}(\Phi_{i,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$ implies polylogarithmic (in M) connectivity for each network $\Phi_{i,\varepsilon}$ and hence connectivity $M\pi(\log(M))$ for the overall network realizing f_M, as desired. By the same token, $\mathcal{B}(\Phi_{i,\varepsilon}) \leq \pi(\varepsilon^{-1}, i)$ guarantees that the weights of $\Phi_{i,\varepsilon}$ are polynomial in M.

There is another aspect to effective representability by neural networks that we would like to illustrate by way of example, namely that of ordering the dictionary elements. Specifically, we consider, for $d = 1$ and $\Omega = [-\pi, \pi)$, the class $\mathcal{C}$ of real-valued even functions in $\mathcal{C} = L^2(\Omega)$, and take the dictionary as $\mathcal{D} = \{\cos(ix), i \in \mathbb{N}_0\}$. As the index i enumerating the dictionary elements corresponds to frequencies, the basis functions in $\mathcal{D}$ are hence ordered according to increasing frequencies. Next, note that the parameter a in Theorem 2 corresponds to the frequency index i in our example. As the network $\Psi_{a,D,\varepsilon}$ in Theorem 2 is of finite width, it hence follows, upon replacing a in the expression for $\mathcal{L}(\Psi_{a,D,\varepsilon})$ by i, that $\mathcal{M}(\Psi_{i,D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$. The condition on the weights for effective representability is satisfied trivially, simply as $\mathcal{B}(\Psi_{i,D,\varepsilon}) \leq 1 \leq \pi(\varepsilon^{-1}, i)$.

We are now ready to state the rate optimality transfer result.

Theorem 5. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ be bounded, and consider the compact function class $\mathcal{C} \subseteq L^2(\Omega)$. Suppose that the dictionary $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ is effectively representable by neural networks. Then, for every $\gamma \in (0, \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}))$, there exist a polynomial π and a map*

$$\Psi : \left(0, \tfrac{1}{2}\right) \times \mathcal{C} \to \mathcal{N}_{d,1},$$

such that for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, the network $\Psi(\varepsilon, f)$ has $(\lceil \pi(\log(\varepsilon^{-1}))\rceil, \varepsilon)$-quantized weights while satisfying

$$\|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon,$$

$$\mathcal{L}(\Psi(\varepsilon, f)) \leq \pi(\log(\varepsilon^{-1})),$$

$$\mathcal{B}(\Psi(\varepsilon, f)) \leq \pi(\varepsilon^{-1}),$$

and we have

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0, \tag{1.41}$$

with the implicit constant in (1.41) being independent of f. In particular, it holds that

$$\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}).$$

Remark 5. *Theorem 5 allows us to draw the following conclusion. If $\mathcal{D}$ optimally represents the function class $\mathcal{C}$ in the sense of Definition 7, i.e., $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C})$, and if it is, in addition, effectively representable by neural networks in the sense of Definition 13, then, due to Theorem 4, which states that $\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) \leq \gamma^*(\mathcal{C})$, we have $\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) = \gamma^*(\mathcal{C})$ and hence $\mathcal{C}$ is optimally representable by neural networks in the sense of Definition 11.*

Proof of Theorem 5. Let $\gamma' \in (\gamma, \gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}))$. According to Definition 6, there exist a constant $C \geq 1$ and a polynomial π_1, such that for every $f \in \mathcal{C}$, $M \in \mathbb{N}$, there is an index set $I_{f,M} \subseteq \{1, \ldots, \pi_1(M)\}$ of cardinality M and coefficients $(c_i)_{i\in I_{f,M}}$ with $|c_i| \leq \pi_1(M)$, such that

$$\left\| f - \sum_{i\in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} \leq \frac{C M^{-\gamma'}}{2}. \tag{1.42}$$

Let $A := \max\{1, |\Omega|^{1/2}\}$. Effective representability of $\mathcal{D}$ according to Definition 13 ensures the existence of a bivariate polynomial π_2 such that for all $M \in \mathbb{N}$, $i \in I_{f,M}$, there is a neural network $\Phi_{i,M} \in \mathcal{N}_{d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,M}\|_{L^2(\Omega)} \leq \tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)} \tag{1.43}$$

with

$$\begin{aligned}
\mathcal{M}(\Phi_{i,M}) &\leq \pi_2\left(\log\left(\left(\tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)}\right)^{-1}\right), \log(i)\right) \\
&= \pi_2\left((\gamma'+1)\log(M) + \log\left(\tfrac{4A\pi_1(M)}{C}\right), \log(i)\right), \\
\mathcal{B}(\Phi_{i,M}) &\leq \pi_2\left(\left(\tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)}\right)^{-1}, i\right) \\
&= \pi_2\left(\tfrac{4A\pi_1(M)}{C} M^{\gamma'+1}, i\right).
\end{aligned} \tag{1.44}$$

Consider now for $f \in \mathcal{C}$, $M \in \mathbb{N}$ the networks given by

$$\Psi_{f,M}(x) := \sum_{i \in I_{f,M}} c_i \Phi_{i,M}(x).$$

Due to $\max(I_{f,M}) \leq \pi_1(M)$, (1.44) and Lemma 19 imply the existence of a polynomial π_3 such that $\mathcal{L}(\Psi_{f,M}) \leq \pi_3(\log(M))$, $\mathcal{M}(\Psi_{f,M}) \leq M\pi_3(\log(M))$, and $\mathcal{B}(\Psi_{f,M}) \leq \pi_3(M)$, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$, and, owing to (1.43), we get

$$\begin{aligned}
&\left\| \Psi_{f,M} - \sum_{i \in I_{f,M}} c_i \varphi_i \right\|_{L^2(\Omega)} \\
&\leq \sum_{i \in I_{f,M}} |c_i| \tfrac{C}{4A\pi_1(M)} M^{-(\gamma'+1)} \\
&\leq \tfrac{CM^{-\gamma'}}{4A} \sum_{i=1}^{|I_{f,M}|} \tfrac{\max_{i \in I_{f,M}} |c_i|}{M\pi_1(M)} \leq \tfrac{CM^{-\gamma'}}{4A}.
\end{aligned} \tag{1.45}$$

Lemma 9 therefore ensures the existence of a polynomial π_4 such that for all $f \in \mathcal{C}$, $M \in \mathbb{N}$, there is a network $\widetilde{\Psi}_{f,M} \in \mathcal{N}_{d,1}$ with $(\lceil \pi_4(\log(\frac{4A}{C}M^{\gamma'}))\rceil, \frac{CM^{-\gamma'}}{4A})$-quantized weights satisfying $\mathcal{L}(\widetilde{\Psi}_{f,M}) = \mathcal{L}(\Psi_{f,M})$, $\mathcal{M}(\widetilde{\Psi}_{f,M}) = \mathcal{M}(\Psi_{f,M})$, $\mathcal{B}(\widetilde{\Psi}_{f,M}) \leq \mathcal{B}(\Psi_{f,M}) + \frac{CM^{-\gamma'}}{4A}$, and

$$\left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4A}. \tag{1.46}$$

As Ω is bounded by assumption, we have

$$\left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} \left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^\infty(\Omega)} \leq \tfrac{CM^{-\gamma'}}{4}, \tag{1.47}$$

for all $f \in \mathcal{C}$, $M \in \mathbb{N}$. Combining (1.47) with (1.42) and (1.45), we get, for all $f \in \mathcal{C}$, $M \in \mathbb{N}$,

$$\begin{aligned}
&\left\|f - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} \\
&\leq \left\|f - \sum_{i \in I_{f,M}} c_i \varphi_i\right\|_{L^2(\Omega)} + \left\|\sum_{i \in I_{f,M}} c_i \varphi_i - \Psi_{f,M}\right\|_{L^2(\Omega)} \\
&+ \left\|\Psi_{f,M} - \widetilde{\Psi}_{f,M}\right\|_{L^2(\Omega)} \\
&\leq CM^{-\gamma'}.
\end{aligned} \tag{1.48}$$

For $\varepsilon \in (0, 1/2)$ and $f \in \mathcal{C}$, we now set $M_\varepsilon := \left\lceil (C/\varepsilon)^{1/\gamma'} \right\rceil$ and

$$\Psi(\varepsilon, f) := \widetilde{\Psi}_{f,M_\varepsilon}.$$

Thus, (1.48) yields

$$\|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq CM_\varepsilon^{-\gamma'} \leq \varepsilon.$$

Next, we note that, for all polynomials π and $0 \leq m < n$,

$$\mathcal{O}(\varepsilon^{-m}\pi(\log(\varepsilon^{-1}))) \subseteq \mathcal{O}(\varepsilon^{-n}),\ \varepsilon \to 0.$$

As $1/\gamma' < 1/\gamma$, this establishes

$$\mathcal{M}(\Psi(\varepsilon, f)) \in \mathcal{O}(M_\varepsilon \pi_3(\log(M_\varepsilon))) \subseteq \mathcal{O}(\varepsilon^{-1/\gamma}),\ \varepsilon \to 0. \quad (1.49)$$

Since M_ε and π_3 are independent of f, the implicit constant in (1.49) does not depend on f.

Next, note that, in general, an (n, η)-quantized network is also (m, δ)-quantized for $n \geq m$ and $\eta \leq \delta$, simply as

$$2^{-m\lceil \log(\delta^{-1})\rceil}\mathbb{Z} \cap [-\delta^{-m}, \delta^{-m}] \subseteq 2^{-n\lceil \log(\eta^{-1})\rceil}\mathbb{Z} \cap [-\eta^{-n}, \eta^{-n}].$$

Since $\frac{CM_\varepsilon^{-\gamma'}}{4A} \leq \varepsilon$ this ensures the existence of a polynomial π such that, for every $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, the network $\Psi(\varepsilon, f)$ is $(\lceil \pi(\log(\varepsilon^{-1}))\rceil, \varepsilon)$-quantized, $\mathcal{L}(\Psi(\varepsilon, f)) \leq \pi(\log(\varepsilon^{-1}))$, and $\mathcal{B}(\Psi(\varepsilon, f)) \leq \pi(\varepsilon^{-1})$. With (1.49) this establishes the first claim of the theorem. In order to verify the second claim, note that $\Psi(\varepsilon, f) \in \mathcal{N}^\pi_{\mathcal{M}(\Psi(\varepsilon,f)),d,1}$, for all $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, which implies

$$\sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{N}^\pi_{M,d,1}} \|f - \Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M \to \infty.$$

Therefore, owing to Definition 10, we get

$$\gamma^{*,\text{eff}}_{\mathcal{N}}(\mathcal{C}) \geq \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D}),$$

which concludes the proof. □

Remark 6. *We note that Theorem 5 continues to hold for $\Omega = \mathbb{R}^n$ if the elements of $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}}$ are compactly supported with the size of their support sets growing no more than polynomially in i. The technical elements required to show this can be found in the context of the approximation of Gabor dictionaries in the proof of Theorem 10, but are omitted here for ease of exposition.*

The last piece needed to complete our program is to establish that the conditions in Definition 13 guaranteeing effective representability in neural networks are, indeed, satisfied by a wide variety of dictionaries.

Inspecting Table 1, we can see that all example function classes provided therein are optimally represented either by affine dictionaries, i.e., wavelets, the Haar basis, and curvelets or Weyl-Heisenberg dictionaries, namely Fourier bases and Wilson bases. The next two sections will be devoted to proving effective representability of affine dictionaries and Weyl-Heisenberg dictionaries by neural networks, thus allowing us to draw the conclusion that neural networks are universally Kolmogorov-Donoho optimal approximators for all function classes listed in Table 1.

1.8. AFFINE DICTIONARIES ARE EFFECTIVELY REPRESENTABLE BY NEURAL NETWORKS

The purpose of this section is to establish that *affine dictionaries*, including wavelets (Daubechies, 1992), ridgelets (Candès, 1998), curvelets (Candès and Donoho, 2002), shearlets (Guo et al., 2006), α-shearlets and more generally α-molecules (Grohs et al., 2016a), which contain all aforementioned dictionaries as special cases, are effectively representable by neural networks. Due to Theorem 5 and Theorem 4, this will then allow us to conclude that any function class that is optimally representable—in the sense of Definition 7—by an affine dictionary with a suitable generator function is optimally representable by neural networks in the sense of Definition 11. By "suitable" we mean that the generator function can be approximated well by ReLU networks in a sense to be made precise below.

In order to elucidate the main ideas underlying the general definition of affine dictionaries that are effectively representable by neural networks, we start with a basic example, namely the Haar wavelet dictionary on the unit interval, i.e., the set of functions

$$\psi_{n,k}\colon [0,1] \mapsto \mathbb{R},\ x \mapsto 2^{\frac{n}{2}}\psi(2^n x - k),\ n \in \mathbb{N}_0,\ k = 0, \dots, 2^n - 1,$$

with

$$\psi\colon \mathbb{R} \to \mathbb{R},\ x \mapsto \begin{cases} 1, & x \in [0, 1/2) \\ -1, & x \in [1/2, 1) \\ 0, & \text{else.} \end{cases}$$

We approximate the piecewise constant mother wavelet ψ through a continuous piecewise linear function realized by a neural network as follows

$$\begin{aligned} \Psi_\delta(x) :=\ & \tfrac{1}{2\delta}\rho(x+\delta) - \tfrac{1}{2\delta}\rho(x-\delta) - \tfrac{1}{\delta}\rho(x-(\tfrac{1}{2}-\delta)) \\ & + \tfrac{1}{\delta}\rho(x-(\tfrac{1}{2}+\delta)) + \tfrac{1}{2\delta}\rho(x-(1-\delta)) - \tfrac{1}{2\delta}\rho(x-(1+\delta)) \end{aligned}$$

and, setting $\delta(\varepsilon) := \varepsilon^2$ for $\varepsilon \in (0, 1/2)$, let

$$\Phi_{n,k,\varepsilon}(x) := 2^{\frac{n}{2}}\Psi_{\delta(\varepsilon)}(2^n x - k),\ n \in \mathbb{N}_0,\ k = 0, \dots, 2^n - 1.$$

The basic idea in the approximation of ψ through Ψ_δ is to let the transition regions around $0, 1/2$, and 1 shrink, as a function of ε, sufficiently fast for the construction to realize an approximation error of no more than ε. Now, a direct calculation yields that, indeed, for $\varepsilon \in (0, 1/2)$,

$$\|\psi_{n,k} - \Phi_{n,k,\varepsilon}\|_{L^2([0,1])} \leq \varepsilon.$$

Moreover, we have $\mathcal{M}(\Phi_{n,k,\varepsilon}) = 18$ and $\mathcal{B}(\Phi_{n,k,\varepsilon}) \leq \max\{2^{\frac{n}{2}}\varepsilon^{-2}, 2^n\}$. In order to establish effective representability by neural networks, we need to order the Haar wavelet dictionary suitably. Specifically, we proceed from coarse to fine scales, i.e., we let

$$(\varphi_i)_{i\in\mathbb{N}} = \mathcal{D} = \{\mathcal{D}_0, \mathcal{D}_1, \dots\},$$

with $\mathcal{D}_n := \{\psi_{n,k} \mapsto \mathbb{R}\colon k = 0, \dots, 2^n - 1\}$, where the ordering within the $\mathcal{D}_n$ may be chosen arbitrarily. Next, note that for every pair $n \in \mathbb{N}_0$, $k \in \{0, \dots, 2^n - 1\}$, there exists a unique index $i \in \mathbb{N}$ such that $\varphi_i = \psi_{n,k} = \psi_{n(i),k(i)}$ and, owing to $|\mathcal{D}_n| = 2^n$, we have $2^{n(i)} \leq i$. Finally, taking $\Phi_{i,\varepsilon} := \Phi_{n(i),k(i),\varepsilon}$ and $\pi(a, b) := a^2 b + b + 18$,

the conditions in Definition 13 for effective representability by neural networks are readily verified. A more elaborate example, namely spline wavelets, is considered at the end of this section.

We are now ready to proceed to the general definition of affine dictionaries with canonical ordering.

A. Affine Dictionaries with Canonical Ordering

Definition 14. *Let $d, S \in \mathbb{N}$, $\delta > 0$, $\Omega \subseteq \mathbb{R}^d$ be bounded, and let $g_s \in L^\infty(\mathbb{R}^d)$, $s \in \{1, \dots, S\}$, be compactly supported. Furthermore, for $s \in \{1, \dots, S\}$, let $J_s \subseteq \mathbb{N}$ and $A_{s,j} \in \mathbb{R}^{d\times d}$, $j \in J_s$, be full-rank and with eigenvalues bounded below by 1 in absolute value. We define the* affine dictionary $\mathcal{D} \subseteq L^2(\Omega)$ *with generator functions $(g_s)_{s=1}^S$ as*

$$\mathcal{D} := \Big\{ g_s^{j,e} := \Big(|\det(A_{s,j})|^{\frac{1}{2}} g_s(A_{s,j} \cdot - \delta e) \Big) \big|_\Omega \colon \ s \in \{1, \dots, S\}, \\ e \in \mathbb{Z}^d, \ j \in J_s, \text{ and } g_s^{j,e} \neq 0 \Big\}.$$

Moreover, we define the sub-dictionaries

$$\mathcal{D}_{s,j} := \{ g_s^{j,e} \in \mathcal{D} : e \in \mathbb{Z}^d \text{ and } g_s^{j,e} \neq 0 \}, \\ \text{for } j \in J_s, \ s \in \{1, \dots, S\}$$

$$\mathcal{D}_j := \bigcup_{s \in \{1,\dots,S\} \colon \ j \in J_s} \mathcal{D}_{s,j}, \quad \text{for } j \in \mathbb{N}.$$

We call an affine dictionary canonically ordered if it is arranged according to

$$(\varphi_i)_{i\in\mathbb{N}} = \mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \dots), \tag{1.50}$$

where the elements within each $\mathcal{D}_j$ may be ordered arbitrarily, and there exist constants $a, c > 0$ such that

$$\sum_{k=1}^{j-1} |\det(A_{s,k})| \geq c \|A_{s,j}\|_\infty^a, \ \ \textit{for all } j \in J_s \backslash \{1\}, \ s \in \{1, \dots, S\}. \tag{1.51}$$

We call an affine dictionary nondegenerate if for every $j \in J_s$, $s \in \{1, \dots, S\}$, the sub-dictionary $\mathcal{D}_{s,j}$ contains at least one element.

Note that for sake of greater generality, we associate possibly different sets $J_s \subseteq \mathbb{N}$ with the generator functions g_s and, in particular, also allow these sets to be finite. The Haar wavelet dictionary example above is recovered as a nondegenerate affine dictionary by taking $d = 1$, $\Omega = [0, 1]$, $S = 1$, $J_s = \mathbb{N}$, $g_1 = \psi$, $\delta = 1$, $A_{1,j} = 2^{j-1}$, $a = 1$, $c = 1/2$, and noting that nondegeneracy is verified as for scale j, the sub-dictionary $\mathcal{D}_{s,j}$ contains 2^{j-1} elements. Moreover, the weights of the networks approximating the individual Haar wavelet dictionary elements grow linearly in the index of the dictionary elements. This is a consequence of the weights being determined by the dilation factor 2^n and $2^{n(i)} \leq i$ due to the ordering we chose. As will be shown below, morally this continues to hold for general nondegenerate affine dictionaries, thereby revealing what informed our definition of canonical ordering. Besides, our notion of canonical ordering is also inspired by the ordering employed in the tail compactness considerations for Besov spaces and orthonormal wavelet dictionaries as detailed in Appendix B. We remark that (1.51) constitutes a very weak restriction on how fast the size of dilations may grow; in fact, we are not aware of any affine dictionaries in the literature that would violate this condition. Finally, we note that the dilations $A_{s,j}$ are not required to be ordered in ascending size, as was the case in the Haar wavelet dictionary example. Canonical ordering does, however, ensure a modicum of ordering.

B. Invariance to Affine Transformations

Affine dictionaries consist of dilations and translations of a given generator function. It is therefore important to understand the impact of these operations on the approximability—by neural networks—of a given function. As neural networks realize concatenations of affine functions and nonlinearities, it is clear that translations and dilations can be absorbed into the first layer of the network and the transformed function should inherit the approximability properties of the generator

function. However, what we will have to understand is how the weights, the connectivity, and the domain of approximation of the resulting network are impacted. The following result makes this quantitative.

Proposition 5. *Let $d \in \mathbb{N}$, $p \in [1, \infty]$, and $f \in L^p(\mathbb{R}^d)$. Assume that there exists a bivariate polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon, \tag{1.52}$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, for all full-rank matrices $A \in \mathbb{R}^{d \times d}$, and all $e \in \mathbb{R}^d$, $E \in \mathbb{R}_+$, and $\eta \in (0, 1/2)$, there is a network $\Psi_{A,e,E,\eta} \in \mathcal{N}_{d,1}$ satisfying

$$\left\| |\det(A)|^{\frac{1}{p}} f(A \cdot - e) - \Psi_{A,e,E,\eta} \right\|_{L^p([-E,E]^d)} \leq \eta,$$

with $\mathcal{M}(\Psi_{A,e,E,\eta}) \leq \pi'(\log(\eta^{-1}), \log(\lceil F \rceil))$ and $\mathcal{B}(\Psi_{A,e,E,\eta}) \leq \max\{\mathcal{B}(\Phi_{F,\eta}), |\det(A)|^{\frac{1}{p}}, \|A\|_\infty, \|e\|_\infty\}$, where $F = dE\|A\|_\infty + \|e\|_\infty$ and π' is of the same degree as π.

Proof. By a change of variables, we have for every $\Phi \in \mathcal{N}_{d,1}$,

$$\left\| |\det(A)|^{\frac{1}{p}} f(A \cdot - e) - |\det(A)|^{\frac{1}{p}} \Phi(A \cdot - e) \right\|_{L^p([-E,E]^d)} \tag{1.53}$$

$$= \|f - \Phi\|_{L^p(A \cdot [-E,E]^d - e)}. \tag{1.54}$$

Furthermore, observe that

$$\begin{aligned} A \cdot [-E, E]^d - e &\subseteq [-(dE\|A\|_\infty + \|e\|_\infty), (dE\|A\|_\infty + \|e\|_\infty)]^d \\ &= [-F, F]^d. \end{aligned} \tag{1.55}$$

Next, we consider the affine transformations $W_{A,e}(x) := Ax - e$, $W'_A(x) := |\det(A)|^{\frac{1}{p}} x$ as depth-1 networks and take $\Psi_{A,e,E,\eta} := W'_A \circ \Phi_{F,\eta} \circ W_{A,e}$ according to Lemma 1. Combining (1.53) and

(1.55) yields

$$\begin{aligned}&\big\||\det(A)|^{\frac{1}{p}} f(A\cdot - e) - \Psi_{A,e,E,\eta}\big\|_{L^p([-E,E]^d)}\\ &= \|f - \Phi_{F,\eta}\|_{L^p(A\cdot[-E,E]^d - e)}\\ &\leq \|f - \Phi_{F,\eta}\|_{L^p([-F,F]^d)} \leq \eta.\end{aligned}$$

The desired bounds on $\mathcal{M}(\Psi_{A,e,E,\eta})$ and $\mathcal{B}(\Psi_{A,e,E,\eta})$ follow directly by construction. □

C. Canonically Ordered Affine Dictionaries are Effectively Representable

The next result establishes that canonically ordered affine dictionaries with generator functions that can be approximated well by neural networks are effectively representable by neural networks.

Theorem 6. *Let $d, S \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ be bounded with nonempty interior, $(g_s)_{s=1}^S \in L^\infty(\mathbb{R}^d)$ compactly supported, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ a nondegenerate canonically ordered affine dictionary with generator functions $(g_s)_{s=1}^S$. Assume that there exists a polynomial π such that, for all $s \in \{1,\dots,S\}$, $\varepsilon \in (0,1/2)$, there is a network $\Phi_{s,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g_s - \Phi_{s,\varepsilon}\|_{L^2(\mathbb{R}^d)} \leq \varepsilon, \tag{1.56}$$

with $\mathcal{M}(\Phi_{s,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_{s,\varepsilon}) \leq \pi(\varepsilon^{-1})$. Then, $\mathcal{D}$ is effectively representable by neural networks.

Proof. By Definition 13 we need to establish the existence of a bivariate polynomial π such that for each $i \in \mathbb{N}$, $\eta \in (0,1/2)$, there is a network $\Phi_{i,\eta} \in \mathcal{N}_{d,1}$ satisfying

$$\|\varphi_i - \Phi_{i,\eta}\|_{L^2(\Omega)} \leq \eta, \tag{1.57}$$

with $\mathcal{M}(\Phi_{i,\eta}) \leq \pi(\log(\eta^{-1}), \log(i))$ and $\mathcal{B}(\Phi_{i,\eta}) \leq \pi(\eta^{-1}, i)$. Note that we have

$$\varphi_i = g_{s_i}^{j_i,e_i} = \Big(|\det(A_{s_i,j_i})|^{\frac{1}{2}} g_{s_i}(A_{s_i,j_i}\cdot - \delta e_i)\Big)\big|_\Omega,$$

for $s_i \in \{1, \dots, S\}$, $j_i \in J_{s_i}$, and $e_i \in \mathbb{Z}^d$. In order to devise networks satisfying (1.57), we employ Proposition 5, upon noting that, by virtue of (1.56), the networks $\Phi_{s,\varepsilon}$ satisfy (1.52) with $p = 2$, $f = g_s$, for every $D \in \mathbb{R}_+$. Consequently Proposition 5 yields a connectivity bound that is even slightly stronger than needed, as it is independent of i. It remains to ensure that the desired bound on $\mathcal{B}(\Phi_{i,\eta})$ holds. This is the case for $\|A_{s_i,j_i}\|_\infty$ and $\|e_i\|_\infty$ both bounded polynomially in i. In order to verify this, we first bound $\|e_i\|_\infty$ relative to $\|A_{s_i,j_i}\|_\infty$. As the generators $(g_s)_{s=1}^S$ are compactly supported by assumption, there exists $E \in \mathbb{R}_+$ such that, for every $s \in \{1, \dots, S\}$, the support of g_s is contained in $[-E, E]^d$. We thus get, for all $s \in \{1, \dots, S\}$, $j \in J_s$, and $e \in \mathbb{Z}^d$, that

$$\begin{aligned} &\|\delta e\|_\infty \geq \sup_{x \in \Omega} \|A_{s,j} x\|_\infty + E \\ &\implies g_s^{j,e}(x) = 0, \, \forall x \in \Omega \implies g_s^{j,e} \notin \mathcal{D}_j. \end{aligned}$$

Since Ω is bounded by assumption, there hence exists a constant $c = c(\Omega, (g_s)_{s=1}^S, \delta, d)$ such that, for all $s \in \{1, \dots, S\}$, $j \in J_s$, and $e \in \mathbb{Z}^d$, we have

$$g_s^{j,e} \in \mathcal{D}_j \implies \|e\|_\infty \leq c \|A_{s,j}\|_\infty.$$

It remains to show that $\|A_{s_i,j_i}\|_\infty$ is polynomially bounded in i. We start by claiming that, for every $s \in \{1, \dots, S\}$, there is a constant $c_s := c_s(\Omega, \delta, d) > 0$ such that

$$|\det(A_{s,j})| \leq c_s |\mathcal{D}_{s,j}|, \text{ for all } j \in J_s. \tag{1.58}$$

To verify this claim, first note that $|\mathcal{D}_{s,j}| \geq 1$, for all $s \in \{1, \dots, S\}, j \in J_s$, owing to the nondegeneracy condition. Thus, for every $s \in \{1, \dots, S\}$, $j \in J_s$, there exist $x_0 \in \Omega$ and $e_0 \in \mathbb{Z}^d$ such that $g_s^{j,e_0}(x_0) \neq 0$, which implies

$$\begin{aligned} g_s^{j,e}(x_0 + A_{s,j}^{-1} \delta(e - e_0)) &= |\det(A_{s,j})|^{\frac{1}{2}} g_s(A_{s,j} x_0 - \delta e_0) \\ &= g_s^{j,e_0}(x_0) \neq 0. \end{aligned}$$

We can therefore conclude that $x_0 + A_{s,j}^{-1}\delta(e - e_0) \in \Omega$ implies $g_s^{j,e} \in \mathcal{D}_{s,j}$. Consequently, we have

$$\begin{aligned}|\mathcal{D}_{s,j}| &\geq |\{e \in \mathbb{Z}^d \colon x_0 + A_{s,j}^{-1}\delta(e - e_0) \in \Omega\}| \\ &= |\{e \in \mathbb{Z}^d \colon A_{s,j}^{-1}\delta e \in \Omega - x_0\}| \\ &= |\mathbb{Z}^d \cap \tfrac{1}{\delta} A_{s,j}(\Omega - x_0)|.\end{aligned}$$

As Ω was assumed to have nonempty interior, there exists a constant $C = C(\Omega)$ such that

$$\begin{aligned}|\mathbb{Z}^d \cap \tfrac{1}{\delta} A_{s,j}(\Omega - x_0)| &\geq C\, vol\left(\tfrac{1}{\delta} A_{s,j}(\Omega - x_0)\right) \\ &= C\, \delta^{-d} |\det(A_{s,j})|\, vol(\Omega).\end{aligned}$$

We have hence established the claim (1.58). Combining (1.51) and (1.58), we obtain, for all $s_i \in \{1, \ldots, S\}$, $j \in J_s \backslash \{1\}$,

$$c\|A_{s_i,j_i}\|_\infty^a \leq \sum_{k=1}^{j_i - 1} |\det(A_{s_i,k})| \leq c_{s_i} \sum_{k=1}^{j_i - 1} |\mathcal{D}_{k,s_i}| \leq c_s i,$$

where the last inequality follows from the fact that $\varphi_i \in \mathcal{D}_{j_i,s_i}$ and hence its index i must be larger than the number of elements contained in preceding sub-dictionaries. This ensures that

$$\|A_{s_i,j_i}\|_\infty \leq \left(\frac{1}{c} \max_{s=1,\ldots,S} c_s\right)^{\frac{1}{a}} i^{\frac{1}{a}} + \max_{s=1,\ldots,S} \|A_{s,1}\|_\infty, \text{ for all } i \in \mathbb{N},$$

thereby completing the proof. □

Remark 7. *Theorem 6 is restricted, for ease of exposition, to bounded Ω and compactly supported generator functions g_s. The result can be extended to $\Omega = \mathbb{R}^d$ and to generator functions g_s of unbounded support but sufficiently fast decay. This extension requires additional technical steps and an alternative definition of canonical ordering. For conciseness we do not provide the details here, but instead refer to the proofs of Theorems 10 and 12, which deal with the corresponding technical aspects in the context of approximation of Gabor dictionaries by neural networks.*

We can now put the results together to conclude a remarkable universality and optimality property of neural networks: Consider an affine dictionary generated by functions g_s that can be approximated well by neural networks. If this dictionary provides Kolmogorov-Donoho-optimal approximation for a given function class, then so do neural networks.

Theorem 7. *Let $d, S \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ be bounded with nonempty interior, $(g_s)_{s=1}^S \in L^\infty(\mathbb{R}^d)$ compactly supported, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subseteq L^2(\Omega)$ a nondegenerate canonically ordered affine dictionary with generator functions $(g_s)_{s=1}^S$. Assume that there exists a polynomial π such that, for all $s \in \{1, \dots, S\}$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{s,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying $\|g_s - \Phi_{s,\varepsilon}\|_{L^2(\mathbb{R}^d)} \leq \varepsilon$ with $\mathcal{M}(\Phi_{s,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}))$ and $\mathcal{B}(\Phi_{s,\varepsilon}) \leq \pi(\varepsilon^{-1})$. Then, we have*

$$\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D})$$

for all compact function classes $\mathcal{C} \subseteq L^2(\Omega)$. In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{D}$ (in the sense of Definition 7), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition 11).

Proof. The first statement follows from Theorem 5 and Theorem 6, the second from Theorem 4. □

D. Spline wavelets

We next particularize the results developed above to show that neural networks Kolmogorov-Donoho optimally represent all function classes $\mathcal{C}$ that are optimally representable by spline wavelet dictionaries. As spline wavelet dictionaries have B-splines as generator functions, we start by showing how B-splines can be realized through neural networks. For simplicity of exposition, we restrict ourselves to the univariate case throughout.

Definition 15. *Let $N_1 := \chi_{[0,1]}$ and for $m \in \mathbb{N}$, define*

$$N_{m+1} := N_1 * N_m,$$

where $*$ stands for convolution. We refer to N_m as the univariate cardinal B-spline of order m.

Recognizing that B-splines are piecewise polynomial, we can build on Proposition 3 to get the following statement on the approximation of B-splines by deep neural networks.

Lemma 10. *Let $m \in \mathbb{N}$. There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ satisfying*

$$\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_\varepsilon) \leq C \log(\varepsilon^{-1})$ and $\mathcal{B}(\Phi_\varepsilon) \leq 1$.

Proof. The proof is based on the following representation (Unser, 1997, Eq. 19)

$$N_m(x) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho((x-k)^m). \tag{1.59}$$

While N_m is supported on $[0, m]$, the networks Φ_ε can have support outside $[0, m]$ as well. We only need to ensure that Φ_ε is "close" to N_m on $[0, m]$ and at the same time "small" outside the interval $[0, m]$. To accomplish this, we first approximate N_m on the slightly larger domain $[-1, m+1]$ by a linear combination of networks realizing shifted monomials according to (1.59), and then multiply the resulting network by another one that takes on the value 1 on $[0, m]$ and 0 outside of $[-1, m+1]$. Specifically, we proceed as follows. Proposition 3 ensures the existence of a constant C_1 such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{m+2,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Psi_{m+2,\varepsilon}(x) - x^m\|_{L^\infty([-(m+2),m+2])} \leq \tfrac{\varepsilon}{4(m+2)},$$

with $\mathcal{M}(\Psi_{m+2,\varepsilon}) \leq C_1 \log(\varepsilon^{-1})$ and $\mathcal{B}(\Psi_{m+2,\varepsilon}) \leq 1$. Note that we did not make the dependence of $\mathcal{M}(\Psi_{m+2,\varepsilon})$ on m explicit as we consider m to be fixed. Next, let $T_k(x) := x - k$ and observe that $\rho((x-k)^m)$ can be realized as a neural network according to $\rho \circ \Psi_{m+2,\varepsilon} \circ T_k$, where T_k is taken pursuant to Corollary 2. Next, we define, for $\varepsilon \in (0, 1/2)$, the network

$$\widetilde{\Phi}_\varepsilon := \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho \circ \Psi_{m+2,\varepsilon} \circ T_k$$

and note that

$$\frac{1}{m!}\binom{m+1}{k} = \frac{m+1}{k!(m-k+1)!} \leq 2,$$

for $k = 0, \ldots, m+1$. As ρ is 1-Lipschitz, we have, for all $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}
&\|\widetilde{\Phi}_\varepsilon - N_m\|_{L^\infty([-1,m+1])} \\
&\leq \sum_{k=0}^{m+1} \frac{1}{m!}\binom{m+1}{k} \|\rho \circ \Psi_{m+2,\varepsilon} \circ T_k - \rho \circ T_k^m\|_{L^\infty([-1,m+1])} \\
&\leq 2 \sum_{k=0}^{m+1} \|\Psi_{m+2,\varepsilon}(x) - x^m\|_{L^\infty([-(m+2),m+2])} \leq \tfrac{\varepsilon}{2}.
\end{aligned} \tag{1.60}$$

Let now $\Gamma(x) := \rho(x+1) - \rho(x) - \rho(x-m) + \rho(x-(m+1))$, note that $0 \leq \Gamma(x) \leq 1$, and take $\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2}$ to be the multiplication network from Lemma 2. We define $\Phi_\varepsilon := \Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2} \circ (\widetilde{\Phi}_\varepsilon, \Gamma)$ according to Lemma 1 and Lemma 18 and note that

$$\begin{aligned}
&\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})} \\
&\leq \|\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2} \circ (\widetilde{\Phi}_\varepsilon, \Gamma) - \widetilde{\Phi}_\varepsilon \cdot \Gamma\|_{L^\infty([-1,m+1])} \\
&+ \|\widetilde{\Phi}_\varepsilon \cdot \Gamma - N_m\|_{L^\infty([-1,m+1])}
\end{aligned} \tag{1.61}$$

as both N_m and Γ vanish outside $[-1, m+1]$ and $\Phi^{\mathrm{mult}}_{1+\varepsilon/2,\varepsilon/2}$ delivers zero whenever at least one of its inputs is zero. Note that the first

term on the right-hand-side of (1.61) is upper-bounded by $\frac{\varepsilon}{2}$ as a consequence of $N_m(x) \leq 1$ and hence $\widetilde{\Phi}_\varepsilon(x) \leq 1 + \frac{\varepsilon}{2}$, for $x \in [-1, m+1]$, owing to (1.60). For the second term, we split up the interval $[-1, m+1]$ and first note that, for $x \in [0, m]$, $\Gamma(x) = 1$, which implies $\|\widetilde{\Phi}_\varepsilon \cdot \Gamma - N_m\|_{L^\infty([0,m])} = \|\widetilde{\Phi}_\varepsilon - N_m\|_{L^\infty([0,m])} \leq \varepsilon/2$, again owing to (1.60). For $x \in [-1, m+1] \setminus [0, m]$, we have $N_m(x) = 0$ and $\Gamma(x) \leq 1$, which yields

$$\begin{aligned}
|\widetilde{\Phi}_\varepsilon(x) \cdot \Gamma(x) - N_m(x)| &\leq |\widetilde{\Phi}_\varepsilon(x)| \\
&\leq |\widetilde{\Phi}_\varepsilon(x) - N_m(x)| + |N_m(x)| \\
&= |\widetilde{\Phi}_\varepsilon(x) - N_m(x)| \\
&\leq \varepsilon/2,
\end{aligned}$$

again by (1.60). In summary, (1.60) hence ensures that the second term in (1.61) is also upper-bounded by $\frac{\varepsilon}{2}$ and therefore $\|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})} \leq \varepsilon$. Combining Lemma 1, Proposition 2, Corollary 2, Lemma 15, and Lemma 18 establishes the desired bounds on $\mathcal{M}(\Phi_{D,\varepsilon})$ and $\mathcal{B}(\Phi_{D,\varepsilon})$. □

Remark 8. *As both N_m and the approximating networks Φ_ε we constructed in the proof of Lemma 10 are supported in $[-1, m+1]$, we have $\|\Phi_\varepsilon - N_m\|_{L^2(\mathbb{R})} \leq (m+2)^{1/2} \|\Phi_\varepsilon - N_m\|_{L^\infty(\mathbb{R})}$, which shows that Lemma 10 continues to hold when the approximation error is measured in $L^2(\mathbb{R})$-norm, albeit with a different constant C.*

We are now ready to introduce spline wavelet dictionaries. For $n, j \in \mathbb{Z}$, set

$$V_n := \mathrm{clos}_{L^2}\Big(\mathrm{span}\,\{N_m(2^n x - k) : k \in \mathbb{Z}\}\Big),$$

where clos_{L^2} denotes closure with respect to L^2-norm. Spline spaces V_n, $n \in \mathbb{Z}$, constitute a multiresolution analysis (Mallat, 1989) of $L^2(\mathbb{R})$ according to

$$\{0\} \subseteq \ldots V_{-1} \subseteq V_0 \subseteq V_1 \subseteq \cdots \subseteq L^2(\mathbb{R}).$$

Moreover, with the orthogonal complements $(\dots, W_{-1}, W_0, W_1, \dots)$ such that $V_{n+1} = V_n \oplus W_n$, where $\oplus$ denotes the orthogonal sum, we have

$$L^2(\mathbb{R}) = V_0 \oplus \bigoplus_{k=0}^{\infty} W_k.$$

Theorem 8 ((Chui and Wang, 1992, Theorem 1))**.** *Let $m \in \mathbb{N}$. The m-th order spline*

$$\psi_m(x) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j+1) \frac{d^m}{dx^m} N_{2m}(2x - j), \quad (1.62)$$

with support $[0, 2m-1]$, is a basic wavelet that generates W_0 and thereby all the spaces W_n, $n \in \mathbb{Z}$. Consequently, the set

$$\begin{aligned} \mathcal{W}_m :&= \{\psi_{k,n}(x) = 2^{n/2}\psi_m(2^n x - k) : n \in \mathbb{N}_0, k \in \mathbb{Z}\} \\ &\cup \{\phi_k(x) = N_m(x-k) : k \in \mathbb{Z}\} \end{aligned} \quad (1.63)$$

is a countable complete orthonormal wavelet basis in $L^2(\mathbb{R})$.

Taking $\Omega \subseteq \mathbb{R}$, $S = 2$, $J_1 = \mathbb{N}$, $J_2 = \{1\}$, $A_{1,j} = 2^{j-1}$ for $j \in \mathbb{N}$, and $A_{2,1} = 1$, we get that

$$\begin{aligned} \mathcal{D} := \Big\{ g_s^{j,e}(x) := \Big(|A_j|^{\frac{1}{2}} g_s(A_j \cdot - \delta e) \Big)\Big|_\Omega : s \in \{1,2\}, \\ e \in \mathbb{Z},\ j \in J_s, \text{ and } g_s^{j,e} \neq 0 \Big\} = \mathcal{W}_m \end{aligned} \quad (1.64)$$

is a nondegenerate canonically ordered affine dictionary with generators $g_1 = \psi_m$ and $g_2 = N_m$. The canonical ordering condition (1.51) is satisfied with $a = 1$ and $c = 1/2$. Nondegeneracy follows upon noting that $\operatorname{supp}(\psi_{k,n}) = [2^{-n}k, 2^{-n}(2m-1+k)]$ and $\operatorname{supp}(N_m(\cdot - k)) = [k, m+k]$, which implies that all sub-dictionaries contain at least one element as required.

We have therefore established the following.

Theorem 9. *Let $\Omega \subseteq \mathbb{R}$ be bounded and of nonempty interior and $\mathcal{D} = (\varphi_i)_{i \in \mathbb{N}} \subseteq L^2(\Omega)$ a spline wavelet dictionary according to* (1.64)

ordered per (1.50). *Then, all compact function classes* $\mathcal{C} \subseteq L^2(\Omega)$ *that are optimally representable by* $\mathcal{D}$ *(in the sense of Definition 7) are optimally representable by neural networks (in the sense of Definition 11).*

Proof. As the canonical ordering and the nondegeneracy conditions were already verified, it remains to establish that the generators ψ_m and N_m satisfy the antecedent of Theorem 6. To this end, we first devise an alternative representation of (1.62). Specifically, using the identity (Chui and Wang, 1992, Eq. 2.2)

$$\frac{d^m}{dx^m} N_{2m}(x) = \sum_{j=0}^{m} (-1)^j \binom{m}{j} N_m(x-j),$$

we get

$$\psi_m(x) = \sum_{n=1}^{3m-1} q_n N_m(2x-n+1), \tag{1.65}$$

with

$$q_n = \frac{(-1)^{n+1}}{2^{m-1}} \sum_{j=0}^{m} \binom{m}{j} N_{2m}(n-j).$$

As (1.65) shows that ψ_m is a linear combination of shifts and dilations of N_m, combining Lemma 10 and Remark 8 with Lemma 4 and Proposition 5 ensures that (1.56) is satisfied. Application of Theorem 7 then establishes the claim. □

1.9. WEYL-HEISENBERG DICTIONARIES

In this section, we consider Weyl-Heisenberg a.k.a. Gabor dictionaries (Gröchenig, 2013), which consist of time-frequency translates of a given generator function. Gabor dictionaries play a fundamental role in time-frequency analysis (Gröchenig, 2013) and in the study of partial differential equations (Fefferman, 1983). We start with the formal definition of Gabor dictionaries.

Definition 16 (Gabor dictionaries)**.** *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d)$, and $x, \xi \in \mathbb{R}^d$. We define the translation operator $T_x \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ as*

$$T_x f(t) := f(t - x)$$

and the modulation operator $M_\xi \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \mathbb{C})$ as

$$M_\xi f(t) := e^{2\pi i \langle \xi, t \rangle} f(t).$$

Let $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, and $g \in L^2(\mathbb{R}^d)$. The Gabor dictionary $\mathcal{G}(g, \alpha, \beta, \Omega) \subseteq L^2(\Omega)$ is defined as

$$\mathcal{G}(g, \alpha, \beta, \Omega) := \left\{ M_\xi T_x g \big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \right\}.$$

In order to describe representability in neural networks in the sense of Definition 13, we need to order the elements in $\mathcal{G}(g, \alpha, \beta, \Omega)$. To this end, let $\mathcal{G}_0(g, \alpha, \beta, \Omega) := \{g\big|_\Omega\}$ and define $\mathcal{G}_n(g, \alpha, \beta, \Omega)$, $n \in \mathbb{N}$, recursively according to

$$\mathcal{G}_n(g, \alpha, \beta, \Omega) := \{ M_\xi T_x g \big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d, \|x\|_\infty \leq n\alpha, \\ \|\xi\|_\infty \leq n\beta \} \setminus \bigcup_{k=0}^{n-1} \mathcal{G}_k(g, \alpha, \beta, \Omega).$$

We then organize $\mathcal{G}(g, \alpha, \beta, \Omega)$ as

$$\mathcal{G}(g, \alpha, \beta, \Omega) = (\mathcal{G}_0(g, \alpha, \beta, \Omega), \mathcal{G}_1(g, \alpha, \beta, \Omega), \dots), \tag{1.66}$$

where the ordering within the sets $\mathcal{G}_n(g, \alpha, \beta, \Omega)$ is arbitrary. We hasten to add that the specifics of the overall ordering in (1.66) are irrelevant as long as $\mathcal{G}(g, \alpha, \beta, \Omega) = (\varphi_i)_{i \in \mathbb{N}}$ with $\varphi_i = \mathcal{M}_{\xi(i)} T_{x(i)} g \big|_\Omega$ is such that $\|x(i)\|_\infty$ and $\|\xi(i)\|_\infty$ do not grow faster than polynomially in i; this will become apparent in the proof of Theorem 10. We note that this ordering is also inspired by that employed in the tail compactness considerations for modulation spaces and Wilson bases as detailed in Appendix C.

As Gabor dictionaries are built from time-shifted and modulated versions of the generator function g, and invariance to time-shifts was already established in Proposition 5, we proceed to showing that the approximation-theoretic properties of the generator function are inherited by its modulated versions. This result can be interpreted as an invariance property to frequency shifts akin to that established in Proposition 5 for affine transformations in the context of affine dictionaries. In summary, neural networks exhibit a remarkable invariance property both to the affine group operations of scaling and translation and to the Weyl-Heisenberg group operations of modulation and translation.

Lemma 11. *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and for every $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, let $\Phi_{D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfy*

$$\|f - \Phi_{D,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq \varepsilon.$$

Then, there exists a constant $C > 0$ (which does not depend on f) such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, $\xi \in \mathbb{R}^d$, there are networks $\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}, \Phi^{\mathrm{Im}}_{D,\xi,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\begin{aligned}&\|\mathrm{Re}(M_\xi f) - \Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}\|_{L^\infty([-D,D]^d)} \\ &+ \|\mathrm{Im}(M_\xi f) - \Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}\|_{L^\infty([-D,D]^d)} \\ &\leq 3\varepsilon\end{aligned}$$

with

$$\begin{aligned}\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}), \mathcal{L}(\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}) &\leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) \\ &\quad + (\log(\lceil S_f \rceil))^2) + \mathcal{L}(\Phi_{D,\varepsilon}), \\ \mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}) &\leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) \\ &\quad + (\log(\lceil S_f \rceil))^2 + d) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}),\end{aligned}$$

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 1$, where $S_f := \max\{1, \|f\|_{L^\infty(\mathbb{R}^d)}\}$.

Proof. All statements in the proof involving ε pertain to $\varepsilon \in (0, 1/2)$

without explicitly stating this every time. We start by observing that

$$\begin{aligned}\mathrm{Re}(M_\xi f)(t) &= \cos(2\pi\langle\xi,t\rangle)f(t)\\ \mathrm{Im}(M_\xi f)(t) &= \sin(2\pi\langle\xi,t\rangle)f(t)\end{aligned}$$

due to $f \in \mathbb{R}$. Note that for given $\xi \in \mathbb{R}^d$, the map $t \mapsto \langle\xi,t\rangle = \xi^T t = t_1\xi_1 + \cdots + t_d\xi_d$ is simply a linear transformation. Hence, combining Lemma 1, Theorem 2, and Corollary 2 establishes the existence of a constant C_1 such that for all $D \in \mathbb{R}_+$, $\xi \in \mathbb{R}^d$, $\varepsilon \in (0,1/2)$, there is a network $\Psi_{D,\xi,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\sup_{t\in[-D,D]^d} |\cos(2\pi\langle\xi,t\rangle) - \Psi_{D,\xi,\varepsilon}(t)| \leq \tfrac{\varepsilon}{6S_f} \tag{1.67}$$

with

$$\begin{aligned}\mathcal{L}(\Psi_{D,\xi,\varepsilon}) &\leq C_1((\log(\varepsilon^{-1}))^2 + (\log(S_f))^2\\ &\qquad + \log(\lceil dD\|\xi\|_\infty\rceil)),\\ \mathcal{M}(\Psi_{D,\xi,\varepsilon}) &\leq C_1((\log(\varepsilon^{-1}))^2 + (\log(S_f))^2\\ &\qquad + \log(\lceil dD\|\xi\|_\infty\rceil) + d),\end{aligned} \tag{1.68}$$

and $\mathcal{B}(\Psi_{D,\xi,\varepsilon}) \leq 1$. Moreover, Proposition 2 guarantees the existence of a constant $C_2 > 0$ such that for all $\varepsilon \in (0,1/2)$, there is a network $\mu_\varepsilon \in \mathcal{N}_{2,1}$ satisfying

$$\sup_{x,y\in[-S_f-1/2,S_f+1/2]} |\mu_\varepsilon(x,y) - xy| \leq \tfrac{\varepsilon}{6} \tag{1.69}$$

with

$$\mathcal{L}(\mu_\varepsilon), \mathcal{M}(\mu_\varepsilon) \leq C_2(\log(\varepsilon^{-1}) + \log(\lceil S_f\rceil)) \tag{1.70}$$

and $\mathcal{B}(\mu_\varepsilon) \leq 1$. Using Lemmas 2 and 3, we get that the network $\Gamma_{D,\xi,\varepsilon} := (\Psi_{D,\xi,\varepsilon}, \Phi_{D,\varepsilon}) \in \mathcal{N}_{d,2}$ satisfies

$$\begin{aligned}\mathcal{L}(\Gamma_{D,\xi,\varepsilon}) &\leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\},\\ \mathcal{M}(\Gamma_{D,\xi,\varepsilon}) &\leq 2\,\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{M}(\Phi_{D,\varepsilon})\\ &\qquad + 2\,\mathcal{L}(\Psi_{D,\xi,\varepsilon}) + 2\,\mathcal{L}(\Phi_{D,\varepsilon}),\end{aligned}$$

and $\mathcal{B}(\Gamma_{D,\xi,\varepsilon}) \leq 1$. Finally, applying Lemma 1 to concatenate the networks $\Gamma_{D,\xi,\varepsilon}$ and μ_ε, we obtain the network

$$\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon} := \mu_\varepsilon \circ \Gamma_{D,\xi,\varepsilon} = \mu_\varepsilon \circ (\Psi_{D,\xi,\varepsilon}, \Phi_{D,\varepsilon}) \in \mathcal{N}_{d,1}$$

satisfying

$$\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq \max\{\mathcal{L}(\Psi_{D,\xi,\varepsilon}), \mathcal{L}(\Phi_{D,\varepsilon})\} + \mathcal{L}(\mu_\varepsilon), \tag{1.71}$$

$$\begin{aligned}\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 4\mathcal{M}(\Psi_{D,\xi,\varepsilon}) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Psi_{D,\xi,\varepsilon}) \\ +4\mathcal{L}(\Phi_{D,\varepsilon}) + 2\mathcal{M}(\mu_\varepsilon),\end{aligned} \tag{1.72}$$

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) \leq 1$. Next, observe that (1.67) and (1.69) imply that

$$\begin{aligned}
&\|\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon} - \mathrm{Re}(M_\xi f)\|_{L^\infty([-D,D]^d)} \\
&= \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \|\mu_\varepsilon(\Psi_{D,\xi,\varepsilon}(\,\cdot\,), \Phi_{D,\varepsilon}(\,\cdot\,)) - \Psi_{D,\xi,\varepsilon}(\,\cdot\,)\Phi_{D,\varepsilon}(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,)(\Phi_{D,\varepsilon}(\,\cdot\,) - f(\,\cdot\,))\|_{L^\infty([-D,D]^d)} \\
&\quad + \|\Psi_{D,\xi,\varepsilon}(\,\cdot\,) f(\,\cdot\,) - \cos(2\pi\langle\xi,\,\cdot\,\rangle) f(\,\cdot\,)\|_{L^\infty([-D,D]^d)} \\
&\leq \tfrac{\varepsilon}{6} + (1 + \tfrac{\varepsilon}{6S_f})\varepsilon + \tfrac{\varepsilon}{6} \leq \tfrac{3}{2}\varepsilon.
\end{aligned}$$

Combining (1.68), (1.70), (1.72), and (1.71) we can further see that there exists a constant $C > 0$ such that

$$\begin{aligned}
\mathcal{L}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) &\leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) \\
&\qquad +(\log(\lceil S_f \rceil))^2) + \mathcal{L}(\Phi_{D,\varepsilon}), \\
\mathcal{M}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon}) &\leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil dD\|\xi\|_\infty \rceil) \\
&\quad +(\log(\lceil S_f \rceil))^2 + d) + 4\mathcal{M}(\Phi_{D,\varepsilon}) + 4\mathcal{L}(\Phi_{D,\varepsilon}),
\end{aligned}$$

and $\mathcal{B}(\Phi^{\mathrm{Re}}_{D,\xi,\varepsilon})) \leq 1$. The results for $\Phi^{\mathrm{Im}}_{D,\xi,\varepsilon}$ follow analogously, simply by using $\sin(x) = \cos(x - \pi/2)$. □

Note that Gabor dictionaries necessarily contain complex-valued functions. The theory developed so far was, however, phrased for neural networks with real-valued outputs. As is evident from the proof of Lemma 11, this is not problematic when the generator function g is real-valued. For complex-valued generator functions we would need a version of Proposition 2 that applies to the multiplication of complex numbers. Due to $(a+ib)(a'+ib') = (aa'-bb') + i(ab'+a'b)$ such a network can be constructed by realizing the real and imaginary parts of the product as a sum of real-valued multiplication networks and then proceeding as in the proof above. We omit the details as they are straightforward and would not lead to new conceptual insights. Furthermore, an extension—to the complex-valued case—of the concept of effective representability by neural networks according to Definition 13 would be needed. This can be effected by considering the set of neural networks with 1-dimensional complex-valued output as neural networks with 2-dimensional real-valued output, i.e., by setting

$$\mathcal{N}_{d,1}^{\mathbb{C}} := \mathcal{N}_{d,2},$$

with the convention that the first component represents the real part and the second the imaginary part.

We proceed to establish conditions for effective representability of Gabor dictionaries by neural networks.

Theorem 10. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be the corresponding Gabor dictionary with ordering as defined in* (1.66)*. Assume that Ω is bounded or that $\Omega = \mathbb{R}^d$ and g is compactly supported. Further, suppose that there exists a polynomial π such that for every $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{x,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g - \Phi_{x,\varepsilon}\|_{L^\infty(x+\Omega)} \leq \varepsilon, \tag{1.73}$$

with $\mathcal{M}(\Phi_{x,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\|x\|_\infty))$, $\mathcal{B}(\Phi_{x,\varepsilon}) \leq \pi(\varepsilon^{-1}, \|x\|_\infty)$. Then, $\mathcal{G}(g, \alpha, \beta, \Omega)$ is effectively representable by neural networks.

Proof. We start by noting that owing to (1.66), we have $\mathcal{G}(g,\alpha,\beta,\Omega) = (\varphi_i)_{i\in\mathbb{N}}$ with $\varphi_i = \mathcal{M}_{\xi(i)}T_{x(i)}g \in \mathcal{G}_{n(i)}(g,\alpha,\beta,\Omega)$, where

$$\|\xi(i)\|_\infty \leq n(i)\beta \leq i\beta \quad \text{and} \quad \|x(i)\|_\infty \leq n(i)\alpha \leq i\alpha. \tag{1.74}$$

Next, we take the affine transformation $W_x(y) := y - x$ to be a depth-1 network and observe that, due to (1.73) and Lemma 1, we have, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0,1/2)$,

$$\|T_x g - \Phi_{-x,\varepsilon} \circ W_x\|_{L^\infty(\Omega)} = \|g - \Phi_{-x,\varepsilon}\|_{L^\infty(-x+\Omega)} \leq \varepsilon, \tag{1.75}$$

with

$$\mathcal{M}(\Phi_{-x,\varepsilon} \circ W_x) \leq 2(\pi(\log(\varepsilon^{-1}), \log(\|x\|_\infty)) + 2d)$$

$$\begin{aligned}\mathcal{B}(\mathcal{M}(\Phi_{-x,\varepsilon} \circ W_x)) &\leq \max\{\mathcal{B}(\Phi_{-x,\varepsilon}), \|x\|_\infty\} \\ &\leq \pi(\varepsilon^{-1}, \|x\|_\infty) + \|x\|_\infty.\end{aligned}$$

We first consider the case where Ω is bounded and let $E \in \mathbb{R}_+$ be such that $\Omega \subseteq [-E,E]^d$. Combining (1.75) with Proposition 5 and Lemma 11, we can infer the existence of a multivariate polynomial π_1 such that for all $i \in \mathbb{N}$, $\varepsilon \in (0,1/2)$, there is a network $\Phi_{i,\varepsilon} = (\Phi^{\mathrm{Re}}_{i,\varepsilon}, \Phi^{\mathrm{Im}}_{i,\varepsilon}) \in \mathcal{N}^{\mathbb{C}}_{d,1}$ satisfying

$$\begin{aligned}&\|\mathrm{Re}(\mathcal{M}_{\xi(i)}T_{x(i)}g) - \Phi^{\mathrm{Re}}_{i,\varepsilon}\|_{L^\infty(\Omega)} && (1.76)\\ &+ \|\mathrm{Im}(\mathcal{M}_{\xi(i)}T_{x(i)}g) - \Phi^{\mathrm{Im}}_{i,\varepsilon}\|_{L^\infty(\Omega)} && (1.77)\\ &\leq (2E)^{-\frac{d}{2}}\varepsilon, && (1.78)\end{aligned}$$

with

$$\begin{aligned}\mathcal{M}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{M}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\log(\varepsilon^{-1}), \log(\|\xi(i)\|_\infty), \log(\|x(i)\|_\infty)), \\ \mathcal{B}(\Phi^{\mathrm{Re}}_{i,\varepsilon}), \mathcal{B}(\Phi^{\mathrm{Im}}_{i,\varepsilon}) &\leq \pi_1(\varepsilon^{-1}, \|\xi(i)\|_\infty, \|x(i)\|_\infty).\end{aligned} \tag{1.79}$$

Note that here we did not make the dependence of the connectivity and the weight upper bounds on d and E explicit as these quantities are irrelevant for the purposes of what we want to show, as long as they are finite, of course, which is the case by assumption. Likewise, we did not explicitly indicate the dependence of π_1 on g. As $|z| \leq |\mathrm{Re}(z)|+|\mathrm{Im}(z)|$, it follows from (1.76) that for all $i \in \mathbb{N}, \varepsilon \in (0, 1/2)$,

$$\begin{aligned}
&\|\varphi_i - \Phi_{i,\varepsilon}\|_{L^2(\Omega,\mathbb{C})} \\
&\leq (2E)^{\frac{d}{2}} \|\varphi_i - \Phi_{i,\varepsilon}\|_{L^\infty(\Omega,\mathbb{C})} \\
&\leq (2E)^{\frac{d}{2}} \left(\|\mathrm{Re}(\varphi_i) - \Phi_{i,\varepsilon}^{\mathrm{Re}}\|_{L^\infty(\Omega)} + \|\mathrm{Im}(\varphi_i) - \Phi_{i,\varepsilon}^{\mathrm{Im}}\|_{L^\infty(\Omega)}\right) \\
&\leq \varepsilon.
\end{aligned}$$

Moreover, (1.74) and (1.79) imply the existence of a polynomial π_2 such that

$$\mathcal{M}(\Phi_{i,\varepsilon}^{\mathrm{Re}}), \mathcal{M}(\Phi_{i,\varepsilon}^{\mathrm{Im}}) \leq \pi_2(\log(\varepsilon^{-1}), \log(i)),$$

$$\mathcal{B}(\Phi_{i,\varepsilon}^{\mathrm{Re}}), \mathcal{B}(\Phi_{i,\varepsilon}^{\mathrm{Im}}) \leq \pi_2(\varepsilon^{-1}, i),$$

for all $i \in \mathbb{N}, \varepsilon \in (0, 1/2)$. We can therefore conclude that $\mathcal{G}(g, \alpha, \beta, \Omega)$ is effectively representable by neural networks.

We proceed to proving the statement for the case $\Omega = \mathbb{R}^d$ and g compactly supported, i.e., there exists $E \in \mathbb{R}_+$ such that $\mathrm{supp}(g) \subseteq [-E, E]^d$. This implies

$$\begin{aligned}
\mathrm{supp}(M_\xi T_x g) &= \mathrm{supp}(T_x g) \\
&\subseteq x + [-E, E]^d \\
&\subseteq [-(\|x\|_\infty + E), \|x\|_\infty + E]^d.
\end{aligned}$$

Again, combining (1.75) with Proposition 5 and Lemma 11 establishes the existence of a polynomial π_3 such that for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there are networks $\Psi_{x,\xi,\varepsilon}^{\mathrm{Re}}, \Psi_{x,\xi,\varepsilon}^{\mathrm{Im}} \in \mathcal{N}_{d,1}$ satisfying

$$\begin{aligned}
&\|\mathrm{Re}(M_\xi T_x g) - \Psi_{x,\xi,\varepsilon}^{\mathrm{Re}}\|_{L^\infty(S_x)} \\
&+ \|\mathrm{Im}(M_\xi T_x g) - \Psi_{x,\xi,\varepsilon}^{\mathrm{Im}}\|_{L^\infty(S_x)} \\
&\leq \tfrac{\varepsilon}{2s_x},
\end{aligned} \tag{1.80}$$

with

$$\mathcal{M}(\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{M}(\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}) \leq \pi_3(\log(\varepsilon^{-1}), \log(\|x\|_\infty), \log(\|\xi\|_\infty)),$$
$$\mathcal{B}(\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{B}(\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}) \leq \pi_3(\varepsilon^{-1}, \|x\|_\infty, \|\xi\|_\infty),$$

where we set $S_x := [-(\|x\|_\infty + E + 1), \|x\|_\infty + E + 1]^d$ and $s_x := |S_x|^{1/2}$ to simplify notation. As we want to establish effective representability for $\Omega = \mathbb{R}^d$, the estimate in (1.80) is insufficient. In particular, we have no control over the behavior of the networks $\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}$ outside the set S_x. We can, however, construct networks which exhibit the same scaling behavior in terms of $\mathcal{M}$ and $\mathcal{B}$, are supported in S_x, and realize the same output for all inputs in S_x. To this end let, for $y \in \mathbb{R}_+$, the network $\alpha_y \in \mathcal{N}_{1,1}$ be given by

$$\alpha_y(t) := \rho(t - (-y - 1)) - \rho(t - (-y)) - \rho(t - y) + \rho(t - (y + 1)), t \in \mathbb{R}.$$

Note that $\alpha_y(t) = 1$ for $t \in [-y, y]$, $\alpha_y(t) = 0$ for $t \notin [-y - 1, y + 1]$, and $\alpha_y(t) \in (0, 1)$ else. Next, consider, for $x \in \mathbb{R}^d$, the network given by

$$\chi_x(t) := \rho\left(\left[\sum_{i=1}^{d} \alpha_{\|x\|_\infty + E}(t_i)\right] - (d - 1)\right), \quad t = (t_1, t_2, \ldots, t_d) \in \mathbb{R}^d,$$

and note that

$$\begin{aligned} \chi_x(t) &= 1, \quad \forall t \in [-(\|x\|_\infty + E), \|x\|_\infty + E]^d \\ \chi_x(t) &= 0, \quad \forall t \notin [-(\|x\|_\infty + E + 1), \|x\|_\infty + E + 1]^d \\ 0 \leq \chi_x(t) &\leq 1, \quad \forall t \in \mathbb{R}^d. \end{aligned}$$

As d and E are considered fixed here, there exists a constant C_1 such that, for all $x \in \mathbb{R}^d$, we have $\mathcal{M}(\chi_x) \leq C_1$ and $\mathcal{B}(\chi_x) \leq C_1 \max\{1, \|x\|_\infty\}$. Now, let $B := \max\{1, \|g\|_{L^\infty(\mathbb{R})}\}$. Next, by

Proposition 2 there exists a constant C_2 such that, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\mu_{x,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\sup_{y,z\in[-2B,2B]} |\mu_{x,\varepsilon}(y,z) - yz| \leq \tfrac{\varepsilon}{4s_x}, \tag{1.81}$$

and, for all $y \in \mathbb{R}$,

$$\mu_{x,\varepsilon}(0,y) = \mu_{x,\varepsilon}(y,0) = 0, \tag{1.82}$$

with $\mathcal{M}(\mu_{x,\varepsilon}) \leq C_2(\log(\varepsilon^{-1}) + \log(s_x))$ and $\mathcal{B}(\mu_{x,\varepsilon}) \leq 1$. Note that in the upper bound on $\mathcal{M}(\mu_{x,\varepsilon})$, we did not make the dependence on B explicit as we consider g fixed for the purposes of the proof. Next, as E is fixed, there exists a constant C_3 such that $\mathcal{M}(\mu_{x,\varepsilon}) \leq C_3(\log(\varepsilon^{-1}) + \log(\|x\|_\infty + 1))$, for all $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$.

We now take

$$\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} := \mu_{x,\varepsilon} \circ (\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \chi_x) \quad \text{and} \quad \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon} := \mu_{x,\varepsilon} \circ (\Psi^{\mathrm{Im}}_{x,\xi,\varepsilon}, \chi_x)$$

according to Lemmas 3 and 1, which ensures the existence of a polynomial π_4 such that, for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned} &\mathcal{M}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{M}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) \leq \pi_4(\log(\varepsilon^{-1}), \log(\|x\|_\infty), \log(\|\xi\|_\infty)), \\ &\quad \mathcal{B}(\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}), \mathcal{B}(\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}) \leq \pi_4(\varepsilon^{-1}, \|x\|_\infty, \|\xi\|_\infty). \end{aligned} \tag{1.83}$$

Furthermore,

$$\begin{aligned} &\|\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)} \\ &\leq \|\mu_{x,\varepsilon} \circ (\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}, \chi_x) - \Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} \cdot \chi_x\|_{L^\infty(S_x)} \\ &\quad + \|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} \cdot \chi_x - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)}, \end{aligned} \tag{1.84}$$

where the first term is upper-bounded by $\frac{\varepsilon}{4s_x}$ due to (1.81). The second term on the right-hand side of (1.84) is upper-bounded as follows. First, note that for $t \in S_x \setminus [-(\|x\|_\infty + E), \|x\|_\infty + E]^d$, we have

$\mathrm{Re}(M_\xi T_x g)(t) = 0$ and $|\chi_x(t)| \leq 1$, which implies

$$\begin{aligned}
&|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) \cdot \chi_x(t) - \mathrm{Re}(M_\xi T_x g)(t)| \\
&\leq |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t)| \\
&\leq |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) - \mathrm{Re}(M_\xi T_x g)(t)| + |\mathrm{Re}(M_\xi T_x g)(t)| \\
&= |\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon}(t) - \mathrm{Re}(M_\xi T_x g)(t)|.
\end{aligned}$$

As $|\chi_x(t)| = 1$ for $t \in [-(\|x\|_\infty + E), \|x\|_\infty + E]^d$, together with (1.84), this yields

$$\begin{aligned}
&\|\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)} \\
&\leq \tfrac{\varepsilon}{4s_x} + \|\Psi^{\mathrm{Re}}_{x,\xi,\varepsilon} - \mathrm{Re}(M_\xi T_x g)\|_{L^\infty(S_x)}.
\end{aligned}$$

The analogous estimate for $\|\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon} - \mathrm{Im}(M_\xi T_x g)\|_{L^\infty(S_x)}$ is obtained in exactly the same manner. Together with (1.80), we can finally infer that, for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}
&\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \\
&+ \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \\
&\leq \tfrac{\varepsilon}{s_x}.
\end{aligned}$$

As $M_\xi T_x g$, $\Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}$, and $\Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}$ are supported in S_x for all $x, \xi \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, using (1.82), we get

$$\begin{aligned}
&\|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(\mathbb{R}^d)} \\
&= \|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} + \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^2(S_x)} \\
&\leq s_x \|\mathrm{Re}(M_\xi T_x g) - \Gamma^{\mathrm{Re}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \\
&+ s_x \|\mathrm{Im}(M_\xi T_x g) - \Gamma^{\mathrm{Im}}_{x,\xi,\varepsilon}\|_{L^\infty(S_x)} \\
&\leq \varepsilon.
\end{aligned} \tag{1.85}$$

Consider now, for $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, the complex-valued network $\Gamma_{i,\varepsilon} \in \mathcal{N}^{\mathbb{C}}_{d,1}$ given by

$$\Gamma_{i,\varepsilon} := (\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}, \Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon})$$

and note that, for $f \in L^2(\Omega, \mathbb{C})$,

$$\begin{aligned}\|f\|_{L^2(\Omega,\mathbb{C})} &= \left(\int_\Omega |f(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} \\ &= \left(\int_\Omega |\mathrm{Re}(f(t))|^2 + |\mathrm{Im}(f(t))|^2 \mathrm{d}t\right)^{\frac{1}{2}} \\ &= \left(\|\mathrm{Re}(f)\|_{L^2(\Omega)}^2 + \|\mathrm{Im}(f)\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}} \\ &\leq \|\mathrm{Re}(f)\|_{L^2(\Omega)} + \|\mathrm{Im}(f)\|_{L^2(\Omega)}.\end{aligned}$$

Hence, (1.85) implies that, for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}&\|\varphi_i - \Gamma_{i,\varepsilon}\|_{L^2(\mathbb{R}^d,\mathbb{C})} \\ &= \|M_{\xi(i)} T_{x(i)} g - (\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}, \Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon})\|_{L^2(\mathbb{R}^d,\mathbb{C})} \\ &\leq \varepsilon.\end{aligned}$$

Finally, using (1.74) in (1.83), it follows that there exists a polynomial π_5 such that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, we have

$$\mathcal{M}(\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}), \mathcal{M}(\Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon}) \leq \pi_5(\log(\varepsilon^{-1}), \log(i))$$

and

$$\mathcal{B}(\Gamma^{\mathrm{Re}}_{x(i),\xi(i),\varepsilon}), \mathcal{B}(\Gamma^{\mathrm{Im}}_{x(i),\xi(i),\varepsilon}) \leq \pi_5(\varepsilon^{-1}, i),$$

which finalizes the proof. □

Next, we establish the central result of this section. To this end, we first recall that according to Theorem 7 neural networks provide optimal approximations for all function classes that are optimally approximated by affine dictionaries (generated by functions f that can be approximated well by neural networks). While this universality property is significant as it applies to all affine dictionaries, it is perhaps not completely surprising as affine dictionaries are generated by affine transformations and neural networks consist of concatenations of affine transformations and nonlinearities. Gabor dictionaries, on the other

hand, exhibit a fundamentally different mathematical structure. The next result shows that neural networks also provide optimal approximations for all function classes that are optimally approximated by Gabor dictionaries (again, with generator functions that can be approximated well by neural networks).

Theorem 11. *Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be the corresponding Gabor dictionary with ordering as defined in* (1.66)*. Assume that Ω is bounded or that $\Omega = \mathbb{R}^d$ and g is compactly supported. Further, suppose that there exists a polynomial π such that for every $x \in \mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{x,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying*

$$\|g - \Phi_{x,\varepsilon}\|_{L^\infty(x+\Omega)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{x,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\|x\|_\infty))$, $\mathcal{B}(\Phi_{x,\varepsilon}) \leq \pi(\varepsilon^{-1}, \|x\|_\infty)$. Then, for all compact function classes $\mathcal{C} \subseteq L^2(\Omega)$, we have

$$\gamma^{*,\mathit{eff}}_{\mathcal{N}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{G}(g, \alpha, \beta, \Omega)).$$

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{G}(g, \alpha, \beta, \Omega)$ (in the sense of Definition 7), then $\mathcal{C}$ is optimally representable by neural networks (in the sense of Definition 11).

Proof. The first statement follows from Theorem 5 and Theorem 10, the second is by Theorem 4. □

We complete the program in this section by showing that the Gaussian function satisfies the conditions on the generator g in Theorem 10 for bounded Ω. Gaussian functions are widely used generator functions for Gabor dictionaries owing to their excellent time-frequency localization and their frame-theoretic optimality properties (Gröchenig, 2013). We hasten to add that the result below can be extended to any generator function g of sufficiently fast decay and sufficient smoothness.

Lemma 12. *For $d \in \mathbb{N}$, let $g_d \in L^2(\mathbb{R}^d)$ be given by*

$$g_d(x) := e^{-\|x\|_2^2}.$$

There exists a constant $C > 0$ such that, for all $d \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{d,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\|\Phi_{d,\varepsilon} - g\|_{L^\infty(\mathbb{R}^d)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{d,\varepsilon}) \leq Cd(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$, $\mathcal{B}(\Phi_{d,\varepsilon}) \leq 1$.

Proof. Observe that g_d can be written as the composition $h \circ f_d$ of the functions $f_d \colon \mathbb{R}^d \to \mathbb{R}_+$ and $h \colon \mathbb{R}_+ \to \mathbb{R}$ given by

$$f_d(x) := \|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \quad \text{and} \quad h(y) := e^{-y}.$$

By Proposition 2 and Lemma 4, there exists a constant $C_1 > 0$ such that, for every $d \in \mathbb{N}$, $D \in [1, \infty)$, $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{d,D,\varepsilon} \in \mathcal{N}_{d,1}$ satisfying

$$\sup_{x \in [-D,D]^d} |\Psi_{d,D,\varepsilon}(x) - \|x\|_2^2| \leq \tfrac{\varepsilon}{2}, \tag{1.86}$$

$$\mathcal{M}(\Psi_{d,D,\varepsilon}) \leq C_1 d(\log(\varepsilon^{-1}) + \log(\lceil D \rceil)), \quad \mathcal{B}(\Psi_{d,D,\varepsilon}) \leq 1. \tag{1.87}$$

Moreover, as $|\frac{d^n}{dy^n} e^{-y}| = |e^{-y}| \leq 1$ for all $n \in \mathbb{N}$, $y \geq 0$, Lemma 17 implies the existence of a constant $C_2 > 0$ such that for every $d \in \mathbb{N}$, $D \in [1, \infty)$, $\varepsilon \in (0, 1/2)$, there is a network $\Gamma_{d,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\sup_{y \in [0, dD^2]} |\Gamma_{d,D,\varepsilon}(y) - e^{-y}| \leq \tfrac{\varepsilon}{2}, \tag{1.88}$$

$$\begin{aligned} &\mathcal{M}(\Gamma_{d,D,\varepsilon}) \leq C_2 dD^2((\log(\varepsilon^{-1}))^2 + \log(d) + \log(\lceil D \rceil)), \\ &\mathcal{B}(\Gamma_{D,\varepsilon}) \leq 1. \end{aligned} \tag{1.89}$$

Now, let $D_\varepsilon := \log(\varepsilon^{-1})$ and take $\widetilde{\Phi}_{d,\varepsilon} := \Gamma_{d,D_\varepsilon,\varepsilon} \circ \Psi_{d,D_\varepsilon,\varepsilon}$ according to Lemma 1. Consequently, it follows from (1.87) and (1.89) that there exists a constant $C_2 > 0$ such that for all $d \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, we have $\mathcal{M}(\widetilde{\Phi}_{d,\varepsilon}) \leq C_2 d(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$ and $\mathcal{B}(\widetilde{\Phi}_{d,\varepsilon}) \leq 1$. Moreover, as $|e^{-y}| \leq 1$ for all $y \geq 0$, combining (1.86) and (1.88) yields for all $\varepsilon \in (0, 1/2)$, $x \in [-D_\varepsilon, D_\varepsilon]^d$,

$$\begin{aligned}
|g(x) - \widetilde{\Phi}_{d,\varepsilon}(x)| &= |e^{-\|x\|_2^2} - \Gamma_{d,D_\varepsilon,\varepsilon}(\Psi_{d,D_\varepsilon,\varepsilon}(x))| \\
&\leq |e^{-\|x\|_2^2} - e^{-\Psi_{d,D_\varepsilon,\varepsilon}(x)}| \\
&\quad + |e^{-\Psi_{d,D_\varepsilon,\varepsilon}(x)} - \Gamma_{d,D_\varepsilon,\varepsilon}(\Psi_{d,D_\varepsilon,\varepsilon}(x))| \\
&\leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

We can now use the same approach as in the proof of Theorem 10 to construct networks $\Phi_{d,\varepsilon}$ supported on the interval $[-D_\varepsilon, D_\varepsilon]^d$ over which they approximate g to within error ε, and obey $\mathcal{M}(\Phi_\varepsilon) \leq Cd(\log(\varepsilon^{-1}))^2((\log(\varepsilon^{-1}))^2 + \log(d))$, $\mathcal{B}(\Phi_{d,\varepsilon}) \leq 1$ for some absolute constant C. Together with $|g(x)| \leq \varepsilon$, for all $x \in \mathbb{R}^d \backslash [-D_\varepsilon, D_\varepsilon]^d$, this completes the proof. □

Remark 9. *Note that Lemma 12 establishes an approximation result that is even stronger than what is required by Theorem 10. Specifically, we achieve ε-approximation over all of $\mathbb{R}^d$ with a network that does not depend on the shift parameter x, while exhibiting the desired growth rates on $\mathcal{M}$ and $\mathcal{B}$, which consequently do not depend on the shift parameter as well. The idea underlying this construction can be used to strengthen Theorem 10 to apply to $\Omega = \mathbb{R}^d$ and generator functions of unbounded support, but sufficiently rapid decay.*

We conclude this section with a remark on the neural network approximation of the real-valued counterpart of Gabor dictionaries known as Wilson dictionaries (Gröchenig and Samarah, 2000; Gröchenig, 2013) and consisting of cosine-modulated and time-shifted versions of a given generator function, see also Appendix C. The techniques developed in this section, mutatis mutandis, show that neural networks provide

Kolmogorov-Donoho optimal approximation for all function classes that are optimally approximated by Wilson dictionaries (generated by functions that can be approximated well by neural networks). Specifically, we point out that the proofs of Lemma 11 and Theorem 10 explicitly construct neural network approximations of time-shifted and cosine- and sine-modulated versions of the generator g. As identified in Table 1, Wilson bases provide optimal nonlinear approximation of (unit) balls in modulation spaces (Feichtinger, 1981; Gröchenig and Samarah, 2000). Finally, we note that similarly the techniques developed in the proofs of Lemma 11 and Theorem 10 can be used to establish optimal representability of Fourier bases.

1.10. IMPROVING POLYNOMIAL APPROXIMATION RATES TO EXPONENTIAL RATES

Having established that for all function classes listed in Table 1, Kolmogorov-Donoho-optimal approximation through neural networks is possible, this section proceeds to show that neural networks, in addition to their striking Kolmogorov-Donoho universality property, can also do something that has no classical equivalent.

Specifically, as mentioned in the introduction, for the class of oscillatory textures as considered below and for the Weierstrass function, there are no known methods that achieve exponential accuracy, i.e., an approximation error that decays exponentially in the number of parameters employed in the approximant. We establish below that deep networks fill this gap.

Let us start by defining one-dimensional "oscillatory textures" according to (Demanet and Ying, 2007). To this end, we recall the following definition from Lemma 17,

$$\mathcal{S}_{[a,b]} = \Big\{ f \in C^\infty([a,b],\mathbb{R})\colon \|f^{(n)}(x)\|_{L^\infty([a,b])} \leq n!,$$

for all $n \in \mathbb{N}_0 \Big\}$.

Definition 17. *Let the sets* $\mathcal{F}_{D,a}$, $D, a \in \mathbb{R}_+$, *be given by*

$$\mathcal{F}_{D,a} = \left\{ \cos(ag)h \colon g, h \in \mathcal{S}_{[-D,D]} \right\}.$$

The efficient approximation of functions in $\mathcal{F}_{D,a}$ with a large represents a notoriously difficult problem due to the combination of the rapidly oscillating cosine term and the warping function g. The best approximation results available in the literature (Demanet and Ying, 2007) are based on wave-atom dictionaries[11] and yield low-order polynomial approximation rates. In what follows we show that finite-width deep networks drastically improve these results to exponential approximation rates.

We start with our statement on the neural network approximation of oscillatory textures.

Proposition 6. *There exists a constant* $C > 0$ *such that for all* $D, a \in \mathbb{R}_+, f \in \mathcal{F}_{D,a}$, *and* $\varepsilon \in (0, 1/2)$, *there is a network* $\Gamma_{f,\varepsilon} \in \mathcal{N}_{1,1}$ *satisfying*

$$\|f - \Gamma_{f,\varepsilon}\|_{L^\infty([-D,D])} \leq \varepsilon,$$

with $\mathcal{L}(\Gamma_{f,\varepsilon}) \leq C\lceil D \rceil((\log(\varepsilon^{-1}) + \log(\lceil a \rceil))^2 + \log(\lceil D \rceil) + \log(\lceil D^{-1} \rceil))$, $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 32$, $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq 1$.

Proof. For $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, let $g_f, h_f \in \mathcal{S}_{[-D,D]}$ be functions such that $f = \cos(ag_f)h_f$. Note that Lemma 17 guarantees the existence of a constant $C_1 > 0$ such that for all $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there are networks $\Psi_{g_f,\varepsilon}, \Psi_{h_f,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\begin{aligned} \|\Psi_{g_f,\varepsilon} - g_f\|_{L^\infty([-D,D])} &\leq \tfrac{\varepsilon}{12\lceil a \rceil}, \\ \|\Psi_{h_f,\varepsilon} - h_f\|_{L^\infty([-D,D])} &\leq \tfrac{\varepsilon}{12\lceil a \rceil} \end{aligned} \tag{1.90}$$

[11] To be precise, the results of (Demanet and Ying, 2007) are concerned with the two-dimensional case, whereas here we focus on the one-dimensional case. Note, however, that all our results are readily extended to the multi-dimensional case.

with $\mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon}) \leq C_1\lceil D\rceil(\log((\frac{\varepsilon}{12\lceil a\rceil})^{-1})^2 + \log(\lceil D\rceil) + \log(\lceil D^{-1}\rceil))$, $\mathcal{W}(\Psi_{g_f,\varepsilon}), \mathcal{W}(\Psi_{h_f,\varepsilon}) \leq 16$, and $\mathcal{B}(\Psi_{g_f,\varepsilon}), \mathcal{B}(\Psi_{h_f,\varepsilon}) \leq 1$. Furthermore, Theorem 2 ensures the existence of a constant $C_2 > 0$ such that for all $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Phi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-3/2,3/2])} \leq \tfrac{\varepsilon}{3}, \tag{1.91}$$

with $\mathcal{L}(\Phi_{a,D,\varepsilon}) \leq C_2((\log(\varepsilon^{-1}))^2 + \log(\lceil 3a/2\rceil))$, $\mathcal{W}(\Phi_{a,D,\varepsilon}) \leq 9$, and $\mathcal{B}(\Phi_{a,D,\varepsilon}) \leq 1$. Moreover, due to Proposition 2, there exists a constant $C_3 > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\mu_\varepsilon \in \mathcal{N}_{2,1}$ satisfying

$$\sup_{x,y\in[-3/2,3/2]} |\mu_\varepsilon(x, y) - xy| \leq \tfrac{\varepsilon}{3}, \tag{1.92}$$

with $\mathcal{L}(\mu_\varepsilon) \leq C_3 \log(\varepsilon^{-1})$, $\mathcal{W}(\mu_\varepsilon) \leq 5$, and $\mathcal{B}(\mu_\varepsilon) \leq 1$. By Lemma 1 there exists a network Ψ^1 satisfying $\Psi^1 = \Phi_{a,D,\varepsilon} \circ \Psi_{g_f,\varepsilon}$ with $\mathcal{W}(\Psi^1) \leq 16$, $\mathcal{L}(\Psi^1) = \mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon})$, and $\mathcal{B}(\Psi^1) \leq 1$. Furthermore, combining Lemma 2 and Lemma 18, we can conclude the existence of a network

$$\Psi^2(x) = (\Psi^1(x), \Psi_{h_f,\varepsilon}(x)) = (\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x))$$

with $\mathcal{W}(\Psi^2) \leq 32$, $\mathcal{L}(\Psi^2) = \max\{\mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon})\}$, and $\mathcal{B}(\Psi^2) \leq 1$. Next, for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, we define the network $\Gamma_{f,\varepsilon} := \mu_\varepsilon \circ \Psi^2$. By (1.90), (1.91), and

$$\sup_{x\in\mathbb{R}} |\tfrac{d}{dx}\cos(ax)| = a,$$

we have, for all $x \in [-D, D]$,

$$\begin{aligned}
&|\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)) - \cos(ag_f(x))| \\
&\leq |\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)) - \cos(a\Psi_{g_f,\varepsilon}(x))| \\
&\quad + |\cos(a\Psi_{g_f,\varepsilon}(x)) - \cos(ag_f(x))| \\
&\leq \tfrac{\varepsilon}{3} + a\tfrac{\varepsilon}{12\lceil a\rceil} \leq \tfrac{5\varepsilon}{12}.
\end{aligned}$$

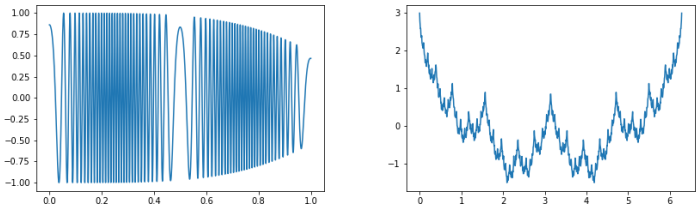


Fig. 1.4: Left: A function in $\mathcal{F}_{1,100}$. Right: The function $W_{\frac{1}{\sqrt{2}},2}$.

Combining this with (1.90), (1.92), and $\|\cos\|_{L^\infty([-D,D])}$, $\|f\|_{L^\infty([-D,D])} \leq 1$ yields for all $x \in [-D, D]$,

$$
\begin{aligned}
&|\Gamma_{f,\varepsilon}(x) - f(x)| \\
&= |\mu_\varepsilon(\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x)) - \cos(ag_f(x))h_f(x)| \\
&\leq |\mu_\varepsilon(\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x)), \Psi_{h_f,\varepsilon}(x)) - \Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x))\Psi_{h_f,\varepsilon}(x)| \\
&\quad + |\Phi_{a,D,\varepsilon}(\Psi_{g_f,\varepsilon}(x))\Psi_{h_f,\varepsilon}(x) - \cos(ag_f(x))\Psi_{h_f,\varepsilon}(x)| \\
&\quad + |\cos(ag_f(x))\Psi_{h_f,\varepsilon}(x) - \cos(ag_f(x))h_f(x)| \\
&\leq \tfrac{\varepsilon}{3} + \tfrac{5\varepsilon}{12}\left(1 + \tfrac{\varepsilon}{12\lceil a\rceil}\right) + \tfrac{\varepsilon}{12\lceil a\rceil} \leq \varepsilon.
\end{aligned}
$$

Finally, by Lemma 1 there exists a constant C_4 such that for all $D, a \in \mathbb{R}_+$, $f \in \mathcal{F}_{D,a}$, $\varepsilon \in (0, 1/2)$, it holds that $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 32$,

$$
\begin{aligned}
&\mathcal{L}(\Gamma_{f,\varepsilon}) \\
&\leq \mathcal{L}(\mu_\varepsilon) + \max\{\mathcal{L}(\Phi_{a,D,\varepsilon}) + \mathcal{L}(\Psi_{g_f,\varepsilon}), \mathcal{L}(\Psi_{h_f,\varepsilon})\} \\
&\leq C_4\lceil D\rceil((\log(\varepsilon^{-1}) + \log(\lceil a\rceil))^2 + \log(\lceil D\rceil) + \log(\lceil D^{-1}\rceil)),
\end{aligned}
$$

and $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq 1$. □

Finally, we show how the Weierstrass function—a fractal function, which is continuous everywhere but differentiable nowhere—can be approximated with exponential accuracy by deep ReLU networks. Specif-

ically, we consider

$$W_{p,a}(x) = \sum_{k=0}^{\infty} p^k \cos(a^k \pi x), \text{ for } p \in (0, 1/2),\ a \in \mathbb{R}_+, \text{ with } ap \geq 1,$$

and let $\alpha = -\frac{\log(p)}{\log(a)}$, see Figure 1.4 right for an example. It is well known (Zygmund, 2002) that $W_{p,a}$ possesses Hölder smoothness α which may be made arbitrarily small by suitable choice of a. While classical approximation methods achieve polynomial approximation rates only, it turns out that finite-width deep networks yield exponential approximation rates. This is formalized as follows.

Proposition 7. *There exists a constant $C > 0$ such that for all $\varepsilon, p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, there is a network $\Psi_{p,a,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying*

$$\|\Psi_{p,a,D,\varepsilon} - W_{p,a}\|_{L^\infty([-D,D])} \leq \varepsilon,$$

with $\mathcal{L}(\Psi_{p,a,D,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^3 + (\log(\varepsilon^{-1}))^2 \log(\lceil a \rceil) + \log(\varepsilon^{-1}) \log(\lceil D \rceil))$, $\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 13$, $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq 1$.

Proof. For every $N \in \mathbb{N}$, $p \in (0, 1/2)$, $a \in \mathbb{R}_+$, $x \in \mathbb{R}$, let $S_{N,p,a}(x) = \sum_{k=0}^{N} p^k \cos(a^k \pi x)$ and note that

$$\begin{aligned} |S_{N,p,a}(x) - W_{p,a}(x)| &\leq \sum_{k=N+1}^{\infty} |p^k \cos(a^k \pi x)| \\ &\leq \sum_{k=N+1}^{\infty} p^k = \tfrac{1}{1-p} - \tfrac{1-p^{N+1}}{1-p} \\ &\leq 2^{-N}. \end{aligned} \tag{1.93}$$

Let $N_\varepsilon := \lceil \log(2/\varepsilon) \rceil$ for $\varepsilon \in (0, 1/2)$. Next, note that Theorem 2 ensures the existence of a constant $C_1 > 0$ such that for all $D, a \in \mathbb{R}_+$, $k \in \mathbb{N}_0$, $\varepsilon \in (0, 1/2)$, there is a network $\phi_{a^k,D,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\phi_{a^k,D,\varepsilon} - \cos(a^k \pi \cdot)\|_{L^\infty([-D,D])} \leq \tfrac{\varepsilon}{4}, \tag{1.94}$$

with $\mathcal{L}(\phi_{a^k,D,\varepsilon}) \leq C_1((\log(\varepsilon^{-1}))^2 + \log(\lceil a^k \pi D \rceil))$, $\mathcal{W}(\phi_{a^k,D,\varepsilon}) \leq 9$, $\mathcal{B}(\phi_{a^k,D,\varepsilon}) \leq 1$. Let $A\colon \mathbb{R}^3 \to \mathbb{R}^3$ and $B\colon \mathbb{R}^3 \to \mathbb{R}$ be the affine transformations given by $A(x_1, x_2, x_3) = (x_1, x_1, x_2 + x_3)^T$ and $B(x_1, x_2, x_3) = x_2 + x_3$, respectively. We now define, for all $p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, $k \in \mathbb{N}_0$, $\varepsilon \in (0, 1/2)$, the networks

$$\psi_{D,\varepsilon}^{p,a,0}(x) = \begin{pmatrix} x \\ p^0 \phi_{a^0,D,\varepsilon}(x) \\ 0 \end{pmatrix}$$

and

$$\psi_{D,\varepsilon}^{p,a,k}(x_1, x_2, x_3) = \begin{pmatrix} x_1 \\ p^k \phi_{a^k,D,\varepsilon}(x_2) \\ x_3 \end{pmatrix},\ k > 0,$$

and, for all $p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, the network

$$\Psi_{p,a,D,\varepsilon} := B \circ \psi_{D,\varepsilon}^{p,a,N_\varepsilon} \circ A \circ \psi_{D,\varepsilon}^{p,a,N_\varepsilon - 1} \circ \cdots \circ A \circ \psi_{D,\varepsilon}^{p,a,0}.$$

Due to (1.94) we get, for all $p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, $x \in [-D, D]$, that

$$\begin{aligned} &|\Psi_{p,a,D,\varepsilon}(x) - S_{N_\varepsilon,p,a}(x)| \\ &= \left| \sum_{k=0}^{N_\varepsilon} p^k \phi_{a^k,D,\varepsilon}(x) - \sum_{k=0}^{N_\varepsilon} p^k \cos(a^k \pi x) \right| \\ &\leq \sum_{k=0}^{N_\varepsilon} p^k |\phi_{a^k,D,\varepsilon}(x) - \cos(a^k \pi x)| \\ &\leq \tfrac{\varepsilon}{4} \sum_{k=0}^{N_\varepsilon} 2^{-k} \leq \tfrac{\varepsilon}{2}. \end{aligned}$$

Combining this with (1.93) establishes, for all $p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, $x \in [-D, D]$,

$$|\Psi_{p,a,D,\varepsilon}(x) - W_{p,a}(x)| \leq 2^{-\lceil \log(\frac{2}{\varepsilon}) \rceil} + \tfrac{\varepsilon}{2} \leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

Applying Lemmas 1, 2, and 3 establishes the existence of a constant C_2 such that for all $p \in (0, 1/2)$, $D, a \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}
&\mathcal{L}(\Psi_{p,a,D,\varepsilon}) \\
&\leq \sum_{k=0}^{N_\varepsilon} (\mathcal{L}(\phi_{a^k,D,\varepsilon}) + 1) \\
&\leq N_\varepsilon + 1 + (N_\varepsilon + 1) C_1((\log(\varepsilon^{-1}))^2 + \log(\lceil a^{N_\varepsilon} \pi D \rceil)) \\
&\leq C_2((\log(\varepsilon^{-1}))^3 + (\log(\varepsilon^{-1}))^2 \log(\lceil a \rceil) + \log(\varepsilon^{-1}) \log(\lceil D \rceil)),
\end{aligned}$$

$\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 13$, and $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq 1$. □

We finally note that the restriction $p \in (0, 1/2)$ in Proposition 7 was made for simplicity of exposition and can be relaxed to $p \in (0, r)$, with $r < 1$, while only changing the constant C.

1.11. IMPOSSIBILITY RESULTS FOR FINITE-DEPTH NETWORKS

The recent successes of neural networks in machine learning applications have been enabled by various technological factors, but they all have in common the use of deep networks as opposed to shallow networks studied intensely in the 1990s. It is hence of interest to understand whether the use of depth offers fundamental advantages. In this spirit, the goal of this section is to make a formal case for depth in neural network approximation by establishing that, for nonconstant periodic functions, finite-width deep networks require asymptotically—in the function's "highest frequency"—smaller connectivity than finite-depth wide networks. This statement is then extended to sufficiently smooth nonperiodic functions, thereby formalizing the benefit of deep networks over shallow networks for the approximation of a broad class of functions.

We start with preparatory material taken from (Telgarsky, 2015).

Definition 18 ((Telgarsky, 2015))**.** *Let $k \in \mathbb{N}$. A function $f : \mathbb{R} \to \mathbb{R}$ is called k-sawtooth if it is piecewise linear with no more than k pieces, i.e., its domain $\mathbb{R}$ can be partitioned into k intervals such that f is linear on each of these intervals.*

Lemma 13 ((Telgarsky, 2015))**.** *Every $\Phi \in \mathcal{N}_{1,1}$ is $(2\mathcal{W}(\Phi))^{\mathcal{L}(\Phi)}$-sawtooth.*

Definition 19. *For a u-periodic function $f \in C(\mathbb{R})$, we define*

$$\xi(f) := \sup_{\delta \in [0,u)} \inf_{c,d \in \mathbb{R}} \|f(x) - (cx+d)\|_{L^\infty([\delta, \delta+u])}.$$

The quantity $\xi(f)$ measures the error incurred by the best linear approximation of f on any segment of length equal to the period of f; $\xi(f)$ can hence be interpreted as quantifying the nonlinearity of f. The next result states that finite-depth networks with width and hence also connectivity scaling polylogarithmically in the "highest frequency" of the periodic function to be approximated can not achieve arbitrarily small approximation error.

Proposition 8. *Let $f \in C(\mathbb{R})$ be a nonconstant u-periodic function, $L \in \mathbb{N}$, and π a polynomial. Then, there exists an $a \in \mathbb{N}$ such that for every network $\Phi \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{W}(\Phi) \leq \pi(\log(a))$, we have*

$$\|f(a\,\cdot) - \Phi\|_{L^\infty([0,u])} \geq \xi(f) > 0.$$

Proof. First note that there exists an even $a \in \mathbb{N}$ such that $a/2 > (2\pi(\log(a)))^L$. Lemma 13 now implies that every network $\Phi \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{W}(\Phi) \leq \pi(\log(a))$ is $(2\pi(\log(a)))^L$-sawtooth and therefore consists of no more than $a/2$ different linear pieces. Hence, there exists an interval $[u_1, u_2] \subseteq [0, u]$ with $u_2 - u_1 \geq (2u/a)$ on which Φ is linear. Since $u_2 - u_1 \geq (2u/a)$ the interval supports

two full periods of $f(a\,\cdot\,)$ and we can therefore conclude that

$$\begin{aligned}\|f(a\,\cdot\,) - \Phi\|_{L^\infty([0,u])} &\geq \|f(a\,\cdot\,) - \Phi\|_{L^\infty([u_1,u_2])} \\ &\geq \inf_{c,d\in\mathbb{R}} \|f(x) - (cx+d)\|_{L^\infty([0,2u])} \\ &\geq \sup_{\delta\in[0,u)} \inf_{c,d\in\mathbb{R}} \|f(x) - (cx+d)\|_{L^\infty([\delta,u+\delta])} \\ &= \xi(f).\end{aligned}$$

Finally, note that $\xi(f) > 0$ as $\xi(f) = 0$ for u-periodic $f \in C(\mathbb{R})$ necessarily implies that f is constant, which, however, is ruled out by assumption. □

Application of Proposition 8 to $f(x) = \cos(x)$ shows that finite-depth networks, owing to $\xi(\cos) > 0$, require faster than polylogarithmic growth of connectivity in a to approximate $x \mapsto \cos(ax)$ with arbitrarily small error, whereas finite-width networks, due to Theorem 2, can accomplish this with polylogarithmic connectivity growth.

The following result from (Frenzen et al., 2010) allows a similar observation for functions that are sufficiently smooth.

Theorem 12 ((Frenzen et al., 2010))**.** *Let $[a,b] \subseteq \mathbb{R}$, $f \in C^3([a,b])$, and for $\varepsilon \in (0,1/2)$, let $s(\varepsilon) \in \mathbb{N}$ denote the smallest number such that there exists a piecewise linear approximation of f with $s(\varepsilon)$ pieces and error at most ε in $L^\infty([a,b])$-norm. Then, it holds that*

$$s(\varepsilon) \sim \frac{c}{\sqrt{\varepsilon}},\ \varepsilon \to 0, \ \text{ where } \ c = \frac{1}{4}\int_a^b \sqrt{|f''(x)|}dx.$$

Combining this with Lemma 13 yields the following result on depth-width tradeoff for three-times continuously differentiable functions.

Theorem 13. *Let $f \in C^3([a,b])$ with $\int_a^b \sqrt{|f''(x)|}dx > 0$, $L \in \mathbb{N}$, and π a polynomial. Then, there exists $\varepsilon > 0$ such that for every network $\Phi \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{W}(\Phi) \leq \pi(\log(\varepsilon^{-1}))$, we have*

$$\|f - \Phi\|_{L^\infty([a,b])} > \varepsilon.$$

Proof. The proof will be effected by contradiction. Assume that for every $\varepsilon > 0$, there exists a network $\Phi_\varepsilon \in \mathcal{N}_{1,1}$ with $\mathcal{L}(\Phi_\varepsilon) \leq L$, $\mathcal{W}(\Phi_\varepsilon) \leq \pi(\log(\varepsilon^{-1}))$, and $\|f - \Phi_\varepsilon\|_{L^\infty([a,b])} \leq \varepsilon$. By Lemma 13 every (ReLU) neural network realizes a piecewise linear function. Application of Theorem 12 hence allows us to conclude the existence of a constant C such that, for all $\varepsilon > 0$, the network Φ_ε must have at least $C\varepsilon^{-\frac{1}{2}}$ different linear pieces. This, however, leads to a contradiction as, by Lemma 13, Φ_ε is at most $(2\pi(\log(\varepsilon^{-1})))^L$-sawtooth and $\widetilde{\pi}(\log(\varepsilon^{-1})) \in o(\varepsilon^{-1/2})$, $\varepsilon \to 0$, for every polynomial $\widetilde{\pi}$. □

In summary, we have hence established that any function which is at least three times continuously differentiable (and does not have a vanishing second derivative) cannot be approximated by finite-depth networks with connectivity scaling polylogarithmically in the inverse of the approximation error. Our results in Section 1.3 establish that, in contrast, this "is" possible with finite-width deep networks for various interesting types of smooth functions such as polynomials and sinusoidal functions. Further results on the limitations of finite-depth networks akin to Theorem 13 were reported in (Petersen and Voigtlaender, 2018).

1.12. APPENDICES

A. Auxiliary neural network constructions

The following three results are concerned with the realization of affine transformations of arbitrary weights by neural networks with weights upper-bounded by 1.

Lemma 14. *Let $d \in \mathbb{N}$ and $a \in \mathbb{R}$. There exists a network $\Phi_a \in \mathcal{N}_{d,d}$ satisfying $\Phi_a(x) = ax$, with $\mathcal{L}(\Phi_a) \leq \lfloor \log(|a|) \rfloor + 4$, $\mathcal{W}(\Phi_a) \leq 3d$, $\mathcal{B}(\Phi_a) \leq 1$.*

Proof. First note that for $|a| \leq 1$ the claim holds trivially, which can be seen by taking Φ_a to be the affine transformation $x \mapsto ax$ and

interpreting it according to Definition 1 as a depth-1 neural network. Next, we consider the case $|a| > 1$ for $d = 1$, set $K := \lfloor \log(a) \rfloor$, $\alpha := a2^{-(K+1)}$, and define $A_1 := (1, -1)^T \in \mathbb{R}^{2\times 1}$,

$$A_2 := \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2},$$

$$A_k := \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}, \quad k \in \{3, \dots, K+3\},$$

and $A_{K+4} := (\alpha, 0, -\alpha)$. Note that $(\rho \circ A_2 \circ \rho \circ A_1)(x) = (\rho(x), \rho(x) + \rho(-x), \rho(-x))$ and $\rho(A_k(x, x+y, y)^T) = 2(x, x+y, y)$, for $k \in \{3, \dots, K+3\}$. The network $\Psi_a := A_{K+4} \circ \rho \circ \cdots \circ \rho \circ A_1$ hence satisfies $\Psi_a(x) = ax$, $\mathcal{L}(\Psi_a) = \lfloor \log(a) \rfloor + 4$, $\mathcal{W}(\Psi_a) = 3$, and $\mathcal{B}(\Phi_a) \leq 1$. Applying Lemma 3 to get a parallelization of d copies of Ψ_a completes the proof. □

Corollary 2. *Let $d, d' \in \mathbb{N}$, $a \in \mathbb{R}_+$, $A \in [-a, a]^{d'\times d}$, and $b \in [-a, a]^{d'}$. There exists a network $\Phi_{A,b} \in \mathcal{N}_{d,d'}$ satisfying $\Phi_{A,b}(x) = Ax + b$, with $\mathcal{L}(\Phi_{A,b}) \leq \lfloor \log(|a|) \rfloor + 5$, $\mathcal{W}(\Phi_{A,b}) \leq \max\{d, 3d'\}$, $\mathcal{B}(\Phi_{A,b}) \leq 1$.*

Proof. Let $\Phi_a \in \mathcal{N}_{d',d'}$ be the multiplication network from Lemma 14, consider $W(x) := a^{-1}(Ax + b)$ as a 1-layer network, and take $\Phi_{A,b} := \Phi_a \circ W$ according to Lemma 1. □

Proposition 9. *Let $d, d' \in \mathbb{N}$ and $\Phi \in \mathcal{N}_{d,d'}$. There exists a network $\Psi \in \mathcal{N}_{d,d'}$ satisfying $\Psi(x) = \Phi(x)$, for all $x \in \mathbb{R}^d$, and with $\mathcal{L}(\Psi) \leq (\lceil \log(\mathcal{B}(\Phi)) \rceil + 5)\mathcal{L}(\Phi)$, $\mathcal{W}(\Psi) \leq \max\{3d', \mathcal{W}(\Phi)\}$, $\mathcal{B}(\Psi) \leq 1$.*

Proof. We write $\Phi = W_{\mathcal{L}(\Phi)} \circ \rho \circ \ldots \circ \rho \circ W_1$ and set $\widetilde{W}_\ell := (\mathcal{B}(\Phi))^{-1} W_\ell$, for $\ell \in \{1, \dots, \mathcal{L}(\Phi)\}$, and $a := \mathcal{B}(\Phi)^{\mathcal{L}(\Phi)}$. Let $\Phi_a \in \mathcal{N}_{d',d'}$ be the multiplication network from Lemma 14 and define

$$\widetilde{\Phi} := \widetilde{W}_{\mathcal{L}(\Phi)} \circ \rho \circ \cdots \circ \rho \circ \widetilde{W}_1,$$

and $\Psi := \Phi_a \circ \widetilde{\Phi}$ according to Lemma 1. Note that $\widetilde{\Phi}$ has weights upper-bounded by 1 and is of the same depth and width as Φ. As ρ is positively homogeneous, i.e., $\rho(\lambda x) = \lambda \rho(x)$, for all $\lambda \geq 0$, $x \in \mathbb{R}$, we have $\Psi(x) = \Phi(x)$, for all $x \in \mathbb{R}^d$. Application of Lemma 1 and Lemma 14 completes the proof. □

Next we record a technical Lemma on how to realize a sum of networks with the same input by a network whose width is independent of the number of constituent networks.

Lemma 15. *Let $d, d' \in \mathbb{N}$, $N \in \mathbb{N}$, and $\Phi_i \in \mathcal{N}_{d,d'}$, $i \in \{1, \ldots, N\}$. There exists a network $\Phi \in \mathcal{N}_{d,d'}$ satisfying*

$$\Phi(x) = \sum_{i=1}^{N} \Phi_i(x), \quad \textit{for all } x \in \mathbb{R}^d,$$

with $\mathcal{L}(\Phi) = \sum_{i=1}^{N} \mathcal{L}(\Phi_i)$, $\mathcal{W}(\Phi) \leq 2d + 2d' + \max\{2d, \max_i\{\mathcal{W}(\Phi_i)\}\}$, $\mathcal{B}(\Phi) = \max\{1, \max_i \mathcal{B}(\Phi_i)\}$.

Proof. We set $L_i = \mathcal{L}(\Phi_i)$ and write the networks Φ_i as

$$\Phi_i = W^i_{L_i} \circ \rho \circ W^i_{L_i - 1} \circ \rho \circ \cdots \circ \rho \circ W^i_1,$$

with $W^i_\ell(x) = A^i_\ell x + b^i_\ell$, where $A^i_\ell \in \mathbb{R}^{N^i_\ell \times N^i_{\ell-1}}$ and $b^i_\ell \in \mathbb{R}^{N^i_\ell}$. Next, using Lemma 2, we turn the identity matrices $\mathbb{I}_d$ and $\mathbb{I}_{d'}$ into networks $\mathbb{I}^i_d$ and $\mathbb{I}^i_{d'}$, respectively, of depth L_i and then parallelize these networks, according to Lemma 3, to get $\Psi_i := (\mathbb{I}^i_d, \mathbb{I}^i_{d'}, \Phi_i)$. Let $V^i_1(x) = E^i_1 x + f^i_1$ and $V^i_{L_i}(x) = E^i_{L_i} x + f^i_{L_i}$ denote the first and last, respectively, affine transformation of the network Ψ_i. By construction

we have

$$E_1^i = \begin{pmatrix} \mathbb{I}_d & 0 & 0 \\ -\mathbb{I}_d & 0 & 0 \\ 0 & \mathbb{I}_{d'} & 0 \\ 0 & -\mathbb{I}_{d'} & 0 \\ 0 & 0 & A_1^i \end{pmatrix} \in \mathbb{R}^{(2d+2d'+N_1^i)\times(2d+d')},$$

$$f_1^i = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ b_1^i \end{pmatrix} \in \mathbb{R}^{2d+2d'+N_1^i}$$

and

$$E_{L_i}^i = \begin{pmatrix} \mathbb{I}_d & -\mathbb{I}_d & 0 & 0 & 0 \\ 0 & 0 & \mathbb{I}_{d'} & -\mathbb{I}_{d'} & 0 \\ 0 & 0 & 0 & 0 & A_{L_i}^i \end{pmatrix} \in \mathbb{R}^{(d+2d')\times(2d+2d'+N_{L_i-1}^i)},$$

$$f_{L_i}^i = \begin{pmatrix} 0 \\ 0 \\ b_{L_i}^i \end{pmatrix} \in \mathbb{R}^{d+2d'}.$$

Next, we define the matrices

$$A_{\text{in}} := \begin{pmatrix} \mathbb{I}_d \\ 0 \\ \mathbb{I}_d \end{pmatrix} \in \mathbb{R}^{(2d+d')\times d},$$

$$A := \begin{pmatrix} \mathbb{I}_d & 0 & 0 \\ 0 & \mathbb{I}_{d'} & \mathbb{I}_{d'} \\ \mathbb{I}_d & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(2d+d')\times(d+2d')},$$

$$A_{\text{out}} := \begin{pmatrix} 0 & \mathbb{I}_{d'} & \mathbb{I}_{d'} \end{pmatrix} \in \mathbb{R}^{d'\times(d+2d')},$$

and note that $A_{\text{in}}x = (x, 0, x)$, $A(x, y, z)^T = (x, y + z, x)^T$, and $A_{\text{out}}(x, y, z)^T = y + z$, for $x \in \mathbb{R}^d, y, z \in \mathbb{R}^{d'}$. We construct

- the network $\widetilde{\Psi}_1$ by taking Ψ_1 and replacing E_1^1 with $E_1^1 A_{\text{in}}$, $E_{L_1}^1$ with $AE_{L_1}^1$, and $f_{L_1}^1$ with $Af_{L_1}^1$,

- the network $\widetilde{\Psi}_N$ by taking Ψ_N and replacing $E^N_{L_N}$ with $A_{\text{out}}E^N_{L_N}$ and $f^N_{L_N}$ with $A_{\text{out}}f^N_{L_N}$,
- the networks $\widetilde{\Psi}_i$, $i \in \{2, \dots, N-1\}$ by taking Ψ_i and replacing $E^i_{L_i}$ with $AE^i_{L_i}$ and $f^i_{L_i}$ with $Af^i_{L_i}$.

We can now verify that

$$\Phi = \widetilde{\Psi}_N \circ \widetilde{\Psi}_{N-1} \circ \cdots \circ \widetilde{\Psi}_1,$$

when the compositions are taken in the sense of Lemma 1. Due to Lemmas 2 and 3, we have $\mathcal{L}(\Psi_i) = \mathcal{L}(\Phi_i)$, $\mathcal{W}(\Psi_i) = 2d + 2d' + \mathcal{W}(\Phi_i)$, and $\mathcal{B}(\Psi_i) = \max\{1, \mathcal{B}(\Phi_i)\}$. The proof is finalized by noting that, owing to the structure of the involved matrices, the depth and the weight magnitude remain unchanged by turning Ψ_i into $\widetilde{\Psi}_i$, whereas the width can not increase, but may decrease owing to the replacement of E^1_1 by $E^1_1 A_{\text{in}}$. □

The following lemma shows how to patch together local approximations using multiplication networks and a partition of unity consisting of hat functions. We note that this argument can be extended to higher dimensions using tensor products (which can be realized efficiently through multiplication networks) of the one-dimensional hat function.

Lemma 16. *Let $\varepsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $a_0 < a_1 < \cdots < a_n \in \mathbb{R}$, $f \in L^\infty([a_0, a_n])$, and*

$$A := \left\lceil \max\{|a_0|, |a_n|, 2 \max_{i \in \{2,\dots,n-1\}} \tfrac{1}{|a_i - a_{i-1}|}\} \right\rceil,$$

$$B := \max\{1, \|f\|_{L^\infty([a_0,a_n])}\}.$$

Assume that for every $i \in \{1, \dots, n-1\}$, there exists a network $\Phi_i \in \mathcal{N}_{1,1}$ with $\|f - \Phi_i\|_{L^\infty([a_{i-1},a_{i+1}])} \leq \varepsilon/3$. Then, there is a network $\Phi \in \mathcal{N}_{1,1}$ satisfying

$$\|f - \Phi\|_{L^\infty([a_0,a_n])} \leq \varepsilon,$$

with $\mathcal{L}(\Phi) \leq \sum_{i=1}^{n-1} \mathcal{L}(\Phi_i) + Cn(\log(\varepsilon^{-1}) + \log(B) + \log(A))$, $\mathcal{W}(\Phi) \leq 7 + \max\{2, \max_{i\in\{1,\dots,n-1\}} \mathcal{W}(\Phi_i)\}$, $\mathcal{B}(\Phi) = \max\{1, \max_i \mathcal{B}(\Phi_i)\}$, *and with* $C > 0$ *an absolute constant, i.e., independent of* $\varepsilon, n, f, a_0, \dots, a_n$.

Proof. We first define the neural networks $(\Psi_i)_{i=1}^{n-1} \in \mathcal{N}_{1,1}$ forming a partition of unity according to

$$
\begin{aligned}
&\Psi_1(x) := 1 - \tfrac{1}{a_2-a_1}\,\rho(x-a_1) + \tfrac{1}{a_2-a_1}\,\rho(x-a_2),\\
&\Psi_i(x) := \tfrac{1}{a_i-a_{i-1}}\,\rho(x-a_{i-1})\\
&-(\tfrac{1}{a_i-a_{i-1}} + \tfrac{1}{a_{i+1}-a_i})\,\rho(x-a_i)\\
&+\tfrac{1}{a_{i+1}-a_i}\,\rho(x-a_{i+1}), \quad i \in \{2,\dots,n-2\},\\
&\Psi_{n-1}(x) := \tfrac{1}{a_{n-1}-a_{n-2}}\,\rho(x-a_{n-2}) - \tfrac{1}{a_{n-1}-a_{n-2}}\,\rho(x-a_{n-1}).
\end{aligned}
$$

Note that $\operatorname{supp}(\Psi_1) = (\infty, a_2)$, $\operatorname{supp}(\Psi_{n-1}) = [a_{n-2}, \infty)$, and $\operatorname{supp}(\Psi_i) = [a_{i-1}, a_{i+1}]$. Proposition 9 now ensures that, for all $i \in \{1,\dots,n-1\}$, Ψ_i can be realized as a network with $\mathcal{L}(\Psi_i) \leq 2(\lceil \log(A) \rceil + 5)$, $\mathcal{W}(\Psi_i) \leq 3$, and $\mathcal{B}(\Psi_i) \leq 1$. Next, let $\Phi_{B+1/6,\varepsilon/3} \in \mathcal{N}_{2,1}$ be the multiplication network according to Proposition 2 and define the networks

$$
\widetilde{\Phi}_i(x) := \Phi_{B+1/6,\varepsilon/3}(\Phi_i(x), \Psi_i(x))
$$

according to Lemma 3 and Lemma 1, along with their sum

$$
\Phi(x) := \sum_{i=1}^{n-1} \widetilde{\Phi}_i(x)
$$

according to Lemma 15. Proposition 2 ensures, for all $i \in \{1,\dots,n-1\}$, $x \in [a_{i-1}, a_{i+1}]$, that

$$
\begin{aligned}
|f(x)\Psi_i(x) - \widetilde{\Phi}_i(x)| &\leq |f(x)\Psi_i(x) - \Phi_i(x)\Psi_i(x)|\\
&\quad + |\Phi_i(x)\Psi_i(x) - \Phi_{B+1/6,\varepsilon/3}(\Phi_i(x), \Psi_i(x))|\\
&\leq (\Psi_i(x) + 1)\tfrac{\varepsilon}{3}
\end{aligned}
$$

and $\operatorname{supp}(\widetilde{\Phi}_i) = [a_{i-1}, a_{i+1}]$. In particular, for every $x \in [a_0, a_n]$, the set

$$I(x) := \{i \in \{1, \dots, n-1\} \colon \widetilde{\Phi}_i(x) \neq 0\}$$

of active indices contains at most two elements. Moreover, we have $\sum_{i \in I(x)} \Psi_i(x) = 1$ by construction, which implies that, for all $x \in \mathbb{R}$,

$$\begin{aligned} |f(x) - \Phi(x)| &= \left| \sum_{i \in I(x)} \Psi_i(x) f(x) - \sum_{i \in I(x)} \widetilde{\Phi}_i(x) \right| \\ &\leq \sum_{i \in I(x)} (\Psi_i(x) + 1) \tfrac{\varepsilon}{3} \leq \varepsilon. \end{aligned}$$

Due to Lemma 1, Lemma 3, Proposition 2, and Lemma 15, we can conclude that Φ, indeed, satisfies the claimed properties. □

Next, we present an extension of Lemma 6 to arbitrary (finite) intervals.

Lemma 17. *For $a, b \in \mathbb{R}$ with $a < b$, let*

$$\mathcal{S}_{[a,b]} := \Big\{ f \in C^\infty([a,b], \mathbb{R}) \colon \|f^{(n)}(x)\|_{L^\infty([a,b])} \leq n!, \text{ for all } n \in \mathbb{N}_0 \Big\}.$$

There exists a constant $C > 0$ such that for all $a, b \in \mathbb{R}$ with $a < b$, $f \in \mathcal{S}_{[a,b]}$, and $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying

$$\|\Psi_{f,\varepsilon} - f\|_{L^\infty([a,b])} \leq \varepsilon,$$

with $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C \max\{2, (b-a)\}((\log(\varepsilon^{-1}))^2 + \log(\lceil \max\{|a|, |b|\} \rceil) + \log(\lceil \frac{1}{b-a} \rceil))$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 16$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq 1$.

Proof. We first recall that the case $[a, b] = [-1, 1]$ has already been dealt with in Lemma 6. Here, we will first prove the statement for the interval $[-D, D]$ with $D \in (0, 1)$ and then use this result to establish

the general case through a patching argument according to Lemma 16. We start by noting that for $g \in \mathcal{S}_{[-D,D]}$, the function $f_g \colon [-1,1] \to \mathbb{R}, x \mapsto g(Dx)$ is in $\mathcal{S}_{[-1,1]}$ due to $D < 1$. Hence, by Lemma 6, there exists a constant $C > 0$ such that for all $g \in \mathcal{S}_{[-D,D]}$ and $\varepsilon \in (0, 1/2)$, there is a network $\widetilde{\Psi}_{g,\varepsilon} \in \mathcal{N}_{1,1}$ satisfying $\|\widetilde{\Psi}_{g,\varepsilon} - f_g\|_{L^\infty([-1,1])} \leq \varepsilon$, with $\mathcal{L}(\widetilde{\Psi}_{g,\varepsilon}) \leq C(\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\widetilde{\Psi}_{g,\varepsilon}) \leq 9$, $\mathcal{B}(\widetilde{\Psi}_{g,\varepsilon}) \leq 1$. The claim is then established by taking the network approximating g to be $\Psi_{g,\varepsilon} := \widetilde{\Psi}_{g,\varepsilon} \circ \Phi_{D^{-1}}$, where $\Phi_{D^{-1}}$ is the scalar multiplication network from Lemma 14, and noting that

$$\begin{aligned}\|\Psi_{g,\varepsilon}(x) - g(x)\|_{L^\infty([-D,D])} &= \sup_{x\in[-D,D]} |\widetilde{\Psi}_{g,\varepsilon}(\tfrac{x}{D}) - f_g(\tfrac{x}{D})| \\ &= \sup_{x\in[-1,1]} |\widetilde{\Psi}_{g,\varepsilon}(x) - f_g(x)| \leq \varepsilon.\end{aligned}$$

Due to Lemma 1, we have $\mathcal{L}(\Psi_{g,\varepsilon}) \leq C((\log(\varepsilon^{-1}))^2 + \log(\lceil \frac{1}{D} \rceil))$, $\mathcal{W}(\Psi_{g,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{g,\varepsilon}) \leq 1$. We are now ready to proceed to the proof of the statement for general intervals $[a,b]$. This will be accomplished by approximating f on intervals of length no more than 2 and stitching the resulting approximations together according to Lemma 16. We start with the case $b - a \leq 2$ and note that here we can simply shift the function by $(a+b)/2$ to center its domain around the origin and then use the result above for approximation on $[-D,D]$ with $D \in (0,1)$ or Lemma 6 if $b - a = 2$, both in combination with Corollary 2 to realize the shift through a neural network with weights bounded by 1. Using Lemma 1 to implement the composition of the network realizing this shift with that realizing g, we can conclude the existence of a constant $C' > 0$ such that, for all $[a,b] \subseteq \mathbb{R}$ with $b - a \leq 2$, $g \in \mathcal{S}_{[a,b]}$, $\varepsilon \in (0, 1/2)$, there is a network satisfying $\|g - \Psi_{g,\varepsilon}\|_{L^\infty([a,b])} \leq \varepsilon$ with $\mathcal{L}(\Psi_{g,\varepsilon}) \leq C'((\log(\varepsilon^{-1}))^2 + \log(\lceil \frac{1}{b-a} \rceil))$, $\mathcal{W}(\Psi_{g,\varepsilon}) \leq 9$, and $\mathcal{B}(\Psi_{g,\varepsilon}) \leq 1$. Finally, for $b - a > 2$, we partition the interval $[a,b]$ and apply Lemma 16 as follows. We set $n := \lceil b - a \rceil$ and define

$$a_i := a + i\tfrac{b-a}{n}, \quad i \in \{0, \dots, n\}.$$

Next, for $i \in \{1, \dots, n-1\}$, let $g_i \colon [a_{i-1}, a_{i+1}] \to \mathbb{R}$ be the restriction of g to the interval $[a_{i-1}, a_{i+1}]$, and note that $a_{i+1} - a_{i-1} = \frac{2(b-a)}{n} \in (\frac{4}{3}, 2]$. Furthermore, for $i \in \{1, \dots, n-1\}$, let $\Psi_{g_i, \varepsilon/3}$ be the network approximating g_i with error $\varepsilon/3$ as constructed above. Then, for every $i \in \{1, \dots, n-1\}$, it holds that $\|g - \Psi_{g_i,\varepsilon/3}\|_{L^\infty([a_{i-1},a_{i+1}])} \leq \frac{\varepsilon}{3}$ and application of Lemma 16 yields the desired result. □

We finally record, for technical purposes, slight variations of Lemmas 3 and 4 to account for parallelizations and linear combinations, respectively, of neural networks with shared input.

Lemma 18. *Let $n, d, L \in \mathbb{N}$ and, for $i \in \{1, 2, \dots, n\}$, let $d_i' \in \mathbb{N}$ and $\Phi_i \in \mathcal{N}_{d,d_i'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{d,\sum_{i=1}^n d_i'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) = \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i \mathcal{B}(\Phi_i)$, and satisfying*

$$\Psi(x) = (\Phi_1(x), \Phi_2(x), \dots, \Phi_n(x)) \in \mathbb{R}^{\sum_{i=1}^n d_i'},$$

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

Proof. The claim is established by following the construction in the proof of Lemma 3, but with the matrix $A_1 = \operatorname{diag}(A_1^1, A_1^2, \dots, A_1^n)$ replaced by

$$A_1 = \begin{pmatrix} A_1^1 \\ \vdots \\ A_1^n \end{pmatrix} \in \mathbb{R}^{(\sum_{i=1}^n N_1^i) \times d},$$

where N_1^i is the dimension of the first layer of Φ_i. □

Lemma 19. *Let $n, d, d', L \in \mathbb{N}$ and, for $i \in \{1, 2, \dots, n\}$, let $a_i \in \mathbb{R}$ and $\Phi_i \in \mathcal{N}_{d,d'}$ with $\mathcal{L}(\Phi_i) = L$. Then, there exists a network $\Psi \in \mathcal{N}_{d,d'}$ with $\mathcal{L}(\Psi) = L$, $\mathcal{M}(\Psi) \leq \sum_{i=1}^n \mathcal{M}(\Phi_i)$, $\mathcal{W}(\Psi) \leq \sum_{i=1}^n \mathcal{W}(\Phi_i)$, $\mathcal{B}(\Psi) = \max_i\{|a_i|\mathcal{B}(\Phi_i)\}$, and satisfying*

$$\Psi(x) = \sum_{i=1}^n a_i \Phi_i(x) \in \mathbb{R}^{d'},$$

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

Proof. The proof follows directly from that of Lemma 18 with the same modifications as those needed in the proof of Lemma 4 relative to that of Lemma 3. □

B. Tail compactness for Besov spaces

We consider the Besov space $B^m_{p,q}([0,1])$ (Mallat, 2008) given by the set of functions $f \in L^2([0,1])$ satisfying

$$\|f\|_{m,p,q} := \|(2^{n(m+\frac{1}{2}-\frac{1}{p})}\|(\langle f, \psi_{n,k}\rangle)_{k=0}^{2^n-1}\|_{\ell^p})_{n\in\mathbb{N}_0}\|_{\ell^q} < \infty, \tag{1.95}$$

with $\mathcal{D} = \{\psi_{n,k} \colon n \in \mathbb{N}_0, k = 0, \dots, 2^n - 1\}$ an orthonormal wavelet basis[12] for $L^2([0,1])$ and ℓ^p denoting the usual sequence norm

$$\|(a_i)_{i\in I}\|_{\ell^p} = \begin{cases} \left(\sum_{i\in I} |a_i|^p\right)^{\frac{1}{p}}, & 1 \le p < \infty \\ \sup_{i\in I} |a_i|, & p = \infty \end{cases}.$$

The unit ball in $B^m_{p,q}([0,1])$ is

$$\mathcal{U}(B^m_{p,q}([0,1])) = \{f \in L^2([0,1]) \colon \|f\|_{m,p,q} \le 1\}. \tag{1.96}$$

For simplicity of notation, we set $a_{n,k}(f) := \langle f, \psi_{n,k}\rangle$ and $A_n(f) := (a_{n,k}(f))_{k=0}^{2^n-1} \in \mathbb{R}^{2^n}$, for $n \in \mathbb{N}_0$. We now want to verify that for $q \in [1,2]$ tail compactness holds for the pair $(\mathcal{U}(B^m_{p,q}([0,1])), \mathcal{D})$ under the ordering $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \dots)$, where $\mathcal{D}_n := \{\psi_{n,k} \colon k = 0, \dots, 2^n - 1\}$. To this end, we first note that owing to $\sum_{n=0}^{N} |\mathcal{D}_n| = 2^{N+1} - 1$, we have tail compactness according to (1.26) if there exist $C, \beta > 0$ such that for all $f \in \mathcal{U}(B^m_{p,q}([0,1]))$, $N \in \mathbb{N}$,

$$\left\| f - \sum_{n=0}^{N} \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} \le C(2^{N+1})^{-\beta}. \tag{1.97}$$

[12] The space does not depend on the particular choice of mother wavelet ψ as long as ψ has at least r vanishing moments and is in $C^r([0,1])$ for some $r > m$. For further details we refer to Section 9.2.3 in (Mallat, 2008).

To see that (1.95) implies (1.97), we note that by orthonormality of $\mathcal{D}$,

$$\begin{aligned}
&\left\| f - \sum_{n=0}^{N} \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} \\
&= \left\| \sum_{n=N+1}^{\infty} \sum_{k=0}^{2^n-1} a_{n,k}(f)\psi_{n,k} \right\|_{L^2([0,1])} \\
&= \left(\sum_{n=N+1}^{\infty} \sum_{k=0}^{2^n-1} |a_{n,k}(f)|^2 \right)^{\frac{1}{2}} \\
&= \|(\|A_n(f)\|_{\ell^2})_{n=N+1}^{\infty}\|_{\ell^2}.
\end{aligned}$$

As the $A_n(f)$ are finite sequences of length $|\mathcal{D}_n| = 2^n$, it follows, by application of Hölder's inequality, that $\|A_n(f)\|_{\ell^2} \leq 2^{n(\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p}$. Together with $\|\cdot\|_{\ell^2} \leq \|\cdot\|_{\ell^q}$, for $q \leq 2$, (1.95) then ensures, for all $f \in \mathcal{U}(B^m_{p,q}([0,1]))$ and $q \in [1,2]$, that

$$\begin{aligned}
&\|(\|A_n(f)\|_{\ell^2})_{n=N+1}^{\infty}\|_{\ell^2} \\
&\leq \|(2^{n(\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p})_{n=N+1}^{\infty}\|_{\ell^q} \\
&\leq 2^{-(N+1)m}\|(2^{n(m+\frac{1}{2}-\frac{1}{p})}\|A_n(f)\|_{\ell^p})_{n=N+1}^{\infty}\|_{\ell^q} \\
&\leq 2^{-(N+1)m}\|f\|_{m,p,q} \leq (2^{N+1})^{-m},
\end{aligned}$$

which establishes (1.97) with $C = 1$ and $\beta = m$.

C. Tail compactness for modulation spaces

We consider tail compactness for unit balls in (polynomially) weighted modulation spaces, which, for $p, q \in [1, \infty)$, are defined as follows

$$M^s_{p,q}(\mathbb{R}) := \{f \colon \|f\|_{M^s_{p,q}(\mathbb{R})} < \infty\},$$

with

$$\|f\|_{M^s_{p,q}(\mathbb{R})} := \left(\int_{\mathbb{R}} \left(\int_{\mathbb{R}} |V_w f(x,\xi)|^p (1 + |x| + |\xi|)^{sp} \mathrm{d}x \right)^{\frac{q}{p}} \mathrm{d}\xi \right)^{\frac{1}{q}},$$

where

$$V_w f(x,\xi) := \int_{\mathbb{R}} f(t)\,\overline{w(t-x)}e^{-2\pi i t\xi}\mathrm{d}t, \quad x,\xi \in \mathbb{R},$$

is the short-time Fourier transform of f with respect to the window function[13] $w \in \mathcal{S}(\mathbb{R})$.

Next, let $g \in L^2(\mathbb{R})$ with $\|g\|_{L^2(\mathbb{R})} = 1$ and $g(x) = \overline{g(-x)}$ such that the Gabor dictionary $\mathcal{G}(g, \frac{1}{2}, 1, \mathbb{R})$ is a tight frame (Morgenshtern and Bölcskei, 2012) for $L^2(\mathbb{R})$. Then, the Wilson dictionary $\mathcal{D} = \{\psi_{k,n} \colon (k,n) \in \mathbb{Z} \times \mathbb{N}_0\}$ with

$$\begin{aligned} \psi_{k,0} &= T_k g, && k \in \mathbb{Z}, \\ \psi_{k,n} &= \tfrac{1}{\sqrt{2}} T_{\frac{k}{2}} (M_n + (-1)^{k+n} M_{-n}) g, && (k,n) \in \mathbb{Z} \times \mathbb{N}, \end{aligned}$$

is an orthonormal basis for $L^2(\mathbb{R})$ (see (Gröchenig, 2013, Thm. 8.5.1)). We have, for every $f \in M^s_{p,q}(\mathbb{R})$, the expansion (Gröchenig, 2013, Thm. 12.3.4)

$$\begin{aligned} f &= \sum_{(k,n)\in\mathbb{Z}\times\mathbb{N}_0} c_{k,n}(f)\psi_{k,n}, \\ &\text{where} \quad c_{k,n}(f) = \langle f, \psi_{k,n}\rangle, \quad c(f) \in \ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0), \end{aligned}$$

with $\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)$ the space of sequences $c \in \mathbb{R}^{\mathbb{Z}\times\mathbb{N}_0}$ satisfying

$$\|c\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)} := \left(\sum_{n\in\mathbb{N}_0} \left(\sum_{k\in\mathbb{Z}} |c_{k,n}|^p (1 + |\tfrac{k}{2}| + |n|)^{sp} \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} < \infty.$$

Moreover, there exists (Gröchenig, 2013, Thm. 12.3.1) a constant $D \geq 1$ such that, for all $f \in M^s_{p,q}(\mathbb{R})$,

$$\tfrac{1}{D}\|f\|_{M^s_{p,q}(\mathbb{R})} \leq \|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)} \leq D\|f\|_{M^s_{p,q}(\mathbb{R})}.$$

[13] The resulting modulation space does not depend on the specific choice of window function w as long as w is in the Schwartz space $\mathcal{S}(\mathbb{R}) = \{f \in C^\infty(\mathbb{R}) \colon \sup_{x\in\mathbb{R}} |x^\alpha f^{(\beta)}(x)| < \infty, \text{ for all } \alpha,\beta \in \mathbb{N}_0\}$, where $f^{(n)}$ stands for the n-th derivative of f.

In particular, we can characterize the unit ball of $M^s_{p,q}(\mathbb{R})$ according to

$$\mathcal{U}(M^s_{p,q}(\mathbb{R})) = \{f\colon \|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)} \leq D\}.$$

We now order the Wilson basis dictionary as follows. Define $\mathcal{D}_0 := \{\psi_{0,0}\}$ and

$$\mathcal{D}_\ell := \{\psi_{k,n}\colon |k|, n \leq \ell\} \setminus \bigcup_{i=0}^{\ell-1} \mathcal{D}_i$$

for $\ell \geq 1$, and order the overall dictionary according to $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \dots)$. Owing to $\sum_{\ell=0}^{N} |\mathcal{D}_\ell| = (2N+1)(N+1)$, we have tail compactness for the pair $(\mathcal{U}(M^s_{p,q}(\mathbb{R})), \mathcal{D})$ if there exist $C, \beta > 0$ such that, for all $f \in \mathcal{U}(M^s_{p,q}(\mathbb{R}))$, $N \in \mathbb{N}$,

$$\left\| f - \sum_{n=0}^{N} \sum_{k=-N}^{N} c_{k,n}(f)\psi_{k,n} \right\|_{L^2(\mathbb{R})} \leq CN^{-\beta}. \tag{1.98}$$

We restrict our attention to $p, q \leq 2$ and use orthonormality of $\mathcal{D}$ and the fact that $\|\cdot\|_{\ell^2} \leq \|\cdot\|_{\ell^p}$, for $p \leq 2$, to obtain, for all $f \in \mathcal{U}(M^s_{p,q}(\mathbb{R}))$,

$$\begin{aligned}
&\left\| f - \sum_{n=0}^{N} \sum_{k=-N}^{N} c_{k,n}(f)\psi_{k,n} \right\|_{L^2(\mathbb{R})} = \left\| \sum_{n>N} \sum_{|k|>N} c_{k,n}(f)\psi_{k,n} \right\|_{L^2(\mathbb{R})} \\
&= \left(\sum_{n>N} \sum_{|k|>N} |c_{k,n}(f)|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{n>N} \left(\sum_{|k|>N} |c_{k,n}(f)|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \\
&\leq (1+\tfrac{3}{2}N)^{-s} \left(\sum_{n>N} \left(\sum_{|k|>N} |c_{k,n}(f)|^p (1+|\tfrac{k}{2}|+|n|)^{sp} \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \\
&\leq (1+\tfrac{3}{2}N)^{-s} \|c(f)\|_{\ell^s_{p,q}(\mathbb{Z}\times\mathbb{N}_0)} \leq (3/2)^{-s} D N^{-s},
\end{aligned}$$

which establishes tail compactness with $C = (3/2)^{-s}D$ and $\beta = s$.

CHAPTER 2

High-dimensional distribution generation through deep neural networks

2.1. INTRODUCTION

Deep neural networks have been employed very successfully as generative models for complex natural data such as images Radford et al. (2016); Karras et al. (2019) and natural language Bowman et al. (2016); Xu et al. (2018). Specifically, the idea is to train deep networks so that they realize complex high-dimensional probability distributions by transforming samples taken from simple low-dimensional distributions such as uniform or Gaussian Kingma and Welling (2014); Goodfellow et al. (2014); Arjovsky et al. (2017).

Generative networks with output dimension higher than the input dimension occur, for instance, in language modelling where deep networks are used to predict the next word in a text sequence. Here, the input layer size is determined by the dimension of the word embedding (typically ~ 100) and the output layer, representing a vector of probabilities for each of the words in the vocabulary, is of the size of the vocabulary (typically $\sim 100k$). Another example where the dimension of the output distribution is mandated to be higher than that of the input

distribution is given by explicit Kingma and Welling (2014); Tolstikhin et al. (2018) and implicit Goodfellow et al. (2014); Arjovsky et al. (2017) density generative networks.

Notwithstanding the practical success of deep generative networks, a profound theoretical understanding of their representational capabilities is still lacking. First results along these lines appear in Lee et al. (2017), where it was shown that generative networks can approximate distributions arising from the composition of Barron functions Barron (1993). It remains unclear, however, which distributions can be obtained in such a manner. More recently, it was established Lu and Lu (2020) that for every given target distribution (of finite third moment) and source distribution, both defined on $\mathbb{R}^d$, there exists a ReLU network whose gradient pushes forward the source distribution to an arbitrarily accurate—in terms of Wasserstein distance—approximation of the target distribution. The aspect of dimensionality increase was addressed in Bailey and Telgarsky (2018), where it is shown that a uniform univariate source distribution can be transformed, by a ReLU network, into a uniform target distribution of arbitrary dimension through a space-filling approach. Besides, Bailey and Telgarsky (2018) also demonstrates how a univariate Gaussian target distribution can be obtained from a univariate uniform source distribution and vice versa. In a general context, the problem of optimal transport Villani (2008) between source and target distributions on spaces of different dimensions was studied in McCann and Pass (2020).

The approximation of distributions through generative networks is inherently related to function approximation and hence to the expressivity of neural networks. A classical result along those lines is the universal approximation theorem Cybenko (1989); Hornik (1991), which states that single-hidden-layer neural networks with sigmoidal activation function can approximate continuous functions on compact subsets of $\mathbb{R}^n$ arbitrarily well. More recent developments in this area are concerned with the influence of network depth on attainable approximation quality Telgarsky (2016); Daubechies et al. (2019); Eldan and Shamir (2016). A theory establishing the fundamental limits of

deep neural network expressivity is provided in Bölcskei et al. (2019); Elbrächter et al. (2021).

The aim of the present chapter is to develop a universal approximation result for generative neural networks. Specifically, we show that every target distribution supported on a bounded subset of $\mathbb{R}^d$ can be approximated arbitrarily well in terms of Wasserstein distance by pushing forward a 1-dimensional uniform source distribution through a ReLU network. The result is constructive in the sense of explicitly identifying the corresponding generative network. Concretely, we proceed in two steps. Given a target distribution, we find the histogram distribution that best approximates it—for a given histogram resolution—in Wasserstein distance. This histogram distribution is then realized by a ReLU network driven by a uniform univariate input distribution. The construction of this ReLU network exploits a space-filling property, vastly generalizing the one discovered in Bailey and Telgarsky (2018); Perekrestenko et al. (2020). The main conceptual insight of the present chapter is that generating arbitrary d-dimensional target distributions from a 1-dimensional uniform distribution through a deep ReLU network does not come at a cost—in terms of approximation error measured in Wasserstein distance—relative to generating the target distribution from d independent random variables through, e.g., (for arbitrary d) the normalizing flows method Rezende and Mohamed (2015) and (for $d = 2$) the Box-Muller method Box and Muller (1958). We emphasize that the generating network has to be deep, in fact its depth has to go to infinity to obtain the same Wasserstein-distance error as a construction from d independent random variables would yield. Finally, we find that, for histogram target distributions, the number of bits needed to encode the corresponding generative network equals the fundamental limit for encoding probability distributions as dictated by quantization theory Graf and Luschgy (2000).

2.2. DEFINITIONS AND NOTATION

We start by introducing general notation.

log stands for the logarithm to base 2. For $n_1, n_2 \in \mathbb{N}$, the set of integers in the range $[n_1, n_2]$ is designated as $[n_1 : n_2]$. For $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, we denote the vector obtained by retaining the first $t,\ t \leq d$, entries by $\mathbf{x}_{[1:t]} := (x_1, x_2, \ldots, x_t) \in \mathbb{R}^t$. $U(\Delta)$ stands for the uniform distribution on the interval Δ; when $\Delta = [0, 1]$, we simply write U. Given a probability density function (pdf) p, we denote its push-forward under the function f as $f\#p$. δ_x refers to the Dirac delta distribution. $\mathfrak{B}^d$ stands for the Borel σ-algebra on $\mathbb{R}^d$, i.e., the smallest σ-algebra on $\mathbb{R}^d$ that contains all open subsets of $\mathbb{R}^d$. For a vector $\mathbf{b} \in \mathbb{R}^d$, we let $\|\mathbf{b}\|_\infty := \max_i |b_i|$, similarly we write $\|\mathbf{A}\|_\infty := \max_{i,j} |\mathbf{A}_{i,j}|$ for the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. The Cartesian product of the intervals $\mathcal{I}_i,\ i \in [1 : d]$, is denoted by $\bigtimes_{i=1}^d \mathcal{I}_i = \mathcal{I}_1 \bigtimes \mathcal{I}_2 \bigtimes \cdots \bigtimes \mathcal{I}_d$. Finally, $\chi_{\mathcal{I}}$ stands for the indicator function on the set $\mathcal{I}$.

We proceed to define ReLU neural networks.

Definition 20. *Let $L \in \mathbb{N}$ and $N_0, N_1, \ldots, N_L \in \mathbb{N}$. A* ReLU neural network Φ *is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3, \end{cases}$$

where, for $\ell \in [1 : L]$, $W_\ell \colon \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(\mathbf{x}) := \mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell$ are the associated affine transformations with (weight) matrices $\mathbf{A}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $\mathbf{b}_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho \colon \mathbb{R} \to \mathbb{R},\ \rho(x) := \max(x, 0)$ acts component-wise, i.e., $\rho(x_1, \ldots, x_N) := (\rho(x_1), \ldots, \rho(x_N))$. We denote by $\mathcal{NN}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. Moreover, we define the following quantities related to the notion of size of the network Φ:

- *the* connectivity $\mathcal{M}(\Phi)$ *is the total number of nonzero weights, i.e., entries in the matrices* $\mathbf{A}_\ell$, $\ell \in [1:L]$, *and the vectors* $\mathbf{b}_\ell$, $\ell \in [1:L]$
- depth $\mathcal{L}(\Phi) := L$
- width $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$

The distance between probability measures will be quantified through Wasserstein distance defined as follows.

Definition 21. *Let* μ *and* ν *be probability measures on* $(\mathbb{R}^d, \mathfrak{B}^d)$. *A coupling between* μ *and* ν *is defined as a probability measure* π *on* $(\mathbb{R}^d \times \mathbb{R}^d, \mathfrak{B}^{2d})$ *such that* $\pi(A_1 \times \mathbb{R}^d) = \mu(A_1)$ *and* $\pi(\mathbb{R}^d \times A_2) = \nu(A_2)$, *for all* $A_1, A_2 \in \mathfrak{B}^d$. *Let* $\prod(\mu, \nu)$ *be the set of all couplings between* μ *and* ν. *The Wasserstein distance between* μ *and* ν *is defined as*

$$W(\mu, \nu) := \inf_{\pi \in \prod(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\| d\pi(\mathbf{x}, \mathbf{y}),$$

where $\|\cdot\|$ *denotes Euclidean norm.*

We will frequently use the concept of histogram distributions formalized as follows.

Definition 22. *A random variable* X *is said to have a general histogram distribution of resolution* n *on* $[0, 1]$, *denoted as* $X \sim \mathcal{G}[0, 1]_n^1$, *if for some* $t_0, t_1, \ldots, t_n \in \mathbb{R}$ *such that* $0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1$ *and if* $t_i = t_{i+1}$, *then* $t_{i+2} \neq t_i$, $\forall i \in [0:(n-2)]$, *its pdf is given by*

$$p(x) = \sum_{k=0}^{n-1} w_k \kappa_{[t_k, t_{k+1}]}(x), \text{ with } \sum_{k=0}^{n-1} w_k d(t_k, t_{k+1}) = 1, \qquad (2.1)$$

and $w_k > 0$ *for all* $k \in [0:(n-1)]$. *Here,*

$$\kappa_{[t_k, t_{k+1}]}(x) = \begin{cases} \chi_{[t_k, t_{k+1}]}(x), & \text{if } t_k < t_{k+1} \\ \delta_{x - t_k}, & \text{if } t_k = t_{k+1} \end{cases},$$

and

$$d(t_k, t_{k+1}) = \begin{cases} t_{k+1} - t_k, & \text{if } t_k < t_{k+1} \\ 1, & \text{if } t_k = t_{k+1} \end{cases}. \qquad (2.2)$$

General histogram distributions allow for bins of arbitrary size and for point singularities, see the right-hand plot in Figure 2.4. We will, however, mostly be concerned with histogram distributions of uniform tile size, defined as follows.

Definition 23. *A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$ is said to have a histogram distribution of resolution n on the d-dimensional unit cube, denoted as $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$, if its pdf is given by*

$$p_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{k}} w_{\mathbf{k}} \chi_{c_{\mathbf{k}}}(\mathbf{x}), \text{ with } \sum_{\mathbf{k}} w_{\mathbf{k}} = n^d,$$

and $w_{\mathbf{k}} > 0$ for all vectors $\mathbf{k} = (i_1, i_2, \ldots, i_d) \in [0 : (n-1)]^d$, referred to as index vectors, and $c_{\mathbf{k}} = [i_1/n, (i_1+1)/n] \times [i_2/n, (i_2+1)/n] \times \cdots \times [i_d/n, (i_d+1)/n]$.

Example histogram distributions in the 1- and 2-dimensional case, respectively, are depicted in Figs. 2.1 and 2.2. Throughout the chapter, to indicate that the random vector $\mathbf{X}$ with distribution $p_{\mathbf{X}}(\mathbf{x})$ satisfies $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$, we shall frequently also write $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$.

Remark 10. *For ease of exposition, in Definitions 22 and 23, we let the intervals $[t_k, t_{k+1}]$ and the cubes $c_{\mathbf{k}}$, respectively, be closed, thus allowing the breakpoints to belong to different intervals/cubes. While this comes without loss of generality, for concreteness, it is understood that the value of the pdf at a breakpoint is given by the average across the intervals/cubes containing the breakpoint.*

2.3. SAWTOOTH FUNCTIONS

As mentioned above, our universal generative network construction is based on a new space-filling property of ReLU networks, vastly

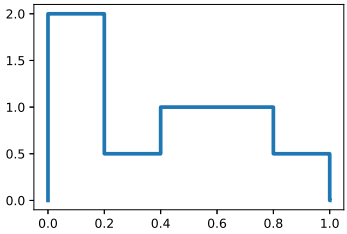


Fig. 2.1: Histogram distribution $\mathcal{E}[0,1]_5^1$

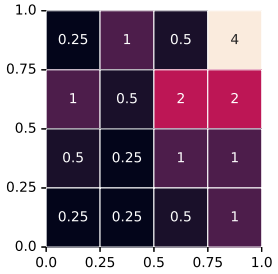


Fig. 2.2: Histogram distribution $\mathcal{E}[0,1]_4^2$

generalizing the one discovered in Bailey and Telgarsky (2018); Perekrestenko et al. (2020). At the heart of this idea are higher-order sawtooth functions obtained as follows. Consider the sawtooth function $g : \mathbb{R} \to [0,1]$,

$$g(x) = \begin{cases} 2x, & \text{if } x \in [0, 1/2), \\ 2(1-x), & \text{if } x \in [1/2, 1], \\ 0, & \text{else}, \end{cases}$$

let $g_1(x) = g(x)$, and define the sawtooth function of order s as the s-fold composition of g with itself according to

$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \quad s \geq 2.$$

Figure 2.3 depicts the sawtooth functions of orders $1, 2$, and 3. Next, we note that g can be realized by a 2-layer ReLU network $\Phi_g \in \mathcal{NN}_{1,1}$ of connectivity $\mathcal{M}(\Phi_g) = 8$ and depth $\mathcal{L}(\Phi_g) = 2$ according to $\Phi_g = W_2 \circ \rho \circ W_1 = g$ with

$$W_1(x) = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix} x - \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}, \quad W_2(\mathbf{x}) = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The sth-order sawtooth function g_s can hence be realized by a ReLU network $\Phi_g^s \in \mathcal{NN}_{1,1}$ of connectivity $\mathcal{M}(\Phi_g^s) = 11s - 3$ and depth $\mathcal{L}(\Phi_g^s) = s + 1$ according to $\Phi_g^s = W_2 \circ \rho \circ \underbrace{W_g \circ \rho \circ \cdots \circ W_g \circ \rho}_{s-1} \circ W_1 = g_s$ with

$$W_g(\mathbf{x}) = \begin{pmatrix} 2 & -2 & 2 \\ 4 & -4 & 4 \\ 2 & -2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}.$$

We close this section with an important technical ingredient of the generalized space-filling idea presented in Section 2.5.

Lemma 20. *Let $f(x)$ be a continuous function on $[0,1]$, with $f(0) = 0$. Then, for all $s \in \mathbb{N}$,*

$$f(g_s(x)) = \sum_{k=0}^{2^{s-1}-1} f\big(g(2^{s-1}x - k)\big), \tag{2.3}$$

and for all $k \in [0 : (2^{s-1} - 1)]$,

$$\operatorname{supp}\big(f\big(g(2^{s-1}x - k)\big)\big) = \left(\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\right). \tag{2.4}$$

Proof. We first note that the sawtooth functions $g_s(x)$ satisfy Telgarsky (2016)

$$g_s(x) = \sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k),$$

with $g(2^{s-1}x - k)$ supported on $\big(\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\big)$. As $f(0) = 0$, the support of $f(g(2^{s-1}x - k))$ coincides with that of $g(2^{s-1}x - k)$, which in turn yields (2.4). To establish (2.3), we note that the supports of $g(2^{s-1}x - k)$ are pairwise disjoint across k and hence

$$f(g_s(x)) = f\left(\sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k)\right) = \sum_{k=0}^{2^{s-1}-1} f\big(g(2^{s-1}x - k)\big).$$

□

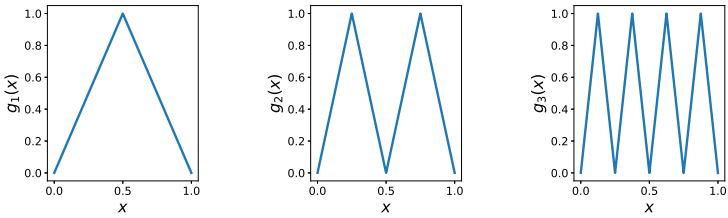


Fig. 2.3: Sawtooth functions

2.4. RELU NETWORKS GENERATE HISTOGRAM DISTRIBUTIONS

This section establishes a systematic connection between ReLU networks and histogram distributions. Specifically, we show that the pushforward of a uniform distribution under a piecewise linear function results in a histogram distribution. We also identify, for a given histogram distribution, the piecewise linear function generating it under pushforward of a uniform distribution. Combined with the insight that ReLU networks realize piecewise linear functions, the desired connection is established.

We start with an auxiliary result.

Lemma 21. *Let $a, b \in \mathbb{R}, a < b, \Delta = [a, b]$, and let $h(x) = mx + s$, for $x \in \mathbb{R}$, with $m \in \mathbb{R}, s \in \mathbb{R}$. Then, $Q = h\#U(\Delta)$ is uniformly distributed on $[ma + s, mb + s]$, for $m > 0$, and on $[mb + s, ma + s]$, for $m < 0$. For $m = 0$, the pdf of Q is given by $\delta_{\cdot - s}$.*

Proof. We start with the case $m \in \mathbb{R} \setminus \{0\}$. The pdf of the pushforward of a random variable with pdf $p(x)$ under the function $f(x)$ is given by

$$q(y) = p(f^{-1}(y)) \left| \frac{d}{dy} f^{-1}(y) \right|.$$

Particularized to $f^{-1}(y) = h^{-1}(y) = \frac{y-s}{m}$ and $p(x) = \frac{1}{b-a}\chi_\Delta(x)$,

this yields

$$q(y) = \begin{cases} \frac{1}{m(b-a)}, & \text{if } y \in [ma+s, mb+s] \\ 0, & \text{otherwise,} \end{cases}$$

for $m > 0$, and

$$q(y) = \begin{cases} \frac{1}{m(a-b)}, & \text{if } y \in [mb+s, ma+s] \\ 0, & \text{otherwise,} \end{cases}$$

for $m < 0$. Finally, if $m = 0$, then the entire interval $[a, b]$ is mapped to the point $y = s$ and the corresponding pdf is given by $q(y) = \delta_{y-s}$. $\square$

We next show that the pushforward of a uniform distribution under a piecewise linear function results in a (general) histogram distribution.

Theorem 14. *For every piecewise linear continuous function $f : \mathbb{R} \to \mathbb{R}$, such that $f(x) \in [0, 1], \forall x \in [0, 1]$, and $f(0) = 0, f(1) = 1$, there exists an n so that $f\#U \sim \mathcal{G}[0, 1]_n^1$.*

Proof. We split the domain of f into $t \in \mathbb{N}$ pairwise disjoint intervals $I_i = [a_i, b_i], i \in [0\!:\!(t-1)]$, each of which f is linear on, specifically $f(x) = m_i x + s_i,\ x \in I_i$. Using the law of total probability and the chain rule, the pdf of $q = f\#U$ can accordingly be represented as

$$q(y) = \sum_{j=0}^{t-1} q(y|u \in I_j)\mathbb{P}(u \in I_j). \tag{2.5}$$

As U is uniform, it is also uniform conditional on being in a given interval I_j. By Lemma 21 it therefore follows that $q(y|u \in I_j)$ can be written as

$$q(y|u \in I_j) = \begin{cases} \frac{\chi_{R_j}(y)}{|R_j|}, & \text{if } m_j \neq 0, \\ \delta_{y-s_j}, & \text{if } m_j = 0, \end{cases}$$

where $R_j = [m_j a_j + s_j, m_j b_j + s_j]$ if $m_j > 0$, and $R_j = [m_j b_j + s_j, m_j a_j + s_j]$ if $m_j < 0$. Noting that by continuity of f and

the boundary conditions $f(0) = 0, f(1) = 1$, we have $\bigcup_j R_j = [0, 1]$, it follows that $q(y)$ in (2.5) corresponds to a general histogram distribution according to (2.1) with $n = t$ and

$$w_j = \begin{cases} \frac{\mathbb{P}(u \in I_j)}{|R_j|}, & \text{if } m_j \neq 0, \\ \mathbb{P}(u \in I_j), & \text{if } m_j = 0. \end{cases} \qquad \square$$

We will also need the converse to the result just established, in particular a constructive version thereof explicitly identifying the piecewise linear function that leads to a given general histogram distribution under pushforward of a uniform distribution on the interval $[0, 1]$.

Theorem 15. *Let $p(x)$ be the pdf of $X \sim \mathcal{G}[0, 1]_n^1$ with weights w_k, $k \in [0 : (n-1)]$, and breakpoints $0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1$, and set $b_0 = 0$, $b_i = \sum_{k=0}^{i-1} w_k \, d(t_k, t_{k+1})$, $i \in [1 : n]$, where*

$$d(t_k, t_{k+1}) = \begin{cases} t_{k+1} - t_k, & \text{if } t_k < t_{k+1} \\ 1, & \text{if } t_k = t_{k+1} \end{cases}.$$

Further, define a_i, $i \in [0 : (n-1)]$, as follows: If $t_0 = t_1$, then $a_0 = 0$ and $a_1 = \frac{1}{w_1}$. If $t_0 \neq t_1$, then $a_0 = \frac{1}{w_0}$. For $k \in [1 : (n-2)]$, if $t_k = t_{k+1}$, then $a_k = -\frac{1}{w_{k-1}}$ and $a_{k+1} = \frac{1}{w_{k+1}}$, and, if $t_{k-1} \neq t_k \neq t_{k+1}$, then $a_k = \frac{1}{w_k} - \frac{1}{w_{k-1}}$. Finally, if $t_{n-1} \neq 1$, then $a_{n-1} = \frac{1}{w_{n-1}} - \frac{1}{w_{n-2}}$. Then,

$$f(x) = \sum_{i=0}^{n-1} a_i \rho(x - b_i) \qquad (2.6)$$

is the piecewise linear function satisfying $f \# U = p$.

Proof. Let $I_i := [b_i, b_{i+1}]$, $i \in [0 : (n-1)]$. Then, $\bigcup_{i \in [0:(n-1)]} I_i = [0, 1]$ and, for all $i \in [0 : (n-1)]$, the function $f(x)$ in (2.6) is linear on I_i with slope given by

$$\sum_{j=0}^{i} a_j = \begin{cases} 1/w_i, & \text{if } t_i \neq t_{i+1} \\ 0, & \text{if } t_i = t_{i+1} \end{cases}.$$

Next, note that under $f(x)$ the interval I_i is mapped to the interval $[t_i, t_{i+1}]$ if $t_i \neq t_{i+1}$ and to the singleton $\{t_i\}$ if $t_i = t_{i+1}$. The proof is finalized upon using $\mathbb{P}(u \in I_i) = b_{i+1} - b_i$ to conclude that (cf. the proof of Theorem 14) $\mathbb{P}(u \in I_i)/|(1/w_i)(b_{i+1} - b_i)| = w_i$, for $t_i \neq t_{i+1}$, and $\mathbb{P}(u \in I_i) = b_{i+1} - b_i = w_i\, d(t_i, t_{i+1}) = w_i$ in the case $t_i = t_{i+1}$. □

An example of a piecewise linear function and the corresponding general histogram distribution according to Theorems 14 and 15 is provided in Figure 2.4. Theorems 14 and 15 are of independent interest as they allow to conclude that ReLU networks, by virtue of always realizing piecewise linear functions, produce general histogram distributions when pushing forward uniform distributions. In the remainder of the chapter, we shall, however, work with histogram distributions $\mathcal{E}[0,1]^d_n$ only, in particular for $d = 1$, in which case Theorem 15 takes on a simpler form spelled out next.

Corollary 3. *Let $p(x)$ be the pdf of $X \sim \mathcal{E}[0,1]^1_n$ with weights w_k, $k \in [0\!:\!n]$, and let $a_0 = \frac{1}{w_0}$, $a_i = \frac{1}{w_i} - \frac{1}{w_{i-1}}$, $i \in [1\!:\!(n-1)]$, $b_0 = 0$, $b_i = \frac{1}{n}\sum_{k=0}^{i-1} w_k$, $i \in [1:n]$. Then, $p = f\#U$ with the piecewise linear function*

$$f(x) = \sum_{i=0}^{n-1} a_i \rho(x - b_i).$$

We shall often need the explicit form of $f(x)$ in Corollary 3 on its intervals of linearity $[b_\ell, b_{\ell+1}]$. A direct calculation reveals that

$$f(x) = \frac{x}{w_\ell} - \frac{\sum_{i=0}^{\ell-1} w_i}{n w_\ell} + \frac{\ell}{n}, \quad x \in [b_\ell, b_{\ell+1}].$$

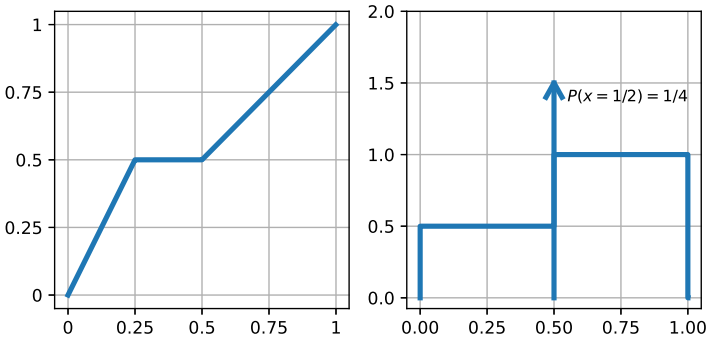


Fig. 2.4: A piecewise linear function f (left) and the corresponding general histogram distribution $f\#U$ (right).

2.5. INCREASING DISTRIBUTION DIMENSIONALITY

This section is devoted to the aspect of distribution dimensionality increase through deep ReLU networks. Specifically, for a given random vector $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$ with (histogram) distribution $p_{\mathbf{X}}(\mathbf{x})$ of resolution n, we construct a piecewise linear map $M : [0,1] \to [0,1]^d$ such that the pushforward $M\#U$ approximates $p_{\mathbf{X}}(\mathbf{x})$ arbitrarily well. The main ingredient of our construction is a vast generalization of the space-filling property of sawtooth functions discovered in Perekrestenko et al. (2020). Informally speaking, the novel space-filling property we describe allows to completely fill the d-dimensional target space according to a prescribed target histogram distribution by transporting probability mass from the 1-dimensional uniform distribution U to d-dimensional space.

We first develop some intuition behind this construction. Specifically, we consider the 2-dimensional case and visualize the idea of approxi-

mating a 2-dimensional target distribution through pushforward of U by the sawtooth functions $g_s(x)$ (depicted in Figure 2.3) as painting the curve $g_s(x)$ with probability mass taken from U. The geodesic distance traveled by the brush distributing probability mass along $g_s(x)$ goes to infinity according to 2^s as $s \to \infty$. This follows by noting that the number of teeth is 2^s and for $s \to \infty$, the length of the individual teeth (in fact their halves) approaches 1. Therefore, as $s \to \infty$ the square $[0,1]^2$ will be filled with paint completely. Moreover, as x traverses from 0 to 1, the speed at which probability mass is allocated to the marginal dimensions, i.e., along the x_1-and x_2-axes, is constant. To see this, simply note that along the x_2-axis the speed of the brush is given by the derivative of $g_s(x)$, which by virtue of $g_s(x)$ consisting of piecewise linear segments, is constant. Likewise, as the inverse of a linear function is again a linear function, the brush moves with constant speed along the x_1-axis as well. This guarantees that the resulting 2-dimensional probability distribution along with its marginals and conditional distributions are all uniform. The rate at which the joint distribution approaches a 2-dimensional uniform distribution can be quantified in terms of Wasserstein distance by defining the transport map $M : x \to (x, g_s(x))$ and noting that $W(M\#U, U([0,1]^2)) \leq \frac{\sqrt{2}}{2^s}$ Bailey and Telgarsky (2018). What is noteworthy here is that the map M takes probability mass from $\mathbb{R}$ to $\mathbb{R}^2$ in a space-filling fashion, i.e., we get a dimensionality increase as $s \to \infty$.

By adjusting the "paint plan", this idea can now be generalized to 2-dimensional histogram target distributions that are constant with respect to one of the dimensions, here, for concreteness, the x_1-dimension. Specifically, we replace $g_s(x)$ in the construction above by $f(g_s(x))$, where the piecewise linear function $f(x)$ determines the paint plan resulting in the desired weights (across the x_2-axis) according to Corollary 3. We refer to Figure 2.5 for an illustration of the idea. While the outer function $f(x)$ determines how much time the paint brush spends in a given interval along the x_2-axis, the inner function $g_s(x)$ takes care of filling the unit square as $s \to \infty$. The larger the slope of $f(x)$ on a given interval along the x_2-axis, the less time the brush spends in

that interval and the smaller the amount of probability mass allocated to the interval. Concretely, by Corollary 3 the amount of probability mass deposited in a given interval is inversely proportional to the slope of $f(x)$ on that interval.

Finally, consider the 2-dimensional histogram distribution $p_{X_1,X_2}(x_1,x_2) \in \mathcal{E}[0,1]^2_n$ and note that it can be represented as follows

$$\begin{aligned} p_{X_1,X_2}(x_1,x_2) &= \sum_{k_1,k_2} w_{k_1,k_2}\chi_{c_{k_1,k_2}}(x_1,x_2) \\ &= \sum_{k_1,k_2} w_{k_1}\, w_{k_2|k_1}\chi_{c_{k_1}}(x_1)\chi_{c_{k_2}}(x_2) \\ &= \sum_{k_1} w_{k_1}\chi_{c_{k_1}}(x_1)\sum_{k_2} w_{k_2|k_1}\chi_{c_{k_2}}(x_2) \qquad (2.7) \\ &= \sum_{i=0}^{n-1} p_{X_1}\big(x_1 \in [i/n,(i+1)/n]\big) \\ &\qquad\qquad p_{X_2|X_1}\big(x_2|x_1 \in [i/n,(i+1)/n]\big), \end{aligned}$$

where $w_{k_1} = \frac{1}{n}\sum_{k_2} w_{k_1,k_2}$, $w_{k_2|k_1} = w_{k_1,k_2}/w_{k_1}$, $p_{X_1}(x_1 \in [i/n,(i+1)/n]) = w_i\chi_{c_i}(x_1)$ denotes the restriction of the marginal histogram distribution p_{X_1} (see Lemma 22 below) to the interval $[i/n,(i+1)/n]$, and $p_{X_2|X_1}\big(x_2|x_1 \in [i/n,(i+1)/n]\big) = \sum_{k_2} w_{k_2|i}\chi_{c_{k_2}}(x_2)$, for each $i \in [0:(n-1)]$, can be viewed as a 2-dimensional histogram distribution that is constant with respect to x_1, and which we assume to be generated by $f^{(i)}_{X_2}(g_s(x))$ according to the procedure described in the previous paragraph. Now, in order to "paint" the general 2-dimensional histogram distribution in (2.7), we have to "squeeze" the space-filling curves $f^{(i)}_{X_2}(g_s(x))$ into the respective boxes $[i/n,(i+1)/n]\times[0,1]$. This is effected by exploiting that $g_s(x)$ is compactly supported on $[0,1]$ for all $s \in \mathbb{N}$, which allows us to realize the desired localization according to $f^{(i)}_{X_2}(g_s(nx-i))$. The resulting localized space-filling curves are then stitched together by adding them up according to $\sum_{i=0}^{n-1} f^{(i)}_{X_2}(g_s(nx-i))$. Denoting the piecewise linear function that generates the marginal histogram

distribution p_{X_1} according to Corollary 3 as[1] $f_{X_1}^{\mathbf{z}_1}$, i.e., $f_{X_1}^{\mathbf{z}_1}\#U = p_{X_1}$, the transport map $M : x \to \left(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nf_{X_1}^{\mathbf{z}_1}(x) - i))\right)$ when applied to U generates p_{X_1,X_2} in (2.7) asymptotically in s. To see this, we first note that the second component of M pushes forward, by $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$, the random variable $f_{X_1}^{\mathbf{z}_1}\#U$ resulting from the pushforward of U by the first component of M. This allows us to read off the conditional distributions $p_{X_2|X_1}$. Specifically, thanks to the individual components in $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$ being disjointly suppported on $[i/n, (i + 1)/n]$, we can conclude that the distribution of the 2-dimensional random variable $M\#U$ satisfies $p_{X_2|X_1}\big(x_2|x_1 \in [i/n, (i + 1)/n]\big) = \sum_{k_2} w_{k_2|i}\chi_{c_{k_2}}(x_2), i \in [0 : (n - 1)]$, as desired. Next, noting that the distribution p_{X_1} of $f_{X_1}^{\mathbf{z}_1}\#U$ has components $p_{X_1}\big(x_1 \in [i/n, (i+1)/n]\big) = w_i\chi_{c_i}(x_1)$ supported disjointly on $[i/n, (i+1)/n]$, it follows from $p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2|X_1}(x_2|x_1)$ that the distribution of $M\#U$ is given by

$$p_{X_1,X_2}(x_1, x_2) = \sum_{i=0}^{n-1} p_{X_1}\big(x_1 \in [i/n, (i + 1)/n]\big) \\ p_{X_2|X_1}\big(x_2|x_1 \in [i/n, (i + 1)/n]\big),$$

which is (2.7). An example illustrating the overall construction can be found in Figure 2.6.

We are now ready to formalize the idea just described and generalize it to target distributions of arbitrary dimension. To this end, we start with a technical lemma stating that all marginal and conditional distributions of a d-dimensional histogram distribution are themselves histogram distributions, a result that was already used implicitly in the description of our main idea in the 2-dimensional case above.

Lemma 22. *Let $p_{\mathbf{X}}(\mathbf{x})$ be the pdf of the random vector $\mathbf{X} \sim \mathcal{E}[0, 1]_n^d$. Then, for all $t \in [1 : (d - 1)]$, its marginal distributions satisfy $p_{X_1,\dots,X_t}(\mathbf{x}_{[1:t]}) \in \mathcal{E}[0, 1]_n^t$. Moreover, for all $t \in [1 : (d - 1)]$ and*

[1] The choice of the superscript $\mathbf{z}_1$ in $f_{X_1}^{\mathbf{z}_1}$ will become clear in Definition 24 below.

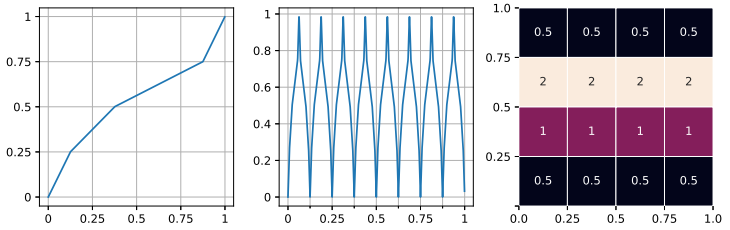


Fig. 2.5: Generating a histogram distribution via the transport map $x \to (x, f(g_s(x)))$. Left—the function $f(x)$, center—$f(g_4(x))$, right—a heatmap of the resulting histogram distribution.

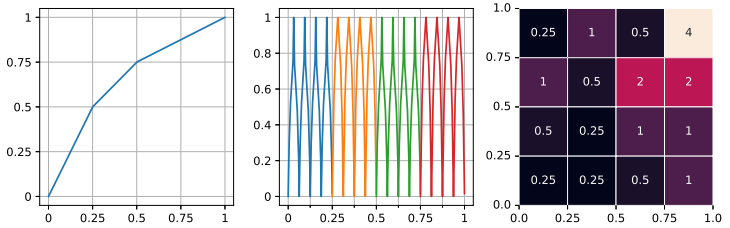


Fig. 2.6: Generating a 2-D histogram distribution via the transport map $x \to (f^{\mathbf{z}_1}_{X_1}(x), \sum_{i=0}^{3} f^{(i)}_{X_2}(g_3(4f^{\mathbf{z}_1}_{X_1}(x) - i)))$. Left—the function $f^{(1)}_{X_2} = f^{(3)}_{X_2} = f^{\mathbf{z}_1}_{X_1}$, center—$\sum_{i=0}^{3} f^{(i)}_{X_2}(g_3(4f^{\mathbf{z}_1}_{X_1}(x) - i))$, right—a heatmap of the resulting histogram distribution. We took $f^{(0)}_{X_2} = f^{(2)}_{X_2}$ to be given by the function depicted on the left in Figure 2.5.

$\mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0 : (n-1)]^t$, *defining* $c_{\mathbf{z}} = [z_1/n, (z_1 + 1)/n] \times [z_2/n, (z_2 + 1)/n] \times \cdots \times [z_t/n, (z_t + 1)/n]$, *the conditional distributions* $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}})$ *are independent of the specific value of* $\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}$ *and obey* $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}) \in \mathcal{E}[0,1]^1_n$.

Proof. The proof of the first statement follows by noting that, for all

$t \in [1:(d-1)]$,

$$\begin{aligned} p_{X_1,\ldots,X_t}(\mathbf{x}_{[1:t]}) &= \int_{[0,1]^{d-t}} \sum_{\mathbf{k}} w_{\mathbf{k}} \chi_{c_{\mathbf{k}}}(\mathbf{x}) dx_{t+1} dx_{t+2} \ldots dx_d \\ &= \sum_{\mathbf{z}} w_{\mathbf{z}} \chi_{c_{\mathbf{z}}}(\mathbf{x}_{[1:t]}), \end{aligned} \tag{2.8}$$

where $\mathbf{z} = \mathbf{k}_{[1:t]} = (z_1, z_2, \ldots, z_t)$ with $\mathbf{k}$ according to Definition 23, and $w_{\mathbf{z}} = (1/n)^{d-t} \sum_{i_{t+1},\ldots,i_d} w_{\mathbf{k}} > 0$. With $\sum_{\mathbf{k}} w_{\mathbf{k}} = n^d$ from Definition 23, we get $\sum_{\mathbf{z}} w_{\mathbf{z}} = n^{t-d} \sum_{\mathbf{k}} w_{\mathbf{k}} = n^t$, which establishes that (2.8) constitutes a valid histogram distribution in $\mathcal{E}[0,1]_n^t$.

To prove the second statement, we first note that for all $t \in [1:(d-1)]$,

$$\begin{aligned} &p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]}) \\ &= \frac{p_{X_1,\ldots,X_{t+1}}(\mathbf{x}_{[1:(t+1)]})}{p_{X_1,\ldots,X_t}(\mathbf{x}_{[1:t]})} \\ &= \frac{\sum_{(\mathbf{z},z_{t+1})} w_{(\mathbf{z},z_{t+1})} \chi_{c_{(\mathbf{z},z_{t+1})}}(\mathbf{x}_{[1:(t+1)]})}{\sum_{\mathbf{z}} w_{\mathbf{z}} \chi_{c_{\mathbf{z}}}(\mathbf{x}_{[1:t]})}. \end{aligned}$$

Next, for a given $\mathbf{z}' \in [1:(n-1)]^t$, we have

$$\begin{aligned} &p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}'}) \\ &= \frac{\sum_{z_{t+1}} w_{(\mathbf{z}',z_{t+1})} \chi_{c_{(\mathbf{z}',z_{t+1})}}(\mathbf{x}_{[1:(t+1)]})}{w_{\mathbf{z}'}} \\ &= \sum_{z_{t+1}} \frac{w_{(\mathbf{z}',z_{t+1})}}{w_{\mathbf{z}'}} \chi_{c_{z_{t+1}}}(x_{t+1}), \end{aligned}$$

which allows us to conclude that, for all $t \in [1:(d-1)]$ and $\mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0:(n-1)]^t$, the conditional distribution $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}})$ is independent of the specific value of $\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}$ and, thanks to $\sum_{z_{t+1}} \frac{w_{(\mathbf{z}',z_{t+1})}}{w_{\mathbf{z}'}} = n$, belongs to the class $\mathcal{E}[0,1]_n^1$. □

As a consequence of Lemma 22 it follows—from the chain rule—that the joint distribution of a histogram distribution $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ is fully specified by the conditional distributions $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}), t \in [1:(d-1)], \mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0:(n-1)]^t$, and the marginal distribution $p_{X_1}(x_1)$.

We next define the auxiliary functions F_r and Z_r needed in the construction of the d-dimensional generalization of the 2-dimensional transport map $M : x \to \Big(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(n f_{X_1}^{\mathbf{z}_1}(x) - i))\Big)$.

Definition 24. *For* $\mathbf{k} = [k_1, \ldots, k_t] \in [0:(n-1)]^t$, $t \in \mathbb{N}$, *define* $c_{\mathbf{k}} = \left[\frac{k_1}{n}, \frac{k_1+1}{n}\right] \times \left[\frac{k_2}{n}, \frac{k_2+1}{n}\right] \times \cdots \times \left[\frac{k_t}{n}, \frac{k_t+1}{n}\right]$. *Let* $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$, *set* $\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$, *for* $i \in [2:d]$, *and fix a histogram distribution* $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ *specified by* $p_{X_i}^{\mathbf{z}_i} = p_{X_i|X_1,\ldots,X_{i-1}}\big(x_i|\mathbf{x}_{[1:(i-1)]} \in c_{\mathbf{z}_i}\big)$, *for* $i \in [2:d]$, *and*[2] $p_{X_1}^{\mathbf{z}_1}(x_1) = p_{X_1}(x_1)$. *For* $i \in [1:d]$, *let* $f_{X_i}^{\mathbf{z}_i}$ *be the piecewise linear function that, according to Corollary 3, satisfies* $f_{X_i}^{\mathbf{z}_i} \# U = p_{X_i}^{\mathbf{z}_i}$, *and define recursively, for all* $s \in \mathbb{N}$,

$$F_r(x, \mathbf{z}_{r+1}, s) := g_s\big(n f_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) - z_r\big), \; r \in [1:(d-1)], \tag{2.9}$$

with the initialization

$$F_0(x, \mathbf{z}_1, s) := x.$$

Further, define the functions Z_r *according to*

$$Z_r(x, s) := \sum_{\mathbf{z}_r} f_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(x, \mathbf{z}_r, s)\big), \; r \in [2:d],$$

and

$$Z_1(x, s) := f_{X_1}^{\mathbf{z}_1}(x).$$

[2]Formally, $\mathbf{z}_1$, albeit not defined, would correspond to a 0-dimensional quantity. It is used throughout the chapter only for notational convenience.

With the quantities just defined, we can write

$$M : x \to \Big(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(n f_{X_1}^{\mathbf{z}_1}(x) - i)) \Big) \\ = (Z_1(x,s), Z_2(x,s)).$$

In the d-dimensional case, the space-filling transport map that takes a 1-dimensional uniform distribution into a given d-dimensional histogram distribution, or more precisely a sufficiently accurate approximation thereof, will be seen to be given by

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s)). \tag{2.10}$$

Theorem 16 below, the central result of this section, makes this formal. The material from here on up to Theorem 16 is all preparatory and technical. We recommend that it be skipped at first reading and suggest to proceed to Theorem 16, in particular the intuition behind the construction of (2.10) provided right after the proof of Theorem 16. We do recommend, however, to first visit Figure 2.7, which illustrates the F_r-functions and their role in generating the target histogram distribution.

The following lemma establishes support properties of the F_r-functions and corresponding consequences for the Z_r-functions.

Lemma 23. *Let $F_i, i \in [0\!:\!(d-1)], \mathbf{z}_i, i \in [2\!:\!d], Z_i, f_{X_i}^{\mathbf{z}_i}, i \in [1\!:\!d]$, be as in Definition 24. Then, for all $r \in [2\!:\!d]$, we have*

$$\bigcap_{\mathbf{z}_r} \operatorname{supp}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) = \varnothing,$$

and hence for every $r \in [2\!:\!d]$, for all $\mathbf{t}_r \in [0\!:\!(n-1)]^{r-1}$, it holds that

$$f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(x, \mathbf{t}_r, s)\big) > 0 \\ \implies Z_r(x,s) = f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(x, \mathbf{t}_r, s)\big), \quad x \in [0,1].$$

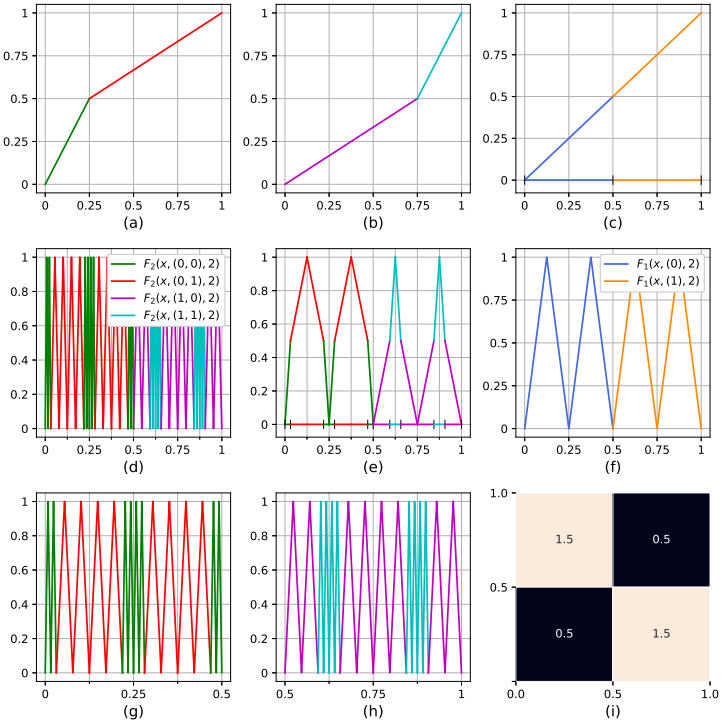


Fig. 2.7: An example illustrating the functions $F_r(x, \mathbf{z}_{r+1}, s)$ for $r = 1$ and $r = 2$. The corresponding target histogram distribution $p_{X_1,X_2}(x_1, x_2)$ is visualized in subplot (i). The function $Z_1(x, s) = f_{X_1}^{\mathbf{z}_1}(x) = x$ characterizing the marginal $p_{X_1} = U$ is shown in subplot (c). The functions $f_{X_2}^{(0)}(x)$ and $f_{X_2}^{(1)}(x)$ characterizing $p_{X_2}^{(0)}$ and $p_{X_2}^{(1)}$, respectively, are depicted in subplots (a) and (b). Subplot (e) shows $Z_2(x, 2)$ and subplot (f) visualizes $F_1(x, (0), 2)$ and $F_1(x, (1), 2)$. The functions $F_2(x, \mathbf{z}_3, 2)$ are shown in subplot (d). Subplots (g) and (h) depict zoomed-in versions of subplot (d). The functions $F_2(x, \mathbf{z}_3, 2)$ have disjoint support sets, but, in contrast to $F_1(x, (0), 2)$ and $F_1(x, (1), 2)$, their support sets are not connected. The support sets of the colored pieces in subplot (d) match those of the respective colored pieces in subplot (e), likewise for subplots (f) and (c).

Proof. The proof is through induction across r. We start with the base case $r = 2$. Since the vector $\mathbf{z}_2$ is, in fact, a scalar, we can write $\mathbf{z}_2 = z_1$ and note that

$$F_1(x, \mathbf{z}_2, s) = F_1(x, z_1, s) = g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)$$

and

$$\begin{aligned} Z_2(x, s) &= \sum_{\mathbf{z}_2} f_{X_2}^{\mathbf{z}_2}\big(g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)\big) \\ &= \sum_{z_1=0}^{n-1} f_{X_2}^{(z_1)}\big(g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)\big). \end{aligned}$$

Fix an arbitrary $\hat{x} \in [0, 1]$. Since $f_{X_1}^{\mathbf{z}_1}(x) \in [0, 1]$, it follows that $f_{X_1}^{\mathbf{z}_1}(\hat{x}) \in [t_1/n, (t_1 + 1)/n]$ for some $t_1 \in [0 : (n - 1)]$. Then, $\big(nf_{X_1}^{\mathbf{z}_1}(\hat{x}) - z_1\big) \in [0, 1]$ if and only if $z_1 = t_1$. Further, as for $x \notin [0, 1]$, $g_s(x) = 0$, we have $F_1(\hat{x}, z_1, s) = g_s\big(nf_{X_1}^{\mathbf{z}_1}(\hat{x}) - z_1\big) = 0$, for all $z_1 \neq t_1$. Combined with the fact that $\hat{x} \in [0, 1]$ was chosen arbitrarily, this implies that for every $x \in [0, 1]$, there exists a $t_1 \in [0 : (n - 1)]$ such that for all $z_1 \neq t_1$, it holds that $F_1(x, z_1, s) = 0$. This can equivalently be expressed as

$$\bigcap_{z_1=0}^{n-1} \operatorname{supp}\big(F_1(x, z_1, s)\big) = \varnothing. \tag{2.11}$$

Next, since $f_{X_2}^{z_1}(0) = 0$, for all $z_1 \in [0 : (n - 1)]$, it follows from (2.11) that, for all $t_1 \in [0 : (n - 1)]$,

$$\begin{aligned} f_{X_2}^{(t_1)}(F_1(x, t_1, s)) > 0 \implies Z_2(x, s) &= \sum_{z_1=0}^{n-1} f_{X_2}^{(z_1)}\big(F_1(x, z_1, s)\big) \\ &= f_{X_2}^{(t_1)}(F_1(x, t_1, s)), \quad x \in [0, 1]. \end{aligned}$$

This establishes the base case. To prove the induction step, we assume that the statement holds for some $r \in [2 : (d - 1)]$. Then, by the

induction assumption,

$$\bigcap_{\mathbf{z}_r} \operatorname{supp}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) = \varnothing,$$

or, equivalently, for every $x \in [0,1]$, there exists a $\mathbf{t}_r \in [0\,{:}\,(n-1)]^{r-1}$ such that for all $\mathbf{z}_r \neq \mathbf{t}_r$, we have $F_{r-1}(x, \mathbf{z}_r, s) = 0$. Now, fix an arbitrary $\hat{x} \in [0,1]$ together with its corresponding $\mathbf{t}_r \in [0\,{:}\,(n-1)]^{r-1}$ such that $F_{r-1}(\hat{x}, \mathbf{z}_r, s) = 0$, for all $\mathbf{z}_r \neq \mathbf{t}_r$, and consider

$$\begin{aligned} &F_r(\hat{x}, \mathbf{z}_{r+1}, s) \\ &= g_s\big(n f_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(\hat{x}, \mathbf{z}_r, s)\big) - z_r\big) \\ &= \begin{cases} 0, & \text{if } \mathbf{z}_r \neq \mathbf{t}_r, \\ g_s\big(n f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - z_r\big), & \text{if } \mathbf{z}_r = \mathbf{t}_r. \end{cases} \end{aligned} \tag{2.12}$$

Again, as $f_{X_r}^{\mathbf{t}_r}(x) \in [0,1]$, we can conclude that, for an arbitrarily fixed $\hat{x} \in [0,1]$, there is a $t_r \in [0\,{:}\,(n-1)]$ such that $f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) \in [t_r/n, (t_r+1)/n]$. Then, $\big(n f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - z_r\big) \in [0,1]$ if and only if $z_r = t_r$. Thanks to $g_s(x) = 0$, for $x \notin [0,1]$, (2.12) becomes

$$\begin{aligned} &F_r(\hat{x}, \mathbf{z}_{r+1}, s) \\ &= \begin{cases} 0, & \text{if } \mathbf{z}_{r+1} \neq \mathbf{t}_{r+1}, \\ g_s\big(n f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - t_r\big), & \text{if } \mathbf{z}_{r+1} = \mathbf{t}_{r+1}. \end{cases} \\ &= \begin{cases} 0, & \text{if } \mathbf{z}_{r+1} \neq \mathbf{t}_{r+1}, \\ F_r(\hat{x}, \mathbf{t}_{r+1}, s), & \text{if } \mathbf{z}_{r+1} = \mathbf{t}_{r+1}. \end{cases} \end{aligned}$$

Combined with the fact that $\hat{x} \in [0,1]$ was chosen arbitrarily, this implies that for every $x \in [0,1]$, there exists a $\mathbf{t}_{r+1} \in [0\,{:}\,(n-1)]^r$ such that for all $\mathbf{z}_{r+1} \neq \mathbf{t}_{r+1}$, it holds that $F_r(x, \mathbf{z}_{r+1}, s) = 0$. This can equivalently be expressed as

$$\bigcap_{\mathbf{z}_{r+1}} \operatorname{supp}\big(F_r(x, \mathbf{z}_{r+1}, s)\big) = \varnothing. \tag{2.13}$$

Next, since $f_{X_{r+1}}^{\mathbf{z}_{r+1}}(0) = 0$, for all $\mathbf{z}_{r+1}$, it follows from (2.13) that, for all $\mathbf{t}_{r+1} \in [0\!:\!(n-1)]^r$,

$$\begin{aligned} f_{X_{r+1}}^{\mathbf{t}_{r+1}}\big(F_r(x, \mathbf{t}_{r+1}, s)\big) > 0 \Rightarrow & Z_{r+1}(x, s) \\ & := \sum_{\mathbf{z}_{r+1}} f_{X_{r+1}}^{\mathbf{z}_{r+1}}\big(F_r(x, \mathbf{z}_{r+1}, s)\big) \\ & = f_{X_{r+1}}^{\mathbf{t}_{r+1}}\big(F_r(x, \mathbf{t}_{r+1}, s)\big),\ x \in [0, 1]. \end{aligned}$$

This concludes the proof. □

Before proceeding, we need to introduce further notation. Specifically, let $P_1 = [a, b], P_2 = [c, d]$ be intervals in $[0, 1]$, i.e., $0 \leq a < b \leq 1$ and $0 \leq c < d \leq 1$. Then, we define $P := P_1 \diamond P_2$ according to $P = [a + c(b - a), a + d(b - a)]$. Note that $|P| = |P_1||P_2|$. The $\diamond$-operation is associative in the sense of $(P_1 \diamond P_2) \diamond P_3 = P_1 \diamond (P_2 \diamond P_3)$. We further define the function $N(x, [a, b]) = a + x(b - a),\ x \in [0, 1]$, which rescales $[0, 1]$ to the interval $[a, b]$, and note that

$$N(N(x, [a, b]), [c, d]) = N(x, [c, d] \diamond [a, b]). \tag{2.14}$$

We will also need the function $N^-(x, [a, b]) = b - x(b - a), x \in [0, 1]$, which, like $N(x, [a, b])$, rescales $[0, 1]$ to the interval $[a, b]$, but does so in reverse manner, i.e., by mapping $x = 0$ to b and $x = 1$ to a. Additionally, we define the operator $S([a, b]) = [1 - b, 1 - a]$, for all $0 \leq a < b \leq 1$, which maps the interval $[a, b] \subseteq [0, 1]$ to the interval $[1 - b, 1 - a]$. Note that S is cardinality-preserving, i.e., $|S([a, b])| = (b - a) = |[a, b]|$. Moreover, we have the relation

$$1 - N(x, S([a, b])) = N^-(x, [a, b]). \tag{2.15}$$

The next lemma establishes that the Z_r-functions indeed realize the per-bin histogram distributions constituting the desired target histogram distribution.

Lemma 24. *Let* $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0 : (n-1)]^d$ *and* $\Delta_h = \left[\frac{h}{2^s}, \frac{h+1}{2^s}\right]$, $h \in [0\!:\!(2^s - 1)]$. *Set* $c_{z_i} = \left[\frac{z_i}{n}, \frac{z_i+1}{n}\right]$, *for* $i \in [1\!:\!d]$. *Let*

$\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$ *and fix* $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ *with* $p_{X_1}^{\mathbf{z}_1} := p_{X_1}$ *and* $p_{X_i}^{\mathbf{z}_i} := p_{X_i|X_1,\ldots,X_{i-1}}\big(x_i|\mathbf{x}_{[1:(i-1)]} \in c_{\mathbf{z}_i}\big)$*, for* $i \in [2:d]$*, where* $p_{X_i}^{\mathbf{z}_i} \in \mathcal{E}[0,1]_n^1$*,* $i \in [1:d]$*, has weights* $w_k^{\mathbf{z}_i}$*, for* $k \in [0:(n-1)]$*. Let* $f_{X_i}^{\mathbf{z}_i}$ *be the piecewise linear function which, according to Corollary 3, satisfies* $f_{X_i}^{\mathbf{z}_i} \# U = p_{X_i}^{\mathbf{z}_i}$*,* $i \in [1:d]$*. Define* $P_r^{\mathbf{z}_i} = \left[\frac{1}{n}\sum_{k=0}^{r-1} w_k^{\mathbf{z}_i}, \frac{1}{n}\sum_{k=0}^{r} w_k^{\mathbf{z}_i}\right]$ *and* $P_r^{\mathbf{z}_i,h} = P_r^{\mathbf{z}_i} \diamond \Delta_h$*,* $r \in [1:(n-1)]$*, and* $P_0^{\mathbf{z}_i} = \left[0, \frac{w_0^{\mathbf{z}_i}}{n}\right]$*,* $P_0^{\mathbf{z}_i,h} = P_0^{\mathbf{z}_i} \diamond \Delta_h$*. Then, for every* $k \in [2:d]$*, for all* $\mathbf{z}_k \in [0:(n-1)]^{k-1}$ *and* $\mathbf{h}_k = (h_1, h_2, \ldots, h_{k-1}) \in [0:(2^s-1)]^{k-1}$*, it holds that*

$$\begin{aligned} &Z_k(N(x, T_{k-1}), s) \\ &= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) \\ &= f_{X_k}^{\mathbf{z}_k}(x), \ \textit{for } x \in [0,1] \textit{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0, \end{aligned}$$

and

$$\begin{aligned} &Z_k(N(x, T_{k-1}), s) \\ &= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) \\ &= f_{X_k}^{\mathbf{z}_k}(1-x), \ \textit{for } x \in [0,1] \textit{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1, \end{aligned}$$

where T_i*,* $i \in [1:d]$*, is defined recursively according to*

$$T_i = \begin{cases} T_{i-1} \diamond P_i^{\mathbf{z}}, & \textit{if } \sum_{\ell=1}^{i-1} h_\ell \in 2\mathbb{N}_0 \\ T_{i-1} \diamond S(P_i^{\mathbf{z}}), & \textit{if } \sum_{\ell=1}^{i-1} h_\ell \in 2\mathbb{N}_0 + 1 \end{cases},$$

for $i \in [2:d]$*, and initialized by* $T_1 = P_0^{\mathbf{z}} \diamond P_1^{\mathbf{z}}$*, with* $P_i^{\mathbf{z}} := P_{z_i}^{\mathbf{z}_i,h_i}$ *and* $P_0^{\mathbf{z}} = [0,1]$*. Moreover,* $|T_k| = \frac{1}{2^{sk}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:k]} \in c_{\mathbf{z}_{k+1}})$*, for all* $k \in [1:(d-1)]$*.*

Proof. The proof is through induction across k. We start with the base case $k = 2$. Since the vectors $\mathbf{z}_2$ and $\mathbf{t}_2$ are, in fact, scalars, we can write $\mathbf{z}_2 = z_1$ and $\mathbf{t}_2 = t_1$. Fix $h_1 \in [0 : (2^s -$

$1)]$, $z_1 \in [0:(n-1)]$, and note that $T_1 = [0,1] \diamond P_1^{\mathbf{z}} = P_{z_1}^{\mathbf{z}_1} = P_{z_1}^{\mathbf{z}_1} \diamond \Delta_{h_1} = \left[\frac{1}{n}\sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1}, \frac{1}{n}\sum_{k=0}^{z_1} w_k^{\mathbf{z}_1}\right] \diamond \left[\frac{h_1}{2^s}, \frac{h_1+1}{2^s}\right] = \left[\frac{1}{n}\sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \frac{h_1 w_{z_1}^{\mathbf{z}_1}}{2^s n}, \frac{1}{n}\sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \frac{(h_1+1)w_{z_1}^{\mathbf{z}_1}}{2^s n}\right]$. Further, note that $|T_1| = \frac{w_{z_1}^{\mathbf{z}_1}}{2^s n} = \frac{1}{2^s} p_{\mathbf{X}}(x_1 \in c_{z_1})$. By Corollary 3, $f_{X_1}^{\mathbf{z}_1}(x)$ is linear on $P_{z_1}^{\mathbf{z}_1}$ with slope $\frac{1}{w_{z_1}^{\mathbf{z}_1}}$ and boundary points $f_{X_1}^{\mathbf{z}_1}(\frac{1}{n}\sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1}) = z_1/n$ and $f_{X_1}^{\mathbf{z}_1}(\frac{1}{n}\sum_{k=0}^{z_1} w_k^{\mathbf{z}_1}) = (z_1+1)/n$. The explicit form of $f_{X_1}^{\mathbf{z}_1}(x)$ on $P_{z_1}^{\mathbf{z}_1}$ follows from the remark after Corollary 3 as

$$f_{X_1}^{\mathbf{z}_1}(x) = \frac{x}{w_{z_1}^{\mathbf{z}_1}} - \frac{\sum_{i=0}^{z_1-1} w_i^{\mathbf{z}_1}}{n w_{z_1}^{\mathbf{z}_1}} + \frac{z_1}{n}, \quad x \in P_{z_1}^{\mathbf{z}_1}.$$

Next, since $N(x,T_1) = \frac{1}{n}\sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \frac{(x+h_1)w_{z_1}^{\mathbf{z}_1}}{2^s n}$, noting that $T_1 \subset P_{z_1}^{\mathbf{z}_1}$, we obtain

$$f_{X_1}^{\mathbf{z}_1}(N(x,T_1)) = \frac{x+h_1}{2^s n} + \frac{z_1}{n}, \quad x \in [0,1], \tag{2.16}$$

and, hence, for $x \in [0,1]$,

$$\begin{aligned}
&f_{X_2}^{(z_1)}\big(F_1(N(x,T_1),z_1,s)\big)\\
&= f_{X_2}^{(z_1)}\big(g_s\big(n f_{X_1}^{\mathbf{z}_1}(N(x,T_1)) - z_1\big)\big)\\
&= f_{X_2}^{(z_1)}(g_s((x+h_1)2^{-s}))\\
&\overset{(a)}{=} \sum_{j=0}^{2^{s-1}-1} f_{X_2}^{(z_1)}(g(2^{s-1}(x+h_1)2^{-s} - j))\\
&= \sum_{j=0}^{2^{s-1}-1} f_{X_2}^{(z_1)}(g(x/2 + h_1/2 - j))\\
&\overset{(b)}{=} f_{X_2}^{(z_1)}(g(x/2 + h_1/2 - \lfloor h_1/2 \rfloor))\\
&\overset{(c)}{=} \begin{cases} f_{X_2}^{(z_1)}(x), & \text{if } h_1 \in 2\mathbb{N}_0,\\ f_{X_2}^{(z_1)}(1-x), & \text{if } h_1 \in 2\mathbb{N}_0+1, \end{cases}
\end{aligned} \tag{2.17}$$

where we used Lemma 20 in (a), the fact that $g(x/2 + h_1/2 - j) = 0$, for all $x \in [0,1]$, for $j \neq \lfloor h_1/2 \rfloor$ in (b), and $h_1/2 - \lfloor h_1/2 \rfloor = 0$ for $h_1 \in 2\mathbb{N}_0$ and $h_1/2 - \lfloor h_1/2 \rfloor = 1/2$ for $h_1 \in 2\mathbb{N}_0 + 1$ along with $g(x) = g(1-x)$, for $x \in [0,1]$, in (c). Finally, as by Corollary 3, $f_{X_2}^{(z_1)}(x) > 0$, for all $x \in (0,1]$, it follows from (2.17) that $f_{X_2}^{(z_1)}\big(F_1(N(x,T_1),z_1,s)\big) > 0$, for all $x \in (0,1]$ for $h_1 \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $h_1 \in 2\mathbb{N}_0 + 1$. Application of Lemma 23 then yields

$$Z_2(N(x,T_1),s) = f_{X_2}^{(z_1)}\big(F_1(N(x,T_1),z_1,s)\big), \tag{2.18}$$

for all $x \in (0,1]$ for $h_1 \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $h_1 \in 2\mathbb{N}_0 + 1$. To see that (2.18) holds for $x = 0$ and $h_1 \in 2\mathbb{N}_0$, simply note that $Z_2(N(0,T_1),s) = \sum_{z_1} f_{X_2}^{(z_1)}(F_1(N(0,T_1),z_1,s))$ and $F_1(N(0,T_1),z_1,s) = g_s(h_1/2^s) = 0$, which thanks to $f_{X_2}^{(z_1)}(0) = 0$ implies $Z_2(N(0,T_1),s) = f_{X_2}^{(z_1)}(F_1(N(0,T_1),z_1,s)) = 0$. The case $x = 1$ and $h_1 \in 2\mathbb{N}_0 + 1$ follows along the exact same lines noting that $F_1(N(1,T_1),z_1,s) = g_s((h_1+1)/2^s) = 0$. This finalizes the proof of the base case.

The proof of the induction step largely follows the arguments underlying the proof of the base case. Fix $k \in \mathbb{N}$, with $k \geq 2$, and assume that for all $\mathbf{z}_k = (z_1, z_2, \ldots, z_{k-1}) \in [0:(n-1)]^{k-1}$ and $\mathbf{h}_k = (h_1, h_2, \ldots, h_{k-1}) \in [0:(2^s-1)]^{k-1}$, it holds that

$$\begin{aligned}
&Z_k(N(x,T_{k-1}),s)\\
&= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_{k-1}),\mathbf{z}_k,s))\\
&= f_{X_k}^{\mathbf{z}_k}(x), \quad \text{for } x \in [0,1] \text{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0,
\end{aligned}$$

and

$$\begin{aligned}
&Z_k(N(x,T_{k-1}),s)\\
&= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_{k-1}),\mathbf{z}_k,s))
\end{aligned}$$

$$= f_{X_k}^{\mathbf{z}_k}(1-x), \quad \text{for } x \in [0,1] \text{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1,$$

with $|T_{k-1}| = \frac{1}{2^{s(k-1)}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k})$. Fix $\mathbf{h}_{k+1} = (h_1, h_2, \ldots, h_k) \in [0 : (2^s - 1)]^k$ and $\mathbf{z}_{k+1} = (z_1, z_2, \ldots, z_k) \in [0 : (n-1)]^k$. Consider $Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s))$ on the interval

$$\begin{aligned} P_k^{\mathbf{z}} &= P_{z_k}^{\mathbf{z}_k} \diamond \Delta_{h_k} \\ &= \left[\frac{1}{n} \sum_{j=0}^{z_k - 1} w_j^{\mathbf{z}_k} + \frac{h_k w_{z_k}^{\mathbf{z}_k}}{2^s n}, \frac{1}{n} \sum_{j=0}^{z_k - 1} w_j^{\mathbf{z}_k} + \frac{(h_k+1) w_{z_k}^{\mathbf{z}_k}}{2^s n} \right]. \end{aligned}$$

We first note that

$$T_k = \begin{cases} T_{k-1} \diamond P_k^{\mathbf{z}}, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ T_{k-1} \diamond S(P_k^{\mathbf{z}}), & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases},$$

and

$$\begin{aligned} |T_k| &= \begin{cases} |T_{k-1}||P_k^{\mathbf{z}}|, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ |T_{k-1}||S(P_k^{\mathbf{z}})|, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases} \\ &= |T_{k-1}| \frac{w_{z_k}^{\mathbf{z}_k}}{2^s n} \\ &= \frac{1}{2^{s(k-1)}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k}) \\ &\qquad \frac{1}{2^s} p_{X_k | X_1, \ldots, X_{k-1}}\left(x_k \in c_{z_k} | \mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k}\right) \\ &= \frac{1}{2^{sk}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:k]} \in c_{\mathbf{z}_{k+1}}). \end{aligned}$$

We first provide the proof for the case $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$. By Corollary 3, $f_{X_k}^{\mathbf{z}_k}(x)$ is linear on $P_{z_k}^{\mathbf{z}_k}$ with slope $1/w_{z_k}^{\mathbf{z}_k}$ and boundary points $f_{X_k}^{\mathbf{z}_k}(\frac{1}{n}\sum_{j=0}^{z_k-1} w_j^{\mathbf{z}_k}) = z_k/n$ and $f_{X_k}^{\mathbf{z}_k}(\frac{1}{n}\sum_{j=0}^{z_k} w_j^{\mathbf{z}_k}) = (z_k+1)/n$. The explicit form of $f_{X_k}^{\mathbf{z}_k}$ on $P_{z_k}^{\mathbf{z}_k}$ follows from the remark after Corollary 3 as

$$f_{X_k}^{\mathbf{z}_k}(x) = \frac{x}{w_{z_k}^{\mathbf{z}_k}} - \frac{\sum_{j=0}^{z_k-1} w_j^{\mathbf{z}_k}}{n w_{z_k}^{\mathbf{z}_k}} + \frac{z_k}{n}.$$

Using $N(x,T_k)=N(N(x,P_k^{\mathbf{z}}),T_{k-1})$, which is thanks to (2.14), in the induction assumption (for $\sum_{i=1}^{k-1}h_i\in 2\mathbb{N}_0$), we get

$$\begin{aligned}
&f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_k),\mathbf{z}_k,s))\\
&=f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(N(x,P_k^{\mathbf{z}}),T_{k-1}),\mathbf{z}_k,s))\\
&=f_{X_k}^{\mathbf{z}_k}(N(x,P_k^{\mathbf{z}}))\\
&=\frac{x+h_k}{2^sn}+\frac{z_k}{n},\quad x\in[0,1].
\end{aligned}\tag{2.19}$$

Next, for $x\in[0,1]$, it follows from (2.9) and (2.19) that

$$\begin{aligned}
&f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big)\\
&=f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s(nf_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_k),\mathbf{z}_k,s))-z_k))\\
&=f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s((x+h_k)2^{-s}))\\
&\overset{(a)}{=}\sum_{j=0}^{2^{s-1}-1}f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(2^{s-1}(x+h_k)2^{-s}-j))\\
&=\sum_{j=0}^{2^{s-1}-1}f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(x/2+h_k/2-j))\\
&\overset{(b)}{=}f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(x/2+h_k/2-\lfloor h_k/2\rfloor))\\
&\overset{(c)}{=}\begin{cases}f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } h_k\in 2\mathbb{N}_0,\\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1-x), & \text{if } h_k\in 2\mathbb{N}_0+1,\end{cases}\\
&\overset{(d)}{=}\begin{cases}f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } \sum_{i=1}^{k}h_i\in 2\mathbb{N}_0,\\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1-x), & \text{if } \sum_{i=1}^{k}h_i\in 2\mathbb{N}_0+1,\end{cases}
\end{aligned}\tag{2.20}$$

where we used Lemma 20 in (a), the fact that $g(x/2+h_k/2-j)=0$, for all $x\in[0,1]$, for $j\neq\lfloor h_k/2\rfloor$ in (b), $h_k/2-\lfloor h_k/2\rfloor=0$ for $h_k\in 2\mathbb{N}_0$ and $h_k/2-\lfloor h_k/2\rfloor=1/2$ for $h_k\in 2\mathbb{N}_0+1$ along with $g(x)=g(1-x)$, for $x\in[0,1]$, in (c), and $\sum_{i=1}^{k-1}h_i\in 2\mathbb{N}_0$ in (d). Finally, as by Corollary 3, $f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x)>0$, for all $x\in(0,1]$, it follows from (2.20) that $f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big)>0$, for all $x\in(0,1]$

for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. Application of Lemma 23 then yields

$$Z_{k+1}(N(x,T_k),s) = f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big),$$

for all $x \in (0,1]$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. The boundary cases i) $x = 0$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$ and ii) $x = 1$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$ follow along the same lines as in the base case upon noting that $F_k(N(0,T_k),\mathbf{z}_{k+1},s) = g_s(h_k/2^s)$ and $F_k(N(1,T_k),\mathbf{z}_{k+1},s) = g_s((h_k+1)/2^s)$.

We proceed to the proof for the case $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$. Using (2.15) and $N(x,T_k) = N(N(x,S(P_k^{\mathbf{z}})),T_{k-1})$, which is thanks to (2.14), in the induction assumption (for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$), we get

$$\begin{aligned}
&f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_k),\mathbf{z}_k,s)) \\
&= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(N(x,S(P_k^{\mathbf{z}})),T_{k-1}),\mathbf{z}_k,s)) \\
&= f_{X_k}^{\mathbf{z}_k}(1 - N(x,S(P_k^{\mathbf{z}}))) \\
&= f_{X_k}^{\mathbf{z}_k}(N^-(x,P_k^{\mathbf{z}})) \\
&= \frac{h_k + 1 - x}{2^s n} + \frac{z_k}{n}, \quad x \in [0,1].
\end{aligned} \tag{2.21}$$

Next, for $x \in [0,1]$, it follows from (2.9) and (2.21) that

$$
\begin{aligned}
& f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big) \\
& = f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s(nf_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x,T_k),\mathbf{z}_k,s))-z_k)) \\
& = f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s((h_k+1-x)2^{-s})) \\
& \overset{(a)}{=} \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(2^{s-1}(h_k+1-x)2^{-s}-j)) \\
& = \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(h_k/2+1/2-x/2-j)) \qquad (2.22) \\
& \overset{(b)}{=} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(h_k/2+1/2-x/2-\lfloor h_k/2 \rfloor)) \\
& \overset{(c)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1-x), & \text{if } h_k \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } h_k \in 2\mathbb{N}_0+1, \end{cases} \\
& \overset{(d)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } \sum_{i=1}^k h_i \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1-x), & \text{if } \sum_{i=1}^k h_i \in 2\mathbb{N}_0+1, \end{cases}
\end{aligned}
$$

where we used Lemma 20 in (a), the fact that $g(h_k/2+1/2-x/2-j) = 0$, for all $x \in [0,1]$, for $j \neq \lfloor h_k/2 \rfloor$ in (b), $h_k/2 - \lfloor h_k/2 \rfloor = 0$ for $h_k \in 2\mathbb{N}_0$ and $h_k/2 - \lfloor h_k/2 \rfloor = 1/2$ for $h_k \in 2\mathbb{N}_0+1$ along with $g(x) = g(1-x)$, for $x \in [0,1]$, in (c), and $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0+1$ in (d). Finally, as by Corollary 3, $f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x) > 0$, for all $x \in (0,1]$, it follows from (2.22) that $f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big) > 0$, for all $x \in (0,1]$ for $\sum_{i=1}^k h_i \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $\sum_{i=1}^k h_i \in 2\mathbb{N}_0+1$. Application of Lemma 23 then yields

$$
Z_{k+1}(N(x,T_k),s) = f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x,T_k),\mathbf{z}_{k+1},s)\big),
$$

for all $x \in (0,1]$ for $\sum_{i=1}^k h_i \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $\sum_{i=1}^k h_i \in 2\mathbb{N}_0+1$. The boundary cases i) $x = 0$ and $\sum_{i=1}^k h_i \in 2\mathbb{N}_0$ and ii) $x = 1$ and $\sum_{i=1}^k h_i \in 2\mathbb{N}_0+1$ follow along the same lines as in the base case upon noting that $F_k(N(0,T_k),\mathbf{z}_{k+1},s) = g_s((h_k+1)/2^s)$ and $F_k(N(1,T_k),\mathbf{z}_{k+1},s) = g_s(h_k/2^s)$.

This concludes the proof of the induction step and thereby the overall proof. □

We continue with a corollary to Lemma 24 complementing the results on $|T_k|, k \in [1:(d-1)]$, by the corresponding expression for $|T_d|$ and specifying the range of the Z_r-functions on the domain T_d.

Corollary 4. *Let* $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$ *and* $\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$. *Fix* $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$, *and for all* $\mathbf{h}_d = (h_1, h_2, \ldots, h_d) \in [0:(2^s-1)]^d$, *let* $T_k, k \in [1:d]$, *be defined as in Lemma 24. Then, it holds that* $|T_d| = \frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. *Moreover, for every* $k \in [1:d]$, *for all* $x \in T_d$, $Z_k(x,s) \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right]$.

Proof. We first prove the statement on $|T_d|$ and start by noting that, owing to Lemma 24,

$$|T_{d-1}| = \frac{1}{2^{s(d-1)}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}).$$

With

$$\begin{aligned} P_{z_d}^{\mathbf{z}_d,h_d} &= P_{z_d}^{\mathbf{z}_d} \diamond \Delta_{h_d} \\ &= \left[\frac{1}{n}\sum_{j=0}^{z_d-1} w_j^{\mathbf{z}_d} + \frac{h_d w_{z_d}^{\mathbf{z}_d}}{2^s n}, \frac{1}{n}\sum_{j=0}^{z_d-1} w_j^{\mathbf{z}_d} + \frac{(h_d+1) w_{z_d}^{\mathbf{z}_d}}{2^s n}\right] \end{aligned}$$

and $|P_{z_d}^{\mathbf{z}_d,h_d}| = |S(P_{z_d}^{\mathbf{z}_d,h_d})|$, we get

$$\begin{aligned} |T_d| &= \begin{cases} |T_{d-1}||P_{z_d}^{\mathbf{z}_d,h_d}|, & \text{if } \sum_{i=1}^{d-1} h_i \in 2\mathbb{N}_0 \\ |T_{d-1}||S(P_{z_d}^{\mathbf{z}_d,h_d})|, & \text{if } \sum_{i=1}^{d-1} h_i \in 2\mathbb{N}_0+1 \end{cases} \\ &= \frac{1}{2^{s(d-1)}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d})\, \frac{w_{z_d}^{\mathbf{z}_d}}{2^s n} \\ &= \frac{1}{2^{s(d-1)}}\, p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}) \\ &\qquad \frac{1}{2^s} p_{X_d|X_1,\ldots,X_{d-1}}\big(x_d \in c_{z_d} | \mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}\big) \\ &= \frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}}). \end{aligned}$$

This establishes the first statement.

To prove the second statement, we first note that, for $k = 1$, by (2.16), $Z_1(x,s) \in \left[\frac{z_1}{n} + \frac{h_1}{2^s n}, \frac{z_1}{n} + \frac{h_1+1}{2^s n}\right]$, for all $x \in T_1$. Next, for every $k \in [2\!:\!d]$, for all T_{k-1}, by Lemma 24, it holds that

$$Z_k(N(x,T_{k-1}),s) = \begin{cases} f_{X_k}^{\mathbf{z}_k}(x), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ f_{X_k}^{\mathbf{z}_k}(1-x), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases},$$

for all $x \in [0,1]$. Now, arbitrarily fix $z_k \in [0:(n-1)], h_k \in [0:(2^s-1)]$ and consider

$$T_k = \begin{cases} T_{k-1} \diamond P_k^{\mathbf{z}}, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ T_{k-1} \diamond S(P_k^{\mathbf{z}}), & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases}$$

with $P_k^{\mathbf{z}} = P_{z_k}^{\mathbf{z}_k, h_k}$. With (2.14), this yields, for all $x \in [0,1]$,

$$\begin{aligned} &Z_k(N(x,T_k),s) \\ &= \begin{cases} Z_k(N(N(x,P_k^{\mathbf{z}}),T_{k-1}),s), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ Z_k(N(N(x,S(P_k^{\mathbf{z}})),T_{k-1}),s), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases} \\ &= \begin{cases} f_{X_k}^{\mathbf{z}_k}(N(x,P_k^{\mathbf{z}})), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ f_{X_k}^{\mathbf{z}_k}(1-N(x,S(P_k^{\mathbf{z}}))), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases}. \end{aligned}$$

Now, by (2.19), it follows for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$ that

$$\begin{aligned} &f_{X_k}^{\mathbf{z}_k}(N(x,P_k^{\mathbf{z}})) \\ &= \frac{x+h_k}{2^s n} + \frac{z_k}{n} \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right], \text{ for } x \in [0,1], \end{aligned}$$

and analogously, for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$, by (2.21),

$$\begin{aligned} &f_{X_k}^{\mathbf{z}_k}(1-N(x,S(P_k^{\mathbf{z}}))) \\ &= f_{X_k}^{\mathbf{z}_k}(N^-(x,P_k^{\mathbf{z}})) \\ &= \frac{h_k+1-x}{2^s n} + \frac{z_k}{n} \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right], \text{ for } x \in [0,1]. \end{aligned}$$

We have hence shown that, for all $k \in [1:d]$, $Z_k(x,s) \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right]$, for all $x \in T_k$. The proof is completed upon noting that $T_d \subseteq T_k$, for all $k \in [1:d]$. $\square$

We are now ready to state the main result of this section, namely that the piecewise linear map

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s))$$

transports a 1-dimensional uniform distribution in a space-filling manner to an arbitrarily close approximation of any high-dimensional histogram distribution.

Theorem 16. *For every distribution $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]^d_n$, the corresponding transport map*

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s)) \qquad (2.23)$$

satisfies

$$W(M\#U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}.$$

Proof. Let $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$, $\Delta_h = \left[\frac{h}{2^s}, \frac{h+1}{2^s}\right]$ with $h \in [0:(2^s-1)]$, and $\mathbf{h} = (h_1, h_2, \ldots, h_d) \in [0:(2^s-1)]^d$. With $c_{z_i} = [\frac{z_i}{n}, \frac{z_i+1}{n}]$, $i \in [1:d]$, let $c^{\mathbf{h}}_{\mathbf{z}} = \bigtimes_{i=1}^{d}(c_{z_i} \diamond \Delta_{h_i})$. Let T_d be defined as in Lemma 24. By Corollary 4, $M : T_d \to c^{\mathbf{h}}_{\mathbf{z}}$ and $|T_d| = \frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. We hence get $(M\#U)(\mathbf{x} \in c^{\mathbf{h}}_{\mathbf{z}}) = |T_d| = \frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. This establishes that the map M transports probability mass $\frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$ to the cube $c^{\mathbf{h}}_{\mathbf{z}}$ of volume $\left(\frac{1}{n2^s}\right)^d$, for all $\mathbf{z}$. As $p_{\mathbf{X}}$ is a histogram distribution, it is uniformly distributed on its constituent cubes $c_{\mathbf{z}}$, which, in turn, implies that the amount of probability mass it exhibits on each subcube $c^{\mathbf{h}}_{\mathbf{z}}$ of $c_{\mathbf{z}}$ is given by $\frac{1}{2^{sd}}\, p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. The map M, when pushing forward U, therefore transports exactly the right amount of probability mass to each cube $c^{\mathbf{h}}_{\mathbf{z}}$ for a coupling between $p_{\mathbf{X}}$

and $M\#U$ to exist. Combining this with $\|\mathbf{x}-\mathbf{y}\| \leq \frac{\sqrt{d}}{n2^s}$, for all points $\mathbf{x}, \mathbf{y}$ in a d-dimensional cube of side length $n^{-1}2^{-s}$, it follows from Definition 21 that

$$W(M\#U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \qquad \square$$

Theorem 16 was proven in Perekrestenko et al. (2020) for $d = 2$. We remark that a space-filling approach for increasing distribution dimensionality was first described by Bailey and Telgarsky in Bailey and Telgarsky (2018). Specifically, the construction in Bailey and Telgarsky (2018) generates uniform target distributions of arbitrary dimension based on the transport map $M : x \rightarrow (x, g_s(x), g_{2s}(x), \dots)$. The generalization introduced in this chapter is capable of producing arbitrary histogram target distributions through space-filling transport maps that build on several key ideas, the first two of which are best illustrated by revisiting the 2-dimensional case with corresponding transport map $M : x \rightarrow (f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nf_{X_1}^{\mathbf{z}_1}(x) - i)))$. First, M in its second component composes the function $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$ with its first component $f_{X_1}^{\mathbf{z}_1}(x)$. Formally, this idea is also present in the Bailey-Telgarsky map, where the second component $g_s(x)$ can be interpreted as a trivial composition of $g_s(\cdot)$ with the first component, x. It is, in fact, this composition idea that leads to the space-filling property. Second, $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nf_{X_1}^{\mathbf{z}_1}(x) - i))$ yields localization through squeezing and shifting of the $f_{X_2}^{(i)}$. This idea allows to realize different marginal distributions for different horizontal histogram bins (see the rightmost subplot in Fig. 2.6) and is not present in the Bailey-Telgarsky construction as, owing to the target distribution being uniform, there is no concept of histogram distributions. Taken together the two ideas just described allow to generate arbitrary marginal histogram distributions $p_{X_2|X_1}(x_2|x_1)$, which are then combined—through the chain rule—with the histogram distribution $p_{X_1}(x_1)$ to the overall target histogram distribution $p_{X_1,X_2}(x_1, x_2)$.

A further idea underlying our transport map construction becomes transparent in the general d-dimensional case. Specifically, in taking

the space-filling idea to higher dimensions, we note that in the Bailey-Telgarsky map $M : x \rightarrow (x, g_s(x), g_{2s}(x), \dots)$, the third component, $g_{2s}(x)$ can actually be interpreted as a composition of $g_s(\cdot)$ with the second component $g_s(x)$, simply as $g_s(g_s(x)) = g_{2s}(x)$. Likewise, as already noted in the previous paragraph, the second component, $g_s(x)$, is a composition of $g_s(\cdot)$ with the first component, x. This insight informs the recursive definition of the F_r-functions according to (2.9), which, modulo the shaping by the localized $f_{X_r}^{\mathbf{z}_r}$-functions, can be seen to exhibit this g_s-composition property as well. The $Z_r(x, s)$-functions constituting the components of our transport map (2.23) are then obtained by applying the localization idea as described above for the 2-dimensional case. There is, however, an important difference between localization in the 2-dimensional case and in the general d-dimensional case. This is best seen by inspecting the 3-dimensional case illustrated in Figure 2.7. Specifically, whereas in the 2-dimensional case the F_r-functions are contiguously supported (see subplot (f)), in the 3-dimensional case, as illustrated in subplot (d), the support sets are disjointed, but exhibit a periodic pattern. Going to higher dimensions yields a fractal-like support set picture. We emphasize that this support set structure is a consequence of interlacing the self-compositions of the g_s-functions with the localized per-bin histogram-distribution shaping functions $f_{X_r}^{\mathbf{z}_r}$.

We finally note that the transport map M in Theorem 16 can be interpreted as a transport operator in the sense of optimal transport theory Peyré and Cuturi (2019); Villani (2008), with the source distribution being 1-dimensional and the target-distribution d-dimensional. What is special here is that the transport operator acts between spaces of different dimensions and does so in a space-filling manner McCann and Pass (2020).

2.6. REALIZATION OF TRANSPORT MAP THROUGH QUANTIZED NETWORKS

This section is concerned with the realization of the transport map M by ReLU networks. In particular, we shall consider networks with quantized weights, for three reasons. First, in practice network weights can not be stored as real numbers on a computer, but rather have to be encoded with a finite number of bits. Second, we want to convince ourselves that the space-filling property of the transport map, brittle as it seems, is, in fact, not dependent on the network weights being real numbers. Third, we will be able to develop a relationship, presented in Section 2.9, between the complexity of target distributions and the complexity of the ReLU networks realizing the corresponding transport maps. Specifically, complexity will be quantified through the number of bits needed to encode the distribution and the network, respectively, to within a prescribed accuracy.

We will see that ReLU networks with quantized weights generate histogram distributions with quantized weights, referred to as quantized histogram distributions in the following. In Section 2.7, we will then study the approximation of general distributions by quantized histogram distributions. Finally, in Section 2.8, we put everything together and characterize the error incurred when approximating arbitrary target distributions by the transportation of a 1-dimensional uniform distribution through a ReLU network with quantized weights.

Before proceeding, we need to define quantized histogram distributions and quantized networks. We start with scalar distributions.

Definition 25. *Let $\delta = 1/A$, for some $A \in \mathbb{N}$. A random variable X is said to have a δ-quantized histogram distribution of resolution n on $[0,1]$, denoted as $X \sim \widetilde{\mathcal{E}}_\delta[0,1]_n^1$, if its pdf is given by*

$$p(x) = \sum_{k=0}^{n-1} w_k \chi_{[k/n,(k+1)/n]}(x), \quad \sum_{k=0}^{n-1} w_k = n,$$
$$w_k = \delta m_k > 0, \;\; m_k \in \mathbb{N}, \;\; \textit{for all} \;\; k \in [0\!:\!(n-1)].$$

We extend this definition to random vectors by saying that a random vector has a δ-quantized histogram distribution, if all its conditional (1-dimensional) distributions $p_{X_i}^{\mathbf{z}_i}$ are δ-quantized histogram distributions.

Definition 26. *Let $\delta = 1/A$, for some $A \in \mathbb{N}$. A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$ is said to have a δ-quantized histogram distribution of resolution n on the d-dimensional unit cube, denoted as $\mathbf{X} \sim \widetilde{\mathcal{E}}_\delta[0,1]_n^d$, if $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$ with $p_{X_i}^{\mathbf{z}_i} \in \widetilde{\mathcal{E}}_\delta[0,1]_n^1$, for every $i \in [1\!:\!d]$, for all $\mathbf{z}_i$.*

We continue with the definition of quantized ReLU networks.

Definition 27. *For $\delta > 0$, we say that a ReLU network is δ-quantized if each of its weights is of one of the following two types. A weight w is of Type 1 if $w \in (\delta\mathbb{Z} \cap [-1/\delta, 1/\delta])$ and of Type 2 if $\frac{1}{w} \in (\delta\mathbb{Z} \cap [-1/\delta, 1/\delta])$.*

Formally, the goal of this section is to find, for fixed $p_{\mathbf{X}} \in \widetilde{\mathcal{E}}_\delta[0,1]_n^d$, a quantized ReLU network Φ such that $\Phi\#U$ approximates $p_{\mathbf{X}}$ to within a prescribed accuracy. To this end, we start with an auxiliary lemma, which constructs the building blocks of such networks.

Lemma 25. *For every δ-quantized $p_{\mathbf{X}} \in \widetilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, the map $M^r : \mathbb{R}^{n^r+r} \to \mathbb{R}^{n^{r+1}+r+1}$, $r \in [0\!:\!(d-1)]$, defined as*

$$M^0 : F_0(x, \mathbf{z}_1, s) \\ \to \Big(F_1(x, \mathbf{z}_2^1, s), F_1(x, \mathbf{z}_2^2, s), \ldots, F_1(x, \mathbf{z}_2^n, s), Z_1(x, s)\Big),$$

and, for $r \in [1\!:\!(d-1)]$,

$$M^r : \Big(F_r(x, \mathbf{z}_{r+1}^1, s), F_r(x, \mathbf{z}_{r+1}^2, s), \ldots, F_r(x, \mathbf{z}_{r+1}^{n^r}, s), Z_1(x, s), \\ Z_2(x, s), \ldots, Z_r(x, s)\Big) \\ \to \Big(F_{r+1}(x, \mathbf{z}_{r+2}^1, s), F_{r+1}(x, \mathbf{z}_{r+2}^2, s), \ldots, F_{r+1}(x, \mathbf{z}_{r+2}^{n^{r+1}}, s), \\ Z_1(x, s), Z_2(x, s), \ldots, Z_{r+1}(x, s)\Big),$$

is realizable through a Δ-quantized ReLU network $\Psi^{M^r} \in \mathcal{NN}_{n^r+r,n^{r+1}+r+1}$ with $\mathcal{M}(\Psi^{M^r}) = \mathcal{O}(n^{r+2} + sn^{r+1})$ and $\mathcal{L}(\Psi^{M^r}) = s+3$. Here, $\Delta = \frac{\delta}{n}$ and the vectors $\mathbf{z}_r^i \in [0\!:\!(n-1)]^{r-1}$, $i \in [1\!:\!n^{r-1}]$, are in natural order[3] with respect to i.

Proof. We start with auxiliary results needed in the proof and then proceed to establish the statement for the cases $r = 0$ and $r \geq 1$ separately. According to Corollary 3, for every $k \in [1:d]$, for all $\mathbf{z}_k \in [0\!:\!(n-1)]^{k-1}$, $f_{X_k}^{\mathbf{z}_k}(x)$ can be realized through a ReLU network $\Phi^{\mathbf{z}_k} : \mathbb{R} \to \mathbb{R} \in \mathcal{NN}_{1,1}$ given by

$$\Phi^{\mathbf{z}_k} : x \to \frac{1}{w_0}\rho(x) + \sum_{i=1}^{n-1}\Big(\frac{1}{w_i} - \frac{1}{w_{i-1}}\Big)\rho\Big(x - \frac{1}{n}\sum_{j=0}^{i-1} w_j\Big),$$

and satisfying $\mathcal{M}(\Phi^{\mathbf{z}_k}) \leq 4n-2$, $\mathcal{L}(\Phi^{\mathbf{z}_k}) = 2$. For $\Delta = \frac{\delta}{n}$, the network $\Phi^{\mathbf{z}_k}$ is Δ-quantized with the weights $\frac{1}{w_0}$, $\frac{1}{w_i}$, and $\frac{1}{w_{i-1}}$ of Type 2, and the weights $\frac{1}{n}\sum_{j=0}^{i-1} w_j$ of Type 1. The networks $\Phi_i^{\mathbf{z}_k}(x)$ implementing $\big(nf_{X_k}^{\mathbf{z}_k}(x) - i\big)$ are in $\mathcal{NN}_{1,1}$ and have $\mathcal{M}(\Phi_i^{\mathbf{z}_k}) \leq 4n-1$, $\mathcal{L}(\Phi_i^{\mathbf{z}_k}) = 2$, with their weights all of either Type 1 or Type 2 w.r.t. Δ-quantization. The network $\Psi_g^s(x)$ realizing $g_s(x)$ (see Section 2.3) is in $\mathcal{NN}_{1,1}$ with $\mathcal{M}(\Psi_g^s) = 11s-3$, $\mathcal{L}(\Psi_g^s) = s+1$, and with all its weights in $\{-4,-2,-1,1,2,4\}$, which are, again, of Type 1 w.r.t. Δ-quantization. It follows from (Elbrächter et al., 2021, Lemma II.3) that the networks $\Psi_{i,s}^{\mathbf{z}_k} = \Psi_g^s(\Phi_i^{\mathbf{z}_k})$ are in $\mathcal{NN}_{1,1}$ with $\mathcal{M}(\Psi_{i,s}^{\mathbf{z}_k}) \leq 8n + 22s - 8$ and $\mathcal{L}(\Psi_{i,s}^{\mathbf{z}_k}) = s+3$.

We are now ready to prove the statement for $r = 0$. Here, $M^0 : \mathbb{R} \to \mathbb{R}^{n+1}$ with

$$\begin{aligned} M^0 : F_0(x, \mathbf{z}_1, s) \\ \to \Big(F_1(x, \mathbf{z}_2^1, s), F_1(x, \mathbf{z}_2^2, s), \ldots, F_1(x, \mathbf{z}_2^n, s), Z_1(x, s)\Big), \end{aligned}$$

[3] e.g., for $n = 2, r = 3$, the order is $\mathbf{z}_3^1 = (0,0), \mathbf{z}_3^2 = (0,1), \mathbf{z}_3^3 = (1,0), \mathbf{z}_3^4 = (1,1)$.

or equivalently

$$M^0 : x \to \Big(g_s\big(nf_{X_1}^{\mathbf{z}_1}(x)\big), g_s\big(nf_{X_1}^{\mathbf{z}_1}(x)-1\big), \ldots, g_s\big(nf_{X_1}^{\mathbf{z}_1}(x)-(n-1)\big), f_{X_1}^{\mathbf{z}_1}(x)\Big).$$

The networks $\Psi_{i,s}^{\mathbf{z}_1}$ realizing the components $g_s\big(nf_{X_1}^{\mathbf{z}_1}(x)-i\big), i \in [0\!:\!(n-1)]$, of the mapping M^0 all have depth $s+3$, whereas the network $\Phi^{\mathbf{z}_1}$ implementing the last component of M^0, $f_{X_1}^{\mathbf{z}_1}(x)$, has depth 2. As we want to apply (Elbrächter et al., 2021, Lemma II.5), we hence need to augment $\Phi^{\mathbf{z}_1}$ to depth $s+3$. This is effected by exploiting that $\Phi^{\mathbf{z}_1}(x) \geq 0,\ \forall x \in \mathbb{R}$, which allows us to retain the input-output relation realized by the network while amending it by multiplications by 1 (acting as affine transformations) interlaced by applications of ρ for an overall depth of $s+3$. This leads to the augmented network $\widetilde{\Phi}^{\mathbf{z}_1} = \rho \circ \ldots \circ \rho \circ \Phi^{\mathbf{z}_1}$, with $\mathcal{M}(\widetilde{\Phi}^{\mathbf{z}_1}) \leq 4n+s-1$, $\mathcal{L}(\widetilde{\Phi}^{\mathbf{z}_1}) = s+3$. Application of (Elbrächter et al., 2021, Lemma II.5) now allows us to conclude that the network $\Psi^{M^0} = \Big(\Psi_{0,s}^{\mathbf{z}_1}, \Psi_{1,s}^{\mathbf{z}_1}, \ldots, \Psi_{n-1,s}^{\mathbf{z}_1}, \widetilde{\Phi}^{\mathbf{z}_1}\Big)$ realizing the map M^0 is in $\mathcal{NN}_{1,1}$ and satisfies $\mathcal{M}(\Psi^{M^0}) = \mathcal{O}(n^2+sn)$, $\mathcal{L}(\Psi^{M^0}) = s+3$. This proves the statement for $r=0$.

We proceed to the proof for the case $r \geq 1$. To this end, we use (2.9) to write the map $M^r : \mathbb{R}^{n^r+r} \to \mathbb{R}^{n^{r+1}+r+1}$, for $r \in [1\!:\!(d-1)]$, as follows

$$\begin{aligned} &M^r : (y_1, y_2, \ldots, y_{n^r+r}) \\ &\to \Big(\Big[g_s\Big(nf_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i) - k\Big)\Big]_{(i,k)\in([1:n^r],[0:(n-1)])}, y_{n^r+1}, \\ &\qquad \ldots, y_{n^r+r}, \sum_{i\in[1:n^r]} f_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i)\Big), \end{aligned}$$

where the notation $[h(i,k)]_{(i,k)\in([1:n^r],[0:(n-1)])}$ designates the sequence $h(i,k)$ with (i,k) ranging over $([1:n^r],[0:(n-1)])$, with ordering according to $\Big([h(1,k)]_{k\in[0:(n-1)]}, [h(2,k)]_{k\in[0:(n-1)]}, \ldots, [h(n^r,k)]_{k\in[0:(n-1)]}\Big)$.

As discussed above, each $g_s\big(nf^{\mathbf{z}^i_{r+1}}_{X_{r+1}}(y_i) - k\big)$ can be realized by a network $\Psi^{\mathbf{z}^i_{r+1}}_{k,s} \in \mathcal{NN}_{1,1}$ with $\mathcal{M}(\Psi^{\mathbf{z}^i_{r+1}}_{k,s}) \leq 8n + 22s - 8$, $\mathcal{L}(\Psi^{\mathbf{z}^i_{r+1}}_{k,s}) = s + 3$. We will also need the identity networks $\Phi^{s+3}_{id}(x) = (\rho \circ \cdots \circ \rho)(x) = x$, for all $x \geq 0$, with $\mathcal{M}(\Phi^{s+3}_{id}) = s+3$, $\mathcal{L}(\Phi^{s+3}_{id}) = s + 3$. Finally, by (Elbrächter et al., 2021, Lemma II.6), there exists a network Ψ^{Σ} realizing the function $\sum_{i\in[1:n^r]} f^{\mathbf{z}^i_{r+1}}_{X_{r+1}}(y_i)$, and with $\Psi^{\Sigma} \in \mathcal{NN}_{n^r,1}$, $\mathcal{M}(\Psi^{\Sigma}) \leq 4n^{r+1}$, $\mathcal{L}(\Psi^{\Sigma}) = 2$. We shall also need the extension of Ψ^{Σ} to a network of depth $s + 3$ according to $\widetilde{\Psi}^{\Sigma} = \rho \circ \ldots \circ \rho \circ \Psi^{\Sigma}$ with $\mathcal{M}(\widetilde{\Psi}^{\Sigma}) \leq 4n^{r+1} + s + 1$, $\mathcal{L}(\widetilde{\Psi}^{\Sigma}) = s + 3$. The proof is now concluded by realizing the map M^r as a ReLU network Ψ^{M^r} according to

$$\begin{aligned}
&\Psi^{M^r}(y_1, y_2, \ldots, y_{n^r+r}) \\
&= \Big(\Big[\Psi^{\mathbf{z}^i_{r+1}}_{k,s}(y_i)\Big]_{(i,k)\in([1:n^r],[0:(n-1)])}, \Phi^{s+3}_{id}(y_{n^r+1}), \ldots, \\
&\qquad\qquad \Phi^{s+3}_{id}(y_{n^r+r}), \widetilde{\Psi}^{\Sigma}(y_1, \ldots, y_{n^r}) \Big).
\end{aligned}$$

Application of (Elbrächter et al., 2021, Lemma II.5) now yields $\Psi^{M^r} \in \mathcal{NN}_{n^r+r,n^{r+1}+r+1}$ with $\mathcal{M}(\Psi^{M^r}) = \mathcal{O}(n^{r+2} + sn^{r+1})$, $\mathcal{L}(\Psi^{M^r}) = s + 3$. □

The next result characterizes the ReLU networks realizing the transport map and quantifies their size in terms of connectivity and depth.

Lemma 26. *For every $p_{\mathbf{X}} \in \widetilde{\mathcal{E}}_{\delta}[0, 1]^d_n$ with $d > 1$, the corresponding transport map*

$$M : x \to (Z_1(x, s), Z_2(x, s), \ldots, Z_d(x, s))$$

can be realized through a Δ-quantized ReLU network $\Psi^M \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi^M) = (s + 3)d - s - 1$, and $\Delta = \frac{\delta}{n}$.

Proof. Consider the map $M' := M^{d-2} \circ M^{d-1} \circ \cdots \circ M^0$,

$$M' : x \to \Big(F_{d-1}(x, \mathbf{z}_d^1, s), F_{d-1}(x, \mathbf{z}_d^2, s), \\ \ldots, F_{d-1}(x, \mathbf{z}_d^{n^{d-1}}, s), Z_1(x, s), Z_2(x, s), \ldots, Z_{d-1}(x, s) \Big),$$

where the M^r, $r \in [0 : (d-2)]$, are as defined in Lemma 25. Since by Lemma 25, $M^r, r \in [0 : (d-2)]$, can be realized by a network with connectivity $\mathcal{O}(n^{r+2} + sn^{r+1})$ and depth $s+3$, it follows from (Elbrächter et al., 2021, Lemma II.3) that the map M' can be implemented by a network $\Psi' \in \mathcal{NN}_{1,n^{d-1}+d-1}$, with $\mathcal{M}(\Psi') = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi') = (s+3)(d-1)$; here, we used $\sum_{k=0}^{d-2} \mathcal{O}(n^{k+2} + sn^{k+1}) = \mathcal{O}(n^d + sn^{d-1})$. Next, consider the map

$$S : \Big(y_1, \ldots, y_{n^{d-1}+d-1} \Big) \to \Big(\rho(y_{n^{d-1}+1}), \rho(y_{n^{d-1}+2}), \ldots, \\ \rho(y_{n^{d-1}+d-1}), \rho(\sum_{i \in [1:n^{d-1}]} y_i) \Big),$$

and note that by (Elbrächter et al., 2021, Lemma II.5), there exists a network $\Psi^S \in \mathcal{NN}_{n^{d-1}+d-1,d}$ with $\mathcal{M}(\Psi^S) \leq n^{d-1} + 2d - 1$, and $\mathcal{L}(\Psi^S) = 2$ realizing S. The proof is concluded by noting that, thanks to (Elbrächter et al., 2021, Lemma II.3), the desired map $M = S \circ M'$ is realized by the network $\Psi^M := \Psi^S(\Psi'(x))$, $\Psi^M \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi') = (s+3)d - s - 1$. Moreover, the weights of Ψ^M are either of Type 1 or Type 2 w.r.t. Δ-quantization. □

We are now ready to state the main result of this section, namely that for every quantized histogram distribution $p_{\mathbf{X}}$ and every $\varepsilon > 0$, there exists a quantized ReLU network Ψ satisfying $W(\Psi \# U, p_{\mathbf{X}}) \leq \varepsilon$. In particular, we also quantify the dependence of ε on the resolution n and the dimension d of $p_{\mathbf{X}}$ as well as the depth of the network Ψ.

Theorem 17. *For every δ-quantized $p_{\mathbf{X}} \in \widetilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, there exists a Δ-quantized ReLU network $\Psi \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Psi) = \mathcal{O}(n^d +$*

sn^{d-1}), $\mathcal{L}(\Psi) = (s+3)d - s - 1$, *and* $\Delta = \frac{\delta}{n}$, *such that*

$$W(\Psi \# U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \tag{2.24}$$

Proof. By Lemma 26, for every $p_{\mathbf{X}} \in \widetilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, the corresponding transport map

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s))$$

can be realized through a Δ-quantized ReLU network $\Psi^M \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi^M) = (s+3)d - s - 1$, and $\Delta = \frac{\delta}{n}$. Moreover, as $\widetilde{\mathcal{E}}_\delta[0,1]_n^d \subset \mathcal{E}[0,1]_n^d$, it follows from Theorem 16 that

$$W(\Psi^M \# U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \qquad \square$$

We note that for fixed histogram resolution n, the upper bound on the approximation error (2.24) decays exponentially in s and hence in network depth $\mathcal{L}(\Psi)$. In particular, choosing $s \sim n$, guarantees that the error in Theorem 17 decays exponentially in n while the connectivity of the network is in $\mathcal{O}(n^d)$; this behavior is asymptotically optimal as the number of parameters in $\widetilde{\mathcal{E}}_\delta[0,1]_n^d$ is of the same order.

2.7. APPROXIMATION OF ARBITRARY DISTRIBUTIONS ON $[0,1]^D$ BY QUANTIZED HISTOGRAM DISTRIBUTIONS

This section is concerned with the approximation of arbitrary distributions ν supported on $[0,1]^d$ by δ-quantized histogram distributions of resolution n as defined in the previous section.

Define the k-dimensional subcube $c_{\mathbf{i}_k} = [i_1/n, (i_1+1)/n] \times [i_2/n, (i_2+1)/n] \times \cdots \times [i_k/n, (i_k+1)/n]$, where $\mathbf{i}_k =$

$(i_1, i_2, \ldots, i_k) \in [0\!:\!(n-1)]^k$, and its corner point

$$p_{\mathbf{i}_k} = \left(\frac{i_1}{n}, \frac{i_2}{n}, \cdots, \frac{i_k}{n}\right).$$

Next, we discretize the domain $[0,1]^d$ into the subcubes $c_{\mathbf{i}_d}$ and characterize the amount of probability mass ν assigns to the individual subcubes. First, set

$$m_{\mathbf{i}_d} := \nu(c_{\mathbf{i}_d}).$$

Then, for $k \in [1:(d-1)]$, we define the projections $P_k : \mathbb{R}^d \to \mathbb{R}^k, (x_1, \ldots, x_k, \ldots, x_d) \mapsto (x_1, \ldots, x_k)$ and the corresponding k-dimensional marginals $\nu_k := P_k \# \nu$ with weights

$$m_{\mathbf{i}_k} := \nu_k(c_{\mathbf{i}_k}).$$

It will also be useful to define conditional masses according to $n_{i_1} = m_{i_1}$ and, for $k \in [2\!:\!d]$, for all[4] $\mathbf{i}_{k-1}$ with $m_{\mathbf{i}_{k-1}} \neq 0$,

$$n_{\mathbf{i}_k} := \frac{m_{\mathbf{i}_k}}{m_{\mathbf{i}_{k-1}}}.$$

For $m_{\mathbf{i}_{k-1}} = 0$, we can, in principle, set the conditional masses arbitrarily, but, for concreteness, we choose

$$n_{\mathbf{i}_k} := \frac{1}{n}.$$

Now that we have defined the masses $m_{\mathbf{i}_k}$ and the conditional masses $n_{\mathbf{i}_k}$ for the distribution ν, we can proceed to derive the masses $\widetilde{m}_{\mathbf{i}_k}$ and $\widetilde{n}_{\mathbf{i}_k}$ of the corresponding δ-quantized histogram distribution. Denote the index of the subcube with the highest (original) mass in the first coordinate as[5]

$$i_1^{*(\mathbf{i}_0)} := \underset{i_1 \in [0:(n-1)]}{\arg\max}\ m_{i_1}. \tag{2.25}$$

[4] Throughout, we use the symbols $\mathbf{i}_1$ and i_1 interchangeably.

[5] Formally, $\mathbf{i}_0$, albeit not defined, would correspond to a 0-dimensional quantity. It is used throughout the chapter only for notational consistency.

If there are multiple subcubes with the same maximal mass, simply pick one of them (it does not matter which one). Now, for $k = 1$ and $i_1 \neq i_1^{*(\mathbf{i}_0)}$, we choose the quantized masses as follows,

$$\widetilde{m}_{i_1} := \widetilde{n}_{i_1} := \begin{cases} \delta\lceil \frac{1}{\delta} m_{i_1} \rceil, & \text{if } m_{i_1} > 0 \\ \delta, & \text{if } m_{i_1} = 0 \end{cases},$$

and for $i_1 = i_1^{*(\mathbf{i}_0)}$,

$$\widetilde{m}_{i_1^{*(\mathbf{i}_0)}} := \widetilde{n}_{i_1^{*(\mathbf{i}_0)}} := 1 - \sum_{i_1 \neq i_1^{*(\mathbf{i}_0)}} \widetilde{m}_{i_1}.$$

Note that with this definition, the quantized masses $\widetilde{m}_{i_1}$ are always nonzero for $i_1 \neq i_1^{*(\mathbf{i}_0)}$, even in subcubes where the original masses m_{i_1} are equal to zero. We will later verify that this is also the case for $i_1 = i_1^{*(\mathbf{i}_0)}$ whenever $\delta < \frac{1}{n(n-1)}$. For $k \geq 2$, we similarly borrow mass from the subcube with maximum mass, and we do so in each coordinate individually. To this end, for each $k \in [2\!:\!d]$, we set for all $\mathbf{i}_{k-1}$,

$$i_k^{*(\mathbf{i}_{k-1})} := \operatorname*{arg\,max}_{i_k \in [0:(n-1)]} m_{\mathbf{i}_k}.$$

As in the assignment (2.25) for the first coordinate, if there are multiple such values, any of them will do. To define the quantized conditional masses, we set for each $\mathbf{i}_{k-1} \in [0\!:\!(n-1)]^{k-1}$ and each $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$,

$$\widetilde{n}_{\mathbf{i}_k} := \begin{cases} \delta\lceil \frac{1}{\delta} n_{\mathbf{i}_k} \rceil = \delta\left\lceil \frac{1}{\delta} \frac{m_{\mathbf{i}_k}}{m_{\mathbf{i}_{k-1}}} \right\rceil, & \text{if } m_{\mathbf{i}_k} > 0 \\ \delta, & \text{if } m_{\mathbf{i}_k} = 0 \end{cases},$$

as long as $m_{\mathbf{i}_{k-1}} > 0$. If $m_{\mathbf{i}_{k-1}} = 0$, we let

$$\widetilde{n}_{\mathbf{i}_k} := \delta\left\lceil \frac{1}{\delta} n_{\mathbf{i}_k} \right\rceil = \delta\left\lceil \frac{1}{\delta} \frac{1}{n} \right\rceil.$$

We can then define the quantized weights according to

$$\widetilde{m}_{\mathbf{i}_k} := \widetilde{m}_{\mathbf{i}_{k-1}} \widetilde{n}_{\mathbf{i}_k} = \widetilde{n}_{\mathbf{i}_k} \cdots \widetilde{n}_{\mathbf{i}_1}.$$

Finally, for $i_k = i_k^{*(\mathbf{i}_{k-1})}$, we set

$$\widetilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} := 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{n}_{(\mathbf{i}_{k-1}, i_k)}$$

and correspondingly

$$\begin{aligned}\widetilde{m}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} &:= \widetilde{m}_{\mathbf{i}_{k-1}} \widetilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} \\ &= \widetilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} \cdots \widetilde{n}_{\mathbf{i}_1}.\end{aligned}$$

We now check that the quantized weights verify the following properties:

1. Correct marginals:

$$\begin{aligned}&\sum_{i_k=1}^{n} \widetilde{m}_{(\mathbf{i}_{k-1}, i_k)} \\ &= \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{m}_{(\mathbf{i}_{k-1}, i_k)} + \widetilde{m}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} \\ &= \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{m}_{\mathbf{i}_{k-1}} \widetilde{n}_{(\mathbf{i}_{k-1}, i_k)} + \widetilde{m}_{\mathbf{i}_{k-1}} \widetilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} \\ &= \widetilde{m}_{\mathbf{i}_{k-1}} \left(\sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{n}_{(\mathbf{i}_{k-1}, i_k)} + \left(1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{n}_{(\mathbf{i}_{k-1}, i_k)} \right) \right) \\ &= \widetilde{m}_{\mathbf{i}_{k-1}}.\end{aligned}$$

2. If $\delta < \frac{1}{n(n-1)}$, then all quantized masses are positive. To this end, we first note that

$$n_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} = \frac{m_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)}}{m_{\mathbf{i}_{k-1}}} \geq \frac{1}{n}.$$

Since for $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$, we have by definition

$$\widetilde{n}_{\mathbf{i}_k} - n_{\mathbf{i}_k} \leq \delta,$$

it follows that

$$\begin{aligned}
\widetilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} &= 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \widetilde{n}_{(\mathbf{i}_{k-1}, i_k)} \\
&\geq 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \left(n_{(\mathbf{i}_{k-1}, i_k)} + \delta\right) \\
&= n_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})}\right)} - (n-1)\delta \\
&> \frac{1}{n} - \frac{n-1}{n(n-1)} = 0.
\end{aligned}$$

We next formalize the procedure for going from the original masses $m_{\mathbf{i}_k}$ to the quantized masses $\widetilde{m}_{\mathbf{i}_k}$ by characterizing a transport map effecting this transition.

Lemma 27. *Let $k \in [1:d]$, ν a distribution supported on $[0,1]^k$ and with masses $m_{\mathbf{i}_k}$ in the subcubes $c_{\mathbf{i}_k}$ and conditional masses $n_{\mathbf{i}_k}$, all as specified above. Let the quantized masses $\widetilde{m}_{\mathbf{i}_k}$ and the conditional quantized masses $\widetilde{n}_{\mathbf{i}_k}$ also be given as above. Then, for all $\mathbf{i}_k$, we have*

$$\begin{aligned}
&\widetilde{m}_{\mathbf{i}_k} \\
&= m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \chi_{[0:(n-1)] \setminus \left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \widetilde{\Upsilon}(\mathbf{i}, k, k'+1) \\
&\qquad\qquad \left(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^{\eta}(\mathbf{i}, k'-1, 1) \\
&\quad - \sum_{k'=1}^{k} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \Upsilon(\mathbf{i}, k, k'+1) \left(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\right) \\
&\qquad\qquad \Upsilon^{\eta}(\mathbf{i}, k'-1, 1),
\end{aligned} \tag{2.26}$$

where

$$\Upsilon(\mathbf{i}, b, a) = \begin{cases} n_{\mathbf{i}_b} \cdots n_{\mathbf{i}_a}, & \text{if } b \geq a \\ 1, & \text{else} \end{cases},$$

$$\widetilde{\Upsilon}(\mathbf{i}, b, a) = \begin{cases} \widetilde{n}_{\mathbf{i}_b} \cdots \widetilde{n}_{\mathbf{i}_a}, & \text{if } b \geq a \\ 1, & \text{else} \end{cases},$$

and

$$\Upsilon^{\eta}(\mathbf{i}, b, a) = \begin{cases} \eta_{\mathbf{i}_b}(i_b) \cdots \eta_{\mathbf{i}_a}(i_a), & \text{if } b \geq a \\ 1, & \text{else} \end{cases},$$

with

$$\eta_{\mathbf{i}_k}(i_k) := \begin{cases} n_{\mathbf{i}_k}, & \text{if } i_k \neq i_k^{*(\mathbf{i}_{k-1})} \\ \widetilde{n}_{\mathbf{i}_k}, & \text{if } i_k = i_k^{*(\mathbf{i}_{k-1})} \end{cases}.$$

The proof of Lemma 27 is provided in the appendix.

We are now ready to state the main result of this section. Specifically, we establish an upper bound on the Wasserstein distance between a given (arbitrary) distribution ν supported on $[0,1]^d$, for any $d \in \mathbb{N}$, and the corresponding δ-quantized histogram distribution of resolution n obtained based on the procedure described above.

Theorem 18. *Let $d \in \mathbb{N}$. For every distribution ν supported on $[0,1]^d$, there exists a δ-quantized histogram distribution μ of resolution n such that*

$$W(\mu, \nu) \leq \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta.$$

Proof. The proof proceeds in three steps as follows.

1. For each $\mathbf{i}_d \in [0\!:\!(n-1)]^d$, we redistribute the mass $m_{\mathbf{i}_d} = \nu(c_{\mathbf{i}_d})$ to a point mass concentrated in the corner point $p_{\mathbf{i}_d}$.

2. We transport the masses $m_{\mathbf{i}_d}$ according to the procedure described above to result in the masses $\widetilde{m}_{\mathbf{i}_d}$, still located at $p_{\mathbf{i}_d}$.

3. For each $\mathbf{i}_d \in [0:(n-1)]^d$, we spread out the mass $\widetilde{m}_{\mathbf{i}_d}$ uniformly across the subcube indexed by $\mathbf{i}_d$.

Step 1. We define the distribution

$$\nu' = \sum_{\mathbf{i}_d \in [0:(n-1)]^d} m_{\mathbf{i}_d} \delta_{p_{\mathbf{i}_d}}$$

and note that transporting ν to ν' incurs transportation cost

$$\begin{aligned} W(\nu, \nu') &\leq \sum_{\mathbf{i}_d} m_{\mathbf{i}_d} \underbrace{\sqrt{\left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{1}{n}\right)^2}}_{\text{maximum distance in each subcube}} \\ &= \sum_{\mathbf{i}_d} m_{\mathbf{i}_d} \frac{\sqrt{d}}{n} \\ &= \frac{\sqrt{d}}{n}. \end{aligned} \tag{2.27}$$

Step 2. To redistribute the masses from the original values $m_{\mathbf{i}_d}$ to the quantized values $\widetilde{m}_{\mathbf{i}_d}$, we proceed coordinate by coordinate. Specifically, in the k-th coordinate, we carry out two (sets of) transportations. The first one moves, for fixed $i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d$, mass from the point $p_{(i_1,\ldots,i_{k-1},i_k^{*(\mathbf{i}_{k-1})},i_{k+1},\ldots,i_d)}$ to the points $p_{(i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_d)}$, for all $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$, and does this for all tuples $i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d$. The second set of transportations reconfigures masses in the coordinates $[1:(k-1)]$ so as to obtain the correct marginals in coordinate k. These reconfigurations moreover preserve the marginals in coordinates $[1:(k-1)]$. We make all this precise through the following claim, proved below after Step 3 has been presented.

Claim: Reconfiguring the masses between the corner points such that

the mass in the point $p_{\mathbf{i}_d}$ is given by

$$
\begin{aligned}
&m_{\mathbf{i}_d} \\
&\quad + \sum_{k'=1}^{k} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \\
&\qquad\qquad \widetilde{\Upsilon}(\mathbf{i},k,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i},k'-1,1) \\
&\quad - \sum_{k'=1}^{k} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \Upsilon(\mathbf{i},d,k'+1)\big(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\big) \\
&\qquad\qquad \Upsilon^{\eta}(\mathbf{i},k'-1,1),
\end{aligned}
$$

where

$$
m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)} := \sum_{i_{k'+1},\ldots,i_k} m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k'+1},\ldots,i_d\right)},
$$

yields the correct marginal masses $\widetilde{m}_{\mathbf{i}_{k'}}$ in all coordinates $k' \in [1:k]$ and comes at a Wasserstein cost of at most $k(n-1)\delta$, i.e., the Wasserstein distance between the configuration of masses before the moves and the configuration after the moves is at most $k(n-1)\delta$. There is a slight complication when $m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)} = 0$ as in this case the fraction in (2.28) is technically undefined. However, analogously to the definition of the $n_{\mathbf{i}_k}$ in the case of zero-masses in the discussion preceding this theorem, we take

$$
\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) := \left(\frac{1}{n}\right)^{d-k}
$$

when $m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)} = 0$. In either case, we have

$$\sum_{i_{k+1},\ldots,i_d} \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) = 1.$$

We note that the transport map in the Claim characterizes, at a high level, the state of the masses at an intermediate step in the transportation, while (2.26) describes the "final state" after all the moves have been completed in coordinate k.

If we accept the claim and apply it for $k = d$ in combination with Lemma 27, it follows that the masses $m_{\mathbf{i}_d}$ are, indeed, redistributed to the masses $\widetilde{m}_{\mathbf{i}_d}$. Moreover, we get that the total cost of the transportations in Step 2 effecting this redistribution is upper-bounded by

$$(n-1)\,\delta + 2\,(n-1)\,\delta + \cdots + d\,(n-1)\,\delta = \frac{d(d+1)}{2}(n-1)\,\delta.$$

Step 3. The Wasserstein cost associated with spreading out the masses $\widetilde{m}_{\mathbf{i}_d}$ uniformly across their associated subcubes follows from (2.27) as

$$W\left(\sum_{\mathbf{i}_d} \widetilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \mu\right) \leq \frac{\sqrt{d}}{n}.$$

Using the fact that Wasserstein distance is a metric, we can put the costs incurred in the individual steps together according to

$$W(\mu,\nu) \leq \underbrace{W\left(\nu, \sum_{\mathbf{i}_d} m_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}\right)}_{\text{Step 1}} + \underbrace{W\left(\sum_{\mathbf{i}_d} m_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \sum_{\mathbf{i}_d} \widetilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}\right)}_{\text{Step 2}} + \underbrace{W\left(\sum_{\mathbf{i}_d} \widetilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \mu\right)}_{\text{Step 3}}$$

$$
\begin{aligned}
&\leq \frac{\sqrt{d}}{n} + \frac{d(d+1)}{2}\,(n-1)\,\delta + \frac{\sqrt{d}}{n} \\
&= \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\,\delta.
\end{aligned}
$$

It remains to prove the claim.

Proof of the Claim. We proceed by induction on k and start with the base case $k = 1$. The statement on the transportation cost associated with (2.28) does not need an induction argument, rather it follows as a byproduct of the proof by induction. Evaluating the transport map for $k = 1$ yields

$$
\begin{aligned}
m_{\mathbf{i}_d} + \chi_{[0:(n-1)]\setminus\left\{i_1^{*(\mathbf{i}_0)}\right\}}(i_1)\left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\widetilde{m}_{i_1} - m_{i_1}) \\
- \chi_{\left\{i_1^{*(\mathbf{i}_0)}\right\}}(i_1)\left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(m_{i_1} - \widetilde{m}_{i_1}).
\end{aligned}
$$

Since masses are moved in the first coordinate only and $(\widetilde{m}_{i_1} - m_{i_1}) \leq \delta$, for $i_1 \neq i_1^{*(\mathbf{i}_0)}$, the Wasserstein cost of the overall transportation satisfies

$$
\sum_{i_1 \neq i_1^{*(\mathbf{i}_0)}} \sum_{i_2,\ldots,i_d} \left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\widetilde{m}_{i_1} - m_{i_1}) \leq (n-1)\,\delta.
$$

Furthermore, we obtain the desired marginal masses in i_1 as a consequence of

$$
\begin{aligned}
&\sum_{i_2,\ldots,i_d} \left(m_{\mathbf{i}_d} + \left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\widetilde{m}_{i_1} - m_{i_1})\right) \\
&= m_{i_1} + (\widetilde{m}_{i_1} - m_{i_1}) \\
&= \widetilde{m}_{i_1}.
\end{aligned}
$$

This completes the proof of the base case.

We proceed to establish the induction step. Assume that transportations were conducted in coordinate k according to (2.28) and that all marginal masses up to and including coordinate k are as desired. We consider the transport equation (2.28) in coordinate $k+1$, i.e., the sums in (2.28) range from 1 to $k+1$ and start by pointing out that

$$
\begin{aligned}
&n_{(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},\ldots,i_d)} n_{(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},\ldots,i_{d-1})} \cdots n_{(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2})} \\
&= \left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}} \right).
\end{aligned}
$$

The first set of transportations (corresponding to the index $k' = k+1$ in the transport equation (2.28) evaluated for coordinate $k+1$) hence amounts to moving, for fixed $i_1,\ldots,i_k,i_{k+2},\ldots,i_d$, the mass

$$
\left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}} \right) \left(n_{\mathbf{i}_{k+1}} - \widetilde{n}_{\mathbf{i}_{k+1}}\right) \Upsilon^{\eta}(\mathbf{i},k,1)
$$

out of the point $p_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}$ and redistributing it across the points $p_{(i_1,\ldots,i_k,i_{k+1},i_{k+2},\ldots,i_d)}$, for $i_{k+1} \neq i_{k+1}^{*(\mathbf{i}_k)}$. Note that for $i_{k+1} = i_{k+1}^{*(\mathbf{i}_k)}$, the quantity $\left(n_{\mathbf{i}_{k+1}} - \widetilde{n}_{\mathbf{i}_{k+1}}\right)$ is positive by definition of $\widetilde{n}$. These transportations are conducted for all possible tuples $i_1,\ldots,i_k,i_{k+2},\ldots,i_d$. The Wasserstein cost associated with the collection of these transportations satisfies

$$
\begin{aligned}
&\sum_{i_1,\ldots,i_k} \sum_{i_{k+1} \neq i_{k+1}^{*(\mathbf{i}_k)}} \sum_{i_{k+2},\ldots,i_d} \left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}} \right) \\
&\qquad\qquad \left(\widetilde{n}_{\mathbf{i}_{k+1}} - n_{\mathbf{i}_{k+1}}\right) \Upsilon^{\eta}(\mathbf{i},k,1) \\
&= \sum_{i_1,\ldots,i_k} \sum_{i_{k+1} \neq i_{k+1}^{*(\mathbf{i}_k)}} \left(\widetilde{n}_{\mathbf{i}_{k+1}} - n_{\mathbf{i}_{k+1}}\right) \Upsilon^{\eta}(\mathbf{i},k,1) \\
&\leq (n-1)\,\delta \sum_{i_1,\ldots,i_k} \Upsilon^{\eta}(\mathbf{i},k,1)
\end{aligned}
$$

$$\leq (n-1)\,\delta \sum_{i_1,\ldots,i_k} \Upsilon(\mathbf{i},k,1)$$
$$= (n-1)\,\delta,$$

where the last inequality follows because $\eta_{\mathbf{i}_k}(i_k) = \widetilde{n}_{\mathbf{i}_k}$ exactly when $i_k = i_k^{*(\mathbf{i}_{k-1})}$, in which case we have $\widetilde{n}_{\mathbf{i}_k} \leq n_{\mathbf{i}_k}$. The second set of transportations reconfigures the masses in the coordinates $k' \leq k$ in order to obtain correct marginals in the $(k+1)$-th coordinate. To this end, we first note that, for each $k' \leq k$, for all $i_{k'}$, the following identity holds

$$\left(\frac{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}\right)$$
$$= \underbrace{\left(\frac{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}\right)}_{\text{independent of } i_{k+1}} \underbrace{\left(\frac{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}\right)}_{\text{sums to 1 in } i_{k+1}},$$

as long as $m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right) \neq 0$. If $m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right) = 0$, then, as usual, we set

$$\left(\frac{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}{m\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}\right) := \frac{1}{n}.$$

Specifically, noting that by the induction assumption transportation according to (2.28) was carried out in coordinate k, we would like to

reconfigure, for each $k' \leq k$, the masses

$$\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}} \right) \tag{2.28}$$

$$\widetilde{\Upsilon}(\mathbf{i},k,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i},k'-1,1)$$

into

$$\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \widetilde{n}_{\mathbf{i}_{k+1}} \widetilde{\Upsilon}(\mathbf{i},k,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)$$

$$\Upsilon^{\eta}(\mathbf{i},k'-1,1). \tag{2.29}$$

This reconfiguration is possible as only one term in each (2.28) and (2.29) depends on i_{k+1} and

$$\sum_{i_{k+1}} \widetilde{n}_{\mathbf{i}_{k+1}} = 1 = \sum_{i_{k+1}} \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}} \right).$$

Masses to be moved in this manner appear for all $i_1,...,i_{k'-1},i_{k'},i_{k'+1},\ldots,i_k,i_{k+2},\ldots,i_d$ with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$. It follows by inspection of the transport map (2.28) that these transportations do not alter the marginals for $k' \leq k$ as, for given k', the mass moved out of the point $p_{\left(i_1,\ldots,i_{k'-1},i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k'+1},\ldots,i_d\right)}$, accounted for by the sum with negative sign in (2.28), equals the total mass moved into the points $p_{(i_1,\ldots,i_{k'-1},i_{k'},i_{k'+1},\ldots,i_d)}$, for $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$, accounted for by the sum with positive sign. These moves hence retain the marginals for $k' \leq k$, which are correct by the induction assumption. Before establishing that the desired marginals

in coordinate $k+1$ are obtained, we compute the Wasserstein cost associated with the mass reconfiguration moves according to

$$\begin{aligned}
&\sum_{i_1,\dots,i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \sum_{i_{k'+1},\dots,i_k} \sum_{i_{k+2},\dots,i_d} \\
&\left(\frac{m_{\left(i_1,\dots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\dots,i_d\right)}}{m_{\left(i_1,\dots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \widetilde{\Upsilon}(\mathbf{i},k,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big) \\
&\qquad\qquad \Upsilon^{\eta}(\mathbf{i},k'-1,1) \\
&= \sum_{i_1,\dots,i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \sum_{i_{k'+1},\dots,i_k} \widetilde{\Upsilon}(\mathbf{i},k,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big) \\
&\qquad\qquad \Upsilon^{\eta}(\mathbf{i},k'-1,1) \\
&= \sum_{i_1,\dots,i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \underbrace{\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)}_{\leq \delta} \Upsilon^{\eta}(\mathbf{i},k'-1,1) \\
&\leq (n-1)\delta \sum_{i_1,\dots,i_{k'-1}} \Upsilon^{\eta}(\mathbf{i},k'-1,1) \\
&\leq (n-1)\delta \sum_{i_1,\dots,i_{k'-1}} \Upsilon(\mathbf{i},k'-1,1) \\
&= (n-1)\delta.
\end{aligned}$$

We must carry this out for all $k' \in [1\!:\!k]$, which results in a Wasserstein cost of $k(n-1)\delta$ for the reconfigurations. Altogether, we have a Wasserstein cost of

$$(k+1)(n-1)\delta$$

incurred by the moves corresponding to coordinate $k+1$.

Next, using $\widetilde{n}_{\mathbf{i}_{k+1}} \widetilde{\Upsilon}(\mathbf{i},k,k'+1) = \widetilde{\Upsilon}(\mathbf{i},k+1,k'+1)$, it follows

that the updated mass in the point $p_{\mathbf{i}_d}$ is given by

$$m_{\mathbf{i}_d} + \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \widetilde{\Upsilon}(\mathbf{i},k+1,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i},k'-1,1)$$
$$- \sum_{k'=1}^{k+1} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \Upsilon(\mathbf{i},d,k'+1)\big(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\big) \Upsilon^{\eta}(\mathbf{i},k'-1,1).$$

Finally, we need to check that the marginals in the $(k+1)$-th coordinate are, indeed, given by $\widetilde{m}_{\mathbf{i}_{k+1}}$. This is accomplished by noting that owing to Lemma 27, we have

$$\sum_{i_{k+2},\ldots,i_d} \Bigg[m_{\mathbf{i}_d} + \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \widetilde{\Upsilon}(\mathbf{i},k+1,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big) \Upsilon^{\eta}(\mathbf{i},k'-1,1)$$
$$- \sum_{k'=1}^{k+1} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \Upsilon(\mathbf{i},d,k'+1)\big(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\big) \Upsilon^{\eta}(\mathbf{i},k'-1,1) \Bigg]$$
$$= m_{\mathbf{i}_{k+1}} + \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \widetilde{\Upsilon}(\mathbf{i},k+1,k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big) \Upsilon^{\eta}(\mathbf{i},k'-1,1)$$

$$
\begin{aligned}
&- \sum_{k'=1}^{k+1} \chi_{\left\{ i_{k'}^{*(\mathbf{i}_{k'-1})} \right\}}(i_{k'})\, \Upsilon(\mathbf{i}, k+1, k'+1)\big(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Upsilon^{\eta}(\mathbf{i}, k'-1, 1) \\
&= \widetilde{m}_{\mathbf{i}_{k+1}}.
\end{aligned}
$$

This concludes the proof of the induction step and hence establishes the claim. □

2.8. APPROXIMATION OF ARBITRARY DISTRIBUTIONS ON BOUNDED SUBSETS OF $\mathbb{R}^D$ WITH GENERATIVE RELU NETWORKS

In this section, we put all the pieces developed together and state the main result of this chapter.

Theorem 19. *For every distribution ν supported on $[0,1]^d$, there exists a $\frac{\delta}{n}$-quantized ReLU network $\Phi \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$ such that*

$$
W(\Phi \# U, \nu) \leq \frac{\sqrt{d}}{n 2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta. \tag{2.30}
$$

Proof. The proof follows by application of the triangle inequality for Wasserstein distance in combination with Theorems 17 and 18 according to

$$
\begin{aligned}
W(\Phi \# U, \nu) &\leq W(\Phi \# U, \mu) + W(\mu, \nu) \\
&\leq \frac{\sqrt{d}}{n 2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta,
\end{aligned}
$$

where μ denotes the δ-quantized histogram distribution of resolution n per Theorem 18. □

When the target distribution is uniform, Theorem 19 recovers (Bailey and Telgarsky, 2018, Theorem 2.1). We can simplify the bound (2.30) by setting $\delta = \frac{1}{\lceil \sqrt{d}(d+1)n(n-1) \rceil}$ (which satisfies the requirement $\delta < \frac{1}{n(n-1)}$ guaranteeing that all quantized weights are positive) to obtain, for $n \geq 2$,

$$W(\Phi \# U, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{\sqrt{d}}{2n} = \frac{\sqrt{d}}{n2^s} + \frac{5\sqrt{d}}{2n}. \tag{2.31}$$

The error bound in (2.31) illustrates the main conceptual insight of this chapter, namely that generating arbitrary d-dimensional distributions from a 1-dimensional uniform distribution by pushforward through a deep ReLU network does not come at a cost—in terms of Wasserstein-distance error—relative to generating the target distribution from d independent random variables. Specifically, if we let the depth s of the generating network go to infinity, the first term in the rightmost expression of (2.31) will go to zero exponentially fast in s—thanks to the space-filling property of the transport map realized by the generating network—leaving us only with the second term, which reflects the error stemming from the histogram approximation of the distribution. Moreover, this second term is inversely proportional to the histogram resolution n and can thus be made arbitrarily small by letting the histogram resolution n approach infinity. The width of the corresponding generating network will grow according to $\mathcal{O}(n^d)$.

Theorem 19 applies to distributions supported on the unit cube $[0, 1]^d$. The extension to distributions supported on bounded subsets of $\mathbb{R}^d$ is, however, fairly straightforward. Before stating this extension, we provide a lemma that will help us deal with the scaling and shifting of distributions.

Lemma 28. *Let μ, ν be distributions on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz-continuous mapping with Lipschitz constant* $\mathrm{Lip}(f) < \infty$. *Then,*

$$W(f\#\mu, f\#\nu) \leq \mathrm{Lip}(f)\, W(\mu, \nu).$$

Proof. Let π be a coupling between μ and ν and let $g : \mathbb{R}^{2d} \to \mathbb{R}^{2d}; (\mathbf{y}_1, \mathbf{y}_2) \mapsto (f(\mathbf{y}_1), f(\mathbf{y}_2))$. Then $g\#\pi$ is a coupling between μ and ν and

$$\begin{aligned} W(f\#\mu, f\#\nu) &\leq \int_{\mathbb{R}^{2d}} \|\mathbf{y}_1 - \mathbf{y}_2\| \, d(g\#\pi)(\mathbf{y}_1, \mathbf{y}_2) \\ &= \int_{\mathbb{R}^{2d}} \|f(\mathbf{y}_1) - f(\mathbf{y}_2)\| \, d\pi(\mathbf{y}_1, \mathbf{y}_2) \\ &\leq \mathrm{Lip}(f) \int_{\mathbb{R}^{2d}} \|\mathbf{y}_1 - \mathbf{y}_2\| \, d\pi(\mathbf{y}_1, \mathbf{y}_2) \\ &= \mathrm{Lip}(f) \, W(\mu, \nu). \end{aligned}$$

□

We are now ready to state the extension announced above.

Theorem 20. *Let ν be a distribution on $\mathbb{R}^d$ supported on $S = \alpha\,[0,1]^d + \boldsymbol{\beta}$ for $\alpha > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. Let $g(\mathbf{x}) = \frac{1}{\alpha}(\mathbf{x} - \boldsymbol{\beta})$ and $\delta = \frac{1}{\lceil \sqrt{d}(d+1)n(n-1) \rceil}$. Then, there exists a $\frac{\delta}{n}$-quantized ReLU network $\Phi \in \mathcal{NN}_{1,d}$ with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$ such that*

$$W(g^{-1}\#\Phi\#U, \nu) \leq \alpha \left(\frac{\sqrt{d}}{n 2^s} + \frac{5\sqrt{d}}{2n} \right).$$

Proof. We first note that g^{-1} is Lipschitz with $\mathrm{Lip}(g^{-1}) = \alpha$. The result then follows immediately from Lemma 28 combined with (2.31) by taking $\Phi \in \mathcal{NN}_{1,d}$ to approximate the distribution $g\#\nu$ according to Theorem 19. □

We finally remark that g^{-1} by virtue of being an affine map can easily be realized by a ReLU network.

2.9. COMPLEXITY OF GENERATIVE NETWORKS

In this section, we compare the complexity of ReLU networks generating a given class of probability distributions to fundamental bounds

on the complexity of encoding classes of probability distributions through discrete approximations, a process commonly referred to as quantization Graf and Luschgy (2000). Specifically, complexity will be measured in terms of the number of bits needed to describe the generative networks and, respectively, distributions. We begin by reviewing a fundamental result on the approximation of (non-singular) distributions.

Definition 28 (Graf and Luschgy (2000))**.** *For $n \in \mathbb{N}$ and the non-singular distribution ν supported on $[0,1]^d$, we define the minimal n-term quantization error as*

$$V_n(\nu) := \inf\{W(\nu, \mu) : |\mathrm{supp}(\mu)| \leq n\}.$$

The quantity $V_n(\nu)$ characterizes the approximation error—in Wasserstein distance—incurred by the best discrete n-point approximation of ν. The next result, taken from Graf and Luschgy (2000), states that this approximation error exhibits the same asymptotics for all (non-singular) distributions satisfying a mild moment constraint.

Theorem 21 ((Graf and Luschgy, 2000, Theorem 6.2))**.** *Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ with $\mathbf{X} \sim \nu$, where ν is non-singular and supported on $[0,1]^d$, and $\mathbb{E}\|\mathbf{X}\|^{1+\delta} < \infty$ for some $\delta > 0$, where $\|\cdot\|$ is any norm on $\mathbb{R}^d$. Then,*

$$\lim_{n\to\infty} n^{1/d} V_n(\nu) = C,$$

where $C > 0$ is a constant depending on d only.

Theorem 21 allows us to conclude that the best-approximating discrete distribution must have at least $n = \mathcal{O}(\varepsilon^{-d})$ points for $V_n(\nu) \leq \varepsilon$ to hold. As Wasserstein distance is a metric, we hence have a covering argument which says that the class of (non-singular) distributions ν supported on $[0,1]^d$ (and satisfying the moment constraint in Theorem 21) has metric entropy lower-bounded by $d\log(\varepsilon^{-1})$ bits. Although this lower bound is very generous, we demonstrate next that it is achieved

for quantized histogram target distributions encoded by their generating ReLU networks.

Lemma 29. *Consider the class of quantized histogram distributions $\widetilde{\mathcal{E}}_\delta[0,1]^d_n$ and let $\varepsilon \in (0, 1/2)$. Then, there exists a set of $\frac{\delta}{n}$-quantized ReLU networks $\Phi(\varepsilon, \cdot)$ of cardinality $2^{\ell(\varepsilon)}$, where $\ell(\varepsilon) \leq C \log(\varepsilon^{-1})$, with C a constant depending on d, δ, n, such that*

$$\sup_{\nu \in \widetilde{\mathcal{E}}_\delta[0,1]^d_n} W(\Phi(\varepsilon, \nu)\#U, \nu) \leq \varepsilon.$$

Proof. By Theorem 17, for every distribution $\nu \in \widetilde{\mathcal{E}}_\delta[0,1]^d_n$, there exists a $\frac{\delta}{n}$-quantized ReLU network Φ, with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$, such that

$$W(\Phi\#U, \nu) \leq \frac{\sqrt{d}}{n2^s} =: \varepsilon.$$

Note that d, n, δ are fixed and ε, as a function of s, can be made arbitrarily small by taking s and hence network depth to be sufficiently large. In particular, network depth needs to scale according to $\mathcal{O}(\log(\varepsilon^{-1}))$. The resulting network Φ will hence depend on ν and ε, indicated by the notation $\Phi(\varepsilon, \nu)$ used henceforth. Next, using $\frac{\delta}{n} \leq 1/2$, which is by assumption, it follows from (Elbrächter et al., 2021, Proposition VI.7) that the number of bits needed to encode $\Phi(\varepsilon, \nu)$ in a uniquely decodable fashion satisfies

$$\begin{aligned} \ell(\varepsilon) &\leq C_0 \left(\mathcal{M}(\Phi(\varepsilon, \nu)) \log(\mathcal{M}(\Phi(\varepsilon, \nu))) + 1\right) \log(n/\delta) \\ &\leq C(d, \delta, n) \log(\varepsilon^{-1}). \end{aligned} \tag{2.32}$$

□

Remark. We note that the quantized networks considered in the present chapter differ slightly from those in Elbrächter et al. (2021) as here we employ two types of quantization, namely Type 1 and Type 2 (see Definition 27), while in Elbrächter et al. (2021) all weights are encoded using Type-1 quantization. This does, however, not have an

impact on the bound on $\ell(\varepsilon)$ in (2.32), in fact, only the constant C_0 changes relative to (Elbrächter et al., 2021, Proposition VI.7). More specifically, to encode the quantized weights in the generative networks considered here, we only need one additional bit per weight signifying whether the weight is quantized according to Type 1 or Type 2.

Lemma 29 tells us that encoding (or quantizing in the sense of Graf and Luschgy (2000)) the class of quantized histogram distributions by pushing forward a scalar uniform distribution through generative ReLU networks achieves the metric entropy limit of $\mathcal{O}(\log(\varepsilon^{-1}))$ as identified in Theorem 21. We hasten to add that a metric entropy scaling of $\mathcal{O}(\log(\varepsilon^{-1}))$ for (quantized) histogram distributions of dimension d, resolution n, and quantization level δ, all fixed, is what one would expect as we essentially have to encode polynomially (in n) many (quantized) real numbers. For general (non-singular) distributions, which constitute a much richer class than (quantized) histogram distributions, we can establish an $\mathcal{O}(\varepsilon^{-d})$ (up to a multiplicative $\log$-term) complexity scaling for their corresponding generative networks, formalized as follows.

Lemma 30. *Consider the class of non-singular distributions supported on $[0,1]^d$, denoted by $\mathcal{F}([0,1]^d)$, and let $\varepsilon \in (0,1/2)$. Then, there exists a set of quantized ReLU networks $\Phi(\varepsilon,\cdot)$ of cardinality $2^{\ell(\varepsilon)}$, where $\ell(\varepsilon) \leq C\varepsilon^{-d}\log^2(\varepsilon^{-1})$, with C a constant depending on d, such that*

$$\sup_{\nu\in\mathcal{F}([0,1]^d)} W(\Phi(\varepsilon,\nu)\#U,\nu) \leq \varepsilon.$$

Proof. By Theorem 19, for every distribution ν supported on $[0,1]^d$, there exists a $\frac{\delta}{n}$-quantized ReLU network Φ, with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$, such that

$$W(\Phi\#U,\nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta.$$

Setting $\delta = \frac{1}{n^2d^2}$ and $s = \log(n)$, we hence get

$$W(\Phi\#U,\nu) \leq \frac{3\sqrt{d}+1}{n} =: \varepsilon,$$

for a $\frac{1}{n^3d^2}$-quantized network $\Phi(\varepsilon,\nu)$, with $\mathcal{M}(\Phi(\varepsilon,\nu)) = \mathcal{O}(n^d)$. Application of (Elbrächter et al., 2021, Proposition VI.7) allows us to conclude that the number of bits needed to encode $\Phi(\varepsilon,\nu)$ in a uniquely decodable fashion satisfies $\ell(\varepsilon) \leq C_0\left(\mathcal{M}(\Phi)\log(\mathcal{M}(\Phi))+1\right)\log(n^3d^2) \leq C(d)\varepsilon^{-d}\log^2(\varepsilon^{-1})$. We note that $C(d)$ scales very unfavorably in d, namely according to $d^{d/2}$. Finally, we remark that the application of (Elbrächter et al., 2021, Proposition VI.7) requires that $\frac{1}{n^3d^2} \leq 1/2$, which is satisfied if at least one of n, d is strictly larger than 1. □

2.10. APPENDIX

A. Proof of Lemma 27

Proof. Note first that

$$\chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) = \begin{cases} 0, & \text{if } i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})} \\ 1, & \text{if } i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})} \end{cases}$$

and

$$\chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) = \begin{cases} 0, & \text{if } i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})} \\ 1, & \text{if } i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})} \end{cases},$$

so for a given $i_{k'}$ only one of the two χ-terms above is active. Terms with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$ correspond to subcubes to which we add mass to get the quantized masses in the k'-th coordinate, while terms with $i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})}$ correspond to the subcube from which we take this extra mass. Correspondingly, we refer to terms with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$ as "+ terms", while we designate terms with $i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})}$ as "– terms". By construction, $\left(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \geq 0$ for + terms, while $\left(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\right) \geq 0$ for – terms. In evaluating the sum (2.26), we consider three different cases.

Case 1: All terms are + terms. In this case, the sum becomes

$$
\begin{aligned}
&m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \widetilde{\Upsilon}(\mathbf{i}, k, k'+1)\big(\widetilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon(\mathbf{i}, k'-1, 1) \\
&= m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \widetilde{\Upsilon}(\mathbf{i}, k, k')\Upsilon(\mathbf{i}, k'-1, 1) \\
&\qquad\qquad - \sum_{k'=1}^{k} \widetilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1) \\
&= m_{\mathbf{i}_k} - \underbrace{\Upsilon(\mathbf{i}, k, 1)}_{=m_{\mathbf{i}_k}} + \underbrace{\widetilde{\Upsilon}(\mathbf{i}, k, 1)}_{=\widetilde{m}_{\mathbf{i}_k}} + \sum_{k'=2}^{k} \widetilde{\Upsilon}(\mathbf{i}, k, k')\Upsilon(\mathbf{i}, k'-1, 1) \\
&\qquad\qquad - \sum_{k'=1}^{k-1} \widetilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1) \\
&= \widetilde{m}_{\mathbf{i}_k} + \sum_{k'=1}^{k-1} \widetilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1) \\
&\qquad\qquad - \sum_{k'=1}^{k-1} \widetilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1) \\
&= \widetilde{m}_{\mathbf{i}_k}.
\end{aligned}
$$

Case 2: All terms are – terms. In this case, the sum is

$$
\begin{aligned}
&m_{\mathbf{i}_k} - \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k'+1)\big(n_{\mathbf{i}_{k'}} - \widetilde{n}_{\mathbf{i}_{k'}}\big)\widetilde{\Upsilon}(\mathbf{i}, k'-1, 1) \\
&= m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k'+1)\widetilde{\Upsilon}(\mathbf{i}, k', 1) \\
&\qquad\qquad - \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k')\widetilde{\Upsilon}(\mathbf{i}, k'-1, 1)
\end{aligned}
$$

$$= m_{\mathbf{i}_k} - \Upsilon(\mathbf{i}, k, 1) + \widetilde{\Upsilon}(\mathbf{i}, k, 1) + \sum_{k'=1}^{k-1} \Upsilon(\mathbf{i}, k, k'+1)\widetilde{\Upsilon}(\mathbf{i}, k', 1)$$
$$- \sum_{k'=2}^{k} \Upsilon(\mathbf{i}, k, k')\widetilde{\Upsilon}(\mathbf{i}, k'-1, 1)$$
$$= \widetilde{m}_{\mathbf{i}_k}.$$

Case 3: There is at least one + term and one – term. Let the indices of the + terms be given by

$$\{k_1^+, \ldots, k_{\ell_1}^+\}$$

and those of the – terms by

$$\{k_1^-, \ldots, k_{\ell_2}^-\},$$

with both sets arranged in increasing order and $\ell_1 + \ell_2 = k$.

We first consider the sum of the – terms given by

$$\sum_{\ell=1}^{\ell_2} \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\left(n_{\mathbf{i}_{k_\ell^-}} - \widetilde{n}_{\mathbf{i}_{k_\ell^-}}\right)\Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1) \tag{2.33}$$

and establish a cancelation property of successive terms in this sum, leaving only the border terms to be considered. Indeed, take ℓ such that $1 < \ell < \ell_2$, with the corresponding term given by

$$\begin{aligned} &\Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, n_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1) \\ &- \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, \widetilde{n}_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1). \end{aligned} \tag{2.34}$$

Next, note that the positive part of the term corresponding to the index $\ell + 1$ is given by

$$\begin{aligned} &\Upsilon(\mathbf{i}, k, k_{\ell+1}^- + 1)\, n_{\mathbf{i}_{k_{\ell+1}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell+1}^- - 1, 1) \\ &= \Upsilon(\mathbf{i}, k, k_{\ell+1}^- + 1)\, n_{\mathbf{i}_{k_{\ell+1}^-}}\, n_{\mathbf{i}_{k_{\ell+1}^- - 1}} \cdots n_{\mathbf{i}_{k_\ell^- + 1}}\, \widetilde{n}_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1), \end{aligned} \tag{2.35}$$

since all indices that lie strictly between k_ℓ^- and $k_{\ell+1}^-$, if there are any, correspond to + terms. Comparing (2.35) with (2.34) reveals that the positive part of the term corresponding to $\ell+1$ cancels out the negative part of the term for ℓ. Similarly, the negative part of the term corresponding to $\ell-1$, given by

$$-\Upsilon(\mathbf{i},k,k_{\ell-1}^-+1)\,\widetilde{n}_{\mathbf{i}_{k_{\ell-1}^-}}\,\Upsilon^\eta(\mathbf{i},k_{\ell-1}^--1,1),$$

cancels out the positive part of the term for index ℓ, which is given by

$$\begin{aligned}&\Upsilon(\mathbf{i},k,k_\ell^-+1)\,n_{\mathbf{i}_{k_\ell^-}}\Upsilon^\eta(\mathbf{i},k_\ell^--1,1)\\&=\Upsilon(\mathbf{i},k,k_\ell^-+1)\,n_{\mathbf{i}_{k_\ell^-}}n_{\mathbf{i}_{k_\ell^--1}}\cdots n_{\mathbf{i}_{k_{\ell-1}^-+1}}\widetilde{n}_{\mathbf{i}_{k_{\ell-1}^-}}\Upsilon^\eta(\mathbf{i},k_{\ell-1}^--1,1).\end{aligned}$$

Hence, the only contributions remaining in the sum (2.33) over all the – terms are the negative part of the term corresponding to the index ℓ_2 and the positive part of the term for the index 1, i.e.,

$$\begin{aligned}&\sum_{\ell=1}^{\ell_2}\Upsilon(\mathbf{i},k,k_\ell^-+1)\Big(n_{\mathbf{i}_{k_\ell^-}}-\widetilde{n}_{\mathbf{i}_{k_\ell^-}}\Big)\Upsilon^\eta(\mathbf{i},k_\ell^--1,1)\\&=\Upsilon(\mathbf{i},k,k_1^-+1)\,n_{\mathbf{i}_{k_1^-}}\,\Upsilon^\eta(\mathbf{i},k_1^--1,1)\\&\qquad-\Upsilon(\mathbf{i},k,k_{\ell_2}^-+1)\,\widetilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\,\Upsilon^\eta(\mathbf{i},k_{\ell_2}^--1,1)\\&=\Upsilon(\mathbf{i},k,k_1^-+1)\,n_{\mathbf{i}_{k_1^-}}\,\Upsilon(\mathbf{i},k_1^--1,1)\\&\qquad-\Upsilon(\mathbf{i},k,k_{\ell_2}^-+1)\,\widetilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\,\Upsilon^\eta(\mathbf{i},k_{\ell_2}^--1,1)\\&=m_{\mathbf{i}_k}-\Upsilon(\mathbf{i},k,k_{\ell_2}^-+1)\,\widetilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\,\Upsilon^\eta(\mathbf{i},k_{\ell_2}^--1,1),\end{aligned}$$

since indices smaller than k_1^- necessarily correspond to + terms.

We proceed to the sum over the + terms. A similar cancelation property between consecutive terms in the sum can be established so that we are again left with contributions from the first and the last term only. Indeed, take ℓ such that $1<\ell<\ell_1$, with the corresponding term

given by

$$\widetilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\, \widetilde{n}_{\mathbf{i}_{k_\ell^+}} \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1)$$
$$- \widetilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\, n_{\mathbf{i}_{k_\ell^+}} \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1).$$

The positive part of the term corresponding to the index $\ell + 1$ is given by

$$\widetilde{\Upsilon}(\mathbf{i}, k, k_{\ell+1}^+ + 1)\, \widetilde{n}_{\mathbf{i}_{k_{\ell+1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell+1}^+ - 1, 1)$$
$$= \widetilde{\Upsilon}(\mathbf{i}, k, k_{\ell+1}^+ + 1)\, \widetilde{n}_{\mathbf{i}_{k_{\ell+1}^+}} \widetilde{n}_{\mathbf{i}_{k_{\ell+1}^+ - 1}} \cdots \widetilde{n}_{\mathbf{i}_{k_\ell^+ + 1}} n_{\mathbf{i}_{k_\ell^+}} \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1),$$

since all indices that lie strictly between k_ℓ^+ and $k_{\ell+1}^+$, if there are any, correspond to – terms. We can hence conclude, as above, that the positive part of the term corresponding to $\ell + 1$ cancels out the negative part of the term for ℓ. By the same argument, the positive part of the term for ℓ is cancelled out by the negative part of the term corresponding to $\ell - 1$. Overall, the only remaining contributions are the negative part of the term corresponding to ℓ_1 and the positive part of the term for the index 1, i.e.,

$$\sum_{\ell=1}^{\ell_1} \widetilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\left(\widetilde{n}_{\mathbf{i}_{k_\ell^+}} - n_{\mathbf{i}_{k_\ell^+}}\right)\Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1)$$
$$= \widetilde{\Upsilon}(\mathbf{i}, k, 1) - \widetilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1)$$
$$= \widetilde{m}_{\mathbf{i}_k} - \widetilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1).$$

Putting pieces together, (2.26) reduces to

$$m_{\mathbf{i}_k} + \widetilde{m}_{\mathbf{i}_k} - m_{\mathbf{i}_k}$$
$$+ \Upsilon(\mathbf{i}, k, k_{\ell_2}^- + 1)\, \widetilde{n}_{\mathbf{i}_{k_{\ell_2}^-}} \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1) \tag{2.36}$$
$$- \widetilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1).$$

There are two possibilities to consider now, either $k^+_{\ell_1} = k$ or $k^-_{\ell_2} = k$. If $k^+_{\ell_1} = k$, then (2.36) becomes

$$\begin{aligned}
&\widetilde{m}_{\mathbf{i}_k} \\
&+ \Upsilon(\mathbf{i}, k^+_{\ell_1}, k^-_{\ell_2} + 1)\, \widetilde{n}_{\mathbf{i}_{k^-_{\ell_2}}}\, \Upsilon^\eta(\mathbf{i}, k^-_{\ell_2} - 1, 1) \\
&- n_{\mathbf{i}_{k^+_{\ell_1}}} \cdots n_{\mathbf{i}_{k^-_{\ell_2}+1}}\, \widetilde{n}_{\mathbf{i}_{k^-_{\ell_2}}}\, \Upsilon^\eta(\mathbf{i}, k^-_{\ell_2} - 1, 1) \\
&= \widetilde{m}_{\mathbf{i}_k},
\end{aligned}$$

since, by definition, $k^-_{\ell_2}$ is the largest index corresponding to – terms. On the other hand, if $k^-_{\ell_2} = k$, then (2.36) reduces to

$$\begin{aligned}
&\widetilde{m}_{\mathbf{i}_k} \\
&+ \widetilde{n}_{\mathbf{i}_{k^-_{\ell_2}}} \cdots \widetilde{n}_{\mathbf{i}_{k^+_{\ell_1}+1}} n_{\mathbf{i}_{k^+_{\ell_1}}}\, \Upsilon^\eta(\mathbf{i}, k^+_{\ell_1} - 1, 1) \\
&- \widetilde{\Upsilon}(\mathbf{i}, k^-_{\ell_2}, k^+_{\ell_1} + 1)\, n_{\mathbf{i}_{k^+_{\ell_1}}}\, \Upsilon^\eta(\mathbf{i}, k^+_{\ell_1} - 1, 1) \\
&= \widetilde{m}_{\mathbf{i}_k},
\end{aligned}$$

since, again by definition, $k^+_{\ell_1}$ is the largest index corresponding to + terms. This concludes the proof. □

CHAPTER 3

Publications

The majority of the results in this thesis have been published during the course of the PhD studies. Specifically, the results in Chapter 1 appear in Elbrächter et al. (2021) and Perekrestenko et al. (2018). The results in Chapter 2 were published in Perekrestenko et al. (2021) and Perekrestenko et al. (2020).

References

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Bailey, B. and Telgarsky, M. J. (2018). Size-noise tradeoffs in generative networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6489–6499. Curran Associates, Inc.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1):115–133.

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

Berner, J., Grohs, P., and Jentzen, A. (2020). Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657.

Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. (2019). Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Proceedings*

of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.*, 29(2):610–611.

Candès, E. J. (1998). *Ridgelets: Theory and Applications*. PhD thesis, Stanford University.

Candès, E. J. and Donoho, D. L. (2002). New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. *Comm. Pure Appl. Math.*, 57:219–266.

Chui, C. K., Li, X., and Mhaskar, H. N. (1994). Neural networks for localized approximation. *Math. Comp.*, 63(208):607–623.

Chui, C. K. and Wang, J.-Z. (1992). On compactly supported spline wavelets and a duality principle. *Trans. Amer. Math. Soc.*, 330(2):903–915.

Cohen, N., Sharir, O., and Shashua, A. (2016). On the expressive power of deep learning: A tensor analysis. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pages 698–728.

Cohen, N. and Shashua, A. (2016). Convolutional rectifier networks as generalized tensor decompositions. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 955–963.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Daubechies, I., DeVore, R., Foucart, S., Hanin, B., and Petrova, G. (2019). Nonlinear approximation and (deep) ReLU networks. *arXiv preprint arXiv:1905.02199*.

Demanet, L. and Ying, L. (2007). Wave atoms and sparsity of oscillatory patterns. *Appl. Comput. Harmon. Anal.*, 23(3):368–387.

DeVore, R., Hanin, B., and Petrova, G. (2020). Neural network approximation. *arXiv:2012.14501*.

DeVore, R., Oskolkov, K., and Petrushev, P. (1996). Approximation by feed-forward neural networks. *Ann. Numer. Math.*, 4:261–287.

DeVore, R. A. (1998). Nonlinear approximation. *Acta Numerica*, 7:51–150.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*. Springer.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.*, 1(1):100 – 115.

Donoho, D. L. (1996). Unconditional bases and bit-level compression. *Appl. Comput. Harm. Anal.*, 3:388–392.

Donoho, D. L. (2001). Sparse components of images and optimal atomic decompositions. *Constr. Approx.*, 17(3):353–382.

Donoho, D. L., Vetterli, M., DeVore, R. A., and Daubechies, I. (1998). Data compression and harmonic analysis. *IEEE Transactions on Information Theory*, 44(6):2435–2476.

Ehler, M. and Filbir, F. (2018). Metric entropy, n-widths, and sampling of functions on manifolds. *Journal of Approximation Theory*, 225:41 – 57.

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

Elbrächter, D., Perekrestenko, D., Grohs, P., and Bölcskei, H. (2021). Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):pp. 2581–2623.

Elbrächter, D. M., Berner, J., and Grohs, P. (2019). How degenerate is the parametrization of neural networks with the ReLU activation function? In *Advances in Neural Information Processing Systems 32*, page 7788–7799. Curran Associates, Inc.

Eldan, R. and Shamir, O. (2016). The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 907–940.

Ellacott, S. (1994). Aspects of the numerical analysis of neural networks. *Acta Numer.*, 3:145–202.

Fefferman, C. L. (1983). The uncertainty principle. *Bull. Amer. Math. Soc. (N.S.)*, 9(2):129–206.

Fefferman, C. L. (1994). Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555.

Feichtinger, H. G. (1981). On a new Segal algebra. *Monatshefte für Mathematik*, 92:269–289.

Fokina, D. and Oseledets, I. (2019). Growing axons: Greedy learning of neural networks with application to function approximation.

Frenzen, C., Sasao, T., and Butler, J. T. (2010). On the number of segments needed in a piecewise linear approximation. *Journal of Computational and Applied Mathematics*, 234(2):437–446.

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192.

Gil, A., Segura, J., and Temme, N. M. (2007). *Numerical Methods for Special Functions*. Society for Industrial and Applied Mathematics.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*. Springer-Verlag, Berlin, Heidelberg.

Gröchenig, K. (2013). *Foundations of time-frequency analysis*. Springer Science & Business Media.

Gröchenig, K. and Samarah, S. (2000). Nonlinear approximation with local Fourier bases. *Constructive Approximation*, 16(3):317–331.

Grohs, P. (2015). Optimally sparse data representations. In *Harmonic and Applied Analysis*, pages 199–248. Springer.

Grohs, P., Hornung, F., Jentzen, A., and von Wurstemberger, P. (2018). A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv e-prints*, page arXiv:1809.02362.

Grohs, P., Keiper, S., Kutyniok, G., and Schäfer, M. (2016a). α-molecules. *Appl. Comput. Harmon. Anal.*, 41(1):297–336.

Grohs, P., Keiper, S., Kutyniok, G., and Schäfer, M. (2016b). Cartoon approximation with α-curvelets. *J. Fourier Anal. Appl.*, 22(6):1235–1293.

Grohs, P., Klotz, A., and Voigtlaender, F. (2020). Phase transitions in rate distortion theory and deep learning. *arxiv:2008.01011*.

Grohs, P. and Kutyniok, G. (2014). Parabolic molecules. *Found. Comput. Math.*, 14:299–337.

Guo, K., Kutyniok, G., and Labate, D. (2006). Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines (Athens, GA, 2005)*, pages 189–201. Nashboro Press, Nashville, TN.

Gühring, I., Kutyniok, G., and Petersen, P. (2020). Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859.

Hanin, B. and Rolnick, D. (2019). Deep ReLU networks have surprisingly

few activation patterns. In *Advances in Neural Information Processing Systems 32*, pages 361–370. Curran Associates, Inc.

Hinrichs, A., Piotrowska-Kurczewski, I., and Piotrowski, M. (2008). On the degree of compactness of embeddings between weighted modulation spaces. *J. Funct. Spaces Appl.*, 6:303–317.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97.

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

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. pages 4396–4405.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kolmogorov, A. and Tikhomirov, V. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk.*, 14(2):3–86.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*, 114(5):953–956.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Müller, U. A., Säckinger, E., Simard, P., and Vapnik, V. (1995). Comparison of learning algorithms for handwritten digit recognition. *International Conference on Artificial Neural Networks*, pages 53–60.

Lee, H., Ge, R., Risteski, A., Ma, T., and Arora, S. (2017). On the ability of neural nets to express distributions. *Proceedings of Machine Learning Research*, 65:1–26.

Liang, S. and Srikant, R. (2017). Why deep neural networks for function approximation? *International Conference on Learning Representations*.

Lu, Y. and Lu, J. (2020). A universal approximation theorem of deep neural

networks for expressing probability distributions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3094–3105. Curran Associates, Inc.

Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.*, 315(1):69–87.

Mallat, S. (2008). *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, Inc., USA, 3rd edition.

McCann, R. J. and Pass, B. (2020). Optimal transportation between unequal dimensions. *Archive for Rational Mechanics and Analysis*.

McCulloch, W. and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133.

Mhaskar, H. (2020). A direct approach for function approximation on data defined manifolds. *Neural Networks*, 132:253 – 268.

Mhaskar, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80.

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

Mhaskar, H. N. and Micchelli, C. A. (1995). Degree of approximation by neural and translation networks with a single hidden layer. *Adv. Appl. Math.*, 16(2):151–183.

Mhaskar, H. N. and Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(6):829–848.

Morgenshtern, V. I. and Bölcskei, H. (2012). *A short course on frame theory*, pages 737–789.

Munkres, J. (2000). *Topology*. Featured Titles for Topology. Prentice Hall, Incorporated.

Nguyen-Thien, T. and Tran-Cong, T. (1999). Approximation of functions and their derivatives: A neural network implementation with applications. *Appl. Math. Model.*, 23(9):687–704.

Opschoor, J. A. A., Petersen, P. C., and Schwab, C. (2020). Deep ReLU networks and high-order finite element methods. *Analysis and Applications*, 18(5):715–770.

Ott, E. (2002). *Chaos in Dynamical Systems*. Cambridge University Press.

Perekrestenko, D., Eberhard, L., and Bölcskei, H. (2021). High-dimensional distribution generation through deep neural networks. *Partial Differential Equations and Applications, Springer*, 2(64).

Perekrestenko, D., Grohs, P., Elbrächter, D., and Bölcskei, H. (2018). The

universal approximation power of finite-width deep ReLU networks. *arXiv:1806.01528*.

Perekrestenko, D., Müller, S., and Bölcskei, H. (2020). Constructive universal high-dimensional distribution generation through deep ReLU networks. In *Proc. of the 37th International Conference on Machine Learning (ICML)*.

Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

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

Prosser, R. T. (1966). The ε-entropy and ε-capacity of certain time-varying channels. *Journal of Mathematical Analysis and Applications*, 16:553–573.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Schmidt-Hieber, J. (2019). Deep ReLU network approximation of functions on a manifold. *arXiv:1908.00695*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897.

Schwab, C. and Zech, J. (2019). Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ. *Analysis and Applications*, 17(1):19–55.

Shaham, U., Cloninger, A., and Coifman, R. R. (2018). Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.*, 44(3):537–557.

Stone, M. H. (1948). The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21:167–184.

Telgarsky, M. (2015). Representation benefits of deep feedforward networks. *arXiv:1509.08101*.

Telgarsky, M. (2016). Benefits of depth in neural networks. *JMLR: Workshop and Conference Proceedings*, 49:1–23.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. (2018). Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR)*.

Unser, M. (1997). Ten good reasons for using spline wavelets. *Wavelet Applications in Signal and Image Processing V*, 3169:422–431.

Villani, C. (2008). *Optimal transport: Old and new*, volume 338. Springer Science & Business Media.

Vlačić, V. and Bölcskei, H. (2021a). Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376(107485):1–72.

Vlačić, V. and Bölcskei, H. (2021b). Neural network identifiability for a family of sigmoidal nonlinearities. *Constructive Approximation*.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Xu, J., Ren, X., Lin, J., and Sun, X. (2018). Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium. Association for Computational Linguistics.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114.

Zygmund, A. (2002). *Trigonometric series*. Cambridge University Press.

Dmytro Perekrestenko

Curriculum Vitae

Experience

Since 2021 **Research Software Engineer**, *Ablacon SA*, Zürich, Switzerland.
Developing an AI-based diagnosis system that enables doctors to cure Atrial Fibrillation.
Tech stack: Python.

2015–2016 **Data Science Intern**, *ABB Corporate Research*, Baden, Switzerland.
Developed a Markov chain based automatic tool for prediction of web visitor behavior and assessment of web page usability based on log data from Google Analytics on abb.com.
Tech stack: Spark, Python, SQL, and Google BigQuery.

Skills

Programming Python, Scala, SQL, MATLAB, C/C++

Frameworks Apache Spark, Tensorflow, PyTorch, Google BigQuery

Knowledge machine learning, deep learning, statistics, natural language understanding, time-series analysis, optimization, audio/image processing

Languages English - fluent, French and German - intermediate, Russian and Ukrainian - native

Education

2016–2021 **Dr.sc.**, *Eidgenössische Technische Hochschule Zürich (ETH Zurich)*, Switzerland.
Specialization - Deep Neural Network Theory
Selected Coursework: *Natural Language Understanding, Entrepreneurial leadership*

2014–2016 **M.sc. in Electrical Engineering and Information Technology**, *École Polytechnique Fédérale de Lausanne (EPFL)*, Switzerland, *GPA: 5.57/6.0*.
Specialization - Information Technology/Signal Processing
Selected Coursework: *Applied Machine Learning, Mathematics of Data, Image and Video Processing, Speech Processing*

2010–2014 **B.sc. with honors in Applied Mathematics and Physics**, *Moscow Institute of Physics and Technology (MIPT)*, Russia, *GPA: 4.84/5.0*.
Specialization - Statistics/Intelligent data analysis
Selected Coursework: *Algorithms and Data Structures, Statistics, Optimization*

+41 76 249 9339 • *dmytro@perekrestenko.ch*
www.perekrestenko.ch

Selected Projects

Convolutional recurrent neural networks for electrocardiogram classification. Proposed two deep neural network architectures for classification of arbitrary-length electrocardiogram (ECG) recordings and achieved top accuracy on the atrial fibrillation (AF) classification data set provided by the PhysioNet/CinC Challenge 2017. Published and presented at Computing in Cardiology 2017.

Deep neural network approximation theory. Proved that deep neural networks are optimal approximators for affine and Gabor function systems. Presented at GAMM 2019.

Constructive universal distribution generation through deep ReLU networks. Proved that deep generative networks with 1D input are capable of optimally approximating high dimensional distributions. Presented at ICML 2020.

Faster optimization through adaptive importance sampling. Introduced new adaptive rules for coordinate descent methods and derived theoretical convergence guarantees for Lasso and SVM. Master thesis. Presented at AISTATS 2017.

Human activity recognizer. Used the smartphone's accelerometer time series and neural networks to classify human activities such as walking, running, standing, etc. Presented at 56th Scientific Conference at MIPT.

Other

Scholarship	In 2010–2013 received MIPT semester scholarships given to the top 10% of students
Coursera	Functional Programming Principles in Scala, Big Data Analysis with Scala and Spark
Hobbies	Skiing, competitive swimming, sci-fi literature, aikido, solving riddles

Selected Publications

[1] Dmytro Perekrestenko, Léandre Eberhard, and Helmut Bölcskei. High-dimensional distribution generation through deep neural networks. *Partial Differential Equations and Applications, Springer*, 2(64), September 2021.

[2] Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, February 2021.

[3] Dmytro Perekrestenko, Stephan Müller, and Helmut Bölcskei. Constructive universal high-dimensional distribution generation through deep ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 7610–7619. PMLR, July 2020.

[4] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *Computing in Cardiology (CinC)*, pages 1–4, September 2017.

[5] Dmytro Perekrestenko, Volkan Cevher, and Martin Jaggi. Faster Coordinate Descent via Adaptive Importance Sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 869–877. PMLR, April 2017.

+41 76 249 9339 • dmytro@perekrestenko.ch
www.perekrestenko.ch