Journal of Machine Learning Research 24 (2023) 1-38 Submitted 12/22; Revised 8/23; Published 11/23

Limitations on approximation by deep and shallow neural networks

Guergana Petrova GPETROVA@TAMU.EDU
Department of Mathematics
Texas A&M University
College Station, TX 77843, USA

Przemysław Wojtaszczyk WOJTASZCZYK@IMPAN.PL
Institut of Mathematics
Polish Academy of Sciences
ul. Śniadeckich 8, 00-656 Warszawa, Poland

Editor: Samory Kpotufe

Abstract

We prove Carl's type inequalities for the error of approximation of compact sets $\mathcal{K}$ by deep and shallow neural networks. This in turn gives estimates from below on how well we can approximate the functions in $\mathcal{K}$ when requiring the approximants to come from outputs of such networks. Our results are obtained as a byproduct of the study of the recently introduced Lipschitz widths.

Keywords: widths, entropy numbers, deep neural networks, shallow neural networks, rate of approximation

1. Introduction

Neural network approximation is the method of approximation used in numerical algorithms in many application areas. Thus, it is important to understand not only how well they approximate the underlying objects, but also what are the limits of their approximation power. In this paper, we study the limitations of deep and shallow neural networks in approximating a compact subset $\mathcal{K} \subset X$ of a Banach space X when it is required that the parameters in the approximation procedure have certain bounds. This is done by proving appropriate Carl's type inequalities that relate the error of neural network approximation of $\mathcal{K}$ to the entropy numbers of this set.

Recall that the classical Carl's inequality relates the entropy numbers $\epsilon_n(\mathcal{K})_X$ of a compact class $\mathcal{K}$ to its Kolmogorov width $d_n(\mathcal{K})_X$. More precisely, see Carl (1981), it states that for every $\alpha > 0$ there is a constant $C(\alpha) > 0$, possibly dependent on α, such that for any $\mathcal{K} \subset X$

$$\epsilon_n(\mathcal{K})_X \leq C(\alpha) n^{-\alpha} \max_{k=1,\dots,n} \{k^\alpha d_{k-1}(\mathcal{K})_X\}.$$

Thus, if we know the rate of decay of the entropy numbers $\epsilon_n(\mathcal{K})_X$ of the class $\mathcal{K}$, we can derive an estimate from below for $d_n(\mathcal{K})_X$. Note that the Kolmogorov width $d_n(\mathcal{K})_X$ describes the best possible approximation rate for the compact set $\mathcal{K}$ if the approximants to $\mathcal{K}$ are coming from linear spaces of dimension n. Therefore, a bound from below for

$d_n(\mathcal{K})_X$ would describe the limitations of such approximation. We adopt this strategy and derive bounds from below for the error of approximation of $\mathcal{K}$ via the outputs of deep and shallow neural networks, using instead of $d_n(\mathcal{K})_X$ the recently introduced Lipschitz widths, see Petrova and Wojtaszczyk (2023). It is worth mentioning that not all widths satisfy Carl's inequality (see Petrova and Wojtaszczyk (2022) where it is shown that the Carl's inequality does not hold for manifold widths).

In this paper, we provide a general framework for obtaining estimates from below for deep and shallow NNs. It can be applied to any compact set $\mathcal{K}$ and any Banach space X, provided we only know a bound from below of the entropy numbers $\epsilon_n(\mathcal{K})_X$ of $\mathcal{K}$. Such bounds are readily available for a wide range of classical and novel classes $\mathcal{K}$ and spaces X. For example, see (Edmunds and Triebel, 1996, Chapters 3,4), (Golitschek and Makovoz, 1996, Chapter 15),(Dominguez and Kuhn, 2018, Section 5), (Siegel and Xu, 2022, Theorem 9), or Cobos and Kuhn (2009); Gao (2008), where all such bounds are of the form $n^{-\alpha}[\log n]^{\beta}$, $\alpha > 0$, $\beta \in \mathbb{R}$, and thus can be treated by our theory. In addition, we provide generalized inverse theorems for neural network approximation via deep and shallow neural networks.

More precisely, we investigate deep feed-forward neural networks (NN) with ReLU or bounded Lipschitz activation function, fixed width $W \geq 2$ and depth n (where we let $n \to \infty$), whose parameters have absolute values bounded by a given function $w(n)$. We prove that the capabilities of these networks to approximate any compact subset $\mathcal{K}$ is limited by the behavior of its entropy numbers. For example, we show that Deep Neural Networks (DNNs) with fixed width W, depth n and parameters bounded by $w(n) = Cn^{\delta}$, $\delta > 0$, cannot approximate better than $C[\log_2 n]^{\beta-\alpha} n^{-2\alpha}$ any compact set of functions $\mathcal{K}$ whose entropy numbers $\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{\beta} n^{-\alpha}$, $n \in \mathbb{N}$, see Corollary 11. We also show that if a class of functions $\mathcal{K}$ is approximated up to accuracy $C[\log_2 n]^{\beta} n^{-\alpha}$, $n \in \mathbb{N}$, by the above mentioned DNNs, then $\epsilon_n(\mathcal{K})_X \leq C' n^{-\frac{\alpha}{2}} [\log_2 n]^{\beta+\frac{\alpha}{2}}$, see Corollary 17. In particular, we obtain estimates for the entropy numbers of the classes of functions that are approximated via DNNs with predetermined rates (approximation classes) as the depth n of the NN grows.

Results of the same type are obtained for shallow (neural networks with one hidden layer) NNs (SNNs) as we let the width $W \to \infty$. For example, we show that SNNs with width W and parameters bounded by $w(W) = CW^{\delta}$, with $\delta \geq 0$, cannot approximate better than $C[\log_2 W]^{\beta-\alpha} W^{-\alpha}$ a compact set of functions $\mathcal{K}$ whose entropy numbers are $\gtrsim [\log_2 W]^{\beta} W^{-\alpha}$, see Corollary 13. Also, if a class $\mathcal{K}$ is approximated up to accuracy $C[\log_2 W]^{\beta} W^{-\alpha}$, $W = 2, 3, \dots$, by a SNN, then it has entropy numbers $\epsilon_W(\mathcal{K})_X \leq C' W^{-\alpha} [\log_2 W]^{\beta+\alpha}$, $W \in \mathbb{N}$, see Corollary 18.

Analogous estimates for DNNs or SNNs with general bounds $w(n)$ or $w(W)$, respectively, on their parameters are also given, including the case $w(n) = C2^{cn^{\nu}}$.

In our analysis of neural network approximation (NNA), we are not concerned with the numerical aspect of the construction of the corresponding DNN or SNN and its stability, but rather with the theoretical bounds from below of the performance of such an approximation. We show that the mapping that assigns to each choice of NN parameters a function, generated by the NN feed-forward architecture with this choice of parameters, is a Lipschitz mapping with a large Lipschitz constant, depending on the upper bound $w(n)$ (or $w(W)$ in the case of SNN) on the NN parameters. Thus, we can view NN approximation as an approximation of a class $\mathcal{K}$ via Lipschitz mappings. This type of approximation is studied via the Lipschitz widths $d_n^\gamma(\mathcal{K})_X$ introduced in Petrova and Wojtaszczyk (2023). These widths join the plethora of classical widths available, see Leviatan and Tikhomirov (1993), and give a theoretical bound on the approximation of $\mathcal{K}$ via γ-Lipschitz mappings defined on unit balls in $\mathbb{R}^n$. An almost complete analysis of these widths with parameter $\gamma = const$ or $\gamma = \gamma_n = \lambda^n$, $\lambda > 1$, was given in Petrova and Wojtaszczyk (2023). Here, we show that DNNs whose parameters are bounded by $w(n)$ are Lipschitz mappings and are associated with Lipschitz widths with constant $\gamma = \gamma_n$ (see Theorem 3 and (4.4)), where

$$2^{c_1 n(1+\log_2 w(n))} < \gamma_n < 2^{c_2 n(1+\log_2 w(n))},$$

(note that when $w(n) = const$, we have $\gamma = \gamma_n = \lambda^n$), and Shallow NNs (SNNs) whose parameters are bounded by $w(W)$ are Lipschitz mappings (see Theorem 5 and (4.5)) associated with Lipschitz widths with constant

$$\gamma = \gamma_W = 2^{c(\log_2 W + \log_2 w(W))}.$$

Thus, the investigation of the approximation power of deep or shallow NNs with a general bound $w(\cdot)$ of their parameters requires a study of Lipschitz widths with Lipschitz constant $\gamma = 2^{\varphi(\cdot)}$ with rather general functions φ. In this paper, we provide such a study and its consequences for NN approximation.

The paper is organized as follows. In §2, we introduce our notation and recall the definitions of NNs, Lipschitz widths and entropy numbers. We show in §3 that the NNs under consideration are Lipschitz mappings. Estimates from below for the Lipschitz widths $d_n^{\gamma_n}(\mathcal{K})_X$ with Lipschitz constants $\gamma_n = 2^{\varphi(n)}$ for a compact class $\mathcal{K}$ and their implication for deep and shallow NN approximation of $\mathcal{K}$ are provided in §4. Generalized inverse theorems for NNA are presented in §5. Further properties of $d_n^{\gamma_n}(\mathcal{K})_X$ are discussed in §6. Finally, our concluding remarks are presented in §7, and some lemmas and their proofs are discussed in the Appendix.

1.1 Previous work

While the expressive power of NNs is an extensively studied topic, involving numerous results, such as Telgarsky (2016); Yang and Barron (1999) and many others, we will focus on those that relate to our specific problem.

Estimates from below for the approximation error for classes $\mathcal{K}$ approximated by the outputs of NNs have been available for certain choices of classes $\mathcal{K}$ (such as Hölder balls of smoothness $s > 0$), the space $X = L_\infty$ and either the ReLU activation function, see for example, Yarotsky (2017, 2018); Yang and Zhang (2022), or other activation functions, see Liaw and Mehrabian (2019); Yarotsky and Zhevnerchuk (2020), or for sets $\mathcal{K}$ that are the unit ball of certain Besov classes, see (Hanin and Petrova, 2021, Section 5.9). These results rely on the technique of using the VC dimension of the outputs of the corresponding NNs, the particular structure of the sets $\mathcal{K}$, and utilize the fact that the error is measured in the $\|\cdot\|_{L_\infty}$ norm. Recently, estimates from below for the approximation error (measured in the $X = L_p$ norm, $1 \le p < \infty$) of DNNs for sets $\mathcal{K}$ have been obtained in Gerchinovitz and Malgouyres (2022). There, the authors study the general problem

$$\sup_{f\in\mathcal{K}} \inf_{g_n\in G} \|f - g_n\|_{L_p(\mu)}$$

of approximating $\mathcal{K}$ by the elements g_n of G (which depend on n parameters), where $\mathcal{K}$ and G are real-valued functions, and all functions in $\mathcal{K}$ have the same fixed range. They derive a lower bound for the above quantity, see (Gerchinovitz and Malgouyres, 2022, Theorem 1), that contains the packing number of $\mathcal{K}$, the fixed range of the functions in $\mathcal{K}$, and the fat-shattering dimension of G, using a key probability result of Mendelson. Finally, they used their general result to derive estimates from below for the approximation power of NNs with piecewise-polynomial activation functions, see (Gerchinovitz and Malgouyres, 2022, Corollary 1). Another recent work is Siegel (2022), where the author gives an optimal bound from below for the error of approximation (measured in the $X = L_p$ norm, $p \geq 1$) of ReLU NNs for the class $\mathcal{K}$ being a Sobolev or a Besov unit ball.

Another recently explored venue are estimates for the approximation power of NNs whose parameters are encoded with a fixed number of bits, or the so-called quantized NNs. We refer the reader to Voigtlaender and Petersen (2019); Petersen and Voigtlaender (2018); Gühring and Raslan (2020), where lower bounds for such networks are obtained.

In the case of SNNs with continuous activation functions, estimates from below for the error of NNA are available for sets $\mathcal{K}$ of functions with smoothness s, defined on a compact in $\mathbb{R}^d$ and the space $X = L_2$, see Maiorov (1999); Meir and Ratsaby (1999). Recently, such estimates for SNNs with bounded parameters and certain activation functions have been derived in (Siegel and Xu, 2022, Section 4.1, Corollaries 2,3) for compact sets $\mathcal{K}$ that are the closures of the symmetric convex hull of certain dictionaries and $X = L_2$, see Remark 14.

2. Preliminaries

In this section, we introduce our notation and recall some known facts about NNs, Lipschitz widths and entropy numbers. In what follows, we will denote by $A \gtrsim B$ the fact that there is an absolute constant $c > 0$ such that $A \geq cB$, where A, B are some expressions that depend on n as $n \to \infty$ (or W as $W \to \infty$). Note that the value of c may change from line to line, but is always independent on n. Similarly, we will use the notation $A \lesssim B$, which is defined in an analogues way, and $A \asymp B$ if $A \gtrsim B$ and $A \lesssim B$.

We also will use the notation $A = A(B)$ to stress the fact that the quantity A depends on B. For example, if C is a constant, the expression $C = C(d, W)$ means that C depends on d and W.

2.1 Neural networks

2.1.1 Deep feed-forward neural networks

We denote by $C(\Omega)$ the set of continuous functions defined on the compact set $\Omega \subset \mathbb{R}^d$, equipped with the uniform norm.

A feed-forward NN with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, constant width W and depth n generates a family $\Sigma_{n,\sigma}$ of continuous functions

$$\Sigma_{n,\sigma} := \{\Phi_\sigma(y) : \; y \in \mathbb{R}^{\tilde{n}}, \; \tilde{n} = \tilde{n}(d, W, n) = C_0(d, W)n\} \subset C(\Omega), \quad \Omega \subset \mathbb{R}^d,$$

that is used to produce an approximant to a given function $f \in \mathcal{K}$ or the whole class $\mathcal{K}$. Each parameter vector $y \in \mathbb{R}^{\tilde{n}}$ determines a continuous function $\Phi_\sigma(y) \in \Sigma_{n,\sigma}$, defined on

Ω, of the form

$$\Phi_\sigma(y) := A^{(n)} \circ \bar{\sigma} \circ A^{(n-1)} \circ \ldots \circ \bar{\sigma} \circ A^{(0)}, \tag{2.1}$$

where $\bar{\sigma} : \mathbb{R}^W \to \mathbb{R}^W$ is given by

$$\bar{\sigma}(z_1, \ldots, z_W) = (\sigma(z_1), \ldots, \sigma(z_W)), \tag{2.2}$$

and $A^{(0)} : \mathbb{R}^d \to \mathbb{R}^W$, $A^{(\ell)} : \mathbb{R}^W \to \mathbb{R}^W$, $\ell = 1, \ldots, n-1$, and $A^{(n)} : \mathbb{R}^W \to \mathbb{R}$ are affine mappings. Note that $y \in \mathbb{R}^{\tilde{n}}$ is the vector with coordinates the entries of the matrices and offset vectors (biases) of the affine mappings $A^{(\ell)}$, $\ell = 0, \ldots, n$. We order them in such a way that the entries of $A^{(\ell)}$ appear before those of $A^{(\ell+1)}$ and the ordering for each $A^{(\ell)}$ is done in the same way. For detailed study of such DNNs we refer the reader to Hanin and Petrova (2021) and the references therein. We investigate the approximation power of $\Sigma_{n,\sigma}$ when the width W is fixed and the depth $n \to \infty$.

2.1.2 Shallow neural networks

A shallow NN with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, and width W generates a family $\Xi_{W,\sigma}$ of continuous functions

$$\Xi_{W,\sigma} := \{\Psi_\sigma(y) : \; y \in \mathbb{R}^{\widetilde{W}}, \; \widetilde{W} = C_0(d)W\} \subset C(\Omega), \quad \Omega \subset \mathbb{R}^d,$$

that is used to produce an approximant to a given function $f \in \mathcal{K}$ or the whole class $\mathcal{K}$. Each parameter vector $y \in \mathbb{R}^{\widetilde{W}}$ determines a continuous function $\Psi_\sigma(y) \in \Xi_{W,\sigma}$, defined on Ω, of the form

$$\Psi_\sigma(y) := A^{(1)} \circ \bar{\sigma} \circ A^{(0)}, \tag{2.3}$$

where $\bar{\sigma} : \mathbb{R}^W \to \mathbb{R}^W$ is given as in (2.2) and $A^{(0)} : \mathbb{R}^d \to \mathbb{R}^W$ and $A^{(1)} : \mathbb{R}^W \to \mathbb{R}$ are affine mappings. We investigate the approximation power of $\Xi_{W,\sigma}$ as the width $W \to \infty$.

2.2 Lipschitz widths

Lipschitz widths $d_n^\gamma(\mathcal{K})_X$ for a compact subset $\mathcal{K} \subset X$ of a Banach space X with a Lipschitz constant $\gamma = C_0 = const$ or $\gamma = \gamma_n = C'\lambda^n$ with $\lambda > 1$ were introduced and analyzed in Petrova and Wojtaszczyk (2023). The latter were used to study bounds from below for ReLU DNNs with weights and biases bounded by 1. However, in practice, the weights and biases used in a DNN may grow. This growth affects the Lipschitz constant associated with the corresponding DNN viewed as a Lipschitz mapping. Thus, providing bounds from below for the approximation power of such networks requires the investigation of Lipschitz widths with varying Lipschitz constants γ that depend on n.

Let us first recall the definition of $d_n^\gamma(\mathcal{K})_X$. We denote by $(\mathbb{R}^n, \|.\|_{Y_n})$, $n \in \mathbb{N}$, the n-dimensional Banach space with a fixed norm $\|\cdot\|_{Y_n}$, by

$$B_{Y_n}(r) := \{y \in \mathbb{R}^n : \; \|y\|_{Y_n} \le r\},$$

its ball with radius r, and by

$$\|y\|_{\ell_\infty^n} := \max_{j=1,\ldots,n} |y_j|, \quad \|y\|_{\ell_p^n} := \Big(\sum_{j=1,\ldots,n} |y_j|^p\Big)^{1/p}, \quad 1 \le p < \infty,$$

the usual ℓ_p norms of $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. For $\gamma \geq 0$ we define the *fixed Lipschitz* width $d^\gamma(\mathcal{K}, Y_n)_X$ as $d^0(\mathcal{K}, Y_n)_X := \mathrm{rad}(\mathcal{K}) := \inf_{g\in X} \sup_{f\in\mathcal{K}} \|g - f\|_X$, and for $\gamma > 0$,

$$d^\gamma(\mathcal{K}, Y_n)_X := \inf_{\mathcal{L}_n} \sup_{f\in\mathcal{K}} \inf_{y\in B_{Y_n}(1)} \|f - \mathcal{L}_n(y)\|_X, \tag{2.4}$$

where the infimum is taken over all Lipschitz mappings $\mathcal{L}_n : (B_{Y_n}(1), \|\cdot\|_{Y_n}) \to X$ that satisfy the Lipschitz condition

$$\sup_{y,y'\in B_{Y_n}(1)} \frac{\|\mathcal{L}_n(y) - \mathcal{L}_n(y')\|_X}{\|y - y'\|_{Y_n}} \leq \gamma, \tag{2.5}$$

with constant γ. Then, the *Lipschitz* width $d_n^\gamma(\mathcal{K})_X$ is defined as

$$d_n^0(\mathcal{K})_X := d^0(\mathcal{K}, Y_n)_X, \quad d_n^\gamma(\mathcal{K})_X := \inf_{\|\cdot\|_{Y_n}} d^\gamma(\mathcal{K}, Y_n)_X,$$

where the infimum is taken over all norms $\|\cdot\|_{Y_n}$ in $\mathbb{R}^n$.

Various notions of widths had been introduced and used in approximation theory to theoretically quantify the limitations of certain types of approximations. We refer the reader to Leviatan and Tikhomirov (1993) or Golitschek and Makovoz (1996), where different widths and their decay rates for common smoothness classes have been discussed. Note that the definition of Lipschitz width is similar to the definition of the manifold n-width $\delta_n(\mathcal{K})_X$, (see e.g. Petrova and Wojtaszczyk (2022)) $\delta_n(\mathcal{K})_X := \inf_{M,a} \sup_{f\in\mathcal{K}} \|f - M(a(f))\|_X$, where the infimum is taken over all continuous mappings $a : \mathcal{K} \to \mathbb{R}^n$, $M : \mathbb{R}^n \to X$. However, in the definition of Lipschitz width, we impose the stronger Lipschitz condition on the approximation mapping.

Before going further, we list some of the properties of the Lipschitz width $d_n^\gamma(\mathcal{K})_X$, proved in Petrova and Wojtaszczyk (2023), which we gather in the following theorem.

Theorem 1 *For any $n \in \mathbb{N}$, any compact set $\mathcal{K} \subset X$, and any constant $\gamma > 0$ we have:*

- *$d_n^\gamma(\mathcal{K})_X$ is a monotone decreasing function of γ and n. More precisely,*
  - *If $\gamma_1 \leq \gamma_2$ then $d_n^{\gamma_2}(\mathcal{K})_X \leq d_n^{\gamma_1}(\mathcal{K})_X$;*
  - *If $n_1 \leq n_2$ then $d_{n_2}^\gamma(\mathcal{K})_X \leq d_{n_1}^\gamma(\mathcal{K})_X$.*
- *there is a norm $\|\cdot\|_{\mathcal{Y}_n}$ on $\mathbb{R}^n$ such that $d_n^\gamma(\mathcal{K})_X = d^\gamma(\mathcal{K}, \mathcal{Y}_n)_X$, where we have the inequalities $\|y\|_{\ell_\infty^n} \leq \|y\|_{\mathcal{Y}_n} \leq \|y\|_{\ell_1^n}$ for every $y \in \mathbb{R}^n$.*

Note that the Lipschitz widths $d_n^\gamma(\mathcal{K})_X$ are defined via Lipschitz mappings with domain the unit balls $B_{Y_n}(1)$, see (2.4). However, having in mind approximation via NNs with parameters that are bounded by $w(n)$, we need to consider Lipschitz mappings whose domain are balls $B_{Y_n}(w(n))$ with radius $w(n)$. The next lemma shows how the Lipschitz width $d_n^\gamma(\mathcal{K})_X$ is related to all γ/r-Lipschitz mappings $\mathcal{L}_n$ with domain $B_{Y_n}(r)$, $r > 0$, whose image is in X. More precisely, we prove that in the definition of fixed Lipschitz widths we can consider mappings defined on balls with changing radiuses as long as the product of the Lipschitz constant of the mappings and the radius of the ball does not exceed γ.

Lemma 2 *For any compact subset $\mathcal{K}$ of X, any $\gamma>0$, any $n\in\mathbb{N}$, and any norm $\|\cdot\|_{Y_n}$ on $\mathbb{R}^n$, we have that*

$$d^{\gamma}(\mathcal{K},Y_n)_X=\inf_{\mathcal{L}_n,r>0}\ \sup_{f\in\mathcal{K}}\ \inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n(y)\|_X, \tag{2.6}$$

where the infimum is taken over all γ/r-Lipschitz mappings $\mathcal{L}_n:(B_{Y_n}(r),\|\cdot\|_{Y_n})\to X$, and all $r>0$.

Proof: We provide the proof in the Appendix. □

2.3 Entropy numbers

We recall, see e.g. Carl (1981); Carl and Stephani (1990); Golitschek and Makovoz (1996), that the *entropy numbers* $\epsilon_n(\mathcal{K})_X$, $n\geq 0$, of a compact set $\mathcal{K}\subset X$ are defined as the infimum of all $\epsilon>0$ for which 2^n balls with centers from X and radius ϵ cover $\mathcal{K}$. Formally, we write

$$\epsilon_n(\mathcal{K})_X=\inf\{\epsilon>0\ :\ \mathcal{K}\subset\bigcup_{j=1}^{2^n}B(g_j,\epsilon),\ g_j\in X,\ j=1,\dots,2^n\}.$$

3. Neural networks are Lipschitz mappings

Our choice of norm when working with NNs is the $\|\cdot\|_{\ell_\infty^n}$ norm of the parameters y of the neural network. This is simply because we are interested in the asymptotic behavior (with respect to the depth n of the DNN or the width W of the SNN) of the approximation error that the network provides for a class $\mathcal{K}$ and not in the best possible constants in the error estimates.

3.1 Deep neural networks

We do not investigate what information about the function f is given or what methods one employs to find an appropriate parameter vector $y^*\in\mathbb{R}^{\tilde{n}}$ such that the function $\Phi(y^*)$ is the (near)best approximant to f from the set $\Sigma_{n,\sigma}$, but rather focus on the properties of the mapping

$$y\to\Phi_\sigma(y),\quad \Phi_\sigma(y)\in\Sigma_{n,\sigma},$$

where all parameters (entries of the matrices and biases) are bounded by $w(n)$, where $w(n)\geq 1$. To illustrate the dependence on $w(n)$, we denote the collection of all such mappings as $\Sigma_{n,\sigma}(w(n))$, namely

$$\Sigma_{n,\sigma}(w(n)):=\Phi_\sigma(B_{\ell_\infty^{\tilde{n}}}(w(n))),$$

with Φ_σ being defined in (2.1). We have shown in Petrova and Wojtaszczyk (2023) that in the case $w(n)\equiv 1$, $\sigma=\text{ReLU}$, and $\Omega=[0,1]^d$, the mapping

$$\Phi_\sigma:(B_{\ell_\infty^{\tilde{n}}}(1),\|\cdot\|_{\ell_\infty^{\tilde{n}}})\to C(\Omega),\quad \tilde{n}=C_0n,\quad C_0=C_0(W),$$

defined in (2.1) is a Lipschitz mapping with Lipschitz constant L_n, where $2^{c_2n}<L_n<2^{c_1n}$ for fixed constants $c_1,c_2>0$ depending on the width W. More precisely, we have

$$\|\Phi_\sigma(z)-\Phi_\sigma(y)\|_{C([0,1]^d)}\leq L_n\|z-y\|_{\ell_\infty^{\tilde{n}}}\quad\text{for all}\quad z,y\in B_{\ell_\infty^{\tilde{n}}}(1).$$

Here, we will investigate what is the Lipschitz constant L_n when the parameters y have components bounded by $w(n)$, namely $y \in B_{\ell_\infty^{\tilde{n}}}(w(n))$, and the activation function σ is either $\mathrm{ReLU}(t) := \max\{0, t\} = t_+$ or a bounded Lipschitz function with Lipschitz constant L. Note that ReLU is a Lipschitz function with a Lipschitz constant $L = 1$.

In what follows, we use the notation $\|g\| := \max\limits_{1 \le i \le W} \|g_i\|_{C(\Omega)}$, when working with vector functions $g = (g_1, \dots, g_W)^T$ whose coordinates $g_i \in C(\Omega)$.

Theorem 3 *Let X be a Banach space such that $C([0,1]^d) \subset X$ is continuously embedded in X. Then the mapping $\Phi_\sigma : (B_{\ell_\infty^{\tilde{n}}}(w(n)), \|\cdot\|_{\ell_\infty^{\tilde{n}}}) \to X$, defined in (2.1), is an L_n-Lipschitz mapping, that is*

$$\|\Phi_\sigma(y) - \Phi_\sigma(y')\|_X \le c_0 \|\Phi_\sigma(y) - \Phi_\sigma(y')\|_{C(\Omega)} \le L_n \|y - y'\|_{\ell_\infty^{\tilde{n}}}, \quad y, y' \in B_{\ell_\infty^{\tilde{n}}}(w(n)),$$

where the constant L_n is bounded by

$$2^{c_1 n(1+\log_2 w(n))} < L_n < 2^{c_2 n(1+\log_2 w(n))},$$

provided σ is a bounded L-Lipschitz function or σ =ReLU and $LWw(n) \ge 2$. The constants c_1, c_2 depend on c_0, d, W, and the function σ.

Proof: The proof follows the arguments from the proof of Theorem 6.1 in Petrova and Wojtaszczyk (2023). Let y, y' be the two parameters from $B_{\ell_\infty^{\tilde{n}}}(w(n))$ that determine the continuous functions $\Phi_\sigma(y), \Phi_\sigma(y') \in \Sigma_{\sigma,n}(w(n))$. They are constructed by ordering in a predetermined way the entries of the affine mappings $A^{(j)}(\cdot) := A_j(\cdot) + b^{(j)}$, $j = 0, \dots, n$, and $A'^{(j)}(\cdot) := A'_j(\cdot) + b'^{(j)}$, $j = 0, \dots, n$, that define $\Phi_\sigma(y)$ and $\Phi_\sigma(y')$, respectively. We fix $x \in \Omega$ and denote by

$$\eta^{(0)}(x) := \overline{\sigma}(A_0 x + b^{(0)}), \quad \eta'^{(0)}(x) := \overline{\sigma}(A'_0 x + b'^{(0)}),$$

$$\eta^{(j)} := \overline{\sigma}(A_j \eta^{(j-1)} + b^{(j)}), \quad \eta'^{(j)} := \overline{\sigma}(A'_j \eta'^{(j-1)} + b'^{(j)}), \quad j = 1, \dots, n-1,$$

$$\eta^{(n)} := A_n \eta^{(n-1)} + b^{(n)}, \quad \eta'^{(n)} := A'_n \eta'^{(n-1)} + b'^{(n)}.$$

Note that $A_0, A'_0 \in \mathbb{R}^{W \times d}$, $A_j, A'_j \in \mathbb{R}^{W \times W}$, $b^{(0)}, b'^{(0)}, b^{(j)}, b'^{(j)} \in \mathbb{R}^W$, for $j = 1, \dots, n-1$, while $A_n, A'_n \in \mathbb{R}^{1 \times W}$, and $b^{(n)}, b'^{(n)} \in \mathbb{R}$. Each of the $\eta^{(j)}, \eta'^{(j)}$, $j = 0, \dots, n-1$, is a continuous vector function with W coordinates and $\eta^{(n)}, \eta'^{(n)}$ are the outputs of the DNN with activation function σ and parameters y, y', respectively.

Case 1: DNN with activation function a bounded Lipschitz function σ.

Observe that in this case

$$|\sigma(t)| \le \tilde{C}, \quad |\sigma(t_1) - \sigma(t_2)| \le L|t_1 - t_2|, \quad t_1, t_2 \in \mathbb{R},$$

and therefore for any m, vectors $\bar{y}, \hat{y}, \eta \in \mathbb{R}^m$ and numbers $y_0, \hat{y}_0 \in \mathbb{R}$, where $\bar{y}, y_0$ and $\hat{y}, \hat{y}_0$ are subsets of the coordinates of $y, y' \in \mathbb{R}^{\tilde{n}}$, respectively, we have

$$|\sigma(\bar{y} \cdot \eta + y_0)| \le \tilde{C}, \quad |\sigma(\bar{y} \cdot \eta + y_0) - \sigma(\hat{y} \cdot \eta + \hat{y}_0)| \le L(m\|\eta\|_{\ell_\infty^m} + 1)\|y - y'\|_{\ell_\infty^{\tilde{n}}}. \quad (3.1)$$

It then follows, see (3.1), that $\|\eta'^{(j)}\| \le \tilde{C}$, $j = 0, \dots, n-1$, and

$$\|\eta^{(0)} - \eta'^{(0)}\| \le L(d+1)\|y - y'\|_{\ell_\infty^{\tilde{n}}} =: C_0 \|y - y'\|_{\ell_\infty^{\tilde{n}}}.$$

Suppose we have proved the inequality $\|\eta^{(j-1)}-\eta'^{(j-1)}\|\le C_{j-1}\|y-y'\|_{\ell_\infty^{\tilde n}}$. Then we have

$$\begin{aligned}\|\eta^{(j)}-\eta'^{(j)}\| &\le L\|A_j\eta^{(j-1)}+b^{(j)}-A'_j\eta'^{(j-1)}-b'^{(j)}\|\\ &\le L\|A_j(\eta^{(j-1)}-\eta'^{(j-1)})\|+L\|(A_j-A'_j)\eta'^{(j-1)}\|+L\|b^{(j)}-b'^{(j)}\|\\ &\le LW\|y\|_{\ell_\infty^{\tilde n}}\|\eta^{(j-1)}-\eta'^{(j-1)}\|+LW\|y-y'\|_{\ell_\infty^{\tilde n}}\|\eta'^{(j-1)}\|+L\|y-y'\|_{\ell_\infty^{\tilde n}}\\ &\le (LWw(n)C_{j-1}+LW\tilde C+L)\|y-y'\|_{\ell_\infty^{\tilde n}}\\ &=: C_j\|y-y'\|_{\ell_\infty^{\tilde n}},\end{aligned}$$

where we have used that $\|y\|_{\ell_\infty^{\tilde n}}\le w(n)$, $\|\eta'^{(j)}\|\le\tilde C$, and the induction hypothesis. Thus, the relation between C_j and C_{j-1} is

$$C_0=L(d+1),\quad C_j=LWw(n)C_{j-1}+LW\tilde C+L,\quad j=1,\dots,n.$$

If we denote by $A:=LWw(n)\ge 2$ and $B:=LW\tilde C+L=L(W\tilde C+1)$, we have that

$$\begin{aligned}C_j &= AC_{j-1}+B=\ldots=A^jC_0+(A^{j-1}+\ldots+1)B\\ &\le (A^j+\ldots+1)L(\max\{W\tilde C,d\}+1)=\frac{A^{j+1}-1}{A-1}L(\max\{W\tilde C,d\}+1)\\ &\le 2A^jL(\max\{W\tilde C,d\}+1)=C'[LWw(n)]^j,\quad C':=2L(\max\{W\tilde C,d\}+1).\end{aligned}$$

Finally, since $\|\Phi_\sigma(y)-\Phi_\sigma(y')\|_{C(\Omega)}=\|\eta^{(n)}-\eta'^{(n)}\|$, we have

$$\|\Phi_\sigma(y)-\Phi_\sigma(y')\|_X\le c_0\|\Phi_\sigma(y)-\Phi_\sigma(y')\|_{C(\Omega)}\le c_0C_n\|y-y'\|_{\ell_\infty^{\tilde n}}<C'[LWw(n)]^n\|y-y'\|_{\ell_\infty^{\tilde n}}.$$

The next case follows the same idea with several slight modifications.

Case 2: DNN with activation function $\sigma=$ReLU.
Observe that for any m, vectors $\bar y,\hat y,\eta\in\mathbb{R}^m$ and numbers $y_0,\hat y_0\in\mathbb{R}$, where $\bar y,y_0$ and $\hat y,\hat y_0$ are subsets of the coordinates of $y,y'\in\mathbb{R}^{\tilde n}$, respectively, we have

$$|(\bar y\cdot\eta+y_0)_+|\le(m\|\eta\|_{\ell_\infty^m}+1)\|y\|_{\ell_\infty^{\tilde n}}\le(m\|\eta\|_{\ell_\infty^m}+1)w(n).\tag{3.2}$$

Note that since $\|y\|_{\ell_\infty^{\tilde n}}\le w(n)$, it follows from (3.2) that $\|\eta'^{(0)}\|\le dw(n)+w(n)$ (when $m=d$ and $\eta=x$), and $\|\eta'^{(j)}\|\le Ww(n)\|\eta'^{(j-1)}\|+w(n)$, $j=1,\dots,n$ (when $m=W$ and $\eta=\eta'^{(j-1)}$). We want to point out that the last two inequalities hold even if we use $\sigma(t)=t$ instead of $\sigma(t)=\mathrm{ReLU}(t)$ for some of the coordinates in the definition (2.2) of $\bar\sigma$. One can show by induction that for $j=1,\dots,n$,

$$\begin{aligned}\|\eta'^{(j)}\| &\le dW^jw(n)^{j+1}+w(n)\sum_{i=0}^{j}[Ww(n)]^i\le dW^jw(n)^{j+1}+2w(n)[Ww(n)]^j\\ &= (d+2)w(n)[Ww(n)]^j,\end{aligned}$$

since $Ww(n)\ge 2$ (note $L=1$ in this case). The above inequality also holds for $j=0$. Clearly, we have

$$\|\eta^{(0)}-\eta'^{(0)}\|\le(d+1)\|y-y'\|_{\ell_\infty^{\tilde n}}=:D_0\|y-y'\|_{\ell_\infty^{\tilde n}}.$$

Suppose we have proved that $\|\eta^{(j-1)}-\eta'^{(j-1)}\| \le D_{j-1}\|y-y'\|_{\ell_\infty^{\tilde{n}}}$. Then, similarly to Case 1 (since ReLU is a Lipschitz function with a Lipschitz constant $L=1$), we obtain that

$$\begin{aligned}\|\eta^{(j)}-\eta'^{(j)}\| &\le \|A_j(\eta^{(j-1)}-\eta'^{(j-1)})\|+\|(A_j-A_j')\eta'^{(j-1)}\|+\|b^{(j)}-b'^{(j)}\|\\ &\le W\|y\|_{\ell_\infty^{\tilde{n}}}\|\eta^{(j-1)}-\eta'^{(j-1)}\|+W\|y-y'\|_{\ell_\infty^{\tilde{n}}}\|\eta'^{(j-1)}\|+\|y-y'\|_{\ell_\infty^{\tilde{n}}}\\ &\le (Ww(n)D_{j-1}+(d+2)[Ww(n)]^j+1)\|y-y'\|_{\ell_\infty^{\tilde{n}}}\\ &=: D_j\|y-y'\|_{\ell_\infty^{\tilde{n}}}.\end{aligned}$$

Thus, $\|\eta^{(j)}-\eta'^{(j)}\| \le D_j\|y-y'\|_{\ell_\infty^{\tilde{n}}}$, where $D_j = Ww(n)D_{j-1}+(d+2)[Ww(n)]^j+1$. Since $D_0 = d+1$ and

$$D_1 = Ww(n)D_0+(d+2)Ww(n)+1 < (d+2)(2Ww(n)+1),$$

we obtain by induction that

$$D_n \;<\; (d+2)\left(n[Ww(n)]^n+\sum_{i=0}^{n}[Ww(n)]^i\right) < (d+2)(n+2)[Ww(n)]^n,$$

where we have used that $2 \le Ww(n)$. Finally, we have

$$\|\Phi_\sigma(y)-\Phi_\sigma(y')\|_{C(\Omega)} = \|\eta^{(n)}-\eta'^{(n)}\| \le D_n\|y-y'\|_{\ell_\infty^{\tilde{n}}} < (d+2)(n+2)[Ww(n)]^n\|y-y'\|_{\ell_\infty^{\tilde{n}}}.$$

In both Case 1 and Case 2, the Lipschitz constant L_n is such that we can find constants $c_1, c_2 > 0$ with the property $2^{c_1 n(1+\log_2 w(n))} < L_n < 2^{c_2 n(1+\log_2 w(n))}$ and the proof is completed. □

Remark 4 *Note that one can follow the same arguments as in Case* 2 *and prove Theorem* 3 *when every coordinate of* $\bar{\sigma}$*, see* (2.2)*, is chosen to be either* $\sigma(t)=\mathrm{ReLU}(t)$ *or* $\sigma(t)=t$*, and this choice can change from layer to layer.*

3.2 Shallow neural networks

In this section, we consider SNNs and prove that they are also Lipschitz mappings. The following theorem holds.

Theorem 5 *Let* X *be a Banach space such that* $C([0,1]^d)\subset X$ *is continuously embedded in* X*. Then the mapping* $\Psi_\sigma : (B_{\ell_\infty^{\widetilde{W}}}(w(W)), \|\cdot\|_{\ell_\infty^{\widetilde{W}}}) \to X$*, defined in* (2.3)*, is an* L_W*-Lipschitz mapping, that is*

$$\|\Psi_\sigma(y)-\Psi_\sigma(y')\|_X \le c_0\|\Psi_\sigma(y)-\Psi_\sigma(y')\|_{C(\Omega)} \le L_W\|y-y'\|_{\ell_\infty^{\widetilde{W}}}, \quad y,y'\in B_{\ell_\infty^{\widetilde{W}}}(w(W)),$$

with constant $L_W = CWw(W)$*, where* $C=C(d,\sigma,c_0)$*,* $w(W)\ge 1$*, and* σ *is either the* ReLU *function or a bounded Lipschitz function.*

Proof: We follow the proof of Theorem 3. In Case 1 we have

$$L_W = c_0C_1 = c_0(L^2Ww(W)(d+1)+LW\tilde{C}+L) \le CWw(W), \quad C = c_0(L^2(d+1)+L\tilde{C}+L),$$

and in Case 2 we have

$$L_W = c_0D_1 = c_0((2d+3)Ww(W)+1) \le CWw(W), \quad C = c_0(2d+4),$$

provided $w(W)\ge 1$. □

4. Bounds for Lipschitz widths and the limitations of NN approximation

In this section, we obtain estimates from below for Lipschitz widths with large Lipschitz constants and apply them in the case of deep and shallow NNs.

4.1 Bounds from below for Lipschitz widths

We start our investigation with the observation, that the Lipschitz width $d_n^\gamma(\mathcal{K})_X$, $\gamma>0$, is bounded from below by the fixed width $d^{(2n+1)\gamma}(\mathcal{K},Z_n)_X$ with respect to any chosen in advance norm $\|\cdot\|_{Z_n}$ on $\mathbb{R}^n$ by paying the price of the slightly bigger Lipschitz constant $(2n+1)\gamma$. Note that $d_n^\gamma(\mathcal{K})_X$ is defined as infimum over all norms in $\mathbb{R}^n$, and thus the lemma below provides a way for calculating lower estimates for the Lipschitz width since we can perform our computations using our favorable (easy to handle) norm Z_n.

Lemma 6 *For any compact subset $\mathcal{K}$ of X, any $\gamma>0$, any $n\in\mathbb{N}$, and any norm $\|\cdot\|_{Z_n}$ on $\mathbb{R}^n$ we have that*

$$d^{(2n+1)\gamma}(\mathcal{K},Z_n)_X \le d_n^\gamma(\mathcal{K})_X.$$

Proof: We provide the proof in the Appendix. □

Remark 7 *In the above inequality we may choose the Z_n norm to be the ℓ_2^n norm, in which case $\rho\le\sqrt{n}$, see (Albiac and Kalton, 2016, Theorem 13.1.5). Thus, for any compact subset $\mathcal{K}$ of X, any $\gamma>0$, and any $n\in\mathbb{N}$, we have that*

$$d_n^{(2\sqrt{n}+1)\gamma}(\mathcal{K})_X \le d^{(2\sqrt{n}+1)\gamma}(\mathcal{K},\ell_2^n)_X \le d_n^\gamma(\mathcal{K})_X.$$

We now proceed with the investigation of other bounds from below for the Lipschitz widths by first recalling the following proposition, see (Petrova and Wojtaszczyk, 2023, Prop 3.8).

Proposition 8 *Let $\mathcal{K}\subset X$ be a compact set and let $\epsilon_n(\mathcal{K})_X>\eta_n$, $n\in\mathbb{N}$, where $(\eta_n)_{n=1}^\infty$ is a sequences of real numbers decreasing to zero. If for some $m\in\mathbb{N}$ and $\delta>0$ we have that $d_m^\gamma(\mathcal{K})_X<\delta$, then*

$$\eta_{m\log_2(3\gamma\delta^{-1})}<2\delta. \tag{4.1}$$

This proposition was used in Petrova and Wojtaszczyk (2023) to prove bounds from below for the Lipschitz widths $d_n^{\gamma_n}(\mathcal{K})_X$ of the compact set $\mathcal{K}$ in the cases $\gamma_n=2^{\varphi(n)}$, $\varphi(n)=$ *const*, see (Petrova and Wojtaszczyk, 2023, Theorem 3.9) and $\varphi(n)=c'n$, see (Petrova and Wojtaszczyk, 2023, Theorem 6.3), provided we have information about the entropy numbers of $\mathcal{K}$. The next theorem is a generalization of Theorem 6.3 from Petrova and Wojtaszczyk (2023) for the case of a general function φ.

Theorem 9 *For any compact set $\mathcal{K}\subset X$, we consider the Lipschitz width $d_n^{\gamma_n}(\mathcal{K})_X$ with Lipschitz constant $\gamma_n=2^{\varphi(n)}$, where $\varphi(n)\ge c\log_2 n$ for some fixed constant $c>0$. Then the following holds:*

$$\text{(i)}\quad \epsilon_n(\mathcal{K})_X \gtrsim \frac{(\log_2 n)^\beta}{n^\alpha},\quad n\in\mathbb{N} \quad\Rightarrow\quad d_n^{\gamma_n}(\mathcal{K})_X \gtrsim [\log_2(n\varphi(n))]^\beta[n\varphi(n)]^{-\alpha},\quad n\in\mathbb{N}, \tag{4.2}$$

where $\alpha > 0$, $\beta \in \mathbb{R}$.

$$\text{(ii)} \qquad \epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha}, \quad n \in \mathbb{N} \Rightarrow \quad d_n^{\gamma_n}(\mathcal{K})_X \gtrsim [\log_2(n\varphi(n))]^{-\alpha}, \quad n \in \mathbb{N}, \tag{4.3}$$

where $\alpha > 0$.

Proof: We provide the proof in the Appendix. □

4.2 Bounds from below for DNN approximation

In this section we consider Banach spaces X such that $C([0,1]^d)$ is continuously embedded in X. Let us denote by

$$E(f, \Sigma_{0,\sigma}(w(n)))_X := \|f\|_X, \quad E(f, \Sigma_{n,\sigma}(w(n)))_X := \inf_{y \in B^{\bar{n}}_{\ell_\infty}(w(n))} \|f - \Phi_\sigma(y)\|_X, \quad n \in \mathbb{N},$$

the error of approximation of a function f by the outputs $\Phi_\sigma(y) \in \Sigma_{n,\sigma}(w(n))$ of the DNN with parameters y, for which $\|y\|_{\ell_\infty^{\bar{n}}} \leq w(n)$, $LWw(n) \geq 2$, measured in $\|\cdot\|_X$. We then denote by

$$E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X := \sup_{f \in \mathcal{K}} E(f, \Sigma_{n,\sigma}(w(n)))_X, \quad n \geq 0,$$

the error for the class $\mathcal{K}$. It follows from Lemma 2 and Theorem 3 that

$$E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \geq d_n^{\gamma_n}(\mathcal{K})_X, \quad \text{with} \quad \gamma_n = 2^{cn(1+\log_2 w(n))} =: 2^{\varphi(n)}, \quad c > 0. \tag{4.4}$$

The latter estimate shows that bounds from below for the error $E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X$ can be obtained by using bounds from below for $d_n^{\gamma_n}(\mathcal{K})_X$, and thus provides a way to study the theoretical limitations of DNNs with ReLU or bounded Lipschitz activation functions and $w(n)$ bounds for their parameters. We next apply the results obtained in §4.1 to the special case of DNNs.

Remark 10 *It follows from* (4.4)*, Lemma* 6 *with* $\gamma_n = 2^{cn(1+\log_2 w(n))}$*, and the monotonicity with respect to* γ *of the fixed Lipschitz width that for any compact subset* $\mathcal{K}$ *of a Banach space* X*,*

$$E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \geq d_n^{\gamma_n}(\mathcal{K})_X \geq d^{(2n+1)\gamma_n}(\mathcal{K}, Z_n)_X \geq d^{2^{c_1 n(1+\log_2 w(n))}}(\mathcal{K}, Z_n)_X,$$

where $\|\cdot\|_{Z_n}$ *is any norm on* $\mathbb{R}^n$*.*

The following Table 1 shows the relation (for sufficiently large n) between the bound $w(n)$ and the parameter $\gamma_n = 2^{\varphi(n)}$ of the Lipschitz width $d_n^{\gamma_n}(\mathcal{K})_X$ from (4.4), where $\varphi(n) = cn(1 + \log_2 w(n))$.

The next corollary follows from (4.4) and Theorem 9 when $\varphi(n) = cn(1 + \log_2 w(n))$, $c > 0$.

Corollary 11 *Let* $\Sigma_{n,\sigma}(w(n))$ *be the set of outputs of a DNN with depth* n*, fixed width* W*, bounded* L*-Lipschitz or ReLU activation function* σ*, and weights and biases bounded by* $w(n)$*, where* $LWw(n) \geq 2$ *(*$L = 1$ *when* $\sigma =$ReLU*). Then, the error of approximation*

$w(n)$	$\varphi(n) = cn(1+\log_2 w(n))$	$\gamma_n = 2^{\varphi(n)}$
$C \geq 1$	$c'n, \quad c' > 0$	$\lambda^n, \quad \lambda > 1$
$Cn^\delta, \quad C, \delta > 0$	$c'n\log_2 n, \quad c' > 0$	$2^{c'n\log_2 n}, \quad c' > 0$
$C2^{cn^\nu}, \quad C, c, \nu > 0$	$c'n^{\nu+1}, \quad c' > 0$	$2^{c'n^{\nu+1}}, \quad c' > 0$

Table 1: Relation between $w(n)$, $\varphi(n)$ and γ_n.

$E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X$ of a compact subset $\mathcal{K}$ of a Banach space X by the outputs $\Sigma_{n,\sigma}(w(n))$ satisfies the following estimates from below, provided we know the following information about the entropy numbers $\epsilon_n(\mathcal{K})_X$ of $\mathcal{K}$:

$$\epsilon_n(\mathcal{K})_X \gtrsim \frac{(\log_2 n)^\beta}{n^\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \gtrsim \frac{[\log_2 n + \log_2(1+\log_2 w(n))]^\beta}{[n^2(1+\log_2 w(n))]^\alpha},\ n \in \mathbb{N},$$

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \gtrsim [\log_2 n + \log_2(1+\log_2 w(n))]^{-\alpha},\ n \in \mathbb{N}.$$

In particular, if $w(n) = Cn^\delta$, with $\delta > 0$, we have:

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^\beta n^{-\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(Cn^\delta))_X \gtrsim [\log_2 n]^{\beta-\alpha} n^{-2\alpha},\ n \in \mathbb{N},$$

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(Cn^\delta))_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N},$$

and when $w(n) = C2^{cn^\nu}$, with $C, c > 0$, $\nu \geq 0$, we have:

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^\beta n^{-\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(C2^{cn^\nu}))_X \gtrsim [\log_2 n]^\beta n^{-(2+\nu)\alpha},\ n \in \mathbb{N},$$

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N} \Rightarrow E(\mathcal{K}, \Sigma_{n,\sigma}(C2^{cn^\nu}))_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N}.$$

4.3 Bounds from below for shallow neural network approximation

In this section, we consider Banach spaces X such that $C([0,1]^d)$ is continuously embedded in X. Let us denote by

$$E(f, \Xi_{W,\sigma}(w(W)))_X := \|f\|_X, \quad W = 0, 1,$$

$$E(f, \Xi_{W,\sigma}(w(W)))_X := \inf_{y \in B_{\ell_\infty}^{\widetilde{W}}(w(W))} \|f - \Psi_\sigma(y)\|_X, \quad W \geq 2,$$

the error of approximation of a function f by the outputs $\Psi_\sigma(y) \in \Xi_{W,\sigma}(w(W))$ of the SNN with width W, parameters y for which $\|y\|_{\ell_\infty^{\widetilde{W}}} \leq w(W)$, $w(W) \geq 1$, measured in the norm of the Banach space X, and by

$$E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X := \sup_{f \in \mathcal{K}} E(f, \Xi_{W,\sigma}(w(W)))_X, \quad W \geq 0,$$

the error for the class $\mathcal{K}$. Note that the sets $\Xi_{W,\sigma}(w(W))$ are nested, namely

$$\Xi_{W,\sigma}(w(W)) \subset \Xi_{W+1,\sigma}(w(W)),$$

and therefore $E(f,\Xi_{W+1,\sigma}(w(W)))_X \le E(f,\Xi_{W,\sigma}(w(W)))_X$.

It follows from Lemma 2 and Theorem 5 that

$$E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X \ge d_W^{\gamma_W}(\mathcal{K})_X, \quad \text{with} \quad \gamma_W = 2^{c(\log_2 W+\log_2 w(W))} =: 2^{\varphi(W)}, \tag{4.5}$$

and thus estimates from below for the error $E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X$ can be obtained by using bounds from below for $d_W^{\gamma_W}(\mathcal{K})_X$. We next apply the results obtained in §4.1 to the special case of SNNs.

Remark 12 *It follows from* (4.5), *Lemma* 6 *with* $\gamma_W = 2^{c(\log_2 W+\log_2 w(W))}$, *and the monotonicity with respect to* γ *of the fixed Lipschitz width that for any compact subset* $\mathcal{K}$ *of a Banach space* X*,*

$$E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X \ge d_W^{\gamma_W}(\mathcal{K})_X \ge d^{(2W+1)\gamma_W}(\mathcal{K},Z_W)_X \ge d^{2^{c_1(\log_2 W+\log_2 w(W))}}(\mathcal{K},Z_W)_X,$$

where $c_1>0$ *is a fixed constant and* $\|\cdot\|_{Z_W}$ *is any norm on* $\mathbb{R}^W$*.*

As in the case of DNNs, we create a table showing the relation (for sufficiently large W) between the bound $w(W)$ and the parameter $\gamma_W = 2^{\varphi(W)}$ of the Lipschitz width $d_W^{\gamma_W}(\mathcal{K})_X$ from (4.5), where $\varphi(W) = c(\log_2 W + \log_2 w(W))$.

$w(W)$	$\varphi(W) = c(\log_2 W+\log_2 w(W))$	$\gamma_W = 2^{\varphi(W)}$
$CW^\delta, \quad C,\delta \ge 0$	$c'\log_2 W, \quad c'>0$	$2^{c'\log_2 W}$
$C2^{cW^\nu}, \quad C,c,\nu>0$	$c'W^\nu, \quad c'>0$	$2^{c'W^\nu}, \quad c'>0$

Table 2: Relation between $w(W)$, $\varphi(W)$ and γ_W, SNNs

The next corollary follows from (4.5) and Theorem 9 when $\varphi(W) = c(\log_2 W + \log_2 w(W))$, $c>0$.

Corollary 13 *Let* $\Xi_{W,\sigma}(w(W))$ *be the set of outputs of a SNN with width* W*, bounded L-Lipschitz or ReLU activation function* σ*, and* $w(W)\ge 1$ *bound on its parameters. Then the error of approximation* $E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X$ *of a compact subset* $\mathcal{K}$ *of a Banach space* X *by* $\Xi_{W,\sigma}(w(W))$ *satisfies the following bounds from below, given the behavior of its entropy numbers* $\epsilon_W(\mathcal{K})_X$*,* $W\in\mathbb{N}$*:*

$$\epsilon_W(\mathcal{K})_X \gtrsim \frac{(\log_2 W)^\beta}{W^\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X \gtrsim \frac{[\log_2 W + \log_2(\log_2 W+\log_2 w(W))]^\beta}{[W(\log_2 W+\log_2 w(W))]^\alpha},$$

$$\epsilon_W(\mathcal{K})_X \gtrsim [\log_2 W]^{-\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(w(W)))_X \gtrsim [W(\log_2 W+\log_2 w(W))]^{-\alpha},\ W\in\mathbb{N}.$$

In particular, if $w(W) = CW^\delta$*, with* $\delta\ge 0$*, we have:*

$$\epsilon_W(\mathcal{K})_X \gtrsim [\log_2 W]^\beta W^{-\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(CW^\delta))_X \gtrsim [\log_2 W]^{\beta-\alpha}W^{-\alpha},\ W\in\mathbb{N},$$

$$\epsilon_W(\mathcal{K})_X \gtrsim [\log_2 W]^{-\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(CW^\delta))_X \gtrsim [\log_2 W]^{-\alpha},\ W\in\mathbb{N},$$

and when $w(W) = C2^{cW^\nu}$*, with* $C\ge 1$*,* $c>0$*,* $\nu>0$*, we have:*

$$\epsilon_W(\mathcal{K})_X \gtrsim [\log_2 W]^\beta W^{-\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(C2^{cW^\nu}))_X \gtrsim [\log_2 W]^\beta W^{-(1+\nu)\alpha},\ W\in\mathbb{N},$$

$$\epsilon_W(\mathcal{K})_X \gtrsim [\log_2 W]^{-\alpha} \Rightarrow E(\mathcal{K},\Xi_{W,\sigma}(C2^{cW^\nu}))_X \gtrsim [\log_2 W]^{-\alpha},\ W\in\mathbb{N}.$$

Remark 14 *Let Ω be the unit Euclidean ball in $\mathbb{R}^d$ and let us denote by*

$$\mathcal{P}_1^d := \{\mathrm{ReLU}(\omega \cdot x + b) : \omega \in S^{d-1}, b \in [c_1, c_2] \subset \mathbb{R}\} \subset L_2(\Omega),$$

$$S^{d-1} := \{w \in \mathbb{R}^d : \|w\|_{\ell_2^d} = 1\},$$

and consider the closure of the convex, symmetric hull of $\mathcal{P}_1^d$, that is

$$\mathcal{K} = \overline{\{\sum_{j=1}^n c_j h_j,\ h_j \in \mathcal{P}_1^d,\ \sum_{j=1}^n |a_j| \leq 1\}}.$$

We also denote by $E_W(\mathcal{K})_{L_2(\Omega)}$ the error $E_W(\mathcal{K})_{L_2(\Omega)} := E(\mathcal{K}, \Xi_{W,\mathrm{ReLU}}(C))_X$ of approximation of the class $\mathcal{K}$ by the outputs $\Xi_{W,\mathrm{ReLU}}(C)$ of a SNN with width W, bounded by C parameters, and a ReLU activation function, measured in the $X = L_2(\Omega)$ norm. Then, Corollary 2 *from* Siegel and Xu (2022) *applied in the case $k = 1$ states that for any $\delta > \frac{1}{2} + \frac{3}{2d}$ we have*

$$\sup_{W \geq 1} W^\delta E_W(\mathcal{K})_{L_2(\Omega)} = \infty.$$

In this particular case, we can apply our theory since we know the entropy numbers of the class $\mathcal{K}$. Indeed, it follows from Theorem 9 in Siegel and Xu (2022) that for $d \geq 2$,

$$\epsilon_W(\mathcal{K})_{L_2(\Omega)} \geq W^{-\frac{1}{2} - \frac{3}{2d}}.$$

It follows then from Corollary 13 with $\beta = 0$ and $\alpha = \frac{1}{2} + \frac{3}{2d}$ that

$$E_W(\mathcal{K})_{L_2(\Omega)} \geq [W \log_2 W]^{-\frac{1}{2} - \frac{3}{2d}},$$

and therefore we arrive at the same conclusion.

5. Bounds for the entropy numbers via Lipschitz widths and the error of neural network approximation

So far, we have been investigating the behavior of the Lipschitz widths given the behavior of the entropy numbers of a class $\mathcal{K}$. We can ask the inverse question, namely, what does the asymptotic behavior of the Lipschitz widths tell us about the entropy numbers of $\mathcal{K}$? Any results in this direction could be viewed as inverse theorems, and in particular, as generalized inverse theorems for NNA. Historically, inverse theorems have been used in approximation theory to characterize approximation spaces (and describe them as certain interpolation spaces), see (DeVore and Lorentz, 1993, Theorem 5.1, Chapter 7) and (DeVore, 1998, Theorem 1). Thus, providing certain inverse theorems for DNNA and SNNA could possibly pave the way to a complete characterization of the spaces of functions that are well approximated via these NNs.

5.1 Upper bounds for entropy numbers via Lipschitz widths

We start with a lemma that is an extension of Lemma 3.7 from Petrova and Wojtaszczyk (2023).

Lemma 15 *If $\mathcal{K} \subset X$ is a compact set, $\gamma_n = 2^{\varphi(n)}$, and $d_n^{\gamma_n}(\mathcal{K})_X < \epsilon/2$, then we have the following bound for the entropy number*

$$\epsilon_r(\mathcal{K})_X \leq \epsilon, \quad \textit{where} \quad r \geq \lceil n(\varphi(n) + \log_2(6/\epsilon))\rceil.$$

In particular:

(i) *Let $d_n^{\gamma_n}(\mathcal{K})_X = 0$ for some $n \in \mathbb{N}$. Then for any $k \in \mathbb{N}$ such that $k > \varphi(n)$, we have*

$$\epsilon_{nk}(\mathcal{K})_X \leq 6 \cdot 2^{\varphi(n)-k} = 6 \cdot \gamma_n 2^{-k}. \tag{5.1}$$

(ii) *Let $\gamma_n = 2^{cn^p[\log_2 n]^q}$ for some $p \geq 0$ and $q \in \mathbb{R}$.*

- *If for some $\alpha > 0$, $\beta \in \mathbb{R}$, we have $0 < d_n^{\gamma_n}(\mathcal{K})_X \lesssim [\log_2 n]^\beta n^{-\alpha}$, $n \in \mathbb{N}$, then*
  - *when $p > 0$ and $q \in \mathbb{R}$, or $p = 0$ and $q \geq 1$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim n^{-\alpha/(1+p)}[\log_2 n]^{\beta + \frac{\alpha q}{1+p}}, \quad n \in \mathbb{N};$$

  - *when $p = 0$ and $q < 1$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim n^{-\alpha}[\log_2 n]^{\alpha+\beta}, \quad n \in \mathbb{N};$$

- *If for some $c > 0$, we have $0 < d_n^{\gamma_n}(\mathcal{K})_X \lesssim 2^{-cn}$, $n \in \mathbb{N}$, then*
  - *when $0 \leq p < 1$ and $q \in \mathbb{R}$, or $p = 1$ and $q \leq 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim 2^{-c\sqrt{n}}, \quad n \in \mathbb{N};$$

  - *when $p > 1$ and $q \in \mathbb{R}$ or $p = 1$, $q > 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim 2^{-c_1 n^{1/(p+1)}[\log_2 n]^{-q/(p+1)}}, \quad n \in \mathbb{N}.$$

Proof: We provide the proof in the Appendix. □

Remark 16 *It follows from* (4.4) *that in Lemma* 15 *we can take $\epsilon = 6E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X$ or $\epsilon = 6E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X$ if both quantities are positive and obtain that*

$$\epsilon_r(\mathcal{K})_X \leq 6E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X, \quad r \geq \lceil cn^2(1 + \log_2 w(n)) - \log_2(E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X))\rceil,$$

or $\epsilon_r(\mathcal{K})_X \leq 6E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X$ for values of r that satisfy the inequality

$$r \geq \lceil cW(\log_2 W + \log_2 w(W)) - \log_2(E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X))\rceil.$$

The case $E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X = 0$ or $E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X = 0$ is the same as when the Lipschitz width $d_n^{\gamma_n}(\mathcal{K})_X = 0$, provided $C([0,1]^d)$ is continuously embedded in X.

5.2 Upper bounds for entropy numbers via DNN approximation rates.

Let us now consider a compact set $\mathcal{K} \subset X$, where $C([0,1]^d)$ is continuously embedded in X. The next corollary is a direct consequence of Lemma 15.

Corollary 17 *Let $\Sigma_{n,\sigma}(w(n))$ be the set of outputs of a DNN with depth n, fixed width W, ReLU or bounded L-Lipschitz activation function σ and a bound w on its parameters, where $LWw(n) \geq 2$. The following holds:*
(i) *Let $E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \lesssim [\log_2 n]^\beta n^{-\alpha}$, $n \in \mathbb{N}$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$.*

- *When $w(n) = Cn^\delta$, $C > 0$, $\delta > 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim n^{-\frac{\alpha}{2}}[\log_2 n]^{\beta + \frac{\alpha}{2}}, \quad n \in \mathbb{N}. \tag{5.2}$$

- *When $w(n) = C2^{cn^\nu}$, $C, c > 0$, $\nu \geq 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim n^{-\frac{\alpha}{\nu+2}}[\log_2 n]^{\beta}, \quad n \in \mathbb{N}. \tag{5.3}$$

(ii) *Let $E(\mathcal{K}, \Sigma_{n,\sigma}(w(n)))_X \lesssim 2^{-cn}$, $n \in \mathbb{N}$ for some $c > 0$.*

- *When $w(n) = Cn^\delta$, with $C, \delta > 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim 2^{-c_1\sqrt{n}/\sqrt{\log_2 n}}, \quad n \in \mathbb{N}. \tag{5.4}$$

- *When $w(n) = C2^{cn^\nu}$, $C, c > 0$, $\nu \geq 0$, we have*

$$\epsilon_n(\mathcal{K})_X \lesssim 2^{-c_1 n^{1/(\nu+2)}}, \quad n \in \mathbb{N}. \tag{5.5}$$

Proof: The proof follows directly from Lemma 15 and inequality (4.4). Note that in Lemma 15 we require that $d_n^{\gamma_n}(\mathcal{K})_X > 0$. However, it could happen that for some n we have $d_n^{\gamma_n}(\mathcal{K})_X = 0$. Then, we proceed as follows:

- When $w(n) = Cn^\delta$, we have that, see Table 1, $\varphi(n) = c'n\log_2 n$, and therefore we can use (5.1) for $k = 2c'n\log_2 n$ to derive that $\epsilon_{2c'n^2\log_2 n}(\mathcal{K})_X \leq 6 \cdot 2^{-c'n\log_2 n}$. For any m large enough, we can find $n = n(m)$, such that

$$2c'n^2\log_2 n \leq m < 2c'(n+1)^2\log_2(n+1) < c_1 n^2 \log_2 n,$$

and therefore

$$\epsilon_m(\mathcal{K})_X \leq \epsilon_{2c'n^2\log_2 n}(\mathcal{K})_X \leq 6 \cdot 2^{-c'n\log_2 n}.$$

Note that for these $n = n(m)$

$$-c'n\log_2 n \leq -\frac{c'}{c_1}\frac{m}{n}, \quad n \lesssim \frac{\sqrt{m}}{\sqrt{\log_2 n}}, \quad \log_2 n \asymp \log_2 m \Rightarrow$$

$$-n\log_2 n \lesssim -\sqrt{m}\sqrt{\log_2 m},$$

and thus

$$\epsilon_m(\mathcal{K})_X \leq 6 \cdot 2^{-\tilde{c}\sqrt{m}\sqrt{\log_2 m}},$$

which agrees with (5.2) and (5.4).

- When $w(n) = C2^{cn^\nu}$, we have that, see Table 1, $\varphi(n) = c'n^{\nu+1}$, and therefore we can use (5.1) for $k = 2c'n^{\nu+1}$, to obtain that

$$\epsilon_{2c'n^{\nu+2}}(\mathcal{K})_X \leq 6 \cdot 2^{-c'n^{\nu+1}} \quad \Rightarrow \quad \epsilon_n(\mathcal{K})_X \leq C \cdot 2^{-c_1\sqrt{n}}.$$

For any m large enough, we can find $n = n(m)$, such that

$$2c'n^{\nu+2} \leq m < 2c'(n+1)^{\nu+2} < c_1 n^{\nu+2} \quad \Rightarrow \quad m^{1/(\nu+2)} \asymp n,$$

and therefore

$$\epsilon_m(\mathcal{K})_X \leq \epsilon_{2c'n^{\nu+2}}(\mathcal{K})_X \leq 6 \cdot 2^{-c'n^{\nu+1}} \lesssim 2^{-\tilde{c}m^{\frac{\nu+1}{\nu+2}}},$$

which agrees with (5.3) and (5.5).

□

5.3 Upper bounds for entropy numbers via SNN approximation rates

In this section, we study classes $\mathcal{K}$ for which $E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X \leq \xi_W$ for general sequences (ξ_W) with non-negative terms and X is a Banach space such that $C([0,1]^d) \subset X$ is continuously embedded in X. Since $\Xi_{W,\sigma}(w(W)) \subset \Xi_{W+1,\sigma}(w(W))$, we require (ξ_W) to be non-increasing sequence with $\lim\limits_{W\to\infty} \xi_W = 0$. We consider the sequences $\xi_W = C[\log_2 W]^\beta W^{-\alpha}$ and $\xi_W = C2^{-cW}$, and SNNs with ReLU or bounded Lipschitz activation function and bounds on the NN parameters $w(W) = CW^\delta$, $\delta \geq 0$ and $w(W) = C2^{cW^\nu}$, $\nu > 0$. Clearly when $\delta = 0$ we have $w(W) = C$. It follows from Table 2 that we have

$$\begin{aligned} d_W^{\gamma_W}(\mathcal{K})_X &\leq \xi_W, \quad \gamma_W = 2^{c'\log_2 W}, \quad \delta \geq 0, \\ d_W^{\gamma_W}(\mathcal{K})_X &\leq \xi_W, \quad \gamma_W = 2^{c'W^\nu}, \end{aligned}$$

and we can apply Lemma 15 with $p = 0$, $q = 1$ (when $w(W) = CW^\delta$), and $p = \nu$, $q = 0$ (when $w(W) = C2^{cW^\nu}$). More precisely, we have the following corollary.

Corollary 18 *Let $\mathcal{K} \subset X$ be a compact subset of a Banach space X, where X is such that $C([0,1]^d) \subset X$ is continuously embedded in X. Let $\Xi_{W,\sigma}(w(W))$ be the set of outputs of a SNN with ReLU or bounded Lipschitz activation function σ and a bound $w(W) \geq 1$ on its weights. Then the following holds:*
(i) *Let $E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X \lesssim [\log_2 W]^\beta W^{-\alpha}$, $W \in \mathbb{N}$, for some $\alpha > 0$ and $\beta \in \mathbb{R}$.*

- *When $w(W) = CW^\delta$, $C > 0$, $\delta \geq 0$, we have*

$$\epsilon_W(\mathcal{K})_X \lesssim W^{-\alpha}[\log_2 W]^{\beta+\alpha}, \quad W \in \mathbb{N}. \tag{5.6}$$

- *When $w(W) = C2^{cW^\nu}$, $C, c > 0$, $\nu > 0$, we have*

$$\epsilon_W(\mathcal{K})_X \lesssim W^{-\frac{\alpha}{\nu+1}}[\log_2 W]^\beta, \quad W \in \mathbb{N}. \tag{5.7}$$

(ii) *Let $E(\mathcal{K}, \Xi_{W,\sigma}(w(W)))_X \lesssim 2^{-cW}$, $W \in \mathbb{N}$, for some $c > 0$.*

- *When* $w(W) = CW^{\delta}$, *with* $C > 0$, $\delta \geq 0$, *we have*

$$\epsilon_W(\mathcal{K})_X \lesssim 2^{-c_1\sqrt{W}}, \quad W \in \mathbb{N}. \tag{5.8}$$

- *When* $w(W) = C2^{cW^{\nu}}$, $C, c, \nu > 0$, *we have*

$$\epsilon_W(\mathcal{K})_X \lesssim 2^{-c_1 W^{1/(\nu+1)}}, \quad W \in \mathbb{N}. \tag{5.9}$$

Proof: The statement follows from Lemma 15 and (4.5). Lemma 15 requires $d_W^{\gamma_W}(\mathcal{K})_X > 0$. If for some W we have $d_W^{\gamma_W}(\mathcal{K})_X = 0$, we proceed as follows:

- When $w(W) = CW^{\delta}$, we have that, see Table 2, $\varphi(W) = c' \log_2 W$, and therefore we can use (5.1) for $k = W + c' \log_2 W$ to derive that

  $$\epsilon_{W(W+c'\log_2 W)}(\mathcal{K})_X \leq 6 \cdot 2^{-W}.$$

  For any m large enough, we can find $W = W(m)$, such that

  $$W^2 < W(W + c' \log_2 W) \leq m < (W+1)(W+1+c'\log_2(W+1)) < c_1 W^2,$$

  and therefore

  $$\epsilon_m(\mathcal{K})_X \leq \epsilon_{W(W+c'\log_2 W)}(\mathcal{K})_X \leq 6 \cdot 2^{-W} < 6 \cdot 2^{-\tilde{c}\sqrt{m}}, \tag{5.10}$$

  which agrees with (5.6) and (5.8).

- When $w(W) = C2^{cW^{\nu}}$, we have $\varphi(W) = c'W^{\nu}$ and we can use (5.1) for values $k = c'W^{\nu} + W^{\nu+1} = W^{\nu}(c' + W)$, to obtain that

  $$\epsilon_{c'W^{\nu}(c'+W)}(\mathcal{K})_X \leq 6 \cdot 2^{-W^{\nu+1}}.$$

  For any m large enough, we can find $W = W(m)$, such that

  $$W^{\nu+1} < W^{\nu}(c' + W) \leq m < (W+1)^{\nu}(c' + W + 1) \leq c_1 W^{\nu+1},$$

  and therefore

  $$\epsilon_m(\mathcal{K})_X \leq \epsilon_{c'W^{\nu}(c'+W)}(\mathcal{K})_X \leq 6 \cdot 2^{-'W^{\nu+1}} \lesssim 2^{-\tilde{c}m},$$

  which agrees with (5.7) and (5.9).

The proof is completed. □

Remark 19 *Similar statement as Corollary* 18 *was proven in Theorem* 10 *from* Siegel and Xu (2022) *in the case* $\sigma_W = CW^{-\alpha}$ *for classes*

$$\mathcal{K} = \overline{\{f = \sum_{j=1}^{n} c_j h_j : \, h_j \in \mathbb{D}, \sum_{j=1}^{n} |a_j| \leq 1\}},$$

which are the closure of the convex, symmetric hull of dictionaries $\mathbb{D}$ *satisfying specific properties and for SNNs with certain activation functions.*

5.4 Approximation classes for DNNs

Recall that if $f \in X$ and A is a subset of the Banach space X, the distance between f and A is defined as $\mathrm{dist}(f,A)_X := \inf_{g\in A}\|f-g\|_X$. Clearly, we have that

$$E(f,\Sigma_{n,\sigma}(w(n)))_X = \mathrm{dist}(f,\Phi_\sigma(B^{\tilde{n}}_{\ell_\infty}(w(n)))_X,$$

where the sets $\Phi_\sigma(B^{\tilde{n}}_{\ell_\infty}(w(n))) \subset X$ are compact with respect to the uniform norm $C(\Omega)$, see Theorem 3.

Let $\xi := (\xi_n)_{n=1}^\infty$ be a sequence of non-negative numbers such that $\inf_n \xi_n = 0$ (in particular, we can have $\lim_{n\to\infty}\xi_n = 0$). We denote by $\mathcal{N}_{\xi,\sigma}(w)$ the set of functions that are approximated by functions from $\Sigma_{n,\sigma}(w(n))$ with accuracy ξ_n for every $n \geq 0$. More precisely,

$$\mathcal{N}_{\xi,\sigma}(w) := \{f \in X: \ E(f,\Sigma_{n,\sigma}(w(n)))_X \leq \xi_n,\ \forall n \geq 0\},$$

which can be written equivalently as

$$\mathcal{N}_{\xi,\sigma}(w) = \bigcap_{n=0}^{\infty} V_n(\xi), \quad V_n(\xi) := \{f \in X: \mathrm{dist}(f,\Phi_\sigma(B^{\tilde{n}}_{\ell_\infty}(w(n)))_X \leq \xi_n\}.$$

Then, if X is such that $C(\Omega)$ is continuously embedded in X, we can apply Lemma 30 (see the Appendix) to obtain that $\mathcal{N}_{\xi,\sigma}(w)$ is a (possibly empty) compact subset of X. In what follows, we show that there are choices of sequences ξ and DNNs with bounds w on their parameters for which the compact sets $\mathcal{N}_{\xi,\sigma}(w) \neq \emptyset$.

Remark 20 *According to Remark* 4*, the set* $\Sigma^s_{n,\sigma}(w(n))$ *of outputs of a DNN where at each layer one uses* $\bar{\sigma} = (\sigma_0,\sigma,\ldots,\sigma,\sigma_0)$ *with* $\sigma_0(t) = t$*, and* $\sigma =$*ReLU, see* (2.1)*, satisfy Theorem* 3. Therefore, all theory developed so far holds for $\Sigma^s_{n,\sigma}(w(n))$.

Let us now consider the case when $\Omega = [0,1]$ *and* $\sigma =$*ReLU. If we denote by* H *the hat function* $H(t) = 2(t-0)_+ - 4(t-\frac{1}{2})_+$ *and by* $H^{\circ k}$ *this function composed with itself* k *times, then, for properly selected* $w(n)$*, we have the inclusion* $\{\psi_n := \sum_{k=1}^n c_k H^{\circ k}\} \subset \Sigma^s_{n,\sigma}(w(n))$*, see Figure* 5.1*.*

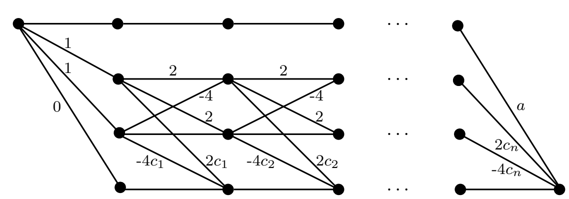

Figure 5.1: Computational graph for ψ_n

Recall that the set $\mathcal{T}$ *defined as*

$$\mathcal{T} := \left\{ f:\ f = \sum_{k=1}^{\infty} c_k H^{\circ k},\ \sum_{k=1}^{\infty}|c_k| < \infty \right\},$$

is called the Takagi class, see Allaart and Kawamurau (2011). *Clearly*

$$\mathcal{R} := \{f_\lambda = \sum_{k=1}^{\infty} \lambda^{-k} H^{\circ k}, \quad |\lambda| > 1\} \subset \mathcal{T},$$

and it is a well known fact that the Takagi function $T = f_2$ and that $t(1-t) = f_4(t)$, see Allaart and Kawamurau (2011). *The elements of $\mathcal{R}$ can be approximated with exponential accuracy by outputs of $\Sigma^s_{n,\sigma}(w(n))$ with $w(n) = C$. Therefore, in this case the set $\mathcal{N}_{\tilde{\xi},\sigma}(w)$, with $\tilde{\xi} = (C2^{-cn})_{n=1}^{\infty}$ is non-empty, and for every $\eta = (\eta_n)_{n=1}^{\infty}$ with the property that* $\inf_n \eta_n = 0$ *and $C2^{-cn} \le \eta_n$ for all $n \ge 0$, we have $\emptyset \ne \mathcal{N}_{\tilde{\xi},\sigma}(w) \subset \mathcal{N}_{\eta,\sigma}(w)$.*

Now, let us return to the relation between the error $E(\mathcal{N}_{\xi,\sigma}(w), \Sigma_{n,\sigma}(w(n)))_X$ and the Lipschitz width. It follows from the definition of $\mathcal{N}_{\xi,\sigma}(w)$ that for every $n \ge 0$, $E(\mathcal{N}_{\xi,\sigma}(w), \Sigma_{n,\sigma}(w(n)))_X \le \xi_n$, and therefore (4.4) gives that

$$d_n^{\gamma_n}(\mathcal{N}_{\xi,\sigma}(w))_X \le \xi_n, \quad \text{with} \quad \gamma_n = 2^{cn(1+\log_2 w(n))}. \tag{5.11}$$

One can now apply Lemma 15 and derive estimates for the entropy numbers of the compact set $\mathcal{N}_{\xi,\sigma}(w)$. Such estimates can be viewed as inverse theorems for DNN approximation.

To simplify the presentation, let us consider the special sequences $\xi_n = C[\log_2 n]^\beta n^{-\alpha}$ and $\xi_n = C2^{-cn}$, and DNNs with depth n, fixed width W, ReLU or bounded Lipschitz activation function and bounds on the NN parameters $w(n) = Cn^\delta$, $\delta \ge 0$ and $w(n) = C2^{cn^\nu}$, $\nu \ge 0$, see Table 1. Note that the case $\nu = 0$ covers the case $w(n) = C$. We apply Corollary 17 to the class $\mathcal{K} = \mathcal{N}_{\xi,\sigma}(w)$ and obtain the following result.

Corollary 21 *The entropy numbers of the compact set $\mathcal{N}_{\xi,\sigma}(w)$ that consists of all functions approximated in the norm of X with accuracy ξ_n by the outputs of a DNN with depth n, fixed width W, ReLU or bounded L-Lipschitz activation function and bounds on the NN parameters w such that $LWw(n) \ge 2$ satisfy the following inequalities for the listed special choices of w and sequences (ξ):*

- *If $C, c, \alpha, \delta > 0$, $\nu \ge 0$, and $\beta \in \mathbb{R}$, we have*

$$\epsilon_n(\mathcal{N}_{\xi,\sigma}(Cn^\delta))_X \lesssim n^{-\frac{\alpha}{2}}[\log_2 n]^{\beta+\frac{\alpha}{2}}, \quad n \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_n) = (C[\log_2 n]^\beta n^{-\alpha}),$$

$$\epsilon_n(\mathcal{N}_{\xi,\sigma}(C2^{cn^\nu}))_X \lesssim n^{-\frac{\alpha}{\nu+2}}[\log_2 n]^{\beta}, \quad n \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_n) = (C[\log_2 n]^\beta n^{-\alpha}).$$

- *If $C, c, \alpha, \delta > 0$, $\nu \ge 0$, and $\beta \in \mathbb{R}$, we have*

$$\epsilon_n(\mathcal{N}_{\xi,\sigma}(Cn^\delta))_X \lesssim 2^{-c_1\sqrt{n}/\sqrt{\log_2 n}}, \quad n \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_n) = (C2^{-cn}),$$

$$\epsilon_n(\mathcal{N}_{\xi,\sigma}(C2^{cn^\nu}))_X \lesssim 2^{-c_1 n^{1/(\nu+2)}}, \quad n \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_n) = (C2^{-cn}).$$

Remark 22 *The above estimates hold also for the sets* $\lambda\mathcal{N}_{\xi,\sigma}(w) = \{\lambda f : f \in \mathcal{N}_{\xi,\sigma}(w)\}$ *with* $\lambda > 1$*, where the constants involved depend on* λ*. Indeed, the fact that* $f \in V_n(\xi)$ *implies the inequality*

$$\mathrm{dist}(\lambda f, \Phi_\sigma(B_{\ell_\infty}^{\tilde{n}}(\lambda w(n))))_X \leq \lambda\xi_n,$$

since $\lambda\Phi_\sigma(B_{\ell_\infty}^{\tilde{n}}(w(n))) \subset \Phi_\sigma(B_{\ell_\infty}^{\tilde{n}}(\lambda w(n)))$*. Then, according to* (4.4) *and the monotonicity of Lipschitz widths, we have*

$$d_n^{\gamma_n'}(\lambda\mathcal{N}_{\xi,\sigma}(w))_X \leq \lambda\xi_n, \quad \textit{with} \quad \gamma_n' := 2^{c(\lambda)n(1+\log_2 w(n))} > 2^{cn(1+\log_2(\lambda w(n)))},$$

and we can apply Lemma 15 *or Corollary* 17*.*

5.5 Approximation classes for SNNs

We consider Banach spaces X such that $C([0,1]^d) \subset X$ is continuously embedded in X. We denote by $\mathcal{A}_{\xi,\sigma}(w) \subset X$ the approximation class

$$\mathcal{A}_{\xi,\sigma}(w) := \{f \in X : \ E(f, \Xi_{W,\sigma}(w(W)))_X \leq \xi_W, \, \forall W \geq 0\},$$

or equivalently written as

$$\mathcal{A}_{\xi,\sigma}(w) = \bigcap_{W=0}^{\infty} \mathcal{V}_W(\xi), \quad \mathcal{V}_W(\xi) := \{f \in X : \, \mathrm{dist}(f, \Psi_\sigma(B_{\ell_\infty}^{\tilde{W}}(w(W)))_X \leq \xi_W\}.$$

As in the case of DNNs, $\mathcal{A}_{\xi,\sigma}(w)$ is a compact subset of X, see Lemma 30. Its Lipschitz widths, see (4.5), satisfy the inequalities

$$d_W^{\gamma_W}(\mathcal{A}_{\xi,\sigma}(w))_X \leq \xi_W, \quad \text{with} \quad \gamma_W = 2^{c(\log_2 W + \log_2 w(W))}, \quad W \in \mathbb{N}. \tag{5.12}$$

We next apply Corollary 18 in the case $\mathcal{K} = \mathcal{A}_{\xi,\sigma}(w)$ and derive the following statement.

Corollary 23 *The entropy numbers of the approximation class* $\mathcal{A}_{\xi,\sigma}(w)$ *that consists of all functions approximated in the norm of* X *with accuracy* ξ_W*,* $W \in \mathbb{N}$*, by the outputs of a SNN with width* W*, ReLU or bounded Lipschitz activation function and bounds on the NN parameters* $w(W) = CW^\delta$*,* $\delta \geq 0$ *or* $w(W) = C2^{cW^\nu}$*,* $C, c > 0$*,* $\nu > 0$*, satisfy the following inequalities:*

- *If* $C, C', \alpha, \nu > 0$*,* $\delta \geq 0$*,* $\beta \in \mathbb{R}$*, we have*

$$\epsilon_W(\mathcal{A}_{\xi,\sigma}(CW^\delta))_X \lesssim W^{-\alpha}[\log_2 W]^{\beta+\alpha}, \, W \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_W) = (C'[\log_2 W]^\beta W^{-\alpha}), \tag{5.13}$$

$$\epsilon_W(\mathcal{A}_{\xi,\sigma}(C2^{cW^\nu}))_X \lesssim W^{-\frac{\alpha}{\nu+1}}[\log_2 W]^\beta, \, W \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_W) = (C'[\log_2 W]^\beta W^{-\alpha}). \tag{5.14}$$

- *If* $C, C', c, \nu > 0$*,* $\delta \geq 0$*, we have*

$$\epsilon_W(\mathcal{A}_{\xi,\sigma}(CW^\delta))_X \lesssim 2^{-c_1\sqrt{W}}, \, W \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_W) = (C'2^{-cW}),$$

$$\epsilon_W(\mathcal{A}_{\xi,\sigma}(C2^{cW^\nu}))_X \lesssim 2^{-c_1W^{1/(\nu+1)}}, \, W \in \mathbb{N}, \quad \textit{where} \quad \xi = (\xi_W) = (C'2^{-cW}).$$

6. Further study of Lipschitz widths with large Lipschitz constants

So far, we have used Lipschitz widths as a tool to obtain estimates from below for the error of approximation of a compact set $\mathcal{K}$ via deep and shallow NNs. However, Lipschitz widths are a subject of interest on their own. We have studied in Petrova and Wojtaszczyk (2023) Lipschitz widths with Lipschitz constants $\gamma = const$ and $\gamma = \gamma_n = \lambda^n$, $\lambda > 1$. In this section, we will complete this study by including Lipschitz widths with constants $\gamma_n = 2^{\varphi(n)}$. We start with the following theorem, which is an application of Theorem 3.3 from Petrova and Wojtaszczyk (2023).

Theorem 24 *Let $\mathcal{K} \subset X$ be a compact subset of a Banach space X, $n \in \mathbb{N}$, and $d_n^{\gamma_n}(\mathcal{K})_X$ be the Lipschitz width for $\mathcal{K}$ with Lipschitz constant $\gamma_n = 2^{\varphi(n)}$, where $\varphi(n) \to \infty$ as $n \to \infty$. Then, we have*

$$d_n^{2^{\varphi(n)}}(\mathcal{K})_X \leq \epsilon_{n\lceil \frac{\varphi(n)}{2} \rceil}(\mathcal{K})_X, \quad \textit{where} \quad n \geq n_0(\mathcal{K}). \tag{6.1}$$

In particular,

$$\text{(i)} \quad \epsilon_n(\mathcal{K})_X \lesssim [\log_2 n]^\beta n^{-\alpha},\ n \in \mathbb{N} \Rightarrow d_n^{\gamma_n}(\mathcal{K})_X \lesssim [\log_2(n\varphi(n))]^\beta [n\varphi(n)]^{-\alpha},\ n \in \mathbb{N}, \tag{6.2}$$

where $\alpha > 0$, $\beta \in \mathbb{R}$;

$$\text{(ii)} \quad \epsilon_n(\mathcal{K})_X \lesssim [\log_2 n]^{-\alpha},\ n \in \mathbb{N},\ \alpha > 0 \Rightarrow d_n^{\gamma_n}(\mathcal{K})_X \lesssim [\log_2(n\varphi(n))]^{-\alpha},\ n \in \mathbb{N}; \tag{6.3}$$

$$\text{(iii)} \quad \epsilon_n(\mathcal{K})_X \lesssim 2^{-cn^\alpha},\ n \in \mathbb{N},\ 0 < \alpha < 1 \Rightarrow d_n^{\gamma_n}(\mathcal{K})_X \lesssim 2^{-c(n\varphi(n))^\alpha},\ n \in \mathbb{N}. \tag{6.4}$$

Proof: Indeed, it follows from (Petrova and Wojtaszczyk, 2023, Theorem 3.3) that for any compact subset $\mathcal{K} \subset X$ of a Banach space X and any $n \in \mathbb{N}$ we have that

$$d_n^{2^k \text{rad}(\mathcal{K})}(\mathcal{K})_X \leq \epsilon_{kn}(\mathcal{K})_X, \quad k = 1, 2, \ldots. \tag{6.5}$$

We choose $k = k(n)$ to be such that

$$2^k \text{rad}(\mathcal{K}) \leq 2^{\varphi(n)} < 2^{k+1} \text{rad}(\mathcal{K}) < 2^{k+\ell},$$

where $\text{rad}(\mathcal{K}) < 2^{\ell-1}$, then $k > \varphi(n) - \ell > \lceil \frac{\varphi(n)}{2} \rceil$ for $n \geq n_0$, $n_0 = n_0(\mathcal{K})$ big enough. Then

$$d_n^{\gamma_n}(\mathcal{K})_X \leq d_n^{2^k \text{rad}(\mathcal{K})}(\mathcal{K})_X \leq \epsilon_{kn}(\mathcal{K})_X \leq \epsilon_{n\lceil \frac{\varphi(n)}{2} \rceil}(\mathcal{K})_X, \quad \text{for} \quad n \geq n_0(\mathcal{K}),$$

and therefore (6.1) holds. Estimates (6.2), (6.3) and (6.4) follow from (6.1). Note that n_0 depends only on $\text{rad}(\mathcal{K})$ and on how fast $\varphi(n)$ grows to infinity. □

Theorem 24 and Theorem 9 can be combined in the next corollary which can be viewed as a generalization of Corollary 6.4 from Petrova and Wojtaszczyk (2023). The latter covers the particular case $\varphi(n) = cn$.

Corollary 25 *Let $\mathcal{K}\subset X$ be a compact set of a Banach space X and the function $\varphi:\mathbb{N}\to\mathbb{R}$ be such that $\varphi(n)\geq c\log_2 n$ for some fixed constant $c>0$. Let $d_n^{\gamma_n}(\mathcal{K})_X$ be the Lipschitz width of $\mathcal{K}$ with Lipschitz constant $\gamma_n=2^{\varphi(n)}$. Then the following holds:*

$$\epsilon_n(\mathcal{K})_X \asymp [\log_2 n]^{\beta} n^{-\alpha},\ n\in\mathbb{N} \Rightarrow d_n^{\gamma_n}(\mathcal{K})_X \asymp [\log_2(n\varphi(n))]^{\beta}[n\varphi(n)]^{-\alpha},\ n\in\mathbb{N},$$

$$\epsilon_n(\mathcal{K})_X \asymp [\log_2 n]^{-\alpha}\ n\in\mathbb{N} \Rightarrow d_n^{\gamma_n}(\mathcal{K})_X \asymp [\log_2(n\varphi(n))]^{-\alpha},\ n\in\mathbb{N},$$

where $\alpha>0$, $\beta\in\mathbb{R}^n$.

It follows from Theorem 3.1 from Petrova and Wojtaszczyk (2023) that when γ is independent on n, i.e $\gamma=const$ (this case is excluded in Corollary 25 because of the condition $\varphi(n)\geq c\log_2 n$), we do not have matching lower and upper bounds for $d_n^{\gamma}(\mathcal{K})_X$ in the case when $\epsilon_n(\mathcal{K})_X\asymp[\log_2 n]^{\beta}n^{-\alpha}$, namely we have

$$\begin{aligned}
\epsilon_n(\mathcal{K})_X &\lesssim [\log_2 n]^{\beta} n^{-\alpha},\ n\in\mathbb{N},\ \alpha>0,\ \beta\in\mathbb{R} \quad\Rightarrow\quad d_n^{\gamma}(\mathcal{K})_X \lesssim [\log_2 n]^{\beta} n^{-\alpha},,\ n\in\mathbb{N},\\
\epsilon_n(\mathcal{K})_X &\gtrsim [\log_2 n]^{\beta} n^{-\alpha},\ n\in\mathbb{N},\ \alpha>0,\ \beta\in\mathbb{R} \quad\Rightarrow\quad d_n^{\gamma}(\mathcal{K})_X \gtrsim [\log_2 n]^{\beta-\alpha} n^{-\alpha},\ n\in\mathbb{N}.
\end{aligned}$$

It is still an open question whether the upper bound for $d_n^{\gamma}(\mathcal{K})_X$ in this case can be improved to $d_n^{\gamma}(\mathcal{K})_X\lesssim[\log_2 n]^{\beta-\alpha}n^{-\alpha}$. The following example, constructed in Petrova and Wojtaszczyk (2023), is in support of this conjecture. The compact subset $\mathcal{K}(\sigma)\subset\mathbf{c_0}$ of the Banach space $\mathbf{c_0}$ of all sequences that converge to 0, equipped with the ℓ_∞ norm, defined as

$$\mathcal{K}(\sigma):=\{\sigma_j e_j\}_{j=1}^{\infty}\cup\{0\},\quad \sigma_j:=(\log_2(j+1))^{-1},$$

where $(e_j)_{j=1}^{\infty}$ are the standard basis in $\mathbf{c_0}$ has entropy numbers $\epsilon_n(\mathcal{K}(\sigma))_{\mathbf{c_0}}\asymp n^{-1}$ and Lipschitz width $d_n^{\gamma}(\mathcal{K}(\sigma))_{\mathbf{c_0}}\asymp n^{-1}[\log_2(n+1)]^{-1}$, $\gamma=const$.

Corollary 26 *It follows from Lemma 6 and Corollary 25 that if a compact subset $\mathcal{K}\subset X$ of a Banach space X has entropy numbers $\epsilon_n(\mathcal{K})_X\asymp[\log_2 n]^{\beta}n^{-\alpha}$ or $\epsilon_n(\mathcal{K})_X\asymp[\log_2 n]^{-\alpha}$, then for any $c_1,\tilde{c}>0$ and $a,a_1>0$, we have*

$$d_n^{2^{\tilde{c}n}}(\mathcal{K})_X \asymp d^{2^{c_1 n}}(\mathcal{K},\ell_\infty^n)_X,\quad d_n^{2^{an\log_2 n}}(\mathcal{K})_X \asymp d^{2^{a_1 n\log_2 n}}(\mathcal{K},\ell_\infty^n)_X,$$

where the constants in $\asymp$ depend on $\tilde{c}$ and c_1, or a and a_1, respectively.

Proof: We consider the case when $\gamma_n=2^{\tilde{c}n}$ (the case $\gamma_n=2^{an\log_2 n}$ is similar). We take any $c_1>0$ and fix $c\leq c_1$. Then, for n big enough we have $c_1 n\geq cn+\log_2(2n+1)$ and it follows from the monotonicity of the fixed Lipschitz widths and Lemma 6 that

$$d_n^{2^{c_1 n}}(\mathcal{K})_X \leq d^{2^{c_1 n}}(\mathcal{K},\ell_\infty^n)_X \leq d^{(2n+1)2^{cn}}(\mathcal{K},\ell_\infty^n)_X \leq d_n^{2^{cn}}(\mathcal{K})_X.$$

On the other hand, $d_n^{2^{cn}}(\mathcal{K})_X\asymp d^{2^{c_1 n}}(\mathcal{K})_X\asymp d^{2^{\tilde{c}n}}(\mathcal{K})_X$, see Corollary 25, and therefore

$$d^{2^{c_1 n}}(\mathcal{K},\ell_\infty^n)_X \asymp d_n^{2^{cn}}(\mathcal{K})_X \asymp d_n^{2^{\tilde{c}n}}(\mathcal{K})_X.$$

The proof is completed. □

We can further refine Theorem 24 in some cases and the following lemma holds.

Lemma 27 *Let $\gamma_n = 2^{\varphi(n)}$ with $\varphi : [1,\infty) \to \mathbb{R}$ an increasing non-negative function such that $\varphi(n) \to \infty$ as $n \to \infty$ and $d_n^{\gamma_n}(\mathcal{K})_X \geq c_0 2^{-\varphi(n)}$ for some $c_0 > 0$. Then, for n sufficiently large, we have that*

$$d_{8n}^{\gamma_{8n}}(\mathcal{K})_X \leq \epsilon_{8n\lceil \frac{\varphi(8n)}{2} \rceil}(\mathcal{K})_X \leq 3d_n^{\gamma_n}(\mathcal{K})_X. \tag{6.6}$$

In particular, if $d_n^{\gamma_n}(\mathcal{K})_X \asymp [\log_2(n\varphi(n))]^{\beta}[n\varphi(n)]^{-\alpha}$, and

- *if the function φ is such that there is a constant $c_1 > 1$ for which $\sup_{n\in\mathbb{N}} \dfrac{\varphi(c_1 n)}{\varphi(n)} < \infty$ then*
$$\epsilon_m(\mathcal{K})_X \asymp [\log_2 m]^{\beta} m^{-\alpha}, \quad m \in \mathbb{N}.$$
- *if for every $c > 1$, $\sup_{n\in\mathbb{N}} \dfrac{\varphi(cn)}{\varphi(n)} = \infty$, then the lower and upper bound in (6.6) are asymptotically different in the sense that*
$$\frac{d_n^{\gamma_n}(\mathcal{K})_X}{d_{8n}^{\gamma_{8n}}(\mathcal{K})_X} \gtrsim \begin{cases} \left[\frac{\varphi(8n)}{\varphi(n)}\right]^{\alpha} \left[\log_2 \frac{\varphi(8n)}{\varphi(n)}\right]^{-\beta}, & \beta > 0, \\ \left[\frac{\varphi(8n)}{\varphi(n)}\right]^{\alpha}, & \beta \leq 0. \end{cases}$$

Proof: Following the proof of Lemma 15 and using the fact that $d_n^{\gamma_n}(\mathcal{K})_X \geq c_0 2^{-\varphi(n)}$, we obtain for the choice $\epsilon = 3d_n^{\gamma_n}(\mathcal{K})_X > 0$ that the entropy numbers $\epsilon_r(\mathcal{K})_X \leq 3d_n^{\gamma_n}(\mathcal{K})_X$, with

$$r = \lceil n(\varphi(n) + \log_2(2[d_n^{\gamma_n}(\mathcal{K})_X]^{-1}))\rceil \leq \lceil n(2\varphi(n) + \tilde{c})\rceil \leq \lceil 4n\varphi(n)\rceil, \tag{6.7}$$

provided n is sufficiently large, and thus

$$\epsilon_{\lceil 4n\varphi(n)\rceil}(\mathcal{K})_X \leq 3d_n^{\gamma_n}(\mathcal{K})_X.$$

On the other hand, Theorem 24, gives that

$$d_{8n}^{\gamma_{8n}}(\mathcal{K})_X \leq \epsilon_{8n\lceil \frac{\varphi(8n)}{2} \rceil}(\mathcal{K})_X.$$

We derive (6.6) from the monotonicity of the entropy numbers, the latter two inequalities and the fact that φ, as an increasing non-negative function, satisfies the condition $8n\lceil \frac{\varphi(8n)}{2}\rceil \geq \lceil 4n\varphi(n)\rceil$.

Now, let $d_n^{\gamma_n}(\mathcal{K})_X \asymp [\log_2(n\varphi(n))]^{\beta}[n\varphi(n)]^{-\alpha}$. If there is a constant $c_1 > 1$ for which the quantity $\sup_{n\in\mathbb{N}} \dfrac{\varphi(c_1 n)}{\varphi(n)} < \infty$, then Lemma 31 (see the Appendix) gives that $\sup_{n\in\mathbb{N}} \dfrac{\varphi(c_0 n)}{\varphi(n)}$ is finite for all $c_0 > 1$ and the conclusion follows from Lemma 32 (see the Appendix) with $c = 8$ and $a_k = \epsilon_k(\mathcal{K})_X$. If for all $c_0 > 1$ we have $\sup_{n\in\mathbb{N}} \dfrac{\varphi(c_0 n)}{\varphi(n)} = \infty$, then we apply again Lemma 32 to complete the proof. $\square$

Remark 28 *All functions φ from Table 1 or Table 2 satisfy the condition* $\sup_{n\in\mathbb{N}} \frac{\varphi(c_1 n)}{\varphi(n)} < \infty$ *for all $c_1 > 1$, and therefore it follows from Lemma 27 that for $\gamma_n = 2^{\varphi(n)}$ with φ being any of these functions,*

$$d_n^{\gamma_n}(\mathcal{K})_X \asymp [\log_2(n\varphi(n))]^\beta [n\varphi(n)]^{-\alpha},\ n \in \mathbb{N},\ \alpha > 0,\ \beta \in \mathbb{R} \Rightarrow \epsilon_n(\mathcal{K})_X \asymp [\log_2 n]^\beta n^{-\alpha},\ n \in \mathbb{N}.$$

Observe that Lemma 27 does not cover the case when $\gamma_n = const$, that is, $\varphi(n) = const$, because the Lipschitz width $d_n^\gamma(\mathcal{K})_X \to 0$ as $n \to \infty$, and therefore would not satisfy the condition $d_n^\gamma(\mathcal{K})_X \geq c_0 2^{-\varphi(n)} = C$. We discuss this case separately in the lemma that follows.

Lemma 29 *Let $\mathcal{K} \subset X$ be a compact set, $\gamma = const$, (ξ_n) be a sequence of positive numbers such that $\xi_n \to 0$ as $n \to \infty$, and $d_n^\gamma(\mathcal{K})_X \geq \xi_n$. Then we have that for some positive constants c, c_1,*

$$d_n^{c_1 \xi_n^{-c}}(\mathcal{K})_X \leq \epsilon_{c\lceil n \log_2(\xi_n^{-1})\rceil}(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X. \tag{6.8}$$

In particular, if $\xi_n = 2^{-n}$, we have $d_n^{c_1 2^{cn}}(\mathcal{K})_X \leq \epsilon_{cn^2}(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X$, and when $\xi_n = n^{-\alpha}$ we have $d_n^{c_1 n^{c\alpha}}(\mathcal{K})_X \leq \epsilon_{c\lceil n \log_2 n\rceil}(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X$.

Proof: Again, it follows from the proof of Lemma 15 that for $\epsilon = 3d_n^\gamma(\mathcal{K})_X \geq 3\xi_n > 0$ the entropy numbers $\epsilon_r(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X$, with

$$r = \lceil n(\log_2 \gamma + \log_2(2[d_n^\gamma(\mathcal{K})_X]^{-1}))\rceil \leq c\lceil n \log_2(\xi_n^{-1})\rceil,$$

if n is large enough, and therefore

$$\epsilon_{c\lceil n \log(\xi_n^{-1})\rceil}(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X. \tag{6.9}$$

We take k_n to be the smallest integer such that $c\lceil n \log(\xi_n^{-1})\rceil \leq nk_n$. From (Petrova and Wojtaszczyk, 2023, Th. 3.3), (6.9), and the monotonicity of entropy numbers and Lipschitz widths we get

$$d_n^{c_1 \xi_n^{-c}}(\mathcal{K})_X \leq d_n^{2^{k_n} \mathrm{rad}(\mathcal{K})}(\mathcal{K})_X \leq \epsilon_{nk_n}(\mathcal{K})_X \leq \epsilon_{c\lceil n \log_2(\xi_n^{-1})\rceil}(\mathcal{K})_X \leq 3d_n^\gamma(\mathcal{K})_X,$$

where we have used in the first inequality the definition of k_n, namely that

$$c\lceil n \log(\xi_n^{-1})\rceil > n(k_n - 1).$$

The proof is completed. □

7. Conclusion

In this paper, we further develop the theory of Lipschitz widths for a compact set $\mathcal{K}$ in a Banach space X to include Lipschitz mappings with large Lipschitz constants. The theory is then applied to NNs to obtain Carl's type inequalities for deep and shallow neural network approximation, where the error is measured in the Banach space norm $\|\cdot\|_X$ and the requirement for X is that $C([0,1]^d) \subset X$ is continuously embedded in X. In fact, this method can be used to obtain Carl's type inequalities for any feed-forward NN with predetermined relation between the width W and depth n of this network.

Our analysis is executed by utilizing the growth of the ℓ_∞^m norm of the parameters of the NN, namely $\|y\|_{\ell_\infty^m} \leq w(m)$ (with $m = \tilde{n}$ or $m = \widetilde{W}$) for a given function w. Note that all results for NN approximation utilize the behavior of the Lipschitz width $d_m^{\gamma_m}(\mathcal{K})_X$, which is defined as the infimum of the fixed Lipschitz width $d^{\gamma_m}(\mathcal{K}, Y_m)$ over all possible norms Y_m in $\mathbb{R}^m$. Therefore, all results will hold no matter what norm we choose to bound the NN parameters, i.e. all statements will hold if instead of $\|y\|_{\ell_\infty^m} \leq w(m)$, we choose $\|y\|_{Y_m} \leq w(m)$, where Y_m is our favorite norm.

Our estimates can be applied to any novel or classical classes of functions as long as we know estimates on their entropy numbers.

Acknowledgments

G.P. was supported by the NSF Grant DMS 2134077, Tripods Grant CCF-1934904, and ONR Contract N00014-20-1-278.
We would like to thank the anonymous referees for their comments and suggestions which improved the quality of our paper.

References

F. Albiac and N. Kalton. *Topics in Banach space theory.* Graduate Texts in Mathematics, Springer International Publishing, 2016.

P. Allaart and K. Kawamurau. The Takagi function: A Survey. *Real Analysis Exchange*, 37(1):1–54, 2011.

B. Carl. Entropy numbers, s-numbers, and eigenvalue problems. *J. Funct. Anal.*, 41:290–306, 1981.

B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators.* Cambridge University Press, 1990.

F. Cobos and T. Kuhn. Approximation and entropy numbers in Besov spaces of generalized smoothness. *J. Approx. Theory*, 160:56–70, 2009.

R. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

R. DeVore and G. Lorentz. *Constructive Approximation.* Springer Verlag, 1993.

F. Cobos O. Dominguez and T. Kuhn. Approximation and entropy numbers of embeddings between approximation spaces. *Constructive Approximation*, 47:453–486, 2018.

D. Edmunds and H. Triebel. *Function spaces, entropy numbers and differnial operators.* Cambridge Tracts in Mathematics 120, 1996.

F. Gao. Entropy estimate for k-monotone functions via small ball probability of integrated Brownian motion. *Elect. Comm. in Probab.*, 13:121–130, 2008.

E-M. Achour A. Foucault S. Gerchinovitz and F. Malgouyres. A general approximation lower bound in L_p norm, with applications to feed-forward neural networks. *arXiv:2206.04360*, 2022.

G. Lorentz M. Golitschek and Y. Makovoz. *Constructive Approximation*. Springer Verlag, Advanced Problems, 1996.

I. Gühring and M. Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2020.

R. DeVore B. Hanin and G. Petrova. Neural network approximation. *Acta Numerica*, 30: 327–444, 2021.

R. DeVore G. Kyriazis D. Leviatan and V. Tikhomirov. Wavelet compression and nonlinear n-widths. *Advances in Computational Mathematics*, 1(2):197–214, 1993.

P. Bartlett N. Harvey C. Liaw and A. Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

V. Maiorov. On best approximation by ridge functions. *J. Approx. Theory*, 99(1):68–94, 1999.

V. Maiorov R. Meir and J. Ratsaby. On the approximation of functional classes equipped with a uniform measure using ridge functions. *J. Approx. Theory*, 99:95–111, 1999.

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

A. Cohen R. DeVore G. Petrova and P. Wojtaszczyk. Optimal stable nonlinear approximation. *Found Comput Math*, 22(3):607–648, 2022.

G. Petrova and P. Wojtaszczyk. Lipschitz widths. *Constructive Approximation*, 7:759–805, 2023.

J. Siegel. Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces. *arXiv:2211.14400*, 2022.

J. Siegel and J. Xu. Sharp bounds on the approximation rates, metric entropy and n widths of shallow neural networks. *Journal of FOCM*, pages to appear, arXiv:2101.12365v9, 2022.

M. Telgarsky. Benefits of depth in neural networks. *In 29th Annual Conferenceon Learning Theory, Proceedings of Machine Learning Research*, 49, 2016.

N. Tomczak-Jaegermann. *Banach-Mazur distances and finite dimensional operator ideals*. Pitman Monographs and Surveys in Pure and Appl. Math, 1989.

F. Voigtlaender and P. Petersen. Approximation in L_p with deep ReLU neural networks. *13th International conference on Sampling Theory and Applications (SampTA)*, 2019.

Y. Yang and A. Barron. Information-theoretic determination of minimaxrates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Z. Shen H. Yang and S. Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathematiques Pures et Appliquees*, 157:101–135, 2022.

D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural networks*, 97:103–114, 2017.

D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. *Proceedings of the 31st Conference On Learning Theory, PMLR*, 75:639–649, 2018.

D. Yarotsky and A. Zhevnerchuk. The phase diagram of approximationrates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Appendix

In this section we provide the proofs of some of the theorems and lemmas we use in the paper.

Proof of Lemma 2: We fix a norm $\|\cdot\|_{Y_n}$ on $\mathbb{R}^n$ and take any $r>0$ and map $\mathcal{L}_n:(B_{Y_n}(r),\|\cdot\|_{Y_n})\to X$ with Lipschitz constant γ/r. We consider the onto r-Lipschitz map $\eta: B_{Y_n}(1)\to B_{Y_n}(r)$, defined as $\eta(y)=ry$. Clearly, $\mathcal{L}_n\circ\eta:(B_{Y_n}(1),\|\cdot\|_{Y_n})\to X$ is γ-Lipschitz. Therefore we have

$$d^\gamma(\mathcal{K},Y_n)_X\le \sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(1)}\|f-\mathcal{L}_n\circ\eta(y)\|_X=\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n(y)\|_X,$$

which gives

$$d^\gamma(\mathcal{K},Y_n)_X\le \inf_{\mathcal{L}_n,r>0}\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n(y)\|_X, \tag{.1}$$

where the infimum is taken over $r>0$ and all γ/r-Lipschitz maps $\mathcal{L}_n:(B_{Y_n}(r),\|\cdot\|_{Y_n})\to X$.

Observe that we can argue in the another direction too. For fixed $\delta>0$, we take a γ-Lipschitz mapping $\mathcal{L}_n':(B_{Y_n}(1),\|\cdot\|_{Y_n})\to X$ such that

$$\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(1)}\|f-\mathcal{L}_n'(y)\|_X\le d^\gamma(\mathcal{K},Y_n)_X+\delta.$$

We fix $r>0$, define the $1/r$-Lipschitz mapping η from $B_{Y_n}(r)$ onto $B_{Y_n}(1)$ as $\eta(y)=y/r$ and consider the γ/r-Lipschitz mapping $\mathcal{L}_n'\circ\eta:(B_{Y_n}(r),\|\cdot\|_{Y_n})\to X$. Since $\mathcal{L}_n'\circ\eta$ is a γ/r-Lipschitz mapping and we have

$$\inf_{\mathcal{L}_n,r>0}\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n(y)\|_X\le\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n'\circ\eta(y)\|=\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(1)}\|f-\mathcal{L}_n'(y)\|,$$

which gives that

$$\inf_{\mathcal{L}_n,r>0}\sup_{f\in\mathcal{K}}\inf_{y\in B_{Y_n}(r)}\|f-\mathcal{L}_n(y)\|_X\le d^\gamma(\mathcal{K},Y_n)_X+\delta.$$

Since δ is arbitrary, the above inequality and (.1) prove the lemma. □

Proof of Lemma 6: It is known, see (Tomczak-Jaegermann, 1989, Prop. 37.1), that for any two n-dimensional Banach spaces Y_1 and Y_2 there exists a constant $\rho \le n$, $\rho = \rho(n)$, and an onto linear map $T : Y_1 \to Y_2$ such that $\|T\| \cdot \|T^{-1}\| = \rho$. By suitable rescaling we can assume that $\|T\| = 1$ and $\|T^{-1}\| = \rho$.

Let $(\mathbb{R}^n, \|.\|_{\mathcal{Y}_n})$ be the space determined from the norm $\| \cdot \|_{\mathcal{Y}_n}$ in Theorem 1. From the above we infer that there exists a linear map T with the properties

$$T : (\mathbb{R}^n, \|.\|_{Z_n}) \to (\mathbb{R}^n, \|.\|_{\mathcal{Y}_n}), \quad \|T\| = 1, \quad \|T^{-1}\| = \rho.$$

We now define the mapping

$$\phi(y) := t(y)Ty, \quad \text{where} \quad t(y) := \frac{\|y\|_{Z_n}}{\|Ty\|_{\mathcal{Y}_n}}.$$

Since $\|\phi(y)\|_{\mathcal{Y}_n} = \|y\|_{Z_n}$, we conclude that $\phi : (B_{Z_n}(1), \|\cdot\|_{Z_n}) \to (B_{\mathcal{Y}_n}(1), \|\cdot\|_{\mathcal{Y}_n})$. Moreover, ϕ is an onto mapping since for every $y' \in B_{\mathcal{Y}_n}(1)$, there is

$$y = \frac{\|y'\|_{\mathcal{Y}_n}}{\|T^{-1}y'\|_{Z_n}} T^{-1}y' \in B_{Z_n}(1), \quad \text{such that} \quad \phi(y) = y'.$$

Note that $t(y) \le \rho$ since $\|y\|_{Z_n} = \|T^{-1}(Ty)\|_{Z_n} \le \|T^{-1}\| \|Ty\|_{\mathcal{Y}_n} = \rho \|Ty\|_{\mathcal{Y}_n}$, and

$$\begin{aligned}
|t(y) - t(z)| &\le \left| \frac{\|y\|_{Z_n}}{\|Ty\|_{\mathcal{Y}_n}} - \frac{\|y\|_{Z_n}}{\|Tz\|_{\mathcal{Y}_n}} \right| + \left| \frac{\|y\|_{Z_n}}{\|Tz\|_{\mathcal{Y}_n}} - \frac{\|z\|_{Z_n}}{\|Tz\|_{\mathcal{Y}_n}} \right| \\
&= \frac{\|y\|_{Z_n}}{\|Ty\|_{\mathcal{Y}_n} \|Tz\|_{\mathcal{Y}_n}} \big| \|Tz\|_{\mathcal{Y}_n} - \|Ty\|_{\mathcal{Y}_n} \big| + \frac{1}{\|Tz\|_{\mathcal{Y}_n}} \big| \|z\|_{Z_n} - \|y\|_{Z_n} \big| \\
&\le \frac{t(y)}{\|Tz\|_{\mathcal{Y}_n}} \|T(z-y)\|_{\mathcal{Y}_n} + \frac{1}{\|Tz\|_{\mathcal{Y}_n}} \|z-y\|_{Z_n} \le \frac{\rho+1}{\|Tz\|_{\mathcal{Y}_n}} \|z-y\|_{Z_n}.
\end{aligned}$$

Therefore, for every $y, z \in B_{Z_n}(1)$, using the above two inequalities, we obtain that

$$\begin{aligned}
\|\phi(y) - \phi(z)\|_{\mathcal{Y}_n} &\le t(y)\|T(y-z)\|_{\mathcal{Y}_n} + |t(y) - t(z)| \|Tz\|_{\mathcal{Y}_n} \\
&\le \rho \|y-z\|_{Z_n} + (\rho+1)\|y-z\|_{Z_n} = (2\rho+1)\|y-z\|_{Z_n},
\end{aligned}$$

which shows that ϕ is a $(2\rho+1)$-Lipschitz mapping.

Now, if $\mathcal{L}_n : (B_{\mathcal{Y}_n}(1), \| \cdot \|_{\mathcal{Y}_n}) \to X$ is any γ-Lipschitz mapping, then the composition $\mathcal{L}_n \circ \phi : (B_{Z_n}(1), \| \cdot \|_Z) \to X$ is a $(2\rho+1)\gamma$-Lipschitz mapping for which

$$d^{(2\rho+1)\gamma}(\mathcal{K}, Z_n)_X \le \sup_{f \in \mathcal{K}} \inf_{y \in B_{Z_n}(1)} \|f - \mathcal{L}_n \circ \phi(y)\|_X = \sup_{f \in \mathcal{K}} \inf_{y \in B_{\mathcal{Y}_n}(1)} \|f - \mathcal{L}_n(y)\|_X.$$

Next, we take the infimum over Φ_n and obtain

$$d^{(2\rho+1)\gamma}(\mathcal{K}, Z_n)_X \le d^{\gamma}(\mathcal{K}, \mathcal{Y}_n)_X = d_n^{\gamma}(\mathcal{K})_X,$$

which completes the proof since $\rho \le n$ and the Lipschitz width is monotone with respect to the Lipschitz constant. $\square$

Proof of Theorem 9: We prove the theorem by contradiction and start with (i). If (4.2) does not hold for some constant C_1, then there exists a strictly increasing sequence of natural numbers $(n_k)_{k=1}^{\infty}$, such that

$$p_k := \frac{d_{n_k}^{\gamma_{n_k}}(\mathcal{K})_X[n_k\varphi(n_k)]^{\alpha}}{[\log_2(n_k\varphi(n_k))]^{\beta}} \to 0 \quad \text{as} \quad k \to \infty.$$

To simplify the notation, we denote $P_k := n_k\varphi(n_k)$ and we then write

$$d_{n_k}^{\gamma_{n_k}}(\mathcal{K})_X = \frac{p_k\,[\log_2 P_k]^{\beta}}{P_k^{\alpha}} < \frac{2p_k\,[\log_2 P_k]^{\beta}}{P_k^{\alpha}} =: \delta_k, \quad k = 1, 2, \dots.$$

We also denote $Q_k := \log_2(3\gamma_{n_k}\delta_k^{-1})$ and apply Proposition 8 with $\eta_n = c_1(\log_2 n)^{\beta} n^{-\alpha}$ and $\delta = \delta_k$ to obtain

$$c_1\,[\log_2(n_kQ_k)]^{\beta}\, n_k{}^{-\alpha}Q_k^{-\alpha} \leq 4\frac{p_k\,[\log_2 P_k]^{\beta}}{[P_k]^{\alpha}}.$$

We rewrite the latter inequality as

$$p_k^{-1}\left[\frac{\log_2(n_kQ_k)}{\log_2 P_k}\right]^{\beta} \leq \frac{4}{c_1}\left[\frac{Q_k}{\varphi(n_k)}\right]^{\alpha}. \tag{.2}$$

In what follows, we denote by C a generic constant whose value may change every time. Observe that

$$\begin{aligned} Q_k &= c + \varphi(n_k) + \log_2(p_k^{-1}) + \alpha\log_2 P_k - \beta\log_2(\log_2 P_k) \\ &\asymp \varphi(n_k) + \log_2(p_k^{-1}) + \log_2 P_k \asymp \varphi(n_k) + \log_2(p_k^{-1}) + \log_2 n_k. \end{aligned}$$

Since for all k's we have $\log_2 n_k \leq c\varphi(n_k)$, then

$$C\varphi(n_k) \leq Q_k \leq C(\varphi(n_k) + \log_2(p_k^{-1})), \tag{.3}$$

and therefore

$$\left[\frac{Q_k}{\varphi(n_k)}\right]^{\alpha} \leq C\left[\frac{\varphi(n_k) + \log_2(p_k^{-1})}{\varphi(n_k)}\right]^{\alpha} \leq C\left[1 + \log_2(p_k^{-1})\right]^{\alpha}. \tag{.4}$$

We now consider the following cases.
Case 1: $\beta \geq 0$. It follows from (.3) that

$$\frac{\log_2(n_kQ_k)}{\log_2 P_k} \geq C\frac{\log_2 P_k}{\log_2 P_k} = C,$$

which combined with (.2) and (.4) gives

$$p_k^{-1} \leq \frac{4}{c_1}\left[\frac{\log_2(n_kQ_k)}{\log_2 P_k}\right]^{-\beta}\left[\frac{Q_k}{\varphi(n_k)}\right]^{\alpha} \leq C\left[1 + \log_2(p_k^{-1})\right]^{\alpha},$$

which contradicts the fact that $p_k^{-1} \to \infty$.

Case 2: $\beta < 0$**.** It follows from (.3) that

$$\frac{\log_2(n_k Q_k)}{\log_2 P_k} \le C\frac{\log_2(n_k(\varphi(n_k)+\log_2(p_k^{-1})))}{\log_2 P_k}, \tag{.5}$$

and we consider two cases.

Case 2.1: If for infinitely many k's we have $\log_2(p_k^{-1}) \le \varphi(n_k)$, then for those k's (.5) becomes

$$\frac{\log_2(n_k Q_k))}{\log_2 P_k} \le C\frac{\log_2(n_k(2\varphi(n_k)))}{\log_2 P_k} \le C,$$

and (.2) gives

$$p_k^{-1} \quad \le \quad \frac{4}{c_1}\left[\frac{\log_2(n_k Q_k))}{\log_2 P_k}\right]^{-\beta}\left[1+\log_2(p_k^{-1})\right]^{\alpha} \le C\left[1+\log_2(p_k^{-1})\right]^{\alpha},$$

which contradicts the fact that $p_k^{-1} \to \infty$.

Case 2.2: If for infinitely many k's we have $\log_2(p_k^{-1}) \ge \varphi(n_k) \ge c\log_2 n_k$, then (.5) becomes

$$\begin{aligned}\frac{\log_2(n_k Q_k)}{\log_2 P_k} \quad &\le \quad C\log_2(n_k\log_2(p_k^{-1})) = C(\log_2 n_k + \log_2(\log_2(p_k^{-1}))) \\ &\le \quad C(\log_2(p_k^{-1}) + \log_2(\log_2(p_k^{-1}))) \le C\log_2(p_k^{-1}),\end{aligned}$$

and leads to

$$p_k^{-1} \quad \le \quad \frac{4}{c_1}\left[\frac{\log_2(n_k Q_k))}{\log_2 P_k}\right]^{-\beta}\left[1+\log_2(p_k^{-1})\right]^{\alpha} \le C\left[\log_2(p_k^{-1})\right]^{-\beta}\left[1+\log_2(p_k^{-1})\right]^{\alpha},$$

which contradicts the fact that $p_k^{-1} \to \infty$.

To prove (ii), we repeat the argument for (i), namely, we assume that (ii) does not hold, and therefore, there exists a strictly increasing sequence of natural numbers $(n_k)_{k=1}^{\infty}$, such that

$$e_k := d_{n_k}^{\gamma_n}(\mathcal{K})_X[D_k]^{\alpha} \to 0 \quad \text{as} \quad k \to \infty,$$

where $D_k := \log_2(n_k\varphi(n_k))$. We write

$$d_{n_k}^{\gamma_{n_k}}(\mathcal{K})_X = e_k[D_k]^{-\alpha} < 2e_k[D_k]^{-\alpha} =: \delta_k, \quad k = 1, 2, \dots, \tag{.6}$$

and use Proposition 8 with $\eta_n = c_1(\log_2 n)^{-\alpha}$ to derive $c_1\left[\log_2(n_k Q_k)\right]^{-\alpha} \le 4e_k D_k^{-\alpha}$, where as in (1) we have the notation $Q_k := \log_2(3\gamma_{n_k}\delta_k^{-1})$. We rewrite this inequality as

$$e_k^{-1} \le \frac{4}{c_1}\left[\frac{\log_2(n_k Q_k)}{D_k}\right]^{\alpha}. \tag{.7}$$

Since $\log_2 n_k \le c\varphi(n_k)$, we have

$$\begin{aligned}Q_k \quad &= \quad \log_2(1.5) + \varphi(n_k) + \log_2(e_k^{-1}) + \alpha\log_2 D_k \\ &\le \quad C[\varphi(n_k) + \log_2(e_k^{-1}) + \log_2(\log_2 n_k + \log_2(\varphi(n_k)))] \\ &\le \quad C[\varphi(n_k) + \log_2(e_k^{-1}) + \log_2(\varphi(n_k) + \log_2(\varphi(n_k)))] \le C[\varphi(n_k) + \log_2(e_k^{-1})],\end{aligned}$$

and (.7) becomes

$$e_k^{-1} \le C \left[\frac{\log_2(n_k(\varphi(n_k) + \log_2(e_k^{-1})))}{D_k} \right]^{\alpha}. \tag{.8}$$

Case 1: If for infinitely many values of k we have $\log_2(e_k^{-1}) \le \varphi(n_k)$, then $e_k^{-1} \le C$, which contradicts with the fact that $e_k^{-1} \to \infty$ as $k \to \infty$.
Case 2: If for infinitely many values of k we have $\log_2(e_k^{-1}) \ge \varphi(n_k) \ge c \log_2 n_k$, then for those k's we get $D_k \ge C$, so (.8) becomes

$$\begin{aligned} e_k^{-1} &\le C \left[\log_2(n_k(\varphi(n_k) + \log_2(e_k^{-1})))\right]^{\alpha} = C \left[\log_2 n_k + \log_2(\varphi(n_k) + \log_2(e_k^{-1}))\right]^{\alpha} \\ &\le C \left[\log_2(e_k^{-1}) + \log_2(\log_2(e_k^{-1}))\right]^{\alpha} \le C \left[\log_2(e_k^{-1})\right]^{\alpha}, \end{aligned}$$

which also contradicts with the fact that $e_k^{-1} \to \infty$ as $k \to \infty$. The proof is completed. □

Proof of Lemma 15: It follows from Proposition 3.6 in Petrova and Wojtaszczyk (2023) that if $d_n^{\gamma_n}(\mathcal{K})_X < \epsilon/2$, then

$$\gamma_n \ge \frac{1}{6} \epsilon N_\epsilon^{1/n}(\mathcal{K}) \quad \Rightarrow \quad N_\epsilon(\mathcal{K}) \le \left[\frac{6 \cdot 2^{\varphi(n)}}{\epsilon}\right]^n \le 2^r,$$

where $N_\epsilon(\mathcal{K})$ is the ϵ covering number of $\mathcal{K}$ (the cardinality of the minimal ϵ-covering of $\mathcal{K}$) and

$$r = r(n) = \lceil n(\varphi(n) + \log_2(6/\epsilon)) \rceil$$

is the smallest integer that this inequality holds. We rewrite the above as $\epsilon_r(\mathcal{K})_X \le \epsilon$.

If $d_n^{\gamma_n}(\mathcal{K})_X = 0$, then for any $k \in \mathbb{N}$ such that $k > \varphi(n)$, we take $\epsilon = 6 \cdot 2^{\varphi(n)-k}$ and obtain $r = nk$ and

$$\epsilon_{nk}(\mathcal{K})_X \le 6 \cdot 2^{\varphi(n)-k} = 6 \cdot \gamma_n 2^{-k}.$$

This is an optimal estimate since for the ball $\mathcal{K}_n := \{y \in \mathbb{R}^n \ : \ \|y\|_{\ell_2^n} \le \gamma_n\}$ of radius γ_n in $\mathbb{R}^n$, we have that $d_n^{\gamma_n}(\mathcal{K}_n)_X = 0$ since for the γ_n-Lipschitz mapping $\Phi(y) = \gamma_n y$ we have $\mathcal{K}_n = \Phi(B_{\ell_2^n}(1))$. On the other hand, it is a well known fact that for any integer s we have, see Carl and Stephani (1990),

$$\gamma_n 2^{-s/n} \le \epsilon_s(\mathcal{K}_n)_X \le 4\gamma_n 2^{-s/n}.$$

Now we continue with the other cases of the behavior of the Lipschitz width $d_n^{\gamma_n}(\mathcal{K})_X$ under the assumption that $\gamma_n = 2^{cn^p[\log_2 n]^q}$. In the case when $\epsilon = 3d_n^{\gamma_n}(\mathcal{K})_X$ with $d_n^{\gamma_n}(\mathcal{K})_X > 0$, we obtain that $\epsilon_r(\mathcal{K})_X \le 3d_n^{\gamma_n}(\mathcal{K})_X$, where

$$r(n) = \lceil n(\varphi(n) + \log_2(2[d_n^{\gamma_n}(\mathcal{K})_X]^{-1})) \rceil \le \lceil n(2\varphi(n) - \log_2(d_n^{\gamma_n}(\mathcal{K})_X)) \rceil. \tag{.9}$$

In the case when $0 < d_n^{\gamma_n}(\mathcal{K})_X \lesssim [\log_2 n]^\beta n^{-\alpha}$, it follows from (.9) that

$$r(n) \lesssim \lceil n^{p+1}[\log_2 n]^q + n\alpha \log_2 n - n\beta \log_2(\log_2 n) \rceil \lesssim \lceil n^{p+1}[\log_2 n]^q + n \log_2 n \rceil$$

provided n is sufficiently large.

Case 1: $p>0$ and $q\in\mathbb{R}$ or $p=0$ and $q\geq 1$. For this range of p and q we have

$$r(n)\lesssim n^{1+p}[\log_2 n]^q.$$

In this case, for each $k\in\mathbb{N}$ large enough, we consider $n=n(k)$, defined as

$$n=\lceil k^{1/(p+1)}[\log_2 k]^{-q/(p+1)}\rceil < 2k^{1/(p+1)}[\log_2 k]^{-q/(p+1)},$$

and for k large enough we have

$$\begin{aligned} r(n) &\lesssim \left[k^{1/(p+1)}[\log_2 k]^{-q/(p+1)}\right]^{p+1}[\log_2(k^{1/(p+1)}[\log_2 k]^{-q/(p+1)})]^q \\ &= k[\log_2 k]^{-q}\left[(p+1)^{-1}\log_2 k-\frac{q}{p+1}\log_2(\log_2 k)\right]^q. \end{aligned}$$

Note that

$$(p+1)^{-1}\log_2 k-\frac{q}{p+1}\log_2\log_2 k\asymp \log_2 k, \tag{.10}$$

and therefore

$$r(n)\lesssim k[\log_2 k]^{-q}\,[\log_2 k]^q=k$$

Since $\epsilon_k(\mathcal{K})_X$ is a monotone decreasing sequence, we have that

$$\begin{aligned} \epsilon_{ck(}\mathcal{K})_X &\leq \epsilon_{r(n)}(\mathcal{K})_X\leq 3d_n^{\gamma_n}(\mathcal{K})_X\lesssim[\log_2 n]^\beta n^{-\alpha} \\ &\lesssim [\log_2(k^{1/(p+1)}[\log_2 k]^{-q/(p+1)})]^\beta[k^{1/(p+1)}[\log_2 k]^{-q/(p+1)}]^{-\alpha} \\ &= \left[\frac{1}{p+1}\log_2 k-\frac{q}{p+1}\log_2(\log_2 k)\right]^\beta k^{-\alpha/(p+1)}[\log_2 k]^{q\alpha/(p+1)} \end{aligned}$$

It follows from (.10) that

$$\epsilon_{ck(}\mathcal{K})_X\lesssim[\log_2 k]^\beta\,k^{-\alpha/(p+1)}[\log_2 k]^{q\alpha/(p+1)}=k^{-\alpha/(p+1)}\,[\log_2 k]^{\beta+\frac{\alpha q}{p+1}}.$$

Case 2: $p=0$ and $q<1$. For this range of p and q we have $r(n)\lesssim n\log_2 n$. In this case, for each $k\in\mathbb{N}$ large enough we consider $n=n(k)$, given as

$$n=\lceil k[\log_2 k]^{-1}\rceil<2k[\log_2 k]^{-1},$$

and for k large enough we have

$$r(n)\lesssim k[\log_2 k]^{-1}\log_2(k[\log_2 k]^{-1})=k[\log_2 k]^{-1}\left[\log_2 k-\log_2(\log_2 k)\right].$$

Using the fact that $\log_2 k-\log_2(\log_2 k)\asymp\log_2 k$, we conclude that $r(n)\lesssim k$, and as in Case 1, we have that

$$\begin{aligned} \epsilon_{ck(}\mathcal{K})_X &\leq \epsilon_{r(n)}(\mathcal{K})_X\leq 3d_n^{\gamma_n}(\mathcal{K})_X\lesssim[\log_2 n]^\beta n^{-\alpha}\lesssim[\log_2(k[\log_2 k]^{-1})]^\beta[k[\log_2 k]^{-1}]^{-\alpha} \\ &= [\log_2 k-\log_2(\log_2 k)]^\beta\,k^{-\alpha}[\log_2 k]^\alpha\lesssim k^{-\alpha}[\log_2 k]^{\beta+\alpha}, \end{aligned}$$

and the proof for this particular case is completed.

Now, when $0 < d_n^{\gamma_n}(\mathcal{K})_X \lesssim 2^{-cn}$, it follows from (.9) that

$$r(n) \lesssim n(\varphi(n)+n) \lesssim \lceil n^{p+1}[\log_2 n]^q + n^2\rceil.$$

Case 1: $0 \leq p < 1$ **and** $q \in \mathbb{R}$ **or** $p = 1$ **and** $q \leq 0$. For this range of p and q we have $r(n) \lesssim n^2$. In this case, for each $k \in \mathbb{N}$ large enough we consider $n = n(k)$ given by $n = \lceil\sqrt{k}\rceil \leq 2\sqrt{k}$, and for k large enough we have $r(n) \lesssim k$ and thus

$$\epsilon_{\tilde{c}k(}\mathcal{K})_X \quad \leq \quad \epsilon_{r(n)}(\mathcal{K})_X \leq 3d_n^{\gamma_n}(\mathcal{K})_X \lesssim 2^{-cn} \lesssim 2^{-c\sqrt{k}}.$$

Case 2: $p > 1$ **and** $q \in \mathbb{R}$ **or** $p = 1$ **and** $q > 0$. In this case $r(n) \lesssim n^{p+1}[\log_2 n]^q$, and for each $k \in \mathbb{N}$ large enough, we define $n = n(k)$ via

$$n = \lceil k^{1/(p+1)}[\log_2 k]^{-q/(p+1)}\rceil < 2k^{1/(p+1)}[\log_2 k]^{-q/(p+1)},$$

and for k large enough, as before, we have

$$\begin{aligned} r(n) \quad &\lesssim \quad \left[k^{1/(p+1)}[\log_2 k]^{-q/(p+1)}\right]^{p+1}[\log_2(k^{1/(p+1)}[\log_2 k]^{-q/(p+1)})]^q \\ &= \quad k[\log_2 k]^{-q}\left[(p+1)^{-1}\log_2 k - \frac{q}{p+1}\log_2(\log_2 k)\right]^q \lesssim k[\log_2 k]^{-q}\,[\log_2 k]^q = k. \end{aligned}$$

Therefore, we obtain

$$\epsilon_{\tilde{c}k(}\mathcal{K})_X \lesssim d_n^{\gamma_n}(\mathcal{K})_X \lesssim 2^{-cn} \leq 2^{-c[k[\log_2 k]^{-q}]^{1/(p+1)}}.$$

This completes the proof of the lemma. □

We now state and proof a lemma that we used in §5.4 and §5.5.

Lemma 30 *Let* $\Sigma_n,\ n \in \mathbb{N}$, *be compact subsets of a Banach space* Y *which is continuously embedded in the Banach space* X *and let* $(\xi_n)_{n=1}^\infty$ *be a sequence of non-negative numbers such that* $\inf_n \xi_n = 0$ *(in particular* $\lim_{n\to\infty}\xi_n = 0$*). If* $V_n := \{f \in X \ :\ \mathrm{dist}(f,\Sigma_n)_X \leq \xi_n\}$, *then the set* $\mathcal{K} := \bigcap_{n=1}^\infty V_n$ *is compact in* X.

Proof: Since each Σ_n is a compact set, each V_n is a closed and bounded subset of X. This implies that $\mathcal{K}$ is closed and bounded (possibly empty) subset of X. To prove that $\mathcal{K}$ is compact, we argue via contradiction. Let us assume that there exists $\delta > 0$ and a sequence of elements $(f_j)_{j=1}^\infty \subset \mathcal{K}$ such that $\|f_j - f_{j'}\|_X \geq \delta$ whenever $j \neq j'$. Let us fix n such that $\xi_n \leq \delta/3$. Since $(f_j)_{j=1}^\infty \subset V_n$, for $j \in N$ we have $f_j = v_j + q_j$, where $v_j \in \Sigma_n$ and $\|q_j\|_X \leq \xi_n$. Thus

$$\|v_j - v_{j'}\|_X = \|f_j - f_{j'} - q_j + q_{j'}\|_X \geq \|f_j - f_{j'}\|_X - \|q_j\|_X - \|q_{j'}\|_X \geq \delta - 2\xi_n \geq \delta/3,$$

which means that $\delta/3 \leq \|v_j - v_{j'}\|_X \leq C\|v_j - v_{j'}\|_Y$, and thus Σ_n is *not compact.* The proof is completed. □

We next state a simple lemma that we use in the proof of Lemma 32.

Lemma 31 *The following holds*

- *The function* $f_{\alpha,\beta} : [1,\infty) \to \mathbb{R}$, *defined as* $f_{\alpha,\beta}(t) := t^{-\alpha}[\log_2 t]^\beta$, *for* $\alpha > 0$, $\beta \in \mathbb{R}$ *is bounded, monotonically decreasing on* $[1,\infty)$ *when* $\beta \le 0$ *and on* $[2^{\beta/\alpha},\infty)$ *when* $\beta \ge 0$.

- *If* $\varphi : [1,\infty) \to \mathbb{R}$ *is an increasing non-negative function then*

$$\exists\, c_1 > 1, \text{ such that } \sup_{n\in\mathbb{N}} \frac{\varphi(c_1 n)}{\varphi(n)} < \infty \quad \Leftrightarrow \quad \forall\, c > 1, \sup_{n\in\mathbb{N}} \frac{\varphi(cn)}{\varphi(n)} < \infty.$$

Proof: The proof is simple calculus and we omit it. □

Lemma 32 *Let* (a_n) *be a monotone non-increasing sequence and* $\varphi : [1,\infty) \to \mathbb{R}$ *be an increasing non-negative function such that* $\varphi(n) \to \infty$ *as* $n \to \infty$. *If there is a constant* $c > 1$ *such that*

$$A_n := \frac{[\log_2(cn\varphi(cn))]^\beta}{[cn\varphi(cn)]^\alpha} \lesssim a_{cn\lceil \frac{1}{2}\varphi(cn)\rceil} \lesssim \frac{[\log_2(n\varphi(n))]^\beta}{[n\varphi(n)]^\alpha} =: B_n, \tag{.11}$$

then

- *if the function* φ *is such that there is a constant* $c_1 > 1$ *for which* $\sup_{n\in\mathbb{N}} \frac{\varphi(c_1 n)}{\varphi(n)} < \infty$ *then* $a_m \asymp (\log_2 m)^\beta m^{-\alpha}$, $m = 1, 2, \ldots$.

- *if there is* $c > 1$ *such that* $\sup_{n\in\mathbb{N}} \frac{\varphi(cn)}{\varphi(n)} = \infty$, *then the lower and upper bound in* (.11) *are asymptotically different in the sense that*

$$\frac{B_n}{A_n} \gtrsim \begin{cases} \left[\frac{\varphi(cn)}{\varphi(n)}\right]^\alpha \left[\log_2 \frac{\varphi(cn)}{\varphi(n)}\right]^{-\beta}, & \beta > 0, \\[2ex] \left[\frac{\varphi(cn)}{\varphi(n)}\right]^\alpha, & \beta \le 0. \end{cases}$$

Proof: It follows from Lemma 31 that the quantity $\sup_{n\in\mathbb{N}} \frac{\varphi(c_1 n)}{\varphi(n)}$ can be either finite for some $c_1 > 1$ (and therefore for all $c > 1$) or infinite for every $c_1 > 1$. Let us first consider the former case, which, according to the same lemma, gives that $\sup_{n\in\mathbb{N}} \frac{\varphi(cn)}{\varphi(n)} := \rho = \rho(c) < \infty$. We next bound the quantity $\frac{B_n}{A_n}$ as

$$\frac{B_n}{A_n} = c^\alpha \left[\frac{\log_2(n\varphi(n))}{\log_2(cn\varphi(cn))}\right]^\beta \left[\frac{\varphi(cn)}{\varphi(n)}\right]^\alpha =: c^\alpha R_n^\beta S_n^\alpha \le \tilde{C},$$

since $S_n \le \rho$ and

$$1 \le R_n^{-1} = \frac{\log_2(cn\varphi(cn))}{\log_2(n\varphi(n))} \le \frac{\log_2(c\rho) + \log_2(n\varphi(n))}{\log_2(n\varphi(n))} = 1 + \frac{\log_2(c\rho)}{\log_2 n\varphi(n)} \le \tilde{c},$$

for n large enough. Thus, (.11) becomes

$$A_n \lesssim a_{cn\lceil \frac{1}{2}\varphi(cn)\rceil} \lesssim A_n. \tag{.12}$$

We now fix k and choose $n = n(k)$ such that $cn\lceil \frac{1}{2}\varphi(cn)\rceil < k \le c(n+1)\lceil \frac{1}{2}\varphi(c(n+1))\rceil$. Clearly, because of the monotonicity of (a_k) and (.12), for such k we have

$$A_{n+1} \lesssim a_k \lesssim A_n. \tag{.13}$$

Next, let us consider the ratio

$$\frac{A_n}{A_{n+1}} = \left[\frac{\log_2(cn\varphi(cn))}{\log_2(c(n+1)\varphi(c(n+1)))}\right]^{\beta} \left[\left(1+\frac{1}{n}\right)\frac{\varphi(c(n+1))}{\varphi(cn)}\right]^{\alpha} =: P_n^{\beta} Q_n^{\alpha}.$$

Note that ${A_n A_{n+1}}^{-1} \le C_1$. Indeed, since

$$\frac{\varphi(c(n+1))}{\varphi(cn)} \le \frac{\varphi(2cn)}{\varphi(cn)} \le \frac{\varphi(2cn)}{\varphi(n)} = \frac{\varphi(2cn)}{\varphi(2n)}\frac{\varphi(2n)}{\varphi(n)} \le \rho(c)\rho(2),$$

we have

$$Q_n = \left(1+\frac{1}{n}\right)\frac{\varphi(c(n+1))}{\varphi(cn)} \le \frac{3}{2}\rho(c)\rho(2),$$

and

$$\begin{aligned} 1 \quad \le \quad P_n^{-1} &= \frac{\log_2(c(n+1)\varphi(c(n+1)))}{\log_2(cn\varphi(cn))} \le \frac{\log_2(2cn\rho(c)\rho(2)\varphi(cn))}{\log_2(cn)+\log_2(\varphi(cn))} \\ &= \frac{1+\log_2(cn)+\log_2(\varphi(cn))+\log_2(\rho(c)\rho(2))}{\log_2(cn)+\log_2\varphi(cn)} = 1+\frac{\log_2(\rho(c)\rho(2))}{\log_2(cn)+\log_2\varphi(cn)} \le 2, \end{aligned}$$

for n large enough. Then, we can rewrite (.13) as

$$A_n \lesssim a_k \lesssim A_n, \quad \text{for} \quad cn\lceil\frac{1}{2}\varphi(cn)\rceil < k \le c(n+1)\lceil\frac{1}{2}\varphi(c(n+1))\rceil.$$

For any k in this interval we have that

$$\frac{c}{2}n\varphi(cn) < k \le c(n+1)\left(\frac{1}{2}\varphi(c(n+1))+1\right) < c(n+1)\varphi(c(n+1))$$

for $n = n(k)$ big enough. Therefore, it follows from Lemma 31 that

$$f_{\alpha,\beta}(2k) \le f_{\alpha,\beta}(cn\varphi(cn)) = A_n \lesssim a_k \lesssim A_n \lesssim A_{n+1} = f_{\alpha,\beta}(c(n+1)\varphi(c(n+1))) \le f_{\alpha,\beta}(k),$$

which implies that $a_k \asymp k^{-\alpha}[\log_2 k]^{\beta}$ with constants depending on ρ, α and β.

Finally, in the case when there is a $c > 1$ such that $\sup_{n\in\mathbb{N}} \dfrac{\varphi(cn)}{\varphi(n)} = \infty$ (and therefore this holds for every $c > 1$), we have that

$$\frac{B_n}{A_n} = c^{\alpha}\left[\frac{\log_2(n\varphi(n))}{\log_2(cn\varphi(cn))}\right]^{\beta}\left[\frac{\varphi(cn)}{\varphi(n)}\right]^{\alpha} \ge c^{\alpha}\left[\frac{\varphi(cn)}{\varphi(n)}\right]^{\alpha}, \quad \text{for} \quad \beta \le 0.$$

When $\beta > 0$ we write $\varphi(cn) = k(n)\varphi(n)$ and have

$$\frac{B_n}{A_n} = c^\alpha [k(n)]^\alpha \left[\frac{\log_2(n\varphi(n))}{\log_2 c + \log_2(k(n)) + \log_2(n\varphi(n))}\right]^\beta.$$

We consider n large enough so that $n\varphi(n) \geq c$.

Case 1: If $k(n) \geq n\varphi(n)$ we have that

$$\frac{B_n}{A_n} \gtrsim [\log_2(n\varphi(n))]^\beta \frac{[k(n)]^\alpha}{[3\log_2(k(n))]^\beta} \gtrsim \frac{k(n)^\alpha}{[\log_2(k(n))]^\beta},$$

for n large enough.

Case 2: If $k(n) < n\varphi(n)$ we obtain that

$$\frac{B_n}{A_n} \gtrsim [k(n)]^\alpha \left[\frac{\log_2(n\varphi(n))}{2\log_2(n\varphi(n)) + \log_2 c}\right]^\beta \geq [k(n)]^\alpha 3^{-\beta} \gtrsim [k(n)]^\alpha,$$

which concludes the proof of the lemma. We do not know whether in the case when $\sup_{n\in\mathbb{N}} \frac{\varphi(cn)}{\varphi(n)} = \infty$, the discrepancy of the behavior of $\frac{B_n}{A_n}$ for positive and negative β's is supported by examples or is due to our approach. $\square$