



**HAL**  
open science

# Approximation speed of quantized vs. unquantized ReLU neural networks and beyond

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti

► **To cite this version:**

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti. Approximation speed of quantized vs. unquantized ReLU neural networks and beyond. *IEEE Transactions on Information Theory*, 2023, 69 (6), pp.3960-3977. 10.1109/TIT.2023.3240360 . hal-03672166v2

**HAL Id: hal-03672166**

**<https://hal.science/hal-03672166v2>**

Submitted on 6 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Approximation speed of quantized *vs.* unquantized ReLU neural networks and beyond

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti

## Abstract

We deal with two complementary questions about approximation properties of ReLU networks. First, we study how the uniform quantization of ReLU networks with real-valued weights impacts their approximation properties. We establish an upper-bound on the minimal number of bits per coordinate needed for uniformly quantized ReLU networks to keep the same polynomial asymptotic approximation speeds as unquantized ones. We also characterize the error of nearest-neighbour uniform quantization of ReLU networks. This is achieved using a new lower-bound on the Lipschitz constant of the map that associates the parameters of ReLU networks to their realization, and an upper-bound generalizing classical results. Second, we investigate when ReLU networks can be expected, or not, to have better approximation properties than other classical approximation families. Indeed, several approximation families share the following common limitation: their polynomial asymptotic approximation speed of any set is bounded from above by the encoding speed of this set. We introduce a new abstract property of approximation families, called  $\infty$ -encodability, which implies this upper-bound. Many classical approximation families, defined with dictionaries or ReLU networks, are shown to be  $\infty$ -encodable. This unifies and generalizes several situations where this upper-bound is known.

## Index Terms

Approximation speed, encoding speed, ReLU neural networks, quantization.

## I. INTRODUCTION

Neural networks are used with success in many applications to approximate functions. In line with the works [5], [8], [10], we are interested in understanding their approximation power in practice and in theory. We deal with two complementary questions. Regarding practical applications, a key question is to be able to compare approximation properties of unquantized versus quantized neural networks, i.e., networks with arbitrary real weights versus networks whose weights are constrained to a prescribed finite set (e.g., floats).

Authors are with Univ Lyon, Ens Lyon, UCBL, CNRS, Inria, LIP, F-69342, LYON Cedex 07, France. This work has been partially presented during a poster session of the winter school "LMS Invited Lectures On The Mathematics Of Deep Learning" at the Isaac Newton Institute, Cambridge, UK (<https://sites.google.com/view/lmslecturesmaths4dl/home>). It has also been partially presented during a talk at the conference "Curve and Surfaces 2022", Arcachon, France (<https://cs2022.sciencesconf.org/>). These two presentations did not lead to publications (there were no proceedings for both meetings). This work was supported in part by the AllegroAssai ANR-19-CHIA-0009 and NuSCAP ANR-20-CE48-0014 projects of the French Agence Nationale de la Recherche.

The results obtained in this direction are described below in Contribution 1. A practical type of quantization on which we will focus is uniform quantization, i.e., when the weights are only allowed to be in a finite subset of a uniform grid of the real line. Another important question is to better understand non-trivial situations where neural networks, quantized or not, can be expected (or not) to have better approximation properties than the best known approximation families<sup>1</sup>. We lay a framework that can be used to identify such situations, as described in Contribution 2.

**Contribution 1 (approximation with quantized networks).** — Consider the parameters  $\theta \in \mathbb{R}^d$  of a ReLU neural network and denote  $R_\theta \in L^p$  the associated function, called its *realization*, cf. Definition II.4. We also say that  $R_\theta$  is *realized* by the network with parameters  $\theta$ . Consider a quantization scheme<sup>2</sup>  $Q : \mathbb{R}^d \rightarrow \mathbb{R}^d$  used to quantize  $\theta$  (e.g.,  $Q(\theta)$  has only float coordinates). We address three main questions.

a) *Quantization error:* A first question is to study how the quantization error  $\|R_\theta - R_{Q(\theta)}\|_p$  (see section II-B for the definition of  $\|\cdot\|_p$ ) depends on the quantization scheme  $Q$ . The following result goes in that direction.

**Informal Theorem I.1 (see Theorem III.1).** — Fix an architecture (see Definition II.2) of feedforward ReLU networks, i.e., fix the number of layers, denoted by  $L$ , and their width. Denote  $W$  the maximal width of the layers. Denote  $\Theta(r)$  the collection of all parameters of such networks having their Euclidean norm bounded by  $r \geq 1$ . Consider  $1 \leq p \leq \infty$ . Then, the Lipschitz constant  $\text{Lips}(W, L, r)$  of the map that associates the parameters  $\theta \in \Theta(r)$  to their realization  $R_\theta \in L^p$  satisfies, with constants  $c, c' > 0$  only depending on the  $L^p$  space:

$$c' L r^{L-1} \leq \text{Lips}(W, L, r) \leq c W L^2 r^{L-1}.$$

To the best of our knowledge, the lower-bound given in this result is new, while the upper-bound generalizes classical results [2, Thm 2.6][13, Lem. 2] to generic  $L^p$  spaces and to more general constraints on the parameters. Thanks to Informal Theorem I.1, the number of bits used by a quantization scheme can be related to the error of this scheme. Our second result does so in the case of a nearest-neighbour uniform quantization scheme.

**Informal Theorem I.2 (see Theorem IV.1 and Theorem IV.2).** — Consider the same setting as in Informal Theorem I.1. Fix a stepsize  $\eta > 0$  and a desired error  $\varepsilon > 0$ . Consider the uniform quantization scheme<sup>3</sup>  $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$  applied coordinate-wise on the parameters  $\theta$  of a ReLU network. Then,  $\|R_\theta - R_{Q_\eta(\theta)}\|_\infty \leq \varepsilon$  holds for every  $\theta \in \Theta(r)$  if, and only if, the number of bits used to store each coordinate of  $Q_\eta(\theta)$ , which is proportional to  $\ln(1/\eta)$ , is linear in the depth  $L$ .

<sup>1</sup>An approximation family is any (often non-decreasing) sequence  $(\Sigma_M)_{M=1,2,\dots}$  of subsets of a metric space  $(\mathcal{F}, d)$ .

<sup>2</sup>A quantization scheme is a function with a finite image.

<sup>3</sup> $\lfloor \cdot \rfloor$  is defined as  $\lfloor x \rfloor := \max\{n \in \mathbb{Z}, n \leq x\}$  for every  $x \in \mathbb{R}$  while  $\lceil \cdot \rceil$  is defined as  $\lceil x \rceil := \min\{n \in \mathbb{Z}, n \geq x\}$  for every  $x \in \mathbb{R}$ .

The generality of our first result suggests that the second one can be generalized to other settings, this will be further discussed in Remark IV.2. As a consequence of our first result, we also prove (cf. Proposition IV.1) that Lemma VI.8 in [8], which controls quantization errors of the type  $\|R_\theta - R_{Q(\theta)}\|_p$  when  $Q$  is a nearest-neighbour uniform quantization scheme, can be improved.

*b) Approximation error:* Given a function  $f \in L^p$ , how "well" can  $f$  be approximated by quantized ReLU networks? If the parameters  $\theta$  of a ReLU network are known to approximate "well"  $f$ , then one can simply quantize  $\theta$  via a quantization scheme  $Q$  and write, using the triangle inequality:  $\|f - R_{Q(\theta)}\|_p \leq \|f - R_\theta\|_p + \|R_\theta - R_{Q(\theta)}\|_p$ . The results above can be used to control the quantization error  $\|R_\theta - R_{Q(\theta)}\|_p$ . Applying this to functions  $f$  in  $L^\infty$ -Sobolev spaces, and using external work [18] to guarantee the existence of parameters  $\theta$  approximating  $f$  "well", we recover Theorem 2 in [7], see Proposition IV.2. Much more generic applications can be envisioned, see Remark IV.3.

*c) Polynomial asymptotic approximation speed:* Consider a function  $f \in L^p$  and a sequence of parameters  $(\theta_M)_{M \in \mathbb{N}}$  (with  $\mathbb{N} = \{1, 2, \dots\}$ ). Can we design a sequence  $(Q_M)$  of quantization schemes such that the realizations of the networks with quantized parameters  $(Q_M(\theta_M))_{M \in \mathbb{N}}$  approximate the function  $f$  at the same asymptotic polynomial rate, with  $M$ , as the unquantized parameters  $(\theta_M)_{M \in \mathbb{N}}$ ? Using the triangle inequality  $\|f - R_{Q_M(\theta_M)}\|_p \leq \|f - R_{\theta_M}\|_p + \|R_{\theta_M} - R_{Q_M(\theta_M)}\|_p$  for each integer  $M$ , it is sufficient to guarantee that  $\|R_{\theta_M} - R_{Q_M(\theta_M)}\|_p$  decreases at the same polynomial asymptotic rate as  $\|f - R_{\theta_M}\|_p$ . Given a subset  $\mathcal{C}$  of a metric function space  $(\mathcal{F}, d)$  and an approximation family  $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$  in  $\mathcal{F}$ , the *polynomial asymptotic approximation speed*  $\gamma^{*approx}(\mathcal{C}|\Sigma)$  of  $\mathcal{C}$  by  $\Sigma$  [8, Def. V.2, Def. VI.1], called simply approximation speed in what follows, is the best polynomial rate at which all functions of  $\mathcal{C}$  are asymptotically approximated by  $\Sigma$ :

$$\gamma^{*approx}(\mathcal{C}|\Sigma) := \sup\{\gamma \in \mathbb{R}, \sup_{f \in \mathcal{C}} \inf_{\Phi \in \Sigma_M} d(f, \Phi) = \mathcal{O}_{M \rightarrow \infty}(M^{-\gamma})\} \in [-\infty, +\infty],$$

with the convention  $\gamma^{*approx}(\mathcal{C}|\Sigma) = -\infty$  if the supremum is over an empty set. In the following result, we exhibit a sufficient number of bits per coordinate that guarantees that nearest-neighbour uniform quantization preserves approximation rates of approximation families defined with ReLU networks.

**Informal Theorem I.3 (see Theorem V.1).** — Consider the approximation family  $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$  in an arbitrary  $L^p$  space, such that  $\Sigma_M$  is the set of functions realized by ReLU networks with depth bounded by  $L_M \in \mathbb{N}$ , with parameters having at most  $M$  non-zero coordinates and with Euclidean norm bounded by  $r_M \geq 1$ . For  $\gamma > 0$ , consider the  $\gamma$ -uniformly quantized sequence  $\mathcal{Q}(\Sigma|\gamma) := (\mathcal{Q}(\Sigma_M|\gamma))_{M \in \mathbb{N}}$ , where  $\mathcal{Q}_M(\Sigma_M|\gamma)$  is the set of functions realized by ReLU networks as above, but with parameters uniformly quantized using the quantization scheme  $Q_{\eta_M}(x) = \lfloor x/\eta_M \rfloor \eta_M$  for a step size  $\eta_M = M^{-\gamma} \text{Lips}(M, L_M, r_M)$ . Then, the  $\gamma$ -uniformly quantized sequence  $\mathcal{Q}(\Sigma|\gamma)$  has, on every set  $\mathcal{C} \subset L^p$ , an approximation speed which is comparable to its unquantized version  $\Sigma$ :

$$\begin{aligned} \gamma^{*approx}(\mathcal{C}|\mathcal{Q}(\Sigma|\gamma)) &= \gamma^{*approx}(\mathcal{C}|\Sigma) && \text{if } \gamma \geq \gamma^{*approx}(\mathcal{C}|\Sigma), \\ \gamma^{*approx}(\mathcal{C}|\mathcal{Q}(\Sigma|\gamma)) &\geq \gamma && \text{otherwise.} \end{aligned}$$

This theorem leads to explicit conditions on the number of bits per coordinate that guarantee quantized ReLU networks to have the same approximation speeds as unquantized ones, see Example V.1. In the proof, approximation speeds are matched by (i) taking unquantized parameters that (almost) achieve the unquantized approximation speed and (ii) quantizing these parameters with a sufficiently large number of bits in order to preserve the approximation speed. Smarter (but computationally more challenging) quantization schemes can be envisioned, such as directly picking the best quantized parameters to approximate the function. If the budget for the number of bits per coordinate is larger than the one given in Informal Theorem I.3, then even the smartest quantization scheme will not beat approach (i) + (ii) in terms of polynomial approximation speed (but it can still have better constants/log-terms etc.). Indeed, (i) + (ii) already yields the same approximation speeds for quantized networks as unquantized ones, and quantized networks cannot do better than unquantized ones. An open question is: what is the minimum number of bits per coordinate needed to keep the same approximation speeds? We partially answer this question: Informal Theorem I.3 gives an upper-bound.

**Contribution 2 (unified and generic framework for a relation between approximation and information theory).** — We investigate generic settings where the approximation speed of a set  $\mathcal{C}$  by an approximation family is bounded from above by the encoding speed of  $\mathcal{C}$ . The encoding speed is an informatic theoretic complexity that measures the best polynomial asymptotic rate at which the number of balls needed to cover the considered set grows as the radius of the balls goes to zero. Given a subset  $\mathcal{C}$  of a metric space  $(\mathcal{F}, d)$  and  $\varepsilon > 0$ , a finite subset  $X \subset \mathcal{C}$  is called an  $\varepsilon$ -covering of  $\mathcal{C}$  if:

$$\mathcal{C} \subset \bigcup_{x \in X} B_d(x, \varepsilon), \quad (1)$$

where  $B_d(x, \varepsilon)$  denotes the closed ball of  $\mathcal{C}$ , with respect to the metric  $d$ , centered in  $x$  and with radius  $\varepsilon$ . The *covering number*  $N(\mathcal{C}, d, \varepsilon)$  is the minimal size of an  $\varepsilon$ -covering of  $\mathcal{C}$ , with the convention that  $N(\mathcal{C}, d, \varepsilon) = +\infty$  if there is no such covering. The *metric entropy* is defined by  $H(\mathcal{C}, d, \varepsilon) := \log_2(N(\mathcal{C}, d, \varepsilon))$ . The encoding speed of  $\mathcal{C}$  is defined [8, Def. IV.1] as:

$$\gamma^{*\text{encod}}(\mathcal{C}) := \sup \left\{ \gamma > 0, H(\mathcal{C}, d, \varepsilon) = \mathcal{O}_{\varepsilon \rightarrow 0}(\varepsilon^{-1/\gamma}) \right\}, \quad (2)$$

with the convention that  $\gamma^{*\text{encod}}(\mathcal{C}) = 0$  if the supremum is over an empty set. The encoding speed is known for many  $\mathcal{C}$ 's, see [8, Table 1]. Consider an approximation family  $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$  and a set  $\mathcal{C}$ . If  $\Sigma_M$  approximates "well"  $\mathcal{C}$  (measured by  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ ), then one can use balls covering  $\Sigma_M$  to cover  $\mathcal{C}$ . This simple observation is at the origin of the following inequality, known in several situations, for instance when  $\Sigma$  is defined with dictionaries [8, Thm. V.3][10, Thm. 5.24] or ReLU networks [8, Thm. VI.4]:

$$\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C}). \quad (3)$$

Inequality (3) happens to be an equality in various cases, see [8, Table 1]. In Definition VI.2, we introduce an abstract property of approximation families  $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$ , called  $\gamma$ -encodability, for  $\gamma > 0$ , that measures how well each  $\Sigma_M$  can be covered by balls. Roughly speaking,  $\Sigma$  is  $\gamma$ -encodable if asymptotically in  $M$ , the

set  $\Sigma_M$  can be covered with nearly of the order of  $M$  balls of radius  $M^{-\gamma}$ . We say that it is  $\infty$ -encodable if it is  $\gamma$ -encodable for every  $\gamma > 0$ . Our next result is to show that  $\gamma$ -encodability can be used to understand several situations where (3) holds.

**Informal Theorem I.4 (see Theorem VI.1).** — Consider an approximation family  $\Sigma$  and  $\gamma > 0$ . If  $\Sigma$  is  $\gamma$ -encodable then for every set  $\mathcal{C}$ :

$$\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leq \gamma^{*\text{encod}}(\mathcal{C}).$$

Note that when an approximation family is  $\infty$ -encodable, this result gives Inequality (3). Inequality (3) is then of particular interest in order to bound the approximation speed  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$  from above without having to look at all at the approximation properties of the set  $\mathcal{C}$  by the sequence  $\Sigma$ . Instead, we can study separately  $\Sigma$  and establish at which speed it can be *encoded*. This lays a framework that we use to unify and generalize several situations where Inequality (3) is known [8, Thm. V.3, Thm. VI.4][10, Thm. 5.24]. Indeed, we show that many approximation families  $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$  are  $\infty$ -encodable: when  $\Sigma_M$  contains  $M$ -terms linear combinations of the first  $\text{poly}(M)$  elements<sup>4</sup> of a bounded dictionary, with boundedness conditions on the coefficients, or when  $\Sigma$  is Lipschitz-parameterized (we say that an approximation family  $(\Sigma_M)_{M \in \mathbb{N}}$  is Lipschitz-parameterized if there is a sequence  $(B_M)_{M \in \mathbb{N}}$  of subsets of finite dimensional spaces and a sequence  $(\varphi_M)_{M \in \mathbb{N}}$  of Lipschitz maps such that  $\Sigma_M = \varphi_M(B_M)$  for every  $M \in \mathbb{N}$ ), which is the case for ReLU neural networks for which we identify "simple" sufficient conditions on the considered architectures for  $\infty$ -encodability to hold, see section VII-C. Another consequence is that  $\infty$ -encodable approximation families  $\Sigma$  defined with ReLU neural networks share a common upper bound on approximation rates with other classical approximation families that we prove to be  $\infty$ -encodable. In particular, given  $\mathcal{C}$ , if an  $\infty$ -encodable sequence  $\Sigma$  is known such that  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = \gamma^{*\text{encod}}(\mathcal{C})$  (examples of such situations can be found in [8, Table 1]), then no improved approximation rate using ReLU networks can be hoped for.

**Organization of the paper.** We recall the definition of feedforward ReLU neural network in section II-A. In section II-B, we describe the  $L^p$  spaces in which approximation is considered. Bounds on the Lipschitz constants of the map that associates the parameters  $\theta$  of ReLU networks to their realization  $R_\theta \in L^p$  are given in section III. The error of nearest-neighbour uniform quantization for ReLU networks is discussed in section IV. Approximation speeds of quantized ReLU networks are established in section V. The notion of  $\infty$ -encodability is introduced in section VI, before discussing its consequences on the relation between the approximation speed and the encoding speed. Examples of  $\infty$ -encodable sequences (e.g., defined with dictionaries or ReLU networks) are then given in section VII. Sections VI and VII are essentially independent of the others. We give some perspectives in section VIII. Some useful definitions, technical results and their proofs are gathered in the appendices.

<sup>4</sup> $M \mapsto \text{poly}(M)$  denotes a positive function that grows at most polynomially in  $M$ .

## II. PRELIMINARIES

We recall the definition of ReLU neural networks, and characterize  $L^p$  spaces containing their realizations.

### A. ReLU neural networks

A ReLU neural network is a parametric description of the alternate composition of affine maps between finite-dimensional spaces and of a non-linear function. The non-linearity consists of the so-called Rectified Linear Unit (ReLU) applied coordinate-wise.

**Definition II.1 (ReLU: Rectified Linear Unit).** — The ReLU function  $\rho$  is defined by [1]:

$$\forall x \in \mathbb{R}, \rho(x) := \max(0, x).$$

For  $d \in \mathbb{N}$ , its  $d$ -dimensional version consists of applying it coordinate-wise:

$$\forall x \in \mathbb{R}^d, \rho(x) := (\rho(x_i))_{i=1\dots d}.$$

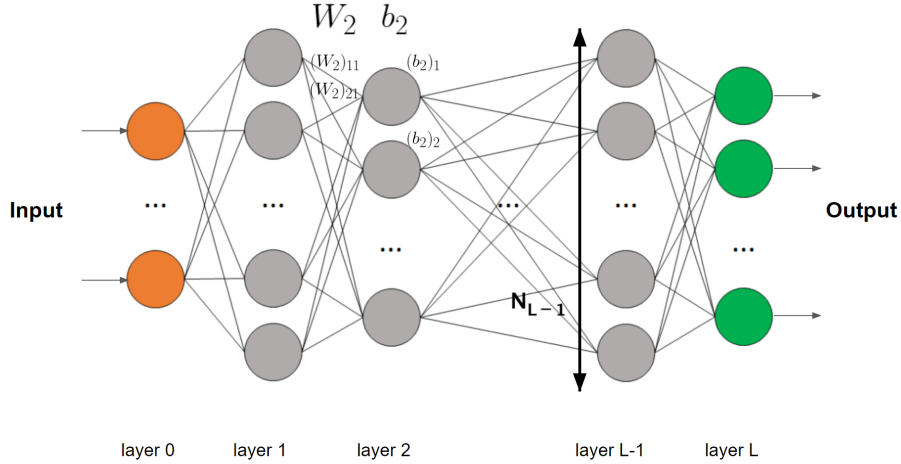


Figure 1. A neural network architecture can be seen as a directed graph. Neurons are represented by vertices, grouped by layers. For each neuron, there are edges going from this neuron to each neuron of the following layer. Coefficient  $(i, j)$  of  $W_\ell$  can be seen as the weight of the edge going from neuron  $j$  of layer  $\ell - 1$  to neuron  $i$  of layer  $\ell$ . Coefficient  $i$  of  $b_\ell$  can be seen as the weight of neuron  $i$  of layer  $\ell$ .

**Definition II.2 (Architecture of a neural network).** — An architecture of a neural network consists of a tuple  $(L, \mathbf{N})$ , with  $L \in \mathbb{N}$  and  $\mathbf{N} = (N_0, \dots, N_L) \in \mathbb{N}^{L+1}$ . We then say that  $L$  is the depth of the network. A network with such an architecture has  $L + 1$  layers of neurons, indexed from  $\ell = 0$  to  $\ell = L$ . Layer  $\ell$  has  $N_\ell$  neurons, we call  $N_\ell$  the width of layer  $\ell$ . The width of the network is  $W := \max_{\ell=0, \dots, L} N_\ell$ . Layer 0 is the input layer while layer  $L$  is the output layer.

An architecture can be represented as a graph, with a vertex for each neuron, and an edge between every pair of neurons within consecutive layers, see Figure 1 (thus, in this work, a layer consists of a set of neurons, not a set of edges).

**Definition II.3 (Parameters associated to a network architecture).** — Let  $(L, \mathbf{N})$  be an architecture. A parameter associated to this architecture consists of a vector  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$ , with  $W_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$  and  $b_\ell \in \mathbb{R}^{N_\ell}$ . Such a parameter  $\theta$  lives in the parameter space

$$\begin{aligned} \Theta_{L, \mathbf{N}} &:= \mathbb{R}^{d_{(L, \mathbf{N})}}, \\ d_{(L, \mathbf{N})} &:= \sum_{\ell=1}^L N_\ell(N_{\ell-1} + 1). \end{aligned} \tag{4}$$

A parameter  $\theta$  can be represented graphically: if neurons on layer  $\ell$  are numbered from 1 to  $N_\ell$ , then  $(W_\ell)_{i,j}$  is the weight on the edge going from neuron  $j$  of layer  $\ell - 1$  to neuron  $i$  of layer  $\ell$ , while  $(b_\ell)_i$  is the weight on neuron  $i$  of layer  $\ell$ , see Figure 1.

**Definition II.4 (ReLU neural network and its realization).** — A ReLU neural network consists of an architecture  $(L, \mathbf{N})$  and an associated parameter  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$ . Its realization is the function denoted  $R_\theta : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_L}$ , given by:

$$\forall x \in \mathbb{R}^{N_0}, R_\theta(x) := \tilde{y}_L(x)$$

with functions  $y_\ell$  and  $\tilde{y}_\ell$  defined by induction on  $\ell = 1, \dots, L$ :

$$\begin{aligned} y_0(x) &= x, \\ \tilde{y}_{\ell+1}(x) &= W_{\ell+1}y_\ell + b_{\ell+1}, \\ y_{\ell+1}(x) &= \rho(\tilde{y}_{\ell+1}(x)). \end{aligned}$$

In words, the input  $x$  goes through each layer sequentially, and when it goes from layer  $\ell$  to  $\ell + 1$ , it first goes through an affine transformation, of linear part  $W_{\ell+1}$  and constant part  $b_{\ell+1}$ , then it goes through the ReLU function  $\rho$  applied coordinate-wise (except on the last layer where the ReLU function is not applied).

### B. Considered functional approximation setting

We consider  $L^p$  spaces that contain all functions realized by ReLU networks (or equivalently all piecewise affine functions). We record the characterization of such spaces in Lemma II.1 (these are the  $L^p$  spaces for which (5) holds true) since we could not find it stated elsewhere. First let us introduce our notations for  $L^p$  spaces. Let  $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$  be input and output dimensions,  $p \in [1, \infty]$  be an exponent,  $\Omega \subset \mathbb{R}^{d_{\text{in}}}$  be the input domain and  $\mu$  be a (non-negative) measure on  $\Omega$ . Given a norm  $\|\cdot\|$  on  $\mathbb{R}^{d_{\text{out}}}$ , we define for every measurable function  $f : \Omega \rightarrow \mathbb{R}^{d_{\text{out}}}$ :

$$\|f\|_{p, \|\cdot\|} := \begin{cases} \left( \int_{x \in \Omega} \|f(x)\|^p d\mu(x) \right)^{\frac{1}{p}} & \text{if } p < \infty, \\ \text{ess sup}_{x \in \Omega} \|f(x)\| & \text{if } p = \infty. \end{cases}$$

We consider approximation in the space  $L^p(\Omega \rightarrow (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$  consisting of all measurable functions  $f$  from  $\Omega$  to  $\mathbb{R}^{d_{\text{out}}}$  such that  $\|f\|_{p, \|\cdot\|} < \infty$ , quotiented by the relation “being equal almost everywhere”. This is a Banach space with respect to the norm  $\|\cdot\|_{p, \|\cdot\|}$ . By the equivalence of norms in  $\mathbb{R}^{d_{\text{out}}}$ , this Banach space



is independent of the choice of norm  $\|\cdot\|$  on  $\mathbb{R}^{d_{\text{out}}}$ , and (for a given  $p$ ) all norms  $\|\cdot\|_{p,\|\cdot\|}$  are equivalent. In light of this fact we will simply denote it  $L^p(\Omega \rightarrow \mathbb{R}^{d_{\text{out}}}, \mu)$ , or even abbreviate it as  $L^p$ . We also denote  $\|\cdot\|_p := \|\cdot\|_{p,\|\cdot\|_\infty}$ . We will stress the dependence on the norm  $\|\cdot\|$  when it plays a role, such as in Theorem III.1.

We now state a necessary and sufficient condition on  $\Omega \subset \mathbb{R}^{d_{\text{in}}}$  and  $\mu$  so that all functions realized by a ReLU neural network with input dimension  $d_{\text{in}}$  and output dimension  $d_{\text{out}}$  are in  $L^p(\Omega \rightarrow \mathbb{R}^{d_{\text{out}}}, \mu)$ . The proof can be found in appendix B.

**Lemma II.1.** — Consider an exponent  $p \in [1, \infty]$ , a dimension  $d_{\text{in}}$ , a domain  $\Omega \subset \mathbb{R}^{d_{\text{in}}}$ , and a measure  $\mu$  on  $\Omega$ . Define

$$C_p(\Omega, \mu) := \begin{cases} \left( \int_{x \in \Omega} (\|x\|_\infty + 1)^p d\mu(x) \right)^{1/p} & \text{if } p < \infty, \\ \text{ess sup}_{x \in \Omega} \|x\|_\infty & \text{if } p = \infty. \end{cases}$$

The condition

$$C_p(\Omega, \mu) < \infty \tag{5}$$

is equivalent to: for every architecture  $(L, \mathbf{N})$  with  $N_0 = d_{\text{in}}$  the realizations of ReLU networks satisfy:

$$\forall \theta \in \Theta_{L, \mathbf{N}}, R_\theta \in L^p(\Omega \rightarrow \mathbb{R}^{N_L}, \mu),$$

where  $\Theta_{L, \mathbf{N}}$  is defined in Equation (4) and  $N_L$  is the width of the output layer (Definition II.2).

Note that the condition  $C_p(\Omega, \mu) < \infty$  holds in particular for every  $p \in [1, \infty]$  when the input domain is bounded and  $\mu$  is the Lebesgue measure.

### III. LIPSCHITZ PARAMETERIZATION OF RELU NEURAL NETWORKS

It is known that some sets of functions realized by ReLU networks are Lipschitz-parameterized<sup>5</sup> [5, Rmk. 9.1]. In Theorem III.1, we give lower- and upper-bounds on the Lipschitz constant depending on the depth, the width and the weight's magnitude of the considered networks. To the best of our knowledge, the lower-bound is new, while the upper-bound generalizes similar upper-bounds established in specific cases [2, Thm 2.6][13, Lem. 2] as discussed below. These bounds will be useful in the next sections to understand how quantization of ReLU networks harms approximation error.

**Definition III.1.** — (Parameter set  $\Theta_{L, \mathbf{N}}^q(r)$ ) Given an architecture  $(L, \mathbf{N})$  and the set of associated parameters  $\Theta_{L, \mathbf{N}}$  (see Equation (4)) we define for each  $r \geq 0$  and  $q \in [1, \infty]$  (the notation  $\|\cdot\|$  refers to the operator norm and is defined in appendix A):

$$\Theta_{L, \mathbf{N}}^q(r) := \{\theta = (W_1, \dots, W_L, b_1, \dots, b_L) \in \Theta_{L, \mathbf{N}} : \|W_\ell\|_q, \|b_\ell\|_q \leq r, \ell = 1, \dots, L\}.$$

<sup>5</sup>A set is Lipschitz-parameterized if it is the image by a Lipschitz map of a subset of a finite dimensional space.

Note that what will play a crucial role in what follows is the Lipschitz constant of the functions realized by the parameters in  $\Theta_{L,\mathbf{N}}^q(r)$ . This Lipschitz constant is bounded by  $r^L$  in the setup of Definition III.1. We do not enforce directly a global constraint on the Lipschitz constant since, to the best of our knowledge, there is no better practical way to enforce this constraint than by enforcing each  $W_\ell$  and  $b_\ell$  to have small norms. A more realistic situation thus corresponds to parameters  $\theta$  with each  $W_\ell$  and  $b_\ell$  bounded for some norms, which is what reflects the definition of the set of parameters  $\Theta_{L,\mathbf{N}}^q(r)$  in Definition III.1.

**Remark III.1.** — Instead of constraints on the operator norms, we may encounter constraints on the Frobenius or the max-norm. Let  $r \geq 0$ , and let  $(L, \mathbf{N})$  be an architecture. Define by  $W := \max_{\ell=0,\dots,L} N_\ell$  the width of the network. Denote  $\|M\|_F = (\sum_{i,j} M_{i,j}^2)^{1/2}$  the Frobenius norm of a matrix  $M$  and  $\|M\|_{\max} = \max_{i,j} |M_{i,j}|$  the max-norm (to be distinguished from  $\|M\|_\infty$  the operator norm defined in appendix A), and define  $\Theta_{L,\mathbf{N}}^F(r)$  (resp.  $\Theta_{L,\mathbf{N}}^{\max}(r)$ ) the set of all  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L) \in \Theta_{L,\mathbf{N}}$  such that for every  $\ell = 1, \dots, L$ :

$$\max(\|W_\ell\|_F, \|b_\ell\|_2) \leq r \quad (\text{resp. } \max(\|W_\ell\|_{\max}, \|b_\ell\|_\infty) \leq r).$$

By standard results about equivalence of norms (see e.g., (19) in the appendix) it holds for every  $q \in [1, \infty]$ :

$$\Theta_{L,\mathbf{N}}^F(r) \subset \Theta_{L,\mathbf{N}}^2(r), \quad \Theta_{L,\mathbf{N}}^{\max}(r) \subset \Theta_{L,\mathbf{N}}^q(Wr) \subset \Theta_{L,\mathbf{N}}^{\max}(Wr).$$

Given an architecture  $(L, \mathbf{N})$ , we now give bounds on the Lipschitz constant of the map associating the parameters to their realization:  $\theta \in \Theta_{L,\mathbf{N}}^q(r) \mapsto R_\theta \in L^p$ . The proof is in appendix D.

**Theorem III.1.** — Consider  $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$ ,  $\Omega \subset \mathbb{R}^{d_{\text{in}}}$ ,  $\mu$  a measure on  $\Omega$  satisfying (5),  $\|\cdot\|$  a norm on  $\mathbb{R}^{d_{\text{out}}}$ ,  $p, q \in [1, \infty]$ , and the space  $\mathcal{F} := L^p(\Omega \rightarrow (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$ . Then there exists a constant  $c > 0$  such that for every architecture  $(L, \mathbf{N})$  with  $N_0 = d_{\text{in}}$  and  $N_L = d_{\text{out}}$ , and every  $r \geq 1$ , denoting by  $W := \max_{\ell=0,\dots,L} N_\ell$  the width of the architecture, the map  $\theta \in \Theta_{L,\mathbf{N}}^q(r) \mapsto R_\theta \in L^p$  for ReLU networks satisfies

$$\|R_\theta - R_{\theta'}\|_{p, \|\cdot\|} \leq cWL^2r^{L-1}\|\theta - \theta'\|_\infty \quad \text{for all } \theta, \theta' \in \Theta_{L,\mathbf{N}}^q(r). \quad (6)$$

In particular, with  $\mu$  the Lebesgue measure on  $\Omega = [-D, D]^d$  for some  $D > 0$ , this holds with:

- $c := Dd^{1/q} + 1$  if  $p = \infty$  and  $\|\cdot\| = \|\cdot\|_q$ ;
- $c := (D+1)(2D)^{d/p}$  if  $\|\cdot\| = \|\cdot\|_q = \|\cdot\|_\infty$ .

Conversely, if  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$  (where  $\mathbb{R}_+ := \mathbb{R}_{\geq 0}$ ),  $\|\cdot\| = \|\cdot\|_q$  and  $p = \infty$  then there exists a constant  $c' > 0$  independent of the architecture, such that, for every  $\varepsilon > 0$ , we can exhibit parameters  $\theta, \theta'$  satisfying

$$\|R_\theta - R_{\theta'}\|_{p, \|\cdot\|} \geq (1 - \varepsilon)c'Lr^{L-1}\|\theta - \theta'\|_\infty. \quad (7)$$

This converse result also holds for  $1 \leq p < \infty$  under the additional assumption that  $N_0 = \min_{0 \leq \ell \leq L} N_\ell$ , i.e., that all layers are at least as wide as the input layer.

It is an open question whether the extra factor  $WL$  in (6) compared to (7) can be improved, and whether the converse result for  $p < \infty$  also holds without the additional assumption. Note that the condition  $r \geq 1$

in Theorem III.1 is reasonable since the realization of every parameter  $\theta \in \Theta_{L,\mathbf{N}}^q(r)$  is a function  $R_\theta$  which is  $r^L$ -Lipschitz with respect to the  $q$ -norm on the input and output spaces. Constraining  $r < 1$  would lead to "very" flat functions, essentially constant, when  $L$  is large. Vice-versa, the stability of a concrete numerical implementation of a neural network probably requires it to have a Lipschitz constant somehow bounded by the format used to represent numbers. Such considerations would probably lead to consider  $r^L \leq C$  for some constant  $C$ , i.e.,  $1 \leq r \leq C^{1/L}$ .

Here is a list of immediate extensions of Theorem III.1:

- *Arbitrary Lipschitz activation:* Theorem III.1 can be extended to the case where the ReLU activation function is replaced by any Lipschitz activation function.
- *Pooling-operation:* Theorem III.1 does not change if we add standard (max- or average-) pooling operations between some layers since they are 1-Lipschitz.
- *Arbitrary  $s$ -norm on the parameters:* since for every exponent  $s \in [1, \infty]$ , it holds  $\|\cdot\|_\infty \leq \|\cdot\|_s$ , Theorem III.1 yields a bound on the Lipschitz constant with arbitrary  $s$ -norm on the parameter space.
- *Generalization error bound:* in the context of learning, for a loss  $\ell(\hat{y}, y)$  that is a Lipschitz function of  $\hat{y}$  with respect to some norm  $\|\cdot\|$  on the support of a distribution  $\mathbb{P}$ , the excess risk  $\mathbb{E}_{(x,y) \sim \mathbb{P}}(\ell(R_\theta(x), y) - \ell(R_{\theta'}(x), y))$  can be bounded from above by  $\mathbb{E}_{(x,y) \sim \mathbb{P}}(\|R_\theta(x) - R_{\theta'}(x)\|)$ , which in turn can be bounded using Theorem III.1. In particular, this is the case when  $\mathbb{P}$  is supported on a compact set and  $\ell(\hat{y}, y)$  is continuously differentiable in  $\hat{y}$ .
- *Skip connections and convolutional layers:* one can also exploit Theorem III.1 to networks with skip connections and/or convolutional layers, since they can be rewritten as networks with fully-connected layers. This rewriting can however artificially inflate the widths of the networks and is unlikely to give sharp bounds. It is left to further work whether an extension of Theorem III.1 with improved tailored bounds may be obtained in these settings.

**Remark III.2 (Related works).** — The fact that some sets of functions realized by ReLU neural networks are Lipschitz-parameterized is already known [5, Rmk. 9.1]. To our knowledge, the lower-bound in Theorem III.1 is new. However the upper-bound is already known in several specific situations: at least for  $d_{\text{out}} = 1$ ,  $L^\infty([0, 1]^{d_{\text{in}}})$  with the Lebesgue measure, and  $q = \max$  [2, Thm 2.6] as well as  $p = \infty$ ,  $q = F$ , and  $\|\cdot\| = \|\cdot\|_2$  [13, Lem. 2]. Theorem III.1 shows that this upper-bound holds true more generally for general constraints on the parameters (arbitrary  $q \in [1, \infty]$ ) and arbitrary  $p \in [1, \infty]$  and  $(\Omega, \mu)$  satisfying condition (5) i.e., in any  $L^p$  space that contains all the functions realized by ReLU neural networks. Let us also mention that Theorem III.1 is based on Lemma C.1 (appendix C), and this lemma is a straightforward generalization of a known inequality for  $q = \infty$  (see for instance [3, Eq. (3.12)] or [8, Eq. (37)]) to arbitrary  $q \in [1, \infty]$ . We prove that the inequality established in Lemma C.1 is optimal. To our knowledge, even in the case  $q = \infty$ , the optimality has not been discussed yet in the literature.

#### IV. NEAREST-NEIGHBOUR UNIFORM QUANTIZATION OF RELU NEURAL NETWORKS

In section IV-A, we characterize the error of nearest-neighbour uniform quantization of ReLU networks in  $L^\infty$ , recovering and improving Lemma VI.8 in [8]. In section IV-B, we show that controlling the error of nearest-neighbour uniform quantization schemes leads to recover existing results [7, Thm. 2] on function approximation by quantized ReLU networks.

##### A. Control of the $L^\infty$ quantization error $\|R_\theta - R_{Q(\theta)}\|_\infty$

The following lemma is a direct consequence of Theorem III.1.

**Lemma IV.1.** — Consider a domain  $[-D, D]^d$ . Fix an architecture  $(L, \mathbf{N}) = (L, (N_0, \dots, N_L))$  with width  $W := \max_{\ell=0, \dots, L} N_\ell$ , a bound  $r \geq 1$  on the norm of the parameters, and an exponent  $q \in [1, \infty]$ . Given  $\eta > 0$ , let  $Q : \Theta_{L, \mathbf{N}} \rightarrow \Theta_{L, \mathbf{N}}$  (recall that  $\Theta_{L, \mathbf{N}}$  is the set of parameters associated with the architecture  $(L, \mathbf{N})$ , see Definition II.2) be such that  $\|Q(\theta) - \theta\|_\infty \leq \eta$  for every parameter  $\theta \in \Theta_{L, \mathbf{N}}^q(r)$ . Consider a subset  $\Theta \subset \Theta_{L, \mathbf{N}}^q(r)$ . Let  $r' \geq 1$ . Assume that:

$$Q(\theta) \in \Theta_{L, \mathbf{N}}^q(r'), \quad \forall \theta \in \Theta. \quad (8)$$

Consider  $\varepsilon > 0$ , and  $0 < \eta \leq \varepsilon (cWL^2(r')^{L-1})^{-1}$ , where  $c := Dd^{1/q} + 1$ . Then, it holds:

$$\max_{\theta \in \Theta} \max_{x \in [-D, D]^d} \|R_\theta(x) - R_{Q(\theta)}(x)\|_q \leq \varepsilon. \quad (9)$$

*Proof of Lemma IV.1.* Fix  $\theta \in \Theta$ . Under assumption (8), it holds  $Q(\theta) \in \Theta_{L, \mathbf{N}}^q(r')$ . This means that we can apply Theorem III.1 with  $p = \infty$ ,  $\|\cdot\| = \|\cdot\|_q$  and with the specific constant  $c = Dd^{1/q} + 1$ . In this situation the essential supremum over  $x \in [-D, D]^d$  in Theorem III.1 is actually a maximum. This yields (9) when  $0 < \eta \leq \varepsilon (cWL^2(r')^{L-1})^{-1}$ .  $\square$

Under mild assumptions on the error  $\varepsilon$ , Property (8) holds for  $r' = 2r$ . This leads to the following theorem.

**Theorem IV.1.** — In the same setting as in Lemma IV.1, consider  $0 < \varepsilon < cL^2(2r)^{L-1}$ . If  $0 < \eta \leq \varepsilon (cWL^2(2r)^{L-1})^{-1}$ , then (9) holds true.

*Proof of Theorem IV.1.* Fix  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L) \in \Theta_{L, \mathbf{N}}^q(r)$ . Assume that  $0 < \varepsilon < cL^2(2r)^{L-1}$  and  $0 < \eta \leq \varepsilon (cWL^2(2r)^{L-1})^{-1}$ . We want to prove that  $Q(\theta) \in \Theta_{L, \mathbf{N}}^q(2r)$ . By assumption on  $\eta$  and  $\varepsilon$ ,  $0 < \eta \leq \varepsilon (cWL^2(2r)^{L-1})^{-1} \leq 1/W \leq r/W$ . Note that for a matrix  $M$  with input and output dimensions bounded by  $W$ , it holds  $\|M\|_q \leq W\|M\|_{\max}$ , see (19). This guarantees that for every layer  $\ell = 1, \dots, L$ , it holds  $\|W_\ell - Q(W_\ell)\|_q \leq W\|W_\ell - Q(W_\ell)\|_{\max} \leq W\eta \leq r$  and  $\|b_\ell - Q(b_\ell)\|_q \leq W^{1/q}\|b_\ell - Q(b_\ell)\|_\infty \leq W\eta \leq r$  so that by the triangle inequality  $Q(\theta) \in \Theta_{L, \mathbf{N}}^q(2r)$ . Then, (9) follows from Lemma IV.1.  $\square$

When  $Q(x) := Q_\eta(x) := \lfloor x/\eta \rfloor \eta$ , we now establish a *necessary* condition for (9) to hold, that almost matches the sufficient condition of Theorem IV.1. This is obtained thanks to the almost matching lower- and upper-bounds of Theorem III.1. The proof is in appendix E.

**Theorem IV.2.** — In the same setting as in Lemma IV.1, consider the function  $Q := Q_\eta$  that acts coordinatewise on vectors and such that for every  $x \in \mathbb{R}$ ,  $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$ . Define  $N_{\min} := \min_{0 \leq \ell \leq L} N_\ell$  and  $c' := DN_{\min}^{1/q}$ . If  $\varepsilon, \eta > 0$  are such that (9) holds true for  $\Theta := \Theta_{L, \mathbf{N}}^q(r)$ , then  $\min(r, \eta) \leq \frac{\varepsilon}{c'r^{L-1}}$ . In particular, if  $\varepsilon < c'r^L$  then  $\eta \leq \frac{\varepsilon}{c'r^{L-1}}$ .

Note that Theorem IV.1 can be applied for every  $\varepsilon \in (0, 1)$  since  $cL^2(2r)^{L-1} > 1$ . Similarly, if the domain  $[-D, D]^d$  is large enough ( $D \geq 1$ ) then  $c'r^L > 1$  and Theorem IV.2 yields that whenever (9) holds true for some  $\varepsilon \in (0, 1)$  and  $\eta > 0$  we must have  $\eta \leq \frac{\varepsilon}{c'r^{L-1}}$ .

**Remark IV.1.** — With  $\eta > 0$  and  $Q_\eta$  the function from Theorem IV.2, the number of bits needed to store one coordinate of  $Q_\eta(\theta)$  is proportional to  $\ln(1/\eta)$ . We just saw that if  $D \geq 1$  and (9) is satisfied with  $\varepsilon \in (0, 1)$ ,  $\eta > 0$ , then  $\eta$  must be exponentially small in  $L$  (as soon as  $r > 1$ ). This means that the number of bits per coordinate must at least grow linearly with the network depth  $L$  to ensure that the worst-case quantization error over networks in  $\Theta_{L, \mathbf{N}}^q(r)$  is controlled. This is essentially due to the fact that there are realizations of parameters in  $\Theta_{L, \mathbf{N}}^q(r)$  that are functions with Lipschitz constant *equal* to  $r^L$ . More optimistic bounds can be envisioned under stronger assumptions on the set of parameters or on the network's architecture.

Another direct consequence of Lemma IV.1 is the following proposition, which is proved in appendix E and yields an improvement of Lemma VI.8 in [8].

**Proposition IV.1 (extension of [8, Lem. VI.8]).** — Consider  $(L, \mathbf{N})$  an architecture with input dimension  $d_{\text{in}}$ , output dimension  $d_{\text{out}}$  and  $L \geq 2$  layers. Consider the space  $\mathcal{F} = L^\infty([-D, D]^{d_{\text{in}}}) \rightarrow (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|_\infty, \mu)$  with  $\mu$  the Lebesgue measure.

Consider  $\varepsilon \in (0, 1/2)$  and  $\theta \in \Theta_{L, \mathbf{N}}$ . Denote  $W = \max_{\ell=0, \dots, L} N_\ell$  the width of the architecture  $(L, \mathbf{N})$ . Let  $k \geq 0$  be the smallest integer such that  $\theta \in \Theta_{L, \mathbf{N}}^{\max}(\varepsilon^{-k})$  and  $\max(W, L) \leq \varepsilon^{-k}$ , i.e.,  $k = \lceil \log_2 \max(\|\theta\|_\infty, W, L) / \log_2(1/\varepsilon) \rceil$ . For every integer  $m \geq 2kL + k + 1 + \log_2(\lceil D \rceil)$ , the weights of  $\theta$  can be rounded up to a closest point in  $\eta\mathbb{Z} \cap [-\varepsilon^{-k}, \varepsilon^{-k}]$  with  $\eta := 2^{-m \lceil \log_2(\varepsilon^{-1}) \rceil} \leq \varepsilon^m$  to obtain  $Q_\eta(\theta) \in \Theta_{L, \mathbf{N}}^{\max}(\varepsilon^{-k}) \cap (\eta\mathbb{Z})^{d(L, \mathbf{N})}$  that satisfies:

$$\|R_\theta - R_{Q_\eta(\theta)}\|_\infty \leq \varepsilon.$$

Let us check that Proposition IV.1 indeed implies the result of Elbrächter et al.<sup>6</sup> [8, Lem. VI.8]. First, since  $\max(W, L) \geq L \geq 2$ , it holds  $k \geq 1$ . Thus, for  $L \geq 2$ , we have  $k(L-1) \geq 1$  so that  $3kL \geq 2kL + k + 1$  and it is thus sufficient to take  $m \geq 3kL + \log_2(\lceil D \rceil)$  (which is the sufficient condition given in [8, Lem. VI.8]). Note however the improved (slower) growth of  $m$  with  $L$  in the sufficient condition of Proposition IV.1 compared to [8, Lem. VI.8].

<sup>6</sup>Lemma VI.8 in [8] is stated for networks having at most  $\varepsilon^{-k}$  non-zero weights. Given such a network, we can always remove neurons having only zero incoming and outgoing weights. This gives another network, with the same realization, but with a width  $W \leq \varepsilon^{-k}$  and a depth  $L \leq \varepsilon^{-k}$ . Then, Proposition IV.1 applies to this new network. (R1-m20)

**Remark IV.2.** — More generally, given bounds on the sparsity (i.e., number of nonzero entries), on the magnitude of the network weights, and an *arbitrary*  $p \in [1, \infty]$ , Theorem III.1 can be used to find an appropriate step size that guarantees that a uniform quantization of the considered network is within error  $\varepsilon > 0$  in  $L^p$ .

*B. Control of the approximation error of a function by quantized networks*  $\|f - R_{Q(\theta)}\|_\infty$

Given a function  $f$ , parameters  $\theta$ , and a quantization function  $Q$ , a simple triangle inequality yields  $\|f - R_{Q(\theta)}\|_\infty \leq \|f - R_\theta\|_\infty + \|R_\theta - R_{Q(\theta)}\|_\infty$ . Theorem III.1 controls the quantization error  $\|R_\theta - R_{Q(\theta)}\|_\infty$  for nearest-neighbour uniform quantization schemes. If, in addition, information about the approximation error  $\|f - R_\theta\|_\infty$  is available, then we can deduce a bound on the approximation error of  $f$  by quantized networks.

We apply this simple observation in the case of functions  $f$  in an  $L^\infty$ -Sobolev space to recover a special case of Theorem 2 in [7] (the other cases can be recovered by combining this special case with Proposition 3 in [7]). The proof is in appendix E.

Let  $n \in \mathbb{N}$  and consider  $\mathcal{W}^{n,\infty}([0, 1]^d)$ , the Sobolev space of real-valued functions on  $[0, 1]^d$  that are in  $L^\infty$  as well as their weak derivatives up to order  $n$  (given  $\mathbf{n} := (n_1, \dots, n_d) \in \mathbb{N}^d$ , the associated weak-derivative of a function  $f$  is denoted  $D^{\mathbf{n}}f$  if it exists). The norm on  $\mathcal{W}^{n,\infty}([0, 1]^d)$  is given by:

$$\|f\|_{\mathcal{W}^{n,\infty}([0,1]^d)} := \max_{\substack{\mathbf{n} := (n_1, \dots, n_d) \in \mathbb{N}^d \\ \sum_i n_i \leq n}} \operatorname{ess\,sup}_{x \in [0,1]^d} |D^{\mathbf{n}}f(x)|.$$

**Proposition IV.2** ([7, Thm. 2]). — Let  $\mathcal{C}_{n,d}$  be the unit ball of  $\mathcal{W}^{n,\infty}([0, 1]^d)$ . There exists a constant  $c > 0$  depending only on  $n$  and  $d$  such that for every  $\varepsilon \in (0, 1)$ , there exists  $\eta > 0$  satisfying  $\ln(1/\eta) \leq c \ln^2(1/\varepsilon)$  and a ReLU network architecture that can approximate every function  $f \in \mathcal{C}_{n,d}$  within error  $\varepsilon > 0$  in  $L^\infty([0, 1]^d)$  using weights in  $\eta\mathbb{Z}$ , with depth bounded by  $c \ln(1/\varepsilon)$ , a number of weights at most equal to  $c\varepsilon^{-d/n} \ln(1/\varepsilon)$ , and with a total number of bits (used to store the network weights) bounded by  $c\varepsilon^{-d/n} \ln^3(1/\varepsilon)$ .

**Remark IV.3.** — Compared to Theorem 2 of [7], Theorem III.1 can also be used to establish similar results, not only for a function  $f$  in the unit ball of an  $L^\infty$ -Sobolev space, but for every  $f \in L^p$  ( $1 \leq p \leq \infty$ ) as soon as it is known how to approximate  $f$  with unquantized ReLU networks, with explicit bounds on the growth of their depth, width and weight's magnitude. For instance, such bounds are known for Hölder spaces [14], classifier functions in  $L^2$  [15] and Besov spaces [16]. The same argument also applies for networks with arbitrary Lipschitz activation (such as the sigmoid function) for which an analog of Theorem III.1 can be derived, and for which we know how to approximate "smooth" functions [9, Table 1].

## V. APPROXIMATION SPEEDS OF QUANTIZED VS. UNQUANTIZED RELU NEURAL NETWORKS

Consider a function  $f$  and a sequence of parameters  $(\theta_M)_{M \in \mathbb{N}}$  such that  $\|f - R_{\theta_M}\|_p$  goes to zero as  $M$  goes to infinity. Given a nearest-neighbour uniform quantization scheme  $Q_M$  with a step size that depends on  $M$ , we saw in the previous sections how to control the quantization error  $\|R_{\theta_M} - R_{Q(\theta_M)}\|_p$ . We used the latter to control  $\|f - R_{Q(\theta_M)}\|_p \leq \|f - R_{\theta_M}\|_p + \|R_{\theta_M} - R_{Q(\theta_M)}\|_p$ , in specific situations such as when

$f$  is in an  $L^\infty$ -Sobolev space, see section IV-B. In this section, we give sufficient conditions on the step size used for quantization with  $Q_M$  to guarantee that the quantization error  $\|R_{\theta_M} - R_{Q(\theta_M)}\|_p$  decreases, with  $M$ , at the same asymptotic polynomial rate as the approximation error  $\|f - R_{\theta_M}\|_p$ . This leads to an explicit sufficient number of bits, depending on the growth with  $M$  of the architecture of the parameters  $\theta_M$ , that guarantees that quantized ReLU networks have the same approximation speeds as unquantized ones, see Example V.1. First, we define the considered approximation families and their uniformly quantized version. Notations are somewhat cumbersome but necessary to introduce the different types of constraints on the considered architectures (depth and width constraints) and on the parameters (sparsity and norm constraints).

**Definition V.1 (Sequence of sets of architectures).** — Consider  $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$  and  $(L_M)_{M \in \mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ . For each  $M \in \mathbb{N}$  define  $\mathbf{A}_M$ , the set of architectures with input dimension  $d_{\text{in}}$ , output dimension  $d_{\text{out}}$ , depth bounded by  $L_M$  and widths of the hidden layers bounded by  $M$ :

$$\mathbf{A}_M := \{(L, (N_0, \dots, N_L)) : L, N_0, \dots, N_L \in \mathbb{N}, L \leq L_M, N_0 = d_{\text{in}}, N_L = d_{\text{out}}, N_\ell \leq M, \ell = 1, \dots, L-1\}.$$

For every  $M \in \mathbb{N}$  and every architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$ , define  $S_{(L, \mathbf{N})}^M$  as the set of all supports  $S \subset \{0, 1\}^{d(L, \mathbf{N})}$  of cardinality at most  $M$ , used to constrain the non-zero entries of a vector  $\theta$  with architecture  $(L, \mathbf{N})$ .

The width constraint  $N_\ell \leq M$  in the definition of the architectures in  $\mathbf{A}_M$  is written for clarity but is superfluous in what follows, given that the realization of a network  $\theta$  (with arbitrary activation function and an architecture of arbitrary width) with at most  $M$  nonzero coefficients can always be written as the realization of a parameter  $\theta'$  on a “pruned” architecture where every hidden layer has width  $N_\ell \leq M$ .

**Definition V.2 (ReLU networks approximation family).** — Consider a sequence  $(\mathbf{A}_M)_{M \in \mathbb{N}}$  of sets of architectures as in Definition V.1. Consider  $L^p(\Omega \rightarrow (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$  satisfying (5),  $q \in [1, \infty] \cup \{F, \max\}$  ( $F$  and  $\max$  refer to the Frobenius norm and the max-norm, see Remark III.1), and a sequence  $(r_M)_{M \in \mathbb{N}}$  of real numbers such that  $r_M \geq 1$ . Define the approximation family  $\mathcal{N} := (\mathcal{N}_M)_{M \in \mathbb{N}}$  of sets  $\mathcal{N}_M \subset L^p(\Omega \rightarrow (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$  of realizations of ReLU neural networks with an architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$  and parameters in  $\Theta_{L, \mathbf{N}}^q(r_M)$ :

$$\mathcal{N}_M := \bigcup_{(L, \mathbf{N}) \in \mathbf{A}_M} \bigcup_{S \in S_{(L, \mathbf{N})}^M} R_{\Theta_{L, \mathbf{N}}^q(r_M), S}$$

where for any parameter set  $\Theta$  and support  $S$  we denote  $R_{\Theta, S} := \{R_\theta, \theta \in \Theta \text{ supported on } S\}$ .

**Definition V.3 (Quantization of parameters on a uniform grid).** — Given a set of parameters  $\Theta$  associated to an architecture  $(L, \mathbf{N})$ , we define, for every  $\eta > 0$  and  $r \in (0, \infty]$ , the quantized version  $\mathcal{Q}(\Theta, \eta, r)$  of  $\Theta$  on a bounded uniform grid  $(\eta\mathbb{Z} \cap [-r, r])^{d(L, \mathbf{N})}$ :

$$\mathcal{Q}(\Theta, \eta, r) := \Theta \cap (\eta\mathbb{Z} \cap [-r, r])^{d(L, \mathbf{N})}.$$

Definition V.3 will essentially be used when  $\Theta$  is a ball. In that case,  $\mathcal{Q}(\Theta, \eta, r)$  is not empty provided that  $r$  and  $\Theta$  are sufficiently large compared to the step size  $\eta$ .

**Definition V.4 (Quantized version of Definition V.2).** — Consider an approximation family  $\mathcal{N} := (\mathcal{N}_M)_{M \in \mathbb{N}}$  as in Definition V.2, with the associated sequences  $(r_M)_{M \in \mathbb{N}}$ ,  $(L_M)_{M \in \mathbb{N}}$  and  $q \in [1, \infty] \cup \{F, \max\}$ . For every  $M \in \mathbb{N}$ , define

$$\text{Lips}(M, q) := \begin{cases} \max(d_{\text{in}}, d_{\text{out}}, M) L_M^2 r_M^{L_M-1} & \text{if } q \in [1, \infty] \cup \{F\}, \\ \text{Lips}(M, 2) \max(d_{\text{in}}, d_{\text{out}}, M)^{L_M-1} & \text{for } q = \max, \end{cases}$$

and observe that  $\text{Lips}(M, q) \geq 1$ . Given any  $\gamma > 0$  define for every  $M \in \mathbb{N}$  the step size  $\eta_M = \eta_M(\gamma, q) := (M^\gamma \text{Lips}(M, q))^{-1}$ . The  $\gamma$ -uniformly quantized version  $\mathcal{Q}(\mathcal{N}|\gamma) := (\mathcal{Q}_M(\mathcal{N}_M|\gamma))_{M \in \mathbb{N}}$  of  $\mathcal{N}$  is

$$\mathcal{Q}_M(\mathcal{N}_M|\gamma) = \bigcup_{(L, \mathbf{N}) \in \mathbf{A}_M} \bigcup_{S \in S_{(L, \mathbf{N})}^M} R_{\mathcal{Q}(\Theta_{L, \mathbf{N}}^q(r_M), \eta_M, r_M), S}$$

with  $\mathbf{A}_M$  and  $S_{(L, \mathbf{N})}^M$  the families of architectures and supports (cf Definition V.1) associated to  $\mathcal{N}_M$ , see Definition V.2 where we recall that the notation  $R_{\Theta, S}$  given a support  $S$  is also introduced.

In general, Lipschitz-parameterized approximation families can be uniformly quantized into sequences having comparable approximations speeds, if a step size sufficiently small is chosen. Theorem V.1 deals only with the case of Lipschitz-parameterized approximation families we are interested in: ReLU neural networks. The upper-bound on the Lipschitz constant established in Theorem III.1 yields explicit conditions on the growth of the depth and the weight's magnitude, that guarantee that the  $\gamma$ -uniformly quantized sequence  $\mathcal{Q}(\mathcal{N}|\gamma)$  has, on every set  $\mathcal{C} \subset L^p$ , an approximation speed which is comparable to its unquantized version  $\mathcal{N}$ . The proof of Theorem V.1 is in appendix F.

**Theorem V.1.** — Consider the context of Definition V.4. Then, for every  $\gamma > 0$ , the  $\gamma$ -uniformly quantized sequence  $\mathcal{Q}(\mathcal{N}|\gamma) := (\mathcal{Q}_M(\mathcal{N}_M|\gamma))_{M \in \mathbb{N}}$  has, on every (non-empty) set  $\mathcal{C} \subset L^p$ , an approximation speed comparable to the unquantized one  $\mathcal{N}$ :

$$\begin{aligned} \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{Q}(\mathcal{N}|\gamma)) &= \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N}) & \text{if } \gamma \geq \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N}), \\ \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{Q}(\mathcal{N}|\gamma)) &\geq \gamma & \text{otherwise.} \end{aligned} \tag{10}$$

We will see in the following sections that the approximation speed  $\gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N})$  can be bounded from above by a quantity denoted  $\gamma^{*\text{encod}}(\mathcal{C})$ , the latter quantity being known for several classical sets  $\mathcal{C}$  (see [8, Table 1]). In such a situation, this guides the choice of  $\gamma$  to define a concrete  $\gamma$ -quantized sequence in the context of Theorem V.1. Indeed, considering  $\mathcal{C} \subset \mathcal{F}$  a classical function class for which the quantity  $\gamma^{*\text{encod}}(\mathcal{C})$  is known, choosing  $\gamma \geq \gamma^{*\text{encod}}(\mathcal{C})$  is sufficient to ensure that  $\gamma \geq \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N})$ . Vice-versa, among all such  $\gamma$ , choosing the smallest one  $\gamma = \gamma^{*\text{encod}}(\mathcal{C})$  is probably the best choice to yield the largest possible step sizes  $\eta_M$  and the best concrete compromise.

**Example V.1 (Comparable approximation speeds with controlled growth of the number of bits).**

— Let  $q \in [1, \infty] \cup \{F, \max\}$  be an exponent and  $\pi$  be a positive polynomial and consider  $\mathcal{N}_M^\pi$  the set of functions parameterized by a ReLU neural network with arbitrary architecture  $(L, \mathbf{N})$  with depth bounded by  $\pi(\log M)$ , with at most  $M$  non-zero parameters and with parameters in  $\Theta_{L, \mathbf{N}}^q(\pi(M))$ . For every  $\gamma > 0$ , there



exists a constant  $c(\gamma) > 0$  such that the  $\gamma$ -uniformly quantized sequence  $Q(\mathcal{N}_M^\pi | \gamma)$  of  $\mathcal{N}_M^\pi$  is obtained with step size  $\eta_M = \mathcal{O}_{M \rightarrow \infty}(M^{-c(\gamma) \log M})$ , i.e., using  $\mathcal{O}_{M \rightarrow \infty}((\log M)^2)$  bits per weight. Theorem V.1 guarantees that this quantized sequence still has approximation speeds comparable to  $\mathcal{N}_M^\pi$ . In the same setup, if we assume in addition that the depths  $L_M$  are uniformly bounded in  $M$ , then for every  $\gamma > 0$ , a step size  $\eta_M = \mathcal{O}_{M \rightarrow \infty}(M^{-c(\gamma)})$  (i.e.,  $\mathcal{O}_{M \rightarrow \infty}(\log M)$  bits per parameter) suffices to get comparable speeds as in Equation (10).

## VI. ENCODING SPEEDS VS APPROXIMATION SPEEDS

We now investigate a fundamental limitation of many approximations families (including ReLU networks): the approximation speed of a set by an approximation family cannot be greater than the encoding speed of this set (see (3)). Section VI and section VII are essentially independent from the others. We introduce an abstract property of approximation families, called "encodability", in Definition VI.2. In Theorem VI.1, we prove that every approximation family satisfying this encodability property must satisfy Inequality (3). As we will see in section VII, this lays a unified and generic framework that captures and recovers different known situations [8, Thm. V.3, Thm. VI.4][10, Thm. 5.24][12, Prop. 11] where (3) holds.

### A. The notion of $\gamma$ -encodability

Let  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  be a sequence of non-empty subsets of a metric space  $(\mathcal{F}, d)$ . Let  $\mathcal{C} \subset \mathcal{F}$  and  $\varepsilon > 0$ . If  $\gamma^{*\text{approx}}(\mathcal{C} | \Sigma) > 0$ , since  $\Sigma$  approximates  $\mathcal{C}$  at speed  $\gamma^{*\text{approx}}(\mathcal{C} | \Sigma)$ , there exists a positive integer  $M$  large enough such that every element  $f \in \mathcal{C}$  can be  $\varepsilon$ -approximated (with respect to the metric  $d$ ) by an element of  $\Sigma_M$ . Since  $\Sigma_M$  can be  $\varepsilon$ -covered (with respect to  $d$ ) with  $N(\Sigma_M, d, \varepsilon)$  elements,  $\mathcal{C}$  can be  $2\varepsilon$ -covered with  $N(\Sigma_M, d, \varepsilon)$  elements. Instances of this simple reasoning can be found in [8, Thm. V.3, Thm. VI.4][10, Thm. 5.24][12, Prop. 11]. This suggests the existence of a relation between the approximation speed  $\gamma^{*\text{approx}}(\mathcal{C} | \Sigma)$  and the encoding speed  $\gamma^{*\text{encod}}(\mathcal{C})$  that depends on the growth with  $M$  of the covering numbers of  $\Sigma_M$ .

We claim that a "reasonable" growth of the covering numbers of  $\Sigma_M$  consists in a situation where, for some  $\gamma > 0$ , the set  $\Sigma_M$  can be  $M^{-\gamma}$ -covered with "roughly"  $2^{M \log M}$  elements. Indeed, this covers the case where each element of  $\Sigma_M$  can be described by  $M$  parameters that can be stored with a number of bits per parameter that grows logarithmically in  $M$ . For instance if  $\Sigma_M$  is a bounded set in dimension  $M$  then it can be uniformly quantized along each dimension with a size step of order  $M^{-\gamma}$ , so that  $\log M$  bits is roughly enough to encode each of the  $M$  coordinates. This "reasonable" growth for the covering numbers of  $\Sigma_M$  is formalized in Definition VI.2, and yields the simple relation  $\min(\gamma^{*\text{approx}}(\mathcal{C} | \Sigma), \gamma) \leq \gamma^{*\text{encod}}(\mathcal{C})$  for every set  $\mathcal{C} \subset \mathcal{F}$ , as shown in Theorem VI.1.

**Definition VI.1** ( $(\gamma, h)$ -encoding). — Let  $(\mathcal{F}, d)$  be a metric space. Let  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  be an arbitrary sequence of (non-empty) subsets of  $\mathcal{F}$ . Let  $\gamma > 0$  and  $h > 0$ . A sequence  $(\Sigma(\gamma, h)_M)_{M \in \mathbb{N}}$  is said to be a  $(\gamma, h)$ -encoding of  $\Sigma$  if there exist constants  $c_1, c_2 > 0$  such that for every  $M \in \mathbb{N}$ , the set  $\Sigma(\gamma, h)_M$  is a  $c_1 M^{-\gamma}$ -covering of  $\Sigma_M$  (recall Equation (1), in particular  $\Sigma(\gamma, h)_M$  must be a subset of  $\Sigma_M$ ) of size satisfying  $\log_2(|\Sigma(\gamma, h)_M|) \leq c_2 M^{1+h}$ .

The following definition captures a "reasonable" growth with  $M$  of the covering numbers of  $\Sigma_M$ .

**Definition VI.2** ( $\gamma$ -*encodable*  $\Sigma$  in  $(\mathcal{F}, d)$ ). — Let  $(\mathcal{F}, d)$  be a metric space. Let  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  be an arbitrary sequence of (by default, non-empty) subsets of  $\mathcal{F}$ . Let  $\gamma > 0$ . We say that  $\Sigma$  is  $\gamma$ -*encodable* in  $(\mathcal{F}, d)$  if for every  $h > 0$ , there exists a  $(\gamma, h)$ -encoding of  $\Sigma$ . We say that  $\Sigma$  is  $\infty$ -*encodable* in  $(\mathcal{F}, d)$  if it is  $\gamma$ -encodable in  $(\mathcal{F}, d)$  for all  $\gamma > 0$ . When the context is clear, we will omit the mention to  $(\mathcal{F}, d)$ .

Note that if  $\Sigma$  is  $\gamma$ -encodable then it is  $\gamma'$ -encodable for every  $\gamma' \leq \gamma$ . Several examples of  $\infty$ -encodable sequences are given in section VII, including classical approximation families defined with dictionaries or ReLU networks.

*B. The encoding speed as a universal upper bound for approximation speeds*

It is known that  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C})$  for various sets  $\mathcal{C}$  when  $\Sigma$  is defined with neural networks [8, Thm. VI.4] or dictionaries [8, Thm. V.3][10, Thm. 5.24]. The following proposition shows that  $\infty$ -encodability implies  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C})$ . This settles a unified and generalized framework for the aforementioned known cases that implicitly use, one way or another, the  $\infty$ -encodability property, as we will detail in section VII-B and section VII-C.

**Theorem VI.1.** — Consider  $(\mathcal{F}, d)$  a metric space and  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  an arbitrary sequence of (non-empty) subsets of  $\mathcal{F}$  which is  $\gamma$ -encodable in  $(\mathcal{F}, d)$ , with  $\gamma \in (0, \infty]$ . Then for every (non-empty)  $\mathcal{C} \subset \mathcal{F}$ :

$$\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leq \gamma^{*\text{encod}}(\mathcal{C}).$$

The proof of Theorem VI.1 is in appendix G. We derive from Theorem VI.1 a generic lower bound on the encoding speed of the set of functions uniformly approximated at a given speed.

**Corollary VI.1.** — Let  $(\mathcal{F}, d)$  be a metric space. Consider  $\gamma \in (0, \infty]$  and  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  an arbitrary sequence of (non-empty) subsets of  $\mathcal{F}$  which is  $\gamma$ -encodable in  $(\mathcal{F}, d)$ . Consider  $\alpha, \beta > 0$  and  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)$  the set of all  $f \in \mathcal{F}$  such that  $\sup_{M \geq 1} M^\alpha d(f, \Sigma_M) \leq \beta$ . This set satisfies

$$\gamma^{*\text{encod}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)) \geq \min(\alpha, \gamma).$$

*Proof.* By the very definition of  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)$ , it holds  $\gamma^{*\text{approx}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)|\Sigma) \geq \alpha$ . Theorem VI.1 then gives the result.  $\square$

The reader may wonder about the role of  $\beta$  in the above result, and whether a similar result can be achieved with  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma) := \cup_{\beta > 0} \mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)$ . While this is left open, a related discussion after Corollary VII.2 suggests this may not be possible without additional assumptions on  $\Sigma$ .

As an immediate corollary of Theorem VI.1 we also obtain the following result.

**Corollary VI.2.** — Consider  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  an arbitrary sequence of (non-empty) subsets of a metric space  $\mathcal{F}$  and a (non-empty) set  $\mathcal{C} \subset \mathcal{F}$ . If  $\Sigma$  is  $\gamma$ -encodable for every  $\gamma < \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$  then:

$$\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C}).$$

*Proof.* For every  $\gamma < \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ , since  $\Sigma$  is  $\gamma$ -encodable, we have  $\gamma = \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leq \gamma^{*\text{encod}}(\mathcal{C})$  by Proposition VI.1. Taking the supremum of such  $\gamma$ , we get the inequality.  $\square$

As we will see in section VII, applying Corollary VI.2 to specific  $\infty$ -encodable sequences allows one to unify and generalize different cases where  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C})$  is known to hold [8, Thm. V.3, Thm. VI.4][10, Thm. 5.24].

Note that the quantity  $\gamma^{*\text{encod}}(\mathcal{C})$  is known in several cases, see [8, Table 1]. In the next section, we discuss concrete examples of  $\infty$ -encodable sequences  $\Sigma$ . For such a sequence  $\Sigma$  and an arbitrary set  $\mathcal{C}$ , independently of the adequation of  $\Sigma$  and  $\mathcal{C}$ , Corollary VI.2 automatically yields an upper bound for the approximation speed of  $\mathcal{C}$  by  $\Sigma$ .

In some situations, the converse of Corollary VI.2 can be established.

**Theorem VI.2.** — Let  $\mathcal{C}$  be a (non-empty) subset of a metric space  $(\mathcal{F}, d)$  and  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  a sequence of (non-empty) subsets of  $\mathcal{F}$  such that  $\Sigma_M \subset \mathcal{C}$  for every  $M$  large enough. If  $\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma^{*\text{encod}}(\mathcal{C})) > 0$  then the sequence  $\Sigma$  is  $\gamma$ -encodable for each  $0 < \gamma < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma^{*\text{encod}}(\mathcal{C}))$ .

In particular, if  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq \gamma^{*\text{encod}}(\mathcal{C})$  then  $\Sigma$  is  $\gamma$ -encodable for every  $0 < \gamma < \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ .

*Proof.* Consider  $0 < \gamma < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma^{*\text{encod}}(\mathcal{C}))$ . By definition of  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ , there exists a constant  $c > 0$  such that for every  $f \in \mathcal{C}$  and every  $M \in \mathbb{N}$ , there exists  $\Phi_M(f) \in \Sigma_M$  such that  $d(f, \Phi_M(f)) \leq cM^{-\gamma}$ . Consider  $\gamma' > 0$  such that  $\gamma < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma^{*\text{encod}}(\mathcal{C}))$ . For  $M \in \mathbb{N}$ , define  $\varepsilon_M := M^{-\gamma}$ . By definition of  $\gamma^{*\text{encod}}(\mathcal{C})$ , there is a constant  $c' > 0$  such that for every  $\varepsilon > 0$ , there exists an  $\varepsilon$ -covering  $\mathcal{C}_\varepsilon$  of  $\mathcal{C}$  of size satisfying  $\log_2(|\mathcal{C}_\varepsilon|) \leq c'\varepsilon^{-1/\gamma'}$ . For  $M$  large enough,  $\Sigma_M \subset \mathcal{C}$ , hence for every such  $M$  and every  $f \in \Sigma_M$ , there exists  $f_{\varepsilon_M} \in \mathcal{C}_{\varepsilon_M}$  such that  $d(f, f_{\varepsilon_M}) \leq \varepsilon_M$ . Using the triangle inequality, we obtain that for every  $M$  large enough and every  $f \in \Sigma_M$ :  $d(f, \Phi_M(f_{\varepsilon_M})) \leq (1+c)M^{-\gamma}$ . This shows that  $\Phi_M(\mathcal{C}_{\varepsilon_M})$  is a  $(1+c)M^{-\gamma}$ -covering of  $\Sigma_M$  of size satisfying  $\log_2(|\Phi_M(\mathcal{C}_{\varepsilon_M})|) \leq c'M^{\gamma/\gamma'}$ , with  $\gamma/\gamma' < 1$ . This shows that  $\Sigma$  is  $\gamma$ -encodable. The rest of the claim follows.  $\square$

## VII. EXAMPLES OF $\infty$ -ENCODABLE APPROXIMATION FAMILIES

We now give several examples of  $\infty$ -encodable sequences  $\Sigma$ . We start with a gentle warmup in section VII-A. It is proven that some sequences of balls (in the sense of the metric space  $\mathcal{F}$ ) of increasing radius and dimension are  $\infty$ -encodable. Quite naturally,  $\infty$ -encodability is preserved under some Lipschitz transformation, as shown in Theorem VII.1 in the specific case of  $\infty$ -encodable sequences of balls (this can be generalized to other  $\infty$ -encodable sequences, but this is not useful here). In section VII-B, we give examples of  $\infty$ -encodable sequences in the context of approximations with dictionaries, see section VII-B, showing that Theorem VI.1 unifies and generalizes Theorem V.3 in [8] and Theorem 5.24 in [10]. Finally, in section VII-C, we give an

example of an  $\infty$ -encodable approximation family defined with ReLU networks. Once again, Theorem VI.1 applied to this  $\infty$ -encodable sequence recovers a known result, see Example VII.1.

#### A. First examples of $\infty$ -encodable sequences

This subsection is a gentle warmup, where basic examples of  $\infty$ -encodable sequences are given in order to manipulate the notion of encodability. Let  $(\mathcal{F}, d)$  be a metric space and  $c > 0$ . Let  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  be a sequence of sets  $\Sigma_M \subset \mathcal{F}$  that can be covered with  $N_M = \mathcal{O}_{M \rightarrow \infty}(2^{cM\pi(\log M)})$  balls (with respect to the ambient metric space) centered in  $\Sigma_M$  of radius  $\varepsilon_M = \mathcal{O}_{M \rightarrow \infty}(M^{-\gamma})$ . Since  $\mathcal{O}_{M \rightarrow \infty}(2^{cM\pi(\log M)}) = \mathcal{O}_{M \rightarrow \infty}(2^{M^{1+h}})$  for every  $h > 0$ , it is clear from the definition that  $\Sigma$  is  $\infty$ -encodable. This is trivially the case when  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$  is a sequence of finite sets  $\Sigma_M \subset \mathcal{F}$  with at most  $2^{cM\pi(\log M)}$  elements since each  $\Sigma_M$  is an exact covering of itself. Another example consists of some sequences of balls (in the sense of the metric space  $\mathcal{F}$ ) of increasing radius and dimension as described in the next lemma. The proof is in appendix H.

**Lemma VII.1.** — Consider  $q \in [1, \infty]$ ,  $(d_M)_{M \in \mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ ,  $(r_M)_{M \in \mathbb{N}}$  a sequence of real numbers satisfying  $r_M \geq 1$  and define  $\Sigma := (\Sigma_M)_{M \in \mathbb{N}}$ , with  $\Sigma_M := B_{d_M, \|\cdot\|_q}(0, r_M)$  being the set of sequences of  $\ell^q(\mathbb{N})$  bounded by  $r_M$  and supported in the first  $d_M$  coordinates. Then,  $\Sigma$  is either  $\infty$ -encodable in  $\ell^q(\mathbb{N})$  or it is never  $\gamma$ -encodable in  $\ell^q(\mathbb{N})$ , whatever  $\gamma > 0$  is. Moreover, it is  $\infty$ -encodable if, and only if,

$$d_M (\log_2(r_M) + 1) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h}), \quad \forall h > 0.$$

Quite naturally,  $\infty$ -encodability can be preserved under Lipschitz maps as shown in the following theorem. The proof is in appendix H.

**Theorem VII.1.** — Consider the same setting as in Lemma VII.1. Consider also a sequence  $\varphi := (\varphi_M)_{M \in \mathbb{N}}$  of maps  $\varphi_M : (\Sigma_M, \|\cdot\|_q) \rightarrow (\mathcal{F}, d)$  that are  $\text{Lips}(\varphi_M)$ -Lipschitz for some constants  $\text{Lips}(\varphi_M) \geq 1$ . Define  $\varphi(\Sigma) := (\varphi_M(\Sigma_M))_{M \in \mathbb{N}}$ . Assume that for every  $h > 0$ :

$$d_M (\log_2(r_M) + \log_2(\text{Lips}(\varphi_M)) + 1) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h}). \quad (11)$$

Then  $\varphi(\Sigma)$  is  $\infty$ -encodable.

#### B. The case of dictionaries

We now consider sequences  $\Sigma$  defined with dictionaries. As detailed below, results of the literature [10, Thm. 5.24][12, Prop. 11] use arguments that implicitly prove  $\gamma$ -encodability. Let us start with the case of approximation in Banach spaces as in [12]. We only explicit the sequence used in [12] which is  $\gamma$ -encodable and we do not delve into more details as results of [12] are out of scope of this paper. A part of the proof of [12, Prop. 11] consists of implicitly showing that some specific sequence  $\Sigma^q$  is  $s$ -encodable, for  $q$  and  $s$  as described below in Proposition VII.1, as shown in appendix I. In particular, the setup of Proposition VII.1 applies when  $\mathcal{F}$  is the  $L^p$  space on  $\mathbb{R}^d$  or  $[0, 1]^d$ ,  $1 < p < \infty$ , and the basis  $B$  is a compactly supported wavelet basis or associated wavelet-tensor product basis.

**Proposition VII.1.** — Let  $\mathcal{F}$  be a Banach space with a basis  $B = (e_i)_{i \in \mathbb{N}}$  satisfying  $\sup_{i \in \mathbb{N}} \|e_i\|_{\mathcal{F}} < \infty$ . Consider  $p \in (0, \infty)$  and assume that  $B$  satisfies the so-called  $p$ -Telmakov property [12, Def. 2], i.e., assume that there exists  $c > 0$  such that for every finite subset  $I$  of  $\mathbb{N}$ :

$$\frac{1}{c} |I|^{1/p} \min_{i \in I} |c_i| \leq \left\| \sum_{i \in I} c_i e_i \right\|_{\mathcal{F}} \leq c |I|^{1/p} \max_{i \in I} |c_i|, \quad \forall (c_i)_{i \in I} \in \mathbb{R}^I. \quad (12)$$

Consider  $0 < q < p$ . For every  $M \in \mathbb{N}$ , define<sup>7</sup>:

$$\Sigma_M^q := \left\{ \sum_{i=1}^M c_i e_i, c_i \in \mathbb{R}, \sup_{0 < \lambda < \infty} \lambda |\{i, |c_i| \geq \lambda\}|^{1/q} \leq 1 \right\}.$$

Define  $s = \frac{1}{q} - \frac{1}{p}$ . Then the sequence  $\Sigma^q := (\Sigma_M^q)_{M \in \mathbb{N}}$  is  $s$ -encodable in  $\mathcal{F}$ .

In the case of Hilbert spaces, much more generic sequences than  $\Sigma^q$  above are in fact  $\infty$ -encodable, as we now discuss. The  $\infty$ -encodability can be used to recover [10, Thm. 5.24] (see Corollary VII.1), and to generalize Corollary VI.1 (see Corollary VII.2). Let  $\mathcal{F}$  be a Hilbert space and  $d$  be the metric associated to the norm on  $\mathcal{F}$ . A dictionary is, by definition [10, Def. 5.19], a subset  $\mathcal{D} = (\phi_i)_{i \in \mathbb{N}}$  of  $\mathcal{F}$  indexed by a countable set, which we assume to be  $\mathbb{N}$  without loss of generality. The dictionary  $\mathcal{D}$  can be used to approach elements of  $\mathcal{F}$  by linear combinations of a growing number  $M$  of its elements.

**Theorem VII.2.** — Let  $\mathcal{F}$  be a Hilbert space. Let  $\mathcal{D} = (\phi_i)_{i \in \mathbb{N}}$  be a dictionary in  $\mathcal{F}$ , and  $\pi : \mathbb{N} \rightarrow \mathbb{N}$  be a function with at most polynomial growth. For every  $I \subset \mathbb{N}$ , define  $(\tilde{\phi}_i^I)_{i \in I}$  as any orthonormalization of  $(\phi_i)_{i \in I}$  (for instance we may consider the Gram-Schmidt orthonormalization). Define for every  $M \in \mathbb{N}$  and  $c > 0$ :

$$\begin{aligned} \Sigma_M^\pi &:= \left\{ \sum_{i \in I} c_i \phi_i, I \subset \{1, \dots, \pi(M)\}, |I| \leq M, (c_i)_{i \in I} \in \mathbb{R}^I \right\}, \\ \tilde{\Sigma}_M^{\pi, c} &:= \left\{ \sum_{i \in I} \tilde{c}_i \tilde{\phi}_i^I, I \subset \{1, \dots, \pi(M)\}, |I| \leq M, (\tilde{c}_i)_{i \in I} \in [-c, c]^I \right\}. \end{aligned}$$

The sequence  $\tilde{\Sigma}^{\pi, c} := (\tilde{\Sigma}_M^{\pi, c})_{M \in \mathbb{N}}$  is  $\infty$ -encodable in  $(\mathcal{F}, d)$ , and for every bounded set  $\mathcal{C} \subset \mathcal{F}$ , it holds:

$$\gamma^{\text{*approx}}(\mathcal{C} | \Sigma^\pi) = \max_{c > 0} \gamma^{\text{*approx}}(\mathcal{C} | \tilde{\Sigma}^{\pi, c}). \quad (13)$$

The proof of Theorem VII.2 is in appendix I. As a consequence of Theorem VII.2, one can recover [10, Thm. 5.24] as we now describe.

**Corollary VII.1 ([10, Thm. 5.24]).** — Let  $(\mathcal{F}, d)$  be a Hilbert space and  $\mathcal{C} \subset \mathcal{F}$ . Under the assumptions of Theorem VII.2, the sequence  $\Sigma^\pi = (\Sigma_M^\pi)_{M \in \mathbb{N}}$  satisfies for every relatively compact<sup>8</sup> set  $\mathcal{C}$ :

$$\gamma^{\text{*approx}}(\mathcal{C} | \Sigma^\pi) \leq \gamma^{\text{*encod}}(\mathcal{C}).$$

<sup>7</sup>In terms of weak- $\ell^q$ -space, the set  $\Sigma_M^q$  is simply the set of linear combinations of elements of  $B$  given by sequences  $(c_i)_{i \in \mathbb{N}}$  in the closed unit ball of  $\ell^{q, \infty}(\mathbb{N})$  with zero coordinates outside the first  $M$  ones.

<sup>8</sup>Recall that a set is relatively compact if its closure is compact. In particular, it must be totally bounded, and in particular bounded.

Actually, instead of stating the previous result with the approximation speed  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ , Theorem 5.24 in [10] considers the following quantity [10, Def. 5.23]:

$$\gamma^*(\mathcal{C}|\Sigma) := \sup\{\gamma \in \mathbb{R}, \forall f \in \mathcal{C}, \exists c > 0, \forall M \in \mathbb{N}, d(f, \Sigma_M) \leq cM^{-\gamma}\},$$

which satisfies  $\gamma^*(\mathcal{C}|\Sigma) \geq \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$  but generally differs from  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$  since in the definition of  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ , the implicit constant  $c > 0$  is not allowed to depend on  $f \in \mathcal{C}$ . However, when  $\mathcal{C}$  is relatively compact (that is, its closure is compact), then  $c > 0$  can be chosen independently of  $f$  [10, Proof of Thm 5.24] so that the two quantities coincide. The proof of Corollary VII.1 that can be found below is essentially a rewriting in the formalism of section VI of the original proof of Theorem 5.24 in [10]. The rewriting makes explicit the use of equality (13) and the  $\infty$ -encodability of the sequences  $\tilde{\Sigma}^{\pi,c}$  for  $c > 0$ , which are only implicitly used in the original proof.

*Proof.* Since  $\mathcal{C}$  is relatively compact, it must be bounded so Equation (13) of Theorem VII.2 holds. For every  $c > 0$ , Theorem VI.1 applied to  $\tilde{\Sigma}^{\pi,c}$  of Theorem VII.2, which is  $\infty$ -encodable, shows that the right hand-side of Equation (13) is bounded from above by  $\gamma^{*\text{encod}}(\mathcal{C})$ . This yields the result.  $\square$

We also obtain a generic lower bound on the encoding speed of balls of approximation spaces [6, Sec. 7.9] (also called maxisets [11]) with general dictionaries.

**Corollary VII.2.** — Let  $(\mathcal{F}, d)$  be a Hilbert space. Under the assumptions of Theorem VII.2, consider  $\alpha, \beta > 0$  and the set<sup>9</sup>  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$  of all  $f \in \mathcal{F}$  such that  $\|f\| \leq \beta$  and  $\sup_{M \geq 1} M^\alpha d(f, \Sigma_M) \leq \beta$ . This set satisfies

$$\gamma^{*\text{encod}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)) \geq \alpha.$$

Corollary VII.2 cannot be generalized to  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi) := \bigcup_{\beta > 0} \mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ : this set is homogeneous (stable by multiplication by any scalar), thus it cannot be encoded at any positive rate. Indeed, a positive encoding rate implies total boundedness of a set, whereas homogeneity implies that the set cannot be totally bounded (at least under the assumption that the metric is induced by a norm; there should, in general, be metrics with respect to which a homogeneous set may be totally bounded).

In some situations, the converse inequality  $\gamma^{*\text{encod}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)) \leq \alpha$  can typically be proven by studying the existence of large enough packing sets of  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ , but this falls out of the scope of this paper. The reader can refer to [12, Sec. 4] for an example.

*Proof of Corollary VII.2.* By the very definition of  $\mathcal{C} := \mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ , this is a bounded set so Equation (13) of Theorem VII.2 holds. For every  $c > 0$ , Theorem VI.1 applied to  $\tilde{\Sigma}^{\pi,c}$  of Theorem VII.2, which is  $\infty$ -encodable, shows that the right hand-side of Equation (13) is bounded from above by  $\gamma^{*\text{encod}}(\mathcal{C})$ , so that

$$\gamma^{*\text{encod}}(\mathcal{C}) \geq \max_{c > 0} \gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c}) = \gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi).$$

<sup>9</sup>This is the ball of radius  $\beta$  of an approximation space [6, Sec. 7.9]/maxiset[11].

<sup>10</sup>Note that compared to the set in Corollary VI.1, we additionally require that  $\|f\| \leq \beta$  so that  $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$  is a bounded set and Equation (13) of Theorem VII.2 holds.

Finally, again by definition of  $\mathcal{C} := \mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ , we have  $\gamma^{\text{approx}}(\mathcal{C}|\Sigma^\pi) \geq \alpha$ .  $\square$

Note that if  $\Sigma^\pi$  was  $\gamma$ -encodable for some  $\gamma > 0$  large enough then Corollary VII.1 would be a special case of Corollary VI.2 whereas Corollary VII.2 would be a special case of Corollary VI.1. But in this situation,  $\Sigma^\pi$  has no reason to be  $\gamma$ -encodable, whatever  $\gamma > 0$  is (since the dictionary is arbitrary and the coefficients of the linear combinations are not bounded). This shows that Corollary VI.2 and Corollary VI.1 actually holds more generally for some sequences  $\Sigma$  that are not  $\gamma$ -encodable, whatever  $\gamma > 0$  is, as soon as  $\Sigma$  can be recovered as a limit of non-decreasing sequences  $\Sigma^c$ ,  $c > 0$ , that are  $\gamma$ -encodable, in the sense that for every  $M \in \mathbb{N}$ , if  $0 < c \leq c'$  then  $\Sigma_M^c \subseteq \Sigma_M^{c'}$  and  $\Sigma_M = \cup_{c>0} \Sigma_M^c$ .

### C. The case of ReLU networks

When  $\Sigma$  is defined with ReLU feed-forward neural networks, we now explicitly study how the property of  $\infty$ -encodability depends on (bounds on) the neural network sparsity, depth, and weights. In particular, Proposition VII.2 establishes a "simple" explicit condition under which Theorem VI.1 generalizes Theorem VI.4 in [8] to other type of constraints. Proposition VII.2 is proven in appendix I.

**Proposition VII.2.** — Consider the context of Definition V.4 and assume that for every  $h > 0$ , it holds:

$$L_M M (1 + \log_2(r_M)) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h}). \quad (14)$$

Then the approximation family  $\mathcal{N}$  (Definition V.2) defined with ReLU networks is  $\infty$ -encodable.

**Example VII.1 ( $\infty$ -encodable sequences of *sparse* neural networks - [8, Thm. VI.4]).** — Let  $\pi$  be a positive polynomial and consider, as in Definition VI.2 of [8],  $\mathcal{N}_M^\pi$  the set of functions parameterized by a ReLU neural network with weights' amplitude bounded by  $\pi(M)$ , depth bounded by  $\pi(\log M)$  and at most  $M$  non-zero parameters. Assumption (14) holds since this corresponds to the case where  $L_M \leq \pi(\log(M))$  and  $1 \leq r_M \leq \max(1, \pi(M))$ . Then, Proposition VII.2 guarantees that  $\mathcal{N}^\pi := (\mathcal{N}_M^\pi)_{M \in \mathbb{N}}$  is  $\infty$ -encodable. Given Theorem VI.1, the fact that  $\mathcal{N}^\pi$  is  $\infty$ -encodable gives  $\gamma^{\text{approx}}(\mathcal{C}|\mathcal{N}^\pi) \leq \gamma^{\text{encod}}(\mathcal{C})$  for arbitrary  $p \in [1, \infty]$  and arbitrary  $\mathcal{C} \subset L^p$ . This is exactly Theorem VI.4 in [8].

## VIII. CONCLUSION

We now summarize our different contributions and discuss perspectives.

**Approximation with quantized ReLU networks** We characterized the error of simple uniform quantization scheme  $Q_\eta$  that acts coordinatewise as  $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$ . We proved in Theorem IV.2 that the number of bits per coordinate must grow linearly with the depth of the network in order to provide  $\varepsilon$ -error in  $L^\infty([-D, D]^d)$ , uniformly on a bounded set of parameters  $\Theta_{L, \mathbf{N}}^q(r)$ . The proof exploits a new lower-bound on the Lipschitz constant of the parameterization of ReLU networks that we established in Theorem III.1. We also proved a generic upper-bound for this Lipschitz constant, which generalizes upper-bounds known in specific situations. As a consequence, we gave explicit conditions on the number of bits per coordinate that guarantees quantized ReLU networks to have the same approximation speeds as unquantized ones in generic

$L^p$  spaces, see Example V.1. We further used in section IV the upper bound on the Lipschitz constant of  $\theta \mapsto R_\theta$  to recover a known approximation result of quantized ReLU networks in  $L^\infty$ -Sobolev spaces [7, Thm. 2] and to improve a result on the error of nearest-neighbour uniform quantization [8, Lem. VI.8].

**Notion of  $\gamma$ -encodability.** This paper introduced in Definition VI.2 a new property of approximation families: being  $\gamma$ -encodable. As soon as  $\Sigma$  is  $\gamma$ -encodable in a metric space  $(\mathcal{F}, d)$ , Theorem VI.1 shows that there is a simple relation between the approximation speed of every set  $\mathcal{C} \subset \mathcal{F}$  and its encoding speed:

$$\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leq \gamma^{*\text{encod}}(\mathcal{C}). \quad (15)$$

As seen in section VII, several classical approximation families  $\Sigma$  are  $\gamma$ -encodable for some  $\gamma > 0$ , including classical families defined with dictionaries (section VII-B) or ReLU neural networks (section VII-C). As a consequence,  $\gamma$ -encodability lays a generic framework that unifies several situations where Inequality (15) is known, such as when doing approximation with dictionaries [10, Thm. 5.24][12, Prop. 11] or ReLU neural networks [8, Thm. VI.4].

**Perspectives.** In Theorem IV.1 and Theorem IV.2, we saw necessary and sufficient conditions on  $\eta > 0$  to guarantee that quantizing coordinatewise by  $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$  provides  $\varepsilon$ -error in  $L^\infty([-D, D]^d)$ , uniformly on a bounded set of parameters  $\Theta_{L, \mathbf{N}}^q(r)$ . In practical applications with post-training quantization, we are only interested in parameters that can be obtained with learning algorithms such as stochastic gradient descent. Moreover, we may not be interested in  $\varepsilon > 0$  arbitrary small. For instance, quantization aware training techniques [4] have been successfully applied for ReLU neural networks with three hidden layers and 1024 neurons per hidden layer [4]. Indeed, the modified learning procedure yields in [4] a network with quantized weights in  $\{-1, 1\}$  that performs similarly, on the MNIST dataset, as the network that would have been obtained with the original learning procedure. Is it possible to have better guarantees if we only care about some prescribed error  $\varepsilon > 0$  and a "small set" of parameters, such as parameters that can indeed be learned in practice?

Another question would be to design schemes to quantize network parameters, in a way that adapts to the architecture. In the quantization schemes covered by Theorem IV.1, the sufficient value of  $\eta > 0$  to ensure a prescribed error  $\varepsilon > 0$  only takes into account the depth and the width of the architecture. However, in practice the network architecture is carefully designed to meet some criterion, such as reducing the inference cost (references can be found in the paragraph "Compact network design" of [19]). Specificities of the architecture could be taken into consideration when designing the quantization scheme.

Another perspective is to take into account functionally equivalent parameters when designing a quantization scheme, as we now detail. Given parameters  $\theta$  of a ReLU neural network (and possibly a finite dataset), we say that  $\theta'$  is functionally equivalent to  $\theta$ , denoted  $\theta' \sim \theta$ , if  $R_\theta = R_{\theta'}$  (resp. equality on the considered dataset). Due to the positive homogeneity of the ReLU function, there are uncountably many equivalent parameters to  $\theta$  that can be obtained by rescaling the coordinates of  $\theta$  (but these are not the only ones since permuting coordinates can also lead to functionally equivalent parameters). When quantizing  $\theta$ , it would be interesting to take these equivalent parameters into account.



Finally, what is the minimum number of bits per coordinate needed to keep the same approximation speeds? While the question remains open, Theorem V.1 makes a first step in that direction by giving an upper-bound.

## REFERENCES

- [1] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding deep neural networks with rectified linear units. *Electron. Colloquium Comput. Complex.*, 24:98, 2017.
- [2] J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. *SIAM J. Math. Data Sci.*, 2:631–657, 2020.
- [3] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.*, 1(1):8–45, 2019.
- [4] M. Courbariaux, Y. Bengio, and J. David. Binaryconnect: Training deep neural networks with binary weights during propagations. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3123–3131, 2015.
- [5] R. DeVore, B. Hanin, and G. Petrova. Neural network approximation. *Acta Numer.*, 30:327–444, 2021.
- [6] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993.
- [7] Y. Ding, J. Liu, J. Xiong, and Y. Shi. On the Universal Approximability and Complexity Bounds of Quantized ReLU Neural Networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [8] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei. Deep neural network approximation theory. *IEEE Trans. Inf. Theory*, 67(5):2581–2623, 2021.
- [9] I. Gühring and M. Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.
- [10] P. Grohs. Optimally sparse data representations. In *Harmonic and applied analysis*, Appl. Numer. Harmon. Anal., pages 199–248. Birkhäuser/Springer, Cham, 2015.
- [11] G. Kerkycharian and D. Picard. Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2):283–344, 2000. With comments, and a rejoinder by the authors.
- [12] G. Kerkycharian and D. Picard. Entropy, universal coding, approximation, and bases properties. *Constr. Approx.*, 20(1):1–37, 2004.
- [13] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. ICLR (Poster), 2018.
- [14] I. Ohn and Y. Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21:627, 2019.
- [15] P. Petersen and F. Voigtländer. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.
- [16] T. Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. ICLR (Poster), 2019.
- [17] M. J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.
- [18] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.
- [19] D. Zhang, J. Yang, D. Ye, and G. Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, volume 11212 of *Lecture Notes in Computer Science*, pages 373–390. Springer, 2018.

APPENDIX A  
NORMS

**Definition A.1** (*p*-norm). — Let  $d \in \mathbb{N}$ . For an exponent  $p \in [1, \infty]$ , the  $p$ -norm on  $\mathbb{R}^d$  is defined by:

$$\forall x = (x_i)_{i=1, \dots, d} \in \mathbb{R}^d, \|x\|_p := \begin{cases} \left( \sum_{i=1}^d |x_i|^p \right)^{\frac{1}{p}} & \text{if } p < \infty, \\ \sup_{i=1, \dots, d} |x_i| & \text{if } p = \infty. \end{cases}$$

**Definition A.2.** — ( $\|\cdot\|_p$ ) Let  $d_1, d_2 \in \mathbb{N}$ . The operator norm  $\|\cdot\|_p$  on  $\mathbb{R}^{d_2 \times d_1}$  associated with the exponent  $p \in [1, \infty]$  is defined by:

$$\forall M \in \mathbb{R}^{d_2 \times d_1}, \|M\|_p := \sup_{\substack{x \in \mathbb{R}^{d_1} \\ x \neq 0}} \frac{\|Mx\|_p}{\|x\|_p}.$$

APPENDIX B

CHARACTERIZATION OF THE  $L^p$  SPACES CONTAINING ALL THE FUNCTIONS REALIZED BY ReLU NETWORKS

*Proof of Lemma II.1.* Assume that  $C_p(\Omega, \mu) < \infty$  and consider the realization  $R_\theta$  of an arbitrary ReLU network on an arbitrary architecture with input dimension  $N_0 = d_{\text{in}}$  and arbitrary output dimension  $N_L$ . It is known [1, Thm. 2.1] that  $R_\theta$  is (continuous and) piecewise linear, so that there is a partition of  $\Omega$  into finitely many  $\Omega_i$ ,  $1 \leq i \leq n$  such that  $R_\theta = \sum_{i=1}^n \chi_{\Omega_i} f_i$  where  $\chi_E(x)$  is the characteristic function of the set  $E$  and each  $f_i$  is an affine function. To prove the result it is thus sufficient to show that  $\chi_E f \in L^p(\Omega \rightarrow \mathbb{R}^{N_L}, \mu)$  for each set  $E \subset \Omega$  and each affine function  $f$ . Since  $\|\chi_E g\|_p \leq \|g\|_p$  for any  $g$  it is enough to prove that any affine function is in the desired space. For this, consider arbitrary  $A \in \mathbb{R}^{N_L \times N_0}$ ,  $b \in \mathbb{R}^{N_L}$ , and  $f : x \mapsto Ax + b$ . Denoting  $c(f) := \max(\|A\|_\infty, \|b\|_\infty)$  (the notation  $\|\cdot\|$  is defined in appendix A) we observe that  $\|f(x)\|_\infty \leq \|A\|_\infty \|x\|_\infty + \|b\|_\infty \leq c(f)(\|x\|_\infty + 1)$  so that  $\|f\|_p \leq c(f)C_p(\Omega, \mu) < \infty$ , showing the result.

Conversely, assume that for every architecture  $(L, \mathbf{N})$  and parameter  $\theta \in \Theta_{L, \mathbf{N}}$  we have  $R_\theta \in L^p(\Omega \rightarrow \mathbb{R}^{N_L}, \mu)$ . Specializing to an architecture with  $L = 1$ ,  $N_1 = N_0 = d_{\text{in}}$ , consider  $\theta = (W_1, b_1)$  with  $W_1$  the identity matrix and  $b_1$  the zero vector,  $\theta' = (W'_1, b'_1)$  with  $W'_1$  the zero matrix and  $b'_1$  any vector with  $\|b'_1\|_\infty = 1$ . We have  $R_\theta(x) = x$  while  $R_{\theta'}(x) = b'_1$ . For  $p < \infty$  we have  $\int_{x \in \Omega} \|x\|_\infty^p d\mu(x) = \|R_\theta\|_p^p < \infty$  and  $\int_{x \in \Omega} 1 d\mu(x) = \|R_{\theta'}\|_p^p < \infty$ . By the triangle inequality we get  $C_p(\Omega, \mu) < \infty$ . The case  $p = \infty$  is similar.  $\square$

APPENDIX C

OPTIMALITY OF A BOUND ON  $\|R_\theta(x) - R_{\theta'}(x)\|_q$

We generalize a known inequality established for  $q = \infty$  [8, Eq. (37)][3, Eq. (3.12)] to arbitrary  $q$ -th norm  $q \in [1, \infty]$ . Moreover, we prove its optimality. This inequality is used in appendix D to bound the Lipschitz constant of the parameterization of ReLU networks. With  $I_{m \times m}$  the identity matrix in dimension  $m$  and  $0_{m \times n}$  the  $m \times n$  matrix full of zeros, we introduce the following notation for ‘‘rectangular identity matrices’’: for  $m < n$ , we set  $I_{m \times n} = (I_{m \times m}; 0_{m \times (n-m)})$ , while for  $m > n$  we set  $I_{m \times n} = I_{n \times m}^\top$ .

**Lemma C.1.** — Let  $(L, \mathbf{N})$  be an architecture with any depth  $L \geq 1$  and  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$ ,  $\theta' = (W'_1, \dots, W'_L, b'_1, \dots, b'_L) \in \Theta_{L, \mathbf{N}}$  (see Equation (4) for the definition of  $\Theta_{L, \mathbf{N}}$ ) be parameters associated to this architecture. For every  $\ell = 1, \dots, L - 1$ , define  $\theta'_\ell$  as the parameter deduced from  $\theta'$ , associated to the architecture  $(\ell, (N_0, \dots, N_\ell))$ :

$$\theta'_\ell = (W'_1, \dots, W'_\ell, b'_1, \dots, b'_\ell).$$

Then for every exponent  $q \in [1, \infty]$  and for every  $x \in \mathbb{R}^{N_0}$ , the realization of neural networks with any 1-Lipschitz activation function  $\varrho$  such that  $\varrho(0) = 0$  satisfy:

$$\begin{aligned} \|R_\theta(x) - R_{\theta'}(x)\|_q &\leq \sum_{\ell=1}^L \left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \times \|W_\ell - W'_\ell\|_q \times \|R_{\theta'_{\ell-1}}(x)\|_q \\ &\quad + \sum_{\ell=1}^L \left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \|b_\ell - b'_\ell\|_q, \end{aligned} \quad (16)$$

where the definition of the  $q$ -th norm and the operator norm of a matrix are recalled in appendix A, and where we set by convention  $R_{\theta'_{\ell-1}}(x) = x$  if  $\ell = 1$ , and  $\prod_{k=\ell+1}^L \|W_k\|_q = 1$  if  $\ell = L$ .

Let  $\lambda_1, \dots, \lambda_L \geq 0$  and  $\varepsilon \geq 0$  and consider an input vector  $x \in \mathbb{R}^{d_{\text{in}}}$  with nonnegative entries and supported on the first  $s := \min_\ell N_\ell$  coordinates. There is equality in (16) for the parameters  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$  and  $\theta' = (W'_1, \dots, W'_L, b'_1, \dots, b'_L)$  defined by, for every  $\ell = 1, \dots, L$ :

$$W_\ell = \lambda_\ell I_{N_\ell \times N_{\ell-1}}, \quad W'_\ell = (1 + \varepsilon)W_\ell, \quad b_\ell = b'_\ell = 0. \quad (17)$$

*Proof.* The proof of Inequality (16) follows by induction on  $L \in \mathbb{N}$  in a similar way as in the case  $q = \infty$  [8, Eq. (37)][3, Eq. (3.12)]. For  $L = 1$ , this is just saying that

$$\begin{aligned} \|R_\theta(x) - R_{\theta'}(x)\|_q &= \|W_1 x + b_1 - W'_1 x - b'_1\|_q \\ &\leq \|W_1 - W'_1\|_q \|x\|_q + \|b_1 - b'_1\|_q. \end{aligned}$$

Assume that the property holds true for  $L \geq 1$ . Then at rank  $L + 1$  (using in the last inequality that the activation function  $\rho$  is 1-Lipschitz and  $\varrho(0) = 0$ ):

$$\begin{aligned} \|R_\theta(x) - R_{\theta'}(x)\|_q &= \|W_{L+1} \rho(R_{\theta_L}(x)) + b_{L+1} - W'_{L+1} \rho(R_{\theta'_L}(x)) - b'_{L+1}\|_q \\ &= \|W_{L+1} \left( \rho(R_{\theta_L}(x)) - \rho(R_{\theta'_L}(x)) \right) \\ &\quad + (W_{L+1} - W'_{L+1}) \rho(R_{\theta'_L}(x)) + b_{L+1} - b'_{L+1}\|_q \\ &\leq \|W_{L+1}\|_q \|\rho(R_{\theta_L}(x)) - \rho(R_{\theta'_L}(x))\|_q \\ &\quad + \|W_{L+1} - W'_{L+1}\|_q \|\rho(R_{\theta'_L}(x))\|_q + \|b_{L+1} - b'_{L+1}\|_q \\ &\leq \|W_{L+1}\|_q \|R_{\theta_L}(x) - R_{\theta'_L}(x)\|_q \\ &\quad + \|W_{L+1} - W'_{L+1}\|_q \|R_{\theta'_L}(x)\|_q + \|b_{L+1} - b'_{L+1}\|_q. \end{aligned}$$

Using the induction hypothesis gives the desired result.

For the equality case, recall the definition of the parameters  $\theta$  and  $\theta'$  in Equation (17). Let  $\lambda = \prod_{\ell=1}^L \lambda_\ell$ . Since  $x = (y^\top, 0_{1 \times (d_{\text{in}} - s)})^\top$  with  $y \in \mathbb{R}_+^s$  we have  $\varrho(W_1 x + b_1) = \lambda_1 (y^\top, 0_{1 \times (N_1 - s)})^\top$ . By induction on  $\ell = 1, \dots, L$ , we can show  $R_\theta(x) = \lambda (y^\top, 0_{1 \times (N_L - s)})^\top$ , and similarly  $R_{\theta'}(x) = (1 + \varepsilon)^L \lambda (y^\top, 0_{1 \times (N_L - s)})^\top$ . This means that:

$$\begin{aligned} \|R_\theta(x) - R_{\theta'}(x)\|_q &= \|\lambda y - (1 + \varepsilon)^L \lambda y\|_q \\ &= ((1 + \varepsilon)^L - 1) \lambda \|x\|_q. \end{aligned}$$

Moreover, for every  $\ell = 1, \dots, L$ , it is easy to check that  $\|W_\ell\|_q = \lambda_\ell$ ,  $\|W'_\ell\|_q = (1 + \varepsilon) \lambda_\ell$  and  $\|W_\ell - W'_\ell\|_q = \varepsilon \lambda_\ell$  so that:

$$\begin{aligned} &\left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \times \|W_\ell - W'_\ell\|_q \times \|R_{\theta'_{\ell-1}}(x)\|_q \\ &= \left( \prod_{k=\ell+1}^L \lambda_k \right) \times \varepsilon \lambda_\ell \times \left( \prod_{k=1}^{\ell-1} (1 + \varepsilon) \lambda_k \right) \|x\|_q \\ &= (1 + \varepsilon)^{\ell-1} \varepsilon \lambda \|x\|_q, \end{aligned}$$

and:

$$\left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \|b_\ell - b'_\ell\|_q = 0.$$

This yields the equality case, since:

$$\begin{aligned} &\sum_{\ell=1}^L \left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \times \|W_\ell - W'_\ell\|_q \times \|R_{\theta'_{\ell-1}}(x)\|_q + \sum_{\ell=1}^L \left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \|b_\ell - b'_\ell\|_q \\ &= \sum_{\ell=1}^L (1 + \varepsilon)^{\ell-1} \varepsilon \lambda \|x\|_q = \frac{(1 + \varepsilon)^L - 1}{1 + \varepsilon - 1} \varepsilon \lambda \|x\|_q = ((1 + \varepsilon)^L - 1) \lambda \|x\|_q. \end{aligned}$$

□

## APPENDIX D

### LIPSCHITZ PARAMETERIZATION OF ReLU NETWORKS (PROOF OF THEOREM III.1)

Recall that we fixed a set  $L^p(\Omega \rightarrow \mathbb{R}^{d_{\text{out}}}, \mu)$  containing all functions realized by ReLU neural networks with input dimension  $d_{\text{in}}$  and output dimension  $d_{\text{out}}$ . The parameter set  $\Theta_{L, \mathbf{N}}^q(r)$  is defined in Definition III.1.

First, Lemma C.1 applied to any  $\theta \in \Theta_{L, \mathbf{N}}$ , and  $\theta' = (0, \dots, 0) \in \Theta_{L, \mathbf{N}}$  yields for every  $x \in \Omega$ :

$$\|R_\theta(x)\|_q \leq \prod_{k=1}^L \|W_k\|_q \|x\|_q + \sum_{\ell=1}^L \left( \prod_{k=\ell+1}^L \|W_k\|_q \right) \|b_\ell\|_q, \quad (18)$$

using that  $\|R_{\theta'_{\ell-1}}(x)\|_q = \|x\|_q$  for  $\ell = 1$  (by convention) and  $\|R_{\theta'_{\ell-1}}(x)\|_q = 0$  for each  $\ell \geq 2$  (since  $\theta' = 0$ ).

Let  $\theta, \theta' \in \Theta_{L, \mathbf{N}}$ . We are going to bound  $\|R_\theta - R_{\theta'}\|_{p, \|\cdot\|}$  from above using Inequality (16) of Lemma C.1. First, we introduce useful notations to write things compactly. Define for every  $i, j \in \mathbb{N}$ :

$$\Pi_{i,j} := \prod_{k=i}^j \|W_k\|_q \text{ and } \Pi'_{i,j} := \prod_{k=i}^j \|W'_k\|_q \text{ if } i \leq j,$$

$$\Pi_{i,j} := \Pi'_{i,j} := 1 \text{ otherwise.}$$

For  $\ell = 2, \dots, L$ , we start by bounding  $\|R_{\theta'_{\ell-1}}(x)\|_q$  by a simple function of  $x \in \Omega$ , since this term appears on the right-handside of Inequality (16). Using (18) for the architecture  $(\ell - 1, (N_0, \dots, N_{\ell-1}))$  we have:

$$\begin{aligned} \|R_{\theta'_{\ell-1}}(x)\|_q &\leq \prod_{k=1}^{\ell-1} \|W'_k\|_q \|x\|_q + \sum_{k=1}^{\ell-1} \left( \prod_{j=k+1}^{\ell-1} \|W'_j\|_q \right) \|b'_k\|_q \\ &= \Pi'_{1,\ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} \Pi'_{k+1,\ell-1} \|b'_k\|_q. \end{aligned}$$

If  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$  and  $N_0 = \min_{0 \leq \ell \leq L} N_\ell$  then for every  $x \in \Omega$ , the parameters defined in Equation (17) are such that the previous inequality is an equality.

Denote  $c_0$  a constant such that for every  $y \in \mathbb{R}^{d_{\text{out}}}$ ,  $\|y\| \leq c_0 \|y\|_q$ . Note that if  $\|\cdot\| = \|\cdot\|_s$  for  $s \in [1, \infty]$ , then we can take  $c_0 = d_{\text{out}}^{\max(0, \frac{1}{s} - \frac{1}{q})}$ . Now, using the previous inequality and integrating both sides of Inequality (16) of Lemma C.1, we get for  $1 \leq p < \infty$ :

$$\begin{aligned} \left( \int_{x \in \Omega} \|R_\theta(x) - R_{\theta'}(x)\|^p d\mu(x) \right)^{\frac{1}{p}} &\leq c_0 \left( \int_{x \in \Omega} \|R_\theta(x) - R_{\theta'}(x)\|_q^p d\mu(x) \right)^{\frac{1}{p}} \\ &\leq c_0 \left( \int_{x \in \Omega} \left[ \sum_{\ell=1}^L \Pi_{\ell+1,L} \left( \Pi'_{1,\ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} \Pi'_{k+1,\ell-1} \|b'_k\|_q \right) \right. \right. \\ &\quad \left. \left. \times \|W_\ell - W'_\ell\|_q + \sum_{\ell=1}^L \Pi_{\ell+1,L} \times \|b_\ell - b'_\ell\|_q \right]^p d\mu(x) \right)^{\frac{1}{p}}. \end{aligned}$$

A trivial adaptation yields a similar result for  $p = \infty$ .

If  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$ ,  $N_0 = \min_{0 \leq \ell \leq L} N_\ell$ , and if  $\|\cdot\| = \|\cdot\|_q$  so that we can take  $c_0 := 1$ , then the previous inequality is an equality for the parameters defined in Equation (17).

Note that in the special case  $p = \infty$ , if we only assume that  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$  and  $\|\cdot\| = \|\cdot\|_q$  (but not that  $N_0 = \min_{0 \leq \ell \leq L} N_\ell$ ), denoting by  $N_{\min} := \min_{0 \leq \ell \leq L} N_\ell$ , then it holds for the parameters of Equation (17) and for every  $x \in \Omega$  supported on the first  $N_{\min}$  coordinates:

$$\|R_\theta(x) - R_{\theta'}(x)\| = \sum_{\ell=1}^L \Pi_{\ell+1,L} \left( \Pi'_{1,\ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} \Pi'_{k+1,\ell-1} \|b'_k\|_q \right) \times \|W_\ell - W'_\ell\|_q + \sum_{\ell=1}^L \Pi_{\ell+1,L} \times \|b_\ell - b'_\ell\|_q.$$

Recall that  $W = \max_{\ell=0, \dots, L} N_\ell$  is the width of the network. For every matrix  $M$  with input/output dimension bounded by  $W$  and every vector  $b$  with dimension bounded by  $W$ , denoting by  $\|M\|_{\max} := \max_{i,j} |M_{i,j}|$ , standard results on equivalence of norms guarantees that for every  $1 \leq q \leq \infty$ , it holds  $\|b\|_q \leq W^{1/q} \|b\|_\infty \leq W \|b\|_\infty$  and  $\max(\|M\|_1, \|M\|_\infty) \leq W \|M\|_{\max}$ . The latter, with Riesz-Thorin theorem [6, Chap.2, Thm 4.3], guarantee that for every  $1 \leq q \leq \infty$ :

$$\|M\|_q \leq W \|M\|_{\max} \text{ and } \|b\|_q \leq W \|b\|_\infty. \quad (19)$$

We deduce that for every  $\ell = 1, \dots, L$ :

$$\max \left( \|W_\ell - W'_\ell\|_q, \|b_\ell - b'_\ell\|_q \right) \leq W \|\theta - \theta'\|_\infty.$$

This time, this is not an equality for the parameters defined in Equation (17). For them it holds instead, assuming that all  $\lambda_\ell$  are equal:

$$\|W_\ell - W'_\ell\|_q = \varepsilon \lambda_\ell = \|W_\ell - W'_\ell\|_{\max} = \|\theta - \theta'\|_\infty, \|b_\ell - b'_\ell\|_q = 0.$$

Using the previous inequalities, we get for  $1 \leq p < \infty$ :

$$\begin{aligned} \|R_\theta - R_{\theta'}\|_{p, \|\cdot\|} &\leq \left( \int_{x \in \Omega} \left[ \sum_{\ell=1}^L \Pi_{\ell+1, L} \left( \Pi'_{1, \ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} \Pi'_{k+1, \ell-1} \|b'_k\|_q \right) \right. \right. \\ &\quad \left. \left. + \sum_{\ell=1}^L \Pi_{\ell+1, L} \right]^p d\mu(x) \right)^{\frac{1}{p}} c_0 W \|\theta - \theta'\|_\infty \end{aligned}$$

with a trivial adaptation for  $p = \infty$ . Now, let's specialize this for  $\theta, \theta' \in \Theta_{L, \mathbf{N}}^q(r)$ . It holds  $\max(\Pi_{i, j}, \Pi'_{i, j}) \leq r^{j-i+1}$  for  $i \leq j$ , and the same also holds for  $i = j + 1$  by definition of  $\Pi_{i, j}$ . Thus:

$$\begin{aligned} &\sum_{\ell=1}^L \Pi_{\ell+1, L} \left( 1 + \Pi'_{1, \ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} \Pi'_{k+1, \ell-1} \|b'_k\|_q \right) \\ &\leq \sum_{\ell=1}^L r^{L-\ell} \left( 1 + r^{\ell-1} \|x\|_q + \sum_{k=1}^{\ell-1} r^{\ell-k} \right) \text{ since } \theta, \theta' \in \Theta_{L, \mathbf{N}}^q(r) \\ &= Lr^{L-1} \|x\|_q + \sum_{\ell=1}^L r^{L-\ell} + \sum_{\ell=1}^L \sum_{k=1}^{\ell-1} r^{L-k} \\ &\leq Lr^{L-1} \|x\|_q + Lr^{L-1} + L(L-1)r^{L-1} \text{ since } r \geq 1 \\ &\leq L^2 r^{L-1} (\|x\|_q + 1) \text{ since } L \geq 1. \end{aligned}$$

If we define:

$$c := \begin{cases} c_0 \left( \int_{x \in \Omega} (\|x\|_q + 1)^p d\mu(x) \right)^{1/p} & \text{if } p < \infty, \\ c_0 \operatorname{ess\,sup}_{x \in \Omega} \|x\|_q + 1 & \text{if } p = \infty. \end{cases}$$

where we recognize in the second factor the constant  $C_p(\Omega, \mu)$  from Lemma II.1 when  $q = \infty$ , then we finally get (6). Let us now explicit  $c$  in specific situations where  $\Omega = [-D, D]^d$  for some  $D > 0$ ,  $\mu$  is the Lebesgue measure and  $\|\cdot\| = \|\cdot\|_q$  so that we can take  $c_0 = 1$ . If  $q = \infty$  we get  $c = C_p(\Omega, \mu) \leq (D+1)(2D)^{d/p}$ . If  $p = \infty$ , then

$$c = \operatorname{ess\,sup}_{x \in \Omega} \|x\|_q + 1 = Dd^{1/q} + 1.$$

Indeed, the essential supremum is actually a maximum in this case and  $\|x\|_q \leq d^{1/q} \|x\|_\infty \leq d^{1/q} D$  for every  $x \in [-D, D]^d$  with equality for  $x = (D, \dots, D)^T$ .

Let us now discuss the optimality of (6). It can be checked that if  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$ ,  $\|\cdot\| = \|\cdot\|_q$ , so that we can take  $c_0 := 1$ , and if  $N_0 = \min_{0 \leq \ell \leq L} N_\ell$ , then the parameters  $\theta, \theta'$  defined in Equation (17) with  $\lambda_1 = \dots = \lambda_L = \frac{r}{1+\varepsilon} \geq 0$  are in  $\Theta_{L, \mathbf{N}}^q(r)$  and satisfy  $\|R_\theta - R_{\theta'}\|_p = c_0 \left( \int_{x \in \Omega} \|x\|_q^p d\mu(x) \right)^{\frac{1}{p}} r^{L-1} \sum_{\ell=1}^L \left( \frac{1}{1+\varepsilon} \right)^{L-\ell} \|\theta - \theta'\|_\infty$ .

In the special case where  $p = \infty$ , if we only assume that  $\Omega \subseteq \mathbb{R}_+^{d_{\text{in}}}$  and  $\|\cdot\| = \|\cdot\|_q$  then the parameters  $\theta, \theta'$  defined in Equation (17) with  $\lambda_1 = \dots = \lambda_L = \frac{r}{1+\varepsilon} \geq 0$  are in  $\Theta_{L, \mathbf{N}}^q(r)$  and if we denote  $N_{\min} := \min_{0 \leq \ell \leq L} N_\ell$  and  $\Omega_{\min}$  the set of  $x \in \Omega$  supported on the first  $N_{\min}$  coordinates:

$$\operatorname{ess\,sup}_{x \in \Omega_{\min}} \|R_\theta(x) - R_{\theta'}(x)\| \geq \left( \operatorname{ess\,sup}_{x \in \Omega_{\min}} \|x\|_q \right) r^{L-1} \sum_{\ell=1}^L \left( \frac{1}{1+\varepsilon} \right)^{L-\ell} \|\theta - \theta'\|_\infty.$$

This yields the conclusion.

## APPENDIX E

### NEAREST-NEIGHBOUR UNIFORM QUANTIZATION ON RELU NETWORKS

*Proof of Theorem IV.2.* Consider  $\varepsilon, \eta > 0$  such that (9) holds true. We must prove that  $\min(r, \eta) \leq \frac{\varepsilon}{c'r^{L-1}}$ . With  $I_{m \times m}$  the identity matrix in dimension  $m$  and  $0_{m \times n}$  the  $m \times n$  matrix full of zeros, we introduce the following notation for “rectangular identity matrices”: for  $m < n$ , we set  $I_{m \times n} = (I_{m \times m}; 0_{m \times (n-m)})$ , while for  $m > n$  we set  $I_{m \times n} = I_{n \times m}^\top$ . Consider  $0 < a < \eta$  and define  $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$  with  $b_1 = \dots = b_L = 0$ ,  $W_1 = \lambda I_{N_1 \times N_0}$  with  $\lambda := \min(r, (\eta - a))$ , and for every layer  $\ell \geq 2$ ,  $W_\ell = r I_{N_\ell \times N_{\ell-1}}$ . Since  $0 < \lambda \leq \eta - a < \eta$ , we have  $Q_\eta(\lambda) = 0$  so that  $Q_\eta(W_1) = 0$ . Since  $b_1 = 0$ , we also have  $Q_\eta(b_1) = 0$  so that  $R_{Q_\eta(\theta)} = 0$ . We deduce that for every  $x \in [0, D]^d$  supported in the first  $N_{\min}$  coordinates:

$$\|R_\theta(x) - R_{Q_\eta(\theta)}(x)\|_q = \|\lambda r^{L-1} x - 0\|_q = \lambda r^{L-1} \|x\|_q.$$

Since the maximum of  $\|x\|_q$  over all  $x \in [0, D]^d$  supported in the first  $N_{\min}$  coordinates is  $c' = D N_{\min}^{1/q}$ , we get:

$$c' \lambda r^{L-1} \leq \max_{x \in [-D, D]^d} \|R_\theta(x) - R_{Q_\eta(\theta)}(x)\|_q$$

As  $\|W_1\|_q = \lambda = \min(r, (\eta - a)) \leq r$ , for every  $\ell \geq 2$ ,  $\|W_\ell\|_q = r$  and for every  $\ell \geq 1$ ,  $\|b_\ell\|_q = 0 \leq r$ , we have  $\theta \in \Theta_{L, \mathbf{N}}^q(r)$  so (9) applies. This implies  $c' \lambda r^{L-1} \leq \varepsilon$ , i.e.,  $\min(r, (\eta - a)) \leq \varepsilon / (c' r^{L-1})$ . This holds for every  $0 < a < \eta$ : taking the limit  $a \rightarrow 0^+$  yields the result.  $\square$

*Proof of Proposition IV.1.* Let us see that Lemma IV.1 applies with  $\Theta = \Theta_{L, \mathbf{N}}^{\max}(\varepsilon^{-k})$ ,  $q = \infty$  and  $r = W \varepsilon^{-k}$ . First, it holds  $\Theta_{L, \mathbf{N}}^{\max}(\varepsilon^{-k}) \subset \Theta_{L, \mathbf{N}}^\infty(W \varepsilon^{-k})$  (see Remark III.1). Since  $\|\theta - \theta'\|_\infty \leq \eta/2 \leq \varepsilon^m/2$ , the  $\eta$  to use for Lemma IV.1 is  $\eta := \varepsilon^m/2 > 0$ . Lemma IV.1 then gives the result if  $\varepsilon^m/2 \leq \varepsilon (c W L^2 r^{L-1})^{-1}$ , where  $c := 1 + D d_{\text{in}}^{1/q} = 1 + D$ . The latter condition is equivalent to  $c W L^2 r^{L-1}/2 \leq \varepsilon^{m-1}$ . Recall that  $r = W \varepsilon^{-k}$  and  $\max(W, L) \leq \varepsilon^{-k}$ . Thus the left hand-side satisfies:  $c W L^2 r^{L-1}/2 \leq ((1+D)/2) \varepsilon^{-2kL-2k}$ . We can conclude if  $((1+D)/2) \varepsilon^{-2kL-2k} \leq \varepsilon^{m-1}$ . By definition of  $m$ , this is true as soon as  $((1+D)/2) \varepsilon^{\lceil \log_2(D) \rceil} \leq 1$ . This is clear when  $0 < D \leq 1$ . While for  $D > 1$ ,  $((1+D)/2) \varepsilon^{\lceil \log_2(D) \rceil} \leq 1$  holds if and only if  $\log_2(\varepsilon) \leq -\frac{\log_2((1+D)/2)}{\log_2(D)}$ . Since  $1 < D$ , it holds  $\frac{1+D}{2} \leq \frac{\lceil D \rceil + \lceil D \rceil}{2} = \lceil D \rceil$  so that  $-\frac{\log_2((1+D)/2)}{\log_2(D)} \geq -1$ . Since  $\varepsilon \in (0, 1/2)$ , it holds  $-1 \geq \log_2(\varepsilon)$ , hence  $-\frac{\log_2((1+D)/2)}{\log_2(D)} \geq \log_2(\varepsilon)$  and the result follows.  $\square$

*Proof of Proposition IV.2.* Using [18, Thm. 1], there exist constants  $c(n, d) > 0$  and  $r(n, d) > 1$  (for instance, a proof examination of [18, Thm. 1] shows that we can take  $r = \max(4, d + n)$ ) such that for every  $\varepsilon \in (0, 1)$ , there exists a ReLU network architecture  $(L, \mathbf{N})$  with depth  $L$  bounded by  $c \ln(1/\varepsilon)$ , a number of

weights at most equal to  $c\varepsilon^{-d/n} \ln(1/\varepsilon)$ , and such that for every  $f \in \mathcal{C}_{n,d}$ , there exists  $\theta \in \Theta_{L,\mathbf{N}}$  such that  $\|f - R_\theta\|_{L^\infty([0,1]^d)} \leq \varepsilon/2$ , and such that  $\theta$  has weight's magnitude bounded by  $r$ . Theorem III.1 can now be used to quantize the weights of  $\theta$ , in order to get a quantized ReLU network  $\varepsilon$ -close to  $f$ . Denote  $W$  the width of this network architecture  $(L, \mathbf{N})$ . Since  $\Theta_{L,\mathbf{N}}^{\max}(r) \subset \Theta_{L,\mathbf{N}}^1(Wr)$  (see Remark III.1) we can use Theorem III.1 with  $q = 1$  to get that there exists a constant  $c' > 0$  that only depends on  $n, d$ , such that the weights of any network  $\theta \in \Theta_{L,\mathbf{N}}^{\max}(r)$  can be uniformly quantized with a step size  $\eta := c'\varepsilon(WL^2(Wr)^{L-1})^{-1}$  to get a quantized network  $\theta'$  such that  $\|R_{\theta'} - R_\theta\|_{L^\infty([0,1]^d)} \leq \varepsilon/2$ . Since the width  $W$  is at most the number of weights, which is at most  $c\varepsilon^{-d/n} \ln(1/\varepsilon)$ , and since the depth  $L$  is at most  $c \ln(1/\varepsilon)$  and  $r$  is a constant that only depends on  $n, d$ , it is straightforward to check that  $\ln(1/\eta) \leq c'' \ln^2(1/\varepsilon)$  for some constant  $c''$  that only depends on  $n$  and  $d$ . Since the weights are bounded in absolute value by  $r(n, d)$ , this means that every quantized weight can be stored using at most  $c''' \ln(1/\eta) \leq c''' \ln^2(1/\varepsilon)$  bits for some constant  $c'''(n, d) > 0$ . Since there are at most  $c\varepsilon^{-d/n} \ln(1/\varepsilon)$  such quantized weights, this yields the result using  $\max(c, c'', c \times c''')$  as the final constant.  $\square$

## APPENDIX F

### APPROXIMATION SPEEDS OF QUANTIZED VERSUS UNQUANTIZED RELU NETWORKS

We first establish two lemmas that will be useful to prove Theorem V.1. Along the way, we also give bounds on the size of the coverings we encounter. These bounds will prove useful in appendix I.

**Lemma F.1.** — Let  $(\mathcal{F}, d)$  be a metric space. Consider  $\gamma$  and two sequences  $\Sigma(\gamma)$  and  $\Sigma$  of subsets of  $\mathcal{F}$ . Assume that there exists a constant  $c > 0$  such that for every  $M \in \mathbb{N}$ , the set  $\Sigma(\gamma)_M$  is a  $cM^{-\gamma}$ -covering of  $\Sigma_M$ . Then for every (non-empty)  $\mathcal{C} \subset \mathcal{F}$ :

$$\begin{aligned} \gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) &= \gamma^{*\text{approx}}(\mathcal{C}|\Sigma) & \text{if } \gamma \geq \gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \\ \gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) &\geq \gamma & \text{otherwise.} \end{aligned}$$

*Proof of Lemma F.1.* For every  $M \in \mathbb{N}$ , the inclusion  $\Sigma(\gamma)_M \subset \Sigma_M$  holds (indeed  $\Sigma(\gamma)_M$  is a covering of  $\Sigma_M$ ) so that  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) \leq \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ . This proves the result when  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = -\infty$ . From now on we assume  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) > -\infty$ . Fix an arbitrary  $-\infty < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$ . By definition of the approximation speed, there exists a constant  $c' > 0$  such that for every  $f \in \mathcal{C}$  and every  $M \in \mathbb{N}$ , there exists a function  $\Phi_M(f) \in \Sigma_M$  that satisfies:

$$d(f, \Phi_M(f)) \leq c'M^{-\gamma'}.$$

The triangle inequality guarantees that for every  $f \in \mathcal{C}$  and every  $M \in \mathbb{N}$ :

$$d(f, \Sigma(\gamma)_M) \leq d(f, \Phi_M(f)) + d(\Phi_M(f), \Sigma(\gamma)_M) \leq c'M^{-\gamma'} + cM^{-\gamma}.$$

Since  $\gamma' \leq \gamma$  (and even if  $\gamma' < 0$ , which can happen if  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) < 0$ ) this means that  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) \geq \gamma'$  for every  $-\infty < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$  so  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) \geq \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$ . Since we also proved that  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma)) \leq \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ , this yields the claim.  $\square$



**Lemma F.2.** — Consider  $q \in [1, \infty]$  and  $\gamma > 0$ . There exists a constant  $c(q, \gamma) > 0$  such that the following holds. Consider arbitrary  $n \in \mathbb{N}$ ,  $r \geq 1$  and consider the set  $B_{n, \|\cdot\|_q}(0, r) \subset \ell^q(\mathbb{N})$  that consists of the sequences bounded in  $\ell^q$ -norm by  $r$ , and with zero coordinates outside the first  $n$  ones. Consider a metric space  $(\mathcal{F}, d)$  and a Lipschitz-map  $\varphi : (B_{n, \|\cdot\|_q}(0, r), \|\cdot\|_q) \rightarrow (\mathcal{F}, d)$  with Lipschitz constant  $\text{Lips}(\varphi) \geq 1$ . For every  $M \in \mathbb{N}$ , define the step size  $\eta_M := (M^\gamma n^{1/q} \text{Lips}(\varphi))^{-1}$  and the "quantized" set  $\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty) := B_{n, \|\cdot\|_q}(0, r) \cap (\eta_M \mathbb{Z})^{\mathbb{N}}$ . Then for every integer  $M \geq 2$ , the set  $\varphi(\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty))$  is an  $M^{-\gamma}$ -covering of  $\varphi(B_{n, \|\cdot\|_q}(0, r))$  of size satisfying:

$$\log_2(|\varphi(\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty))|) \leq c(q, \gamma) \left( n \left[ \log_2(n) + \log_2(r) + \log_2(\text{Lips}(\varphi)) + \log_2(M) \right] \right). \quad (20)$$

*Proof.* When  $q = \infty$ , it is known [17, Examples 5.2 and 5.6] that  $\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty)$  is a  $\eta_M$ -covering of  $B_{n, \|\cdot\|_q}(0, r)$  of size bounded by  $(2r/\eta_M)^n + 1$ . Since  $\varphi$  is  $\text{Lips}(\varphi)$ -Lipschitz, we deduce that the set  $\varphi(\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty))$  is an  $M^{-\gamma}$ -covering of  $\varphi(B_{n, \|\cdot\|_q}(0, r))$  of size satisfying:

$$\log_2(|\varphi(\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty))|) \leq n \left[ 1 + \log_2(r) + \log_2(\text{Lips}(\varphi)) + \gamma \log_2(M) \right]$$

Since  $M \geq 2$ , it holds  $1 + \gamma \log_2(M) \leq (1 + \gamma) \log_2(M)$ , hence Equation (20) for  $c(q, \gamma) = 1 + \gamma \geq 1$ . This settles the case  $q = \infty$ .

When  $q \in [1, \infty)$ , Hölder's inequality yields  $\|x\|_q \leq n^{1/q} \|x\|_\infty$  for every  $x \in \mathbb{R}^n$ . Thus  $B_{n, \|\cdot\|_q}(0, r)$  is a subset of the ball of radius  $rn^{1/q}$  of  $\ell^\infty(\mathbb{N})$ , and the Lipschitz constant of  $\varphi$  with respect to  $\|\cdot\|_\infty$  is bounded by its Lipschitz constant with respect to  $\|\cdot\|_q$ , up to a factor  $n^{1/q}$ . Thus, the case  $q \in [1, \infty)$  can be reduced to the case  $q = \infty$  by replacing  $r$  by  $rn^{1/q}$  and  $\text{Lips}(\varphi)$  by  $n^{1/q} \text{Lips}(\varphi)$ . We get:

$$\log_2(|\varphi(\mathcal{Q}(B_{n, \|\cdot\|_q}(0, r), \eta_M, \infty))|) \leq n \left[ 1 + \frac{2}{q} \log_2(n) + \log_2(r) + \log_2(\text{Lips}(\varphi)) + \gamma \log_2(M) \right]$$

This yields the desired result with  $c(q, \gamma) = \max(\frac{2}{q}, 1 + \gamma)$ .  $\square$

**Lemma F.3.** — In the setting of Theorem V.1, for every  $\gamma > 0$ , there exists a constant  $c > 0$  such that for every  $M \in \mathbb{N}$ , the set  $\mathcal{Q}(\mathcal{N}|\gamma)_M$  is a  $cM^{-\gamma}$ -covering of  $\mathcal{N}$ . Moreover, for every  $M \in \mathbb{N}$ , every architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$  and every support  $S \in S_{(L, \mathbf{N})}^M$ :

$$\log_2(|R_{\mathcal{Q}(\Theta_{L, \mathbf{N}}^q(r_M), \eta_M, r_M), S}|) \leq cM (\log_2(M) + \log_2(r_M) + \log_2(\text{Lips}(M, q))).$$

*Proof.* According to Theorem III.1, there is a constant  $c' > 0$  such that for each  $M \in \mathbb{N}$ , each architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$  and each support  $S \in S_{(L, \mathbf{N})}^M$ , the set  $R_{\Theta_{L, \mathbf{N}}^q(r_M), S}$  is the image under a Lipschitz map of  $(\{\theta \in \Theta_{L, \mathbf{N}}^q(r_M) \text{ supported on } S\}, \|\cdot\|_\infty)$  with a Lipschitz constant bounded by  $c' \text{Lips}(M, q)$ .

Let us now use Lemma F.2 with a Lipschitz constant bounded by  $c' \text{Lips}(M, q)$ ,  $n = |S| \leq M$  the cardinality of the support,  $r = r_M$ , and the same  $q$  as here. This yields the result with  $c := \max(1/c', 2c(q, \gamma))$ .  $\square$

*Proof of Theorem V.1.* Combining Lemma F.1 and Lemma F.3 gives Equality (10).  $\square$

## APPENDIX G

## ENCODABILITY IMPLIES A RELATION BETWEEN APPROXIMATION AND ENCODING SPEEDS

*Proof of Theorem VI.1.* If  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leq 0$  then the result is trivial since we always have  $\gamma^{*\text{encod}}(\mathcal{C}) \geq 0$ . In the rest of the proof we assume  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) > 0$ . Fix  $0 < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$  and  $h > 0$ . First,  $\Sigma$  is  $\gamma$ -encodable so there exists a  $(\gamma, h)$ -encoding of  $\Sigma$  that we denote  $\Sigma(\gamma, h)$ . This means that there exist constants  $c'_1, c'_2 > 0$  such that for every  $M \in \mathbb{N}$ , the set  $\Sigma(\gamma, h)_M$  is a  $c'_1 M^{-\gamma}$ -covering of  $\Sigma_M$  of size  $|\Sigma(\gamma, h)_M| \leq 2^{c'_2 M^{1+h}}$ . Second, since  $0 < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$ , the definition of the approximation speed guarantees that there exists a constant  $c'_3 > 0$  such that for every  $f \in \mathcal{C}$  and every  $M \in \mathbb{N}$ , there exists a function  $\Phi_M(f) \in \Sigma_M$  that satisfies:

$$d(f, \Phi_M(f)) \leq c'_3 M^{-\gamma'}.$$

Since  $0 < \gamma' < \gamma$ , note that for every  $M \in \mathbb{N}$ , it holds  $c'_1 M^{-\gamma} + c'_3 M^{-\gamma'} \leq (c'_1 + c'_3) M^{-\gamma'}$ . Define  $c_1 = c'_1 + c'_3$  and  $c_2 = c'_2$ . We deduce that for every  $M \in \mathbb{N}$ , the set  $\Sigma(\gamma, h)_M$  is a  $c_1 M^{-\gamma'}$ -covering of  $\mathcal{C}$  of size  $|\Sigma(\gamma, h)_M| \leq 2^{c_2 M^{1+h}}$ . Now, for every  $\varepsilon > 0$ , the integer  $M_\varepsilon := \left\lceil \left(\frac{c_1}{\varepsilon}\right)^{1/\gamma'} \right\rceil$  satisfies  $\varepsilon \geq c_1 M_\varepsilon^{-\gamma'}$ . By monotonicity of the metric entropy  $H(\mathcal{C}, d, \cdot)$  we get  $H(\mathcal{C}, d, \varepsilon) \leq H(\mathcal{C}, d, c_1 M_\varepsilon^{-\gamma'}) \leq c_2 M_\varepsilon^{1+h}$ . Note that for  $0 < \varepsilon < c_1$ , denoting by  $c = (2c_1^{1/\gamma'})^{1+h}$  it holds  $M_\varepsilon^{1+h} \leq \left(1 + \left(\frac{c_1}{\varepsilon}\right)^{1/\gamma'}\right)^{1+h} = \left(\frac{c_1}{\varepsilon}\right)^{(1+h)/\gamma'} \left(1 + \left(\frac{\varepsilon}{c_1}\right)^{1/\gamma'}\right)^{1+h} \leq c \varepsilon^{-(1+h)/\gamma'}$ . Finally for every  $0 < \varepsilon < c_1$ , it holds

$$H(\mathcal{C}, d, \varepsilon) \leq c \varepsilon^{-(1+h)/\gamma'},$$

As a direct consequence of Equation (2), this implies  $\gamma^{*\text{encod}}(\mathcal{C}) \geq \frac{\gamma'}{1+h}$  for every  $h > 0$  and every  $0 < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma)$ , hence the desired result.  $\square$

## APPENDIX H

FIRST EXAMPLES OF  $\infty$ -ENCODABLE APPROXIMATION FAMILIES

*Proof of Lemma VII.1.* Each  $\Sigma_M$  can be identified with the closed ball of radius  $r_M$  in dimension  $d_M$  with respect to the  $q$ -th norm, so that standard bounds on covering numbers [17, Eq. (5.9)] yield for every  $0 < \varepsilon \leq r_M$ :

$$d_M \log_2 \left( \frac{r_M}{\varepsilon} \right) \leq H(\Sigma_M, \|\cdot\|_q, \varepsilon) \leq d_M \log_2 \left( \frac{3r_M}{\varepsilon} \right). \quad (21)$$

For  $\varepsilon = M^{-\gamma} (\leq 1 \leq r_M)$ , we get:

$$d_M (\log_2(r_M) + \gamma \log_2(M)) \leq H(\Sigma_M, \|\cdot\|_q, \varepsilon) \leq d_M (\log_2(3r_M) + \gamma \log_2(M)).$$

Everything is non-negative, so if the right hand-side is  $\mathcal{O}_{M \rightarrow \infty}(M^{1+h})$ , for every  $h > 0$ , then so is the left hand-side. The converse is also true since both sides only differ by  $\log_2(3)d_M = \mathcal{O}_{M \rightarrow \infty}(d_M \log M)$ . The non-negativity of the quantities also implies that the condition  $d_M [\log_2(r_M) + \gamma \log_2(M)] = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$ , for every  $h > 0$ , does not depend on  $\gamma$ . As a consequence, either  $\Sigma$  is  $\infty$ -encodable or it is never  $\gamma$ -encodable, whatever  $\gamma > 0$  is. Finally, note that for every  $h > 0$ ,  $d_M (\log_2(r_M) + \log_2(M)) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$  if and only if  $d_M (\log_2(r_M) + 1) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$ . The "only if" part is clear since for  $M \geq 2$ , it holds  $0 \leq$

$d_M (\log_2(r_M) + 1) \leq d_M (\log_2(r_M) + \log_2(M))$ . For the "if" part, use that  $r_M \geq 1$  and the assumption to get  $0 \leq d_M \leq d_M (\log_2(r_M) + 1) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$  so that  $d_M \log_2(M) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h} \log_2(M)) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$ .  $\square$

*Proof of Theorem VII.1.* Fix an arbitrary  $\gamma > 0$ . Lemma F.2 guarantees that for every integer  $M \geq 2$ , the set  $\varphi_M(\mathcal{Q}(\Sigma_M, \eta_M(\gamma), \infty))$  is an  $M^{-\gamma}$ -covering of  $\varphi_M(\Sigma_M)$  of size satisfying:

$$\log_2(|\varphi_M(\mathcal{Q}(\Sigma_M, \eta_M(\gamma), \infty))|) \leq c(q, \gamma) \left( d_M \left[ \log_2(d_M) + \log_2(r_M) + \log_2(\text{Lips}(\varphi_M)) + \log_2(M) \right] \right).$$

Since  $0 \leq d_M \leq d_M (\log_2(r_M) + \log_2(\text{Lips}(\varphi_M)) + 1)$ , Assumption (11) guarantees that for every  $h > 0$ , it holds  $d_M = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$  so that  $d_M (\log_2(d_M) + \log_2(M)) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h} (\log(M^{1+h}) + \log_2(M))) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$ . As a consequence, for every  $h > 0$ , it holds  $\log_2(|\varphi_M(\mathcal{Q}(\Sigma_M, \eta_M(\gamma), \infty))|) = \mathcal{O}_{M \rightarrow \infty}(M^{1+h})$  so that the sequence  $\mathcal{Q}(\varphi(\Sigma)|\gamma)$  is a  $(\gamma, h)$ -encoding of  $\varphi(\Sigma)$ . This shows that  $\varphi(\Sigma)$  is  $\gamma$ -encodable for every  $\gamma > 0$ , so it is  $\infty$ -encodable.  $\square$

## APPENDIX I

### ENCODABILITY OF APPROXIMATION FAMILIES DEFINED WITH DICTIONARIES AND RELU NETWORKS

*Proof of Proposition VII.1.* Fix  $M \in \mathbb{N}$  and  $f = \sum_{i=1}^M c_i e_i \in \Sigma_M^q$ . Let  $0 < \lambda < 1$ . Define  $Q_\lambda(f) := \sum_{i=1}^M \text{sign}(c_i) \lfloor \frac{c_i}{\lambda} \rfloor \lambda e_i$  with  $\text{sign}(x) = 1$  if  $x \geq 0$ ,  $-1$  otherwise. It is proven in [12, Prop. 6] that there exists a constant  $c(p, q) > 0$  that only depends on  $p$  and  $q$  such that:

$$\|f - Q_\lambda(f)\|_{\mathcal{F}} \leq c(p, q) \lambda^{1-q/p} \sup_{i \in \mathbb{N}} \|e_i\|_{\mathcal{F}}.$$

Moreover, it is proven in [12, Lem. 4 and proof of Prop. 11] that the family  $(Q_\lambda(f))_{f \in \Sigma_M^q}$  has at most  $2^{\lambda^{-q}(1 - \log_2(\lambda) + \log_2(M))}$  elements. Setting  $\varepsilon = \lambda^{1-q/p}$ , and observing that  $\lambda^{-q} = \varepsilon^{-1/s}$ , this proves that the family  $(Q_\lambda(f))_{f \in \Sigma_M^q}$  is a  $\mathcal{O}_{\varepsilon \rightarrow 0}(\varepsilon)$ -covering of  $\Sigma_M^q$  of size  $\mathcal{O}_{\varepsilon \rightarrow 0}(2^{\varepsilon^{-1/s}(\log_2 1/\varepsilon + \log_2 M)})$ , with constants independent of  $M$ . For every  $M \in \mathbb{N}$ , using the above result with  $\varepsilon = M^{-s}$  proves that  $\Sigma^q$  is  $s$ -encodable.  $\square$

*Proof of Theorem VII.2.* Consider  $c > 0$ . We first prove that  $\tilde{\Sigma}^{\pi, c}$  is  $\infty$ -encodable. Consider  $M \in \mathbb{N}$ ,  $\mathcal{I}_M := \{I \subset \{1, \dots, \pi(M)\}, |I| \leq M\}$ , and define for each  $I \in \mathcal{I}_M$  the set  $\tilde{\Sigma}^{\pi, c}(I) := \{\sum_{i \in I} \tilde{c}_i \tilde{\phi}_i^I, (\tilde{c}_i)_{i \in I} \in [-c, c]^I\}$ . It holds:

$$\tilde{\Sigma}_M^{\pi, c} = \bigcup_{I \in \mathcal{I}_M} \tilde{\Sigma}^{\pi, c}(I).$$

Since each  $I \in \mathcal{I}_M$  is a set of at most  $M$  integers between 1 and  $\pi(M)$ , one can describe each such set by  $M$  sequences of at most  $\log_2(\pi(M))$  bits so that there are at most  $2^M \pi(M)$  such sets. Moreover, the set  $\tilde{\Sigma}^{\pi, c}(I)$  is the image of  $\varphi_{M, I} : (\tilde{c}_i)_{i \in I} \in ([-c, c]^I, \|\cdot\|_2) \mapsto \sum_{i \in I} \tilde{c}_i \tilde{\phi}_i^I \in \mathcal{F}$ . This map is 1-Lipschitz (since  $(\tilde{\phi}_i^I)_{i \in I}$  is orthonormal). Equation (20) of Lemma F.2 with  $n = |I| \leq M$ ,  $q = \infty$  and  $r = \max(c, 1)$  proves that  $\tilde{\Sigma}^{\pi, c}(I)$  has an  $M^{-\gamma}$ -covering with at most  $2^{2c(q, \gamma)M \log_2(rM)}$  elements. Taking the union of such covers for each  $I \in \mathcal{I}_M$ , we end up with an  $M^{-\gamma}$ -covering of the whole set  $\tilde{\Sigma}_M^{\pi, c}$  with at most  $2^M \pi(M) 2^{2c(q, \gamma)M \log_2(rM)} = \mathcal{O}_{M \rightarrow \infty}(2^{M \log M})$  elements. This proves the  $\infty$ -encodability of  $\tilde{\Sigma}^{\pi, c}$ .

It now remains to prove Equation (13). First, for every  $c > 0$  and every  $M \in \mathbb{N}$ , it holds  $\tilde{\Sigma}_M^{\pi,c} \subset \Sigma_M^\pi$  so that  $\Sigma^\pi$  approximates  $\mathcal{C}$  at least as quickly as  $\tilde{\Sigma}^{\pi,c}$ , that is  $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi) \geq \gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c})$ . As we now prove, there is actually equality for  $c = \sup_{f \in \mathcal{C}} \sup_{M \in \mathbb{N}} \max_{I \in \mathcal{L}_M} \max_{i \in I} |\langle f, \tilde{\phi}_i^I \rangle_{\mathcal{F}}|$  (and thus for any larger  $c$  since  $\gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c})$  is non-decreasing in  $c$ ). Note that by Cauchy-Schwarz,  $c \leq \sup_{f \in \mathcal{C}} \|f\|_{\mathcal{F}}$  which is finite since  $\mathcal{C}$  is bounded. If  $f \in \mathcal{C}$ , then for every  $M \in \mathbb{N}$ , every  $I \subset \{1, \dots, \pi(M)\}$ ,  $|I| \leq M$ , and every  $(c_i)_{i \in I} \in \mathbb{R}^I$ , it holds:

$$d(f, \tilde{\Sigma}_M^{\pi,c}) \leq \|f - \sum_{i \in I} \langle f, \tilde{\phi}_i^I \rangle_{\mathcal{F}} \tilde{\phi}_i^I\|_{\mathcal{F}} \leq \|f - \sum_{i \in I} c_i \phi_i\|_{\mathcal{F}}.$$

This implies that  $d(f, \tilde{\Sigma}_M^{\pi,c}) \leq d(f, \Sigma_M^\pi)$ . As a consequence,  $\tilde{\Sigma}^{\pi,c}$  approximates  $\mathcal{C}$  at least as quickly as  $\Sigma^\pi$ , that is  $\gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c}) \geq \gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi)$ . This yields equality (13).  $\square$

*Proof of Proposition VII.2.* By definition of  $\infty$ -encodability, we have to prove that for every  $\gamma > 0$  and for every  $h > 0$ , the quantized sequence  $\mathcal{Q}(\mathcal{N}|\gamma)$  is a  $(\gamma, h)$ -encoding of  $\mathcal{N}$ . Fix  $\gamma > 0$ . Lemma F.3 proves that there exists a constant  $c > 0$  such that for every  $M \in \mathbb{N}$ , the set  $\mathcal{Q}(\mathcal{N}_M|\gamma)$  is a  $cM^{-\gamma}$ -covering of  $\mathcal{N}_M$ , and for each  $M \in \mathbb{N}$ , each architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$  and each support  $S \in S_{(L, \mathbf{N})}^M$ , the quantized set  $R_{\mathcal{Q}(\Theta_{L, \mathbf{N}}^q(r_M), \eta_M, r_M), S}$  has a number of elements that satisfies:

$$\log_2(|R_{\mathcal{Q}(\Theta_{L, \mathbf{N}}^q(r_M), \eta_M, r_M), S}|) \leq cM (\log_2(r_M) + \log_2(\text{Lips}(M, q)) + \log_2(M)).$$

Fix  $h > 0$ . By assumption (14) of Proposition VII.2, we deduce that there exists  $c' = c'(h) > 0$  such that for every  $M \in \mathbb{N}$ , and every architecture  $(L, \mathbf{N}) \in \mathbf{A}_M$ , we have

$$M (\log_2(r_M) + \log_2(\text{Lips}(M, q)) + \log_2(M)) \leq c' M^{1+h}.$$

Thus, the quantized set  $\mathcal{Q}(\mathcal{N}_M|\gamma)$  is a  $cM^{-\gamma}$ -covering of  $\mathcal{N}_M$  and its cardinality satisfies

$$|\mathcal{Q}(\mathcal{N}_M|\gamma)| \leq \sum_{(L, \mathbf{N}) \in \mathbf{A}_M} \sum_{S \in S_{(L, \mathbf{N})}^M} |R_{\mathcal{Q}(\Theta_{L, \mathbf{N}}^q(r_M), \eta_M, r_M), S}| \leq |\mathbf{A}_M| \cdot |S_{(L, \mathbf{N})}^M| \cdot 2^{cc' M^{1+h}}.$$

Note that for every  $M \in \mathbb{N}$ ,  $|\mathbf{A}_M| \leq L_M M^{L_M - 1}$  (at most  $L_M$  possibilities for the depth and then,  $M$  possibilities for each of the potential  $L_M - 1$  intermediary layers, the size of the input and output being fixed to  $d_{\text{in}}$  and  $d_{\text{out}}$ ). Similarly, since  $S_{(L, \mathbf{N})}^M$  consists at most of all the supports of size  $M$  in dimension  $d_{(L, \mathbf{N})} \leq 2M^2 L_M$ , its cardinality is bounded by  $(2M^2 L_M)^M$ . Overall, we obtain that

$$\log_2(|\mathcal{Q}(\mathcal{N}_M|\gamma)|) \leq \log_2(L_M) + L_M \log_2(M) + M \log_2(2M^2 L_M) + cc' M^{1+h}.$$

Using assumption (14) again, we obtain that there exists  $c'' > 0$  such that  $\log_2(|\mathcal{Q}(\mathcal{N}_M|\gamma)|) \leq c'' M^{1+h}$  for every  $M \in \mathbb{N}^*$ . We deduce that for every  $\gamma > 0$  and for every  $h > 0$ , the sequence  $\mathcal{Q}(\mathcal{N}|\gamma)$  is a  $(\gamma, h)$ -encoding of  $\mathcal{N}$ . This yields the result.  $\square$