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

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti

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

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti

Abstract

We consider general approximation families encompassing ReLU neural networks.

On the one hand, we introduce a new property, that we call ∞-encodability, which lays a framework that we use (i) to guarantee that ReLU networks can be uniformly quantized and still have approximation speeds comparable to unquantized ones, and (ii) to prove that ReLU networks share a common limitation with many other approximation families: the approximation speed of a set $\mathcal{C}$ is bounded from above by an encoding complexity of $\mathcal{C}$ (a complexity well-known for many $\mathcal{C}$'s). The property of ∞-encodability allows us to unify and generalize known results in which it was implicitly used.

On the other hand, we give lower and upper bounds on the Lipschitz constant of the mapping that associates the weights of a network to the function they represent in L^p. It is given in terms of the width, the depth of the network and a bound on the weight's norm, and it is based on well-known upper bounds on the Lipschitz constants of the functions represented by ReLU networks. This allows us to recover known results, to establish new bounds on covering numbers, and to characterize the accuracy of naive uniform quantization of ReLU networks.

Index Terms

Approximation speed, encoding speed, ReLU neural networks, quantization, Lipschitz parameterization, covering numbers.

I. Introduction

Neural networks are used with success in many applications to approximate functions. In line with the works [5], [9], [11], we are interested in understanding their approximation power in practice and in theory. Regarding practical applications, a key question is to be able to compare approximation properties of

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.

quantized versus unquantized neural networks. Another important question is to better understand non-trivial situations where neural networks can be expected (or not) to have better approximation properties than the best known approximation families, quantized or not.

We address these questions by quantitatively characterizing the optimal polynomial speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ at which all functions of a metric space $\mathcal{C}$ can be approximated by a sequence $\Sigma = (\Sigma_M)_{M\in\mathbb{N}^*}$ (with $\mathbb{N}^* = \{1, 2, \dots\}$) of sets Σ_M of "simpler" functions, such as polynomials of degree M, or ReLU neural networks with M non-zero parameters.

Notion of ∞-encodability. We introduce a new property of the sequence Σ, called ∞-encodability, which builds a bridge between approximation and encoding speeds. This property forbids degenerate cases and notably holds for sets Σ_M of linear combinations from a basis in a Banach space, under standard assumptions limiting the growth with M of the coefficients' magnitude and the number of combined basis functions. We show that:

(i) if Σ is ∞-encodable, then $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ is *bounded from above by the Kolmogorov-Donoho complexity* $\gamma^{*\text{encod}}(\mathcal{C})$, which measures the best polynomial speed at which $\mathcal{C}$ can be *encoded* as binary sequences, and which is known for many classical functions sets such as balls of Sobolev spaces;

(ii) many sequences $\Sigma = (\Sigma_M)_{M\in\mathbb{N}^*}$ are ∞-encodable: when Σ_M contains M-terms linear combinations of the first $\text{poly}(M)$ elements[1] of a bounded dictionary, with boundedness conditions on the coefficients, or when Σ_M is Lipschitz-parameterized by some (polynomially) bounded set in finite dimension, (*i.e.*, , it is the image of such a set by a Lipschitz map). The latter includes the case of ReLU neural networks for which we identify "simple" sufficient conditions on the considered architectures for it to hold;

(iii) sequences $\Sigma = (\Sigma_M)_{M\in\mathbb{N}^*}$ of Lipschitz-parameterized sets Σ_M as above are not only ∞-encodable: they can be *uniformly quantized*, in the sense that they can be covered with balls centered on a uniform grid, into sequences that reach comparable approximation speeds on *every* set $\mathcal{C}$ to their unquantized version.

Point (i) is concisely expressed by the following bound on the optimal polynomial approximation speed:

$$\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leqslant \gamma^{*\text{encod}}(\mathcal{C}). \tag{1}$$

In light of point (ii), this bound unifies and generalizes previous results such as [9, Thm. V.3, Thm. VI.4][11, Thm. 5.24] thanks to the notion of ∞-encodability. Inequality (1) is of particular interest in order to bound the approximation speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ from above without looking at all at the approximation properties of the set $\mathcal{C}$ by the sequence Σ. Instead, we can study separately Σ and establish at which speed it can be *encoded*. This inequality happens to be an equality in various cases, see Example 11. Another consequence of this inequality is that approximation families based on an ∞-encodable sequence Σ defined with ReLU neural networks share a common upper bound on approximation rates with other classical approximation families that we prove to be ∞-encodable. In particular, given $\mathcal{C}$, if an ∞-encodable sequence Σ is known

[1] $\text{poly}(M)$ denotes a positive function that grows at most polynomially in M

such that $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = \gamma^{*\text{encod}}(\mathcal{C})$, then no improved approximation rate can be hoped for by using ReLU networks.

The case of ReLU networks. Given a function R_θ represented by a ReLU neural network with parameters θ, our results show how θ can be quantized into $Q(\theta)$ such that $\|R_\theta - R_{Q(\theta)}\|_p$ is small in a given L^p-space. This bounds one of the components of the error that is committed when approximating a function f by a neural network in practice. Indeed, the error committed can be decomposed into (at least) three terms: the *approximation error* (related to the question of finding θ such that R_θ, a ReLU neural network with continuous real parameters, approximates well f), the *quantization error* (related to the need of using an approximation $R_{Q(\theta)}$ of R_θ, for instance the representation of the ReLU neural network with floating-point parameters $Q(\theta)$) and the *evaluation error* (due to finite-precision computations and propagation of the errors when evaluating $R_{Q(\theta)}(x)$).

In light of (iii), in order to analyze the quantization error, we use the fact that (as already noticed in the literature [5, Rmk. 9.1]) the set of functions represented by ReLU neural networks is Lipschitz-parameterized under conditions on the sparsity, the depth and the weight's magnitude.

Lipschitz parameterization of ReLU networks. We give an explicit new upper bound for the Lipschitz constant of this parameterization $\theta \in \Theta \mapsto R_\theta$ in Proposition 36, for some bounded set Θ in finite dimension. The bound is based on the Lipschitz constant of the functions R_θ themselves. For instance, for real-valued functions with input in $\mathbb{R}^d$, $d \in \mathbb{N}^*$, given a width $W \in \mathbb{N}^*$, a depth $L \in \mathbb{N}^*$ and a bound $r \geqslant 1$ on the Euclidean norm of the matrices and biases vectors of the considered parameters, the associated ReLU networks define a parameterization $\theta \mapsto R_\theta \in L^p([0,1]^d)$, $p \geqslant 1$, such that for every parameters θ, θ' satisfying the above conditions:

$$\|R_\theta - R_{\theta'}\|_{L^p} \leqslant c_p W L^2 r^{L-1} \|\theta - \theta'\|_\infty,$$

where the only term that depends on L^p is the constant c_p, defined as the L^p-norm of the function $x \in [0,1]^d \mapsto \|x\|_2 + 1$. Conversely, denoting by c'_p the L^p-norm of the function $x \in [0,1]^d \mapsto \|x\|_2$, then for every $\varepsilon > 0$, we can exhibit parameters θ, θ' satisfying the above conditions and such that $\|R_\theta - R_{\theta'}\|_p \geqslant (1-\varepsilon) c'_p L r^{L-1} \|\theta - \theta'\|_\infty$.

This Lipschitz property, combined with (i), (ii) and (iii), yields guarantees on the approximation power of families of quantized ReLU networks, see Proposition 49. For instance, let $p \geqslant 1$ and $d \in \mathbb{N}^*$, and define $(\Sigma_M)_{M \in \mathbb{N}^*}$ such that $\Sigma_M \subset L^p([0,1]^d)$ is the set of functions represented by ReLU networks with at most $\mathcal{O}_{M\to\infty}(M)$ non-zero parameters, at most[2] $\mathcal{O}_{M\to\infty}(\text{polylog}(M))$ layers and with weights bounded in absolute value by $\mathcal{O}_{M\to\infty}(\text{poly}(M))$. Then for every set $\mathcal{C} \subset L^p([0,1]^d)$, and for every $\gamma \geqslant \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$, the networks representing the functions in Σ_M can be uniformly quantized with a step size of order $\mathcal{O}_{M\to\infty}(M^{-c \log M})$, for some $c = c(\gamma) > 0$ that does not depend on M (so that $\mathcal{O}_{M\to\infty}((\log M)^2)$ bits are enough to store each parameter), and still achieve the same polynomial rate as their unquantized version, see Example 51.

[2] $\text{polylog}(M)$ denotes any positive function of M that grows at most polynomially in $\log M$

We further use this upper bound on the Lipschitz constant of $\theta \mapsto R_\theta$ to bound from above the covering numbers of sets of functions represented by ReLU networks, and to characterize the minimax accuracy of uniformly quantized ReLU networks, see Corollary 39 and Section V-C.

Organization of the paper. The definition of ReLU neural networks is recalled in section II-A. The definitions of the approximation speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ and the encoding speed $\gamma^{*\text{encod}}(\mathcal{C})$ are recalled in section II-B. The notion of ∞-encodability is introduced in section III, it is then proven that being ∞-encodable implies Inequality (1). Examples of ∞-encodable sequences Σ are then given in section IV. Some sequences Σ Lipschitz-parameterized discussed in section IV are not only ∞-encodable: they can be *uniformly quantized*, in the sense that they can be covered with balls centered on a uniform grid, into sequences that reach comparable approximation speeds on *every* set $\mathcal{C}$ to their unquantized version, see Proposition 24. In Section V, ReLU neural networks are shown to be Lipschitz-parameterized. As a consequence, results of section IV apply: explicit conditions on the considered ReLU networks yield sequences Σ that are ∞-encodable and even uniformly quantizable into sequences that reach approximation speeds comparable to their unquantized version, see Proposition 49. We recall our main contributions and give some perspectives in section VI. Some useful definitions, technical results and their proofs are gathered in appendix.

II. Preliminaries

We recall the definition of ReLU neural networks in section II-A. Then, in section II-B we recall the definitions of (i) the optimal polynomial speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ at which all functions of a set $\mathcal{C}$, subset of a metric space $\mathcal{F}$, can be approximated by a sequence Σ, (ii) the Kolmogorov-Donoho complexity $\gamma^{*\text{encod}}(\mathcal{C})$, which measures the best polynomial asymptotic speed at which $\mathcal{C}$ can be encoded into binary sequences.

A. 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 1 (ReLU: Rectified Linear Unit). — The ReLU function ρ is defined by [2]:

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

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

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

Definition 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 ℓ has N_ℓ neurons, we call N_ℓ the width of layer ℓ. Layer 0 is the input layer while layer L is the output

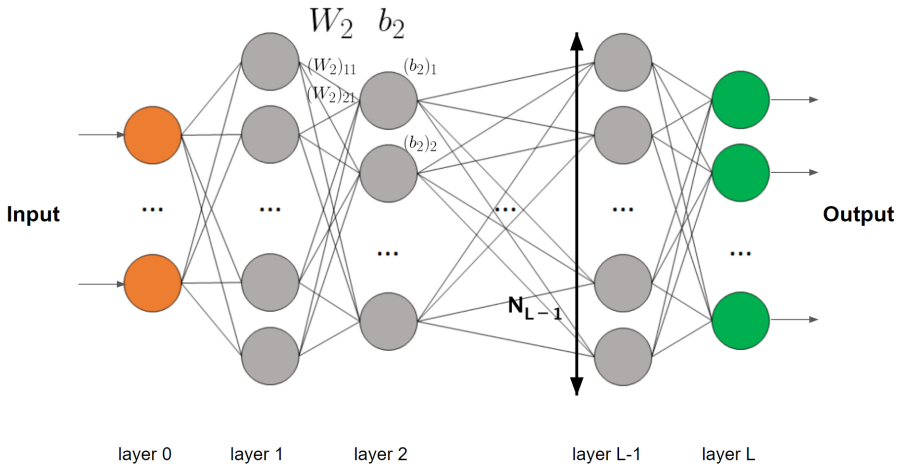

Figure 1. A neural network architecture can be seen as a directed graph. Neurons are represented by vertices, grouped by layers. For each neuron, there are edges going from this neuron to each neuron of the following layer. Coefficient (i, j) of W_ℓ can be seen as the weight of the edge going from neuron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

layer. Such an architecture can be represented as a graph, with a vertex for each neuron, and an arrow 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 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 $b_\ell \in \mathbb{R}^{N_\ell}$. It can be represented graphically: if neurons on layer ℓ are numbered from 1 to N_ℓ, then $(W_\ell)_{i,j}$ is the weight on the arrow going from neuron j of layer $\ell - 1$ to neuron i of layer ℓ, while $(b_\ell)_i$ is the weight bias on neuron i of layer ℓ, see Figure 1. Such a parameter θ lives in the parameter space

$$\begin{aligned} \Theta_{L,\mathbf{N}} &:= \mathbb{R}^{d_{(L,\mathbf{N})}}, \\ d_{(L,\mathbf{N})} &:= \sum_{\ell=1}^{L} N_\ell (N_{\ell-1} + 1). \end{aligned} \tag{2}$$

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

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

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

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

In words, the input x goes through each layer sequentially, and when it goes from layer ℓ 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 ρ applied coordinate-wise (except on the last layer where the ReLU function is not applied).

B. Approximation and encoding speeds

In the following, all subsets are non-empty by default. Let $\Sigma := (\Sigma_M)_{M\in\mathbb{N}^*}$ be a sequence of subsets of a metric space $(\mathcal{F}, d)$ (for example the set of all real-valued continuous functions on $[0,1]$, with d the uniform distance). Generally speaking, Σ is chosen so that Σ_M approximates functions of $\mathcal{F}$ better and better as M grows, and a typical example consists of a nested sequence $\Sigma_1 \subset \Sigma_2 \subset \dots$, where Σ_M is a function set parameterized by $\mathcal{O}_{M\to\infty}(M)$ parameters (for example the set of all polynomial functions of degree M). However, since quantizing the parameters associated to each Σ_M yields a new sequence that has no reason to remain nested, it is indeed important to deal with arbitrary sequences Σ. Considering a subset $\mathcal{C}$ of $\mathcal{F}$ (for example the set of 1-Lipschitz functions on $[0,1]$) one can wonder at which polynomial speed the sequence Σ approximates all functions of $\mathcal{C}$. This is captured by the notion of optimal polynomial speed $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma)$ defined as follows [9, Def. V.2, Def. VI.1].

Definition 5 (Approximation theoretic complexity $\gamma^{*\mathbf{approx}}(\mathcal{C}|\Sigma)$). — Let $\Sigma := (\Sigma_M)_{M\in\mathbb{N}^*}$ be an arbitrary sequence of (non-empty) subsets of $\mathcal{F}$. For any (non-empty) class of functions $\mathcal{C} \subset \mathcal{F}$, we can define the error $\varepsilon_M(\mathcal{C})$ of approximation of $\mathcal{C}$ by Σ_M as follows [9, Def. V.2, Def. VI.1]:

$$\varepsilon_M(\mathcal{C}) := \sup_{f\in\mathcal{C}} \inf_{\Phi\in\Sigma_M} d(f,\Phi) \in [0,+\infty].$$

Then, we define $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma)$ to be the supremum of the polynomial speeds at which all functions of $\mathcal{C}$ are approximated by Σ [9, Def. V.2, Def. VI.1]:

$$\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma) := \sup\{\gamma\in\mathbb{R}, \varepsilon_M(\mathcal{C}) = \mathcal{O}_{M\to\infty}\left(M^{-\gamma}\right)\} \in [-\infty,+\infty].$$

with the convention $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma) = -\infty$ if the supremum is over an empty set.

Remark 6. — In the following, the considered $\mathcal{C}$ will have bounded diameter $\sup_{f,g\in\mathcal{C}} d(f,g) < \infty$ (when $(\mathcal{F}, d)$ is a normed vector space, this is equivalent to assuming that $\mathcal{C}$ is bounded) ensuring that $\varepsilon_M(\mathcal{C})$ is finite for every M. In the classical case of nested sets Σ_M, the error $\varepsilon_M(\mathcal{C})$ is non-increasing with M hence $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma) \geqslant 0$. This also holds as soon as the error remains bounded, but examples with $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma) < 0$ can be built.

Remark 7. — In [11, Def. 5.23], the following quantity is considered

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

which satisfies $\gamma^*(\mathcal{C}|\Sigma) \geqslant \gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma)$ but generally differs from $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma)$ since in the definition of $\gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma)$, 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 [11, Proof of Thm 5.24] so that the two quantities coincide.

The speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ is known in some cases, see Example 11. We shall give a general framework under which $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ can be bounded from above by another measure of complexity of the set $\mathcal{C}$, which measures the best polynomial asymptotic speed at which $\mathcal{C}$ can be *encoded into binary sequences*. We first begin by recalling the notion of covering numbers and metric entropy [8].

Definition 8 (Covering, covering numbers and metric entropy). — Let $(\mathcal{C}, d)$ be a metric space. Consider $\varepsilon > 0$. A finite subset $X \subset \mathcal{C}$ is called an *ε-covering* of $\mathcal{C}$ if $\mathcal{C} \subset \bigcup_{x\in X} B_d(x, \varepsilon)$, where $B_d(x, \varepsilon)$ denotes the closed ball of $\mathcal{C}$, with respect to the metric d, centered in x and with radius ε. The *covering number* $N(\mathcal{C}, d, \varepsilon)$ is the minimal size of an ε-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 $H(\mathcal{C}, d, \varepsilon) := \log_2(N(\mathcal{C}, d, \varepsilon))$.

Definition 9 (Information theoretic complexity $\gamma^{*\text{encod}}(\mathcal{C})$). — Consider $\mathcal{F}$ a set equipped with a metric d and $\mathcal{C} \subset \mathcal{F}$. The Kolmogorov-Donoho complexity is defined [9, Def. IV.1] as:

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

with the convention that $\gamma^{*\text{encod}}(\mathcal{C}) = 0$ if the supremum is over an empty set.

Remark 10. — The quantity $\gamma^{*\text{encod}}(\mathcal{C})$ can also be defined in terms of encoder-decoder pairs. Define the optimal coding length [9, Def. IV.1] of $\mathcal{C}$ with accuracy ε by:

$$L(\mathcal{C}, d, \varepsilon) := \inf\left\{\ell \in \mathbb{N}, \exists E : \mathcal{C} \to \{0,1\}^\ell, \exists D : \{0,1\}^\ell \to \mathcal{F}, \sup_{f\in\mathcal{C}} d\left(f, D(E(f))\right) \leqslant \varepsilon\right\} \in \mathbb{N} \cup \{+\infty\} \tag{3}$$

with the convention that $L(\mathcal{C}, d, \varepsilon) = +\infty$ if the infimum is over an empty set. The metric entropy and the coding length satifies [8]:

$$H(\mathcal{C}, d, \varepsilon) \leqslant L(\mathcal{C}, d, \varepsilon) \leqslant H(\mathcal{C}, d, \varepsilon) + 1, \tag{4}$$

but they are not equal in general since the coding length is an integer. Inequality (4) shows that the metric entropy and the coding length have essentially the same behaviour when $\varepsilon \to 0$. Thus, the encoding speed $\gamma^{*\text{encod}}(\mathcal{C})$ can also be defined as the supremum over $\gamma \geqslant 0$ such that $L(\mathcal{C}, d, \varepsilon) = \mathcal{O}_{\varepsilon\to 0}\left(\varepsilon^{-1/\gamma}\right)$ as it is done in [9, Def. IV.1].

Example 11. — It is known [9, Table 1] that:

$\mathcal{C} :=$ unit ball of		Σ	$\gamma^{*\text{approx}}(\mathcal{C}\|\Sigma) = \gamma^{*\text{encod}}(\mathcal{C})$
α-Hölder	$C^\alpha([0,1])$	Wavelet basis	α
L^p-Sobolev[a]	$W_p^m([0,1]^d)$	Wavelet frame	$\frac{m}{d}$
Besov[b]	$B_{p,q}^m([0,1]^d)$	Wavelet frame	$\frac{m}{d}$

[a] where $p \in [1, \infty], m > d\max(1/p - 1/2, 0)$.

[b] where $p, q \in (0, \infty], m > d\max(1/p - 1/2, 0)$.

III. Encoding speeds *vs* approximation speeds

We start by a definition that captures, as shown in Proposition 16 below, the essence of several known results ([9, Thm. V.3, Thm. VI.4][11, Thm. 5.24][13, Prop. 11]).

A. The notion of γ-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 Σ 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 ε-approximated (with respect to d) by an element of Σ_M. But Σ_M can be ε-covered (with respect to the metric d) with $N(\Sigma_M, d, \varepsilon)$ elements. Hence, $\mathcal{C}$ can be 2ε-covered with $N(\Sigma_M, d, \varepsilon)$ elements. Instances of this simple reasoning can be found in [9, Thm. V.3, Thm. VI.4][11, Thm. 5.24][13, 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 Σ_M.

We claim that a "reasonable" growth of the covering numbers of Σ_M consists in a situation where, for some $\gamma > 0$, the set Σ_M can be $M^{-\gamma}$-covered with "roughly" $2^{M \log M}$ elements. Indeed, this covers the case where each element of Σ_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 Σ_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 Σ_M is formalized in Definition 13, and yields the simple relation $\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leqslant \gamma^{*\text{encod}}(\mathcal{C})$ for every set $\mathcal{C} \subset \mathcal{F}$, as shown in Proposition 16.

Definition 12 ((γ, 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 (γ, h)-*encoding* of Σ if there exists 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 Σ_M (recall Definition 8, in particular $\Sigma(\gamma, h)_M$ must be a subset of Σ_M) of size satisfying $\log_2(|\Sigma(\gamma, h)_M|) \leqslant c_2 M^{1+h}$.

The following definition captures a "reasonable" growth for the covering numbers of Σ_M.

Definition 13 (γ-encodable Σ 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 Σ is γ-*encodable* in $(\mathcal{F}, d)$ if for every $h > 0$, there exists a (γ, h)-encoding of Σ. We say that Σ is ∞-*encodable* in $(\mathcal{F}, d)$ if it is γ-encodable in $(\mathcal{F}, d)$ for all $\gamma > 0$. When the context is clear, we will omit the mention to $(\mathcal{F}, d)$.

Remark 14. — If Σ is γ-encodable then it is γ'-encodable for every $\gamma' \leqslant \gamma$.

Example 15. — Several examples of ∞-encodable sequences are given in section IV. In particular, consider sequences $(d_M)_{M\in\mathbb{N}^*}$ of positive integers, $(r_M)_M$ of real numbers at least equal to one, and $\Sigma := (\Sigma_M)_M$ of

subsets of $\mathcal{F} := \ell^q(\mathbb{N})$ with each Σ_M defined as the subset of sequences of $\ell^q(\mathbb{N})$ with ℓ^q-norm ($1 \leqslant q \leqslant \infty$) bounded by r_M and with support included in the first d_M coordinates. It turns out that Σ, as a sequence of subsets of $\mathcal{F}$, is either ∞-encodable or never γ-encodable, whatever $\gamma > 0$ is (see Lemma 22 and the discussion above it). It is ∞-encodable if and only if for every $h > 0$, $d_M(\log(r_M) + 1) = \mathcal{O}_{M\to\infty}(M^{1+h})$. With the same sequence Σ, if $\phi := (\varphi_M)_M$ is a sequence of Lipschitz maps from $\mathcal{F} = \ell^q(\mathbb{N})$ to a metric space $\mathcal{F}'$, then $\phi(\Sigma) := (\varphi_M(\Sigma_M))_M$ stays ∞-encodable under some conditions on ϕ and Σ, given in Proposition 24. This includes situations with ReLU neural networks, when $\mathcal{F}'$ is some L^p space, see Proposition 49. Moreover, when Σ_M contains linear combinations of (a polynomial number in) M elements of a fixed dictionary, with boundedness conditions on the coefficients, then Σ is ∞-encodable, see section IV-C.

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

It is known that $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leqslant \gamma^{*\text{encod}}(\mathcal{C})$ for various sets $\mathcal{C}$ when Σ is defined with neural networks [9, Thm. VI.4] or dictionaries [9, Thm. V.3][11, Thm. 5.24]. The following proposition shows that ∞-encodability implies $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leqslant \gamma^{*\text{encod}}(\mathcal{C})$. This settles a unified and generalized framework for the aforementioned known cases that implicitly use, one way or another, the ∞-encodability property, as we will detail in section IV-C and Example 50.

Proposition 16. — 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 γ-encodable in $(\mathcal{F}, d)$, with $\gamma \in (0, \infty]$. Then for every (non-empty) $\mathcal{C} \subset \mathcal{F}$:

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

We also gather in the next lemma a useful result that we use in particular to prove Proposition 16.

Lemma 17. — Let $(\mathcal{F}, d)$ be a metric space. Consider $\gamma, h > 0$ and two sequences $\Sigma(\gamma, h)$ and Σ of subsets of $\mathcal{F}$. If $\Sigma(\gamma, h)$ is a (γ, h)-encoding of Σ then for every (non-empty) $\mathcal{C} \subset \mathcal{F}$:

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

Proof of Lemma 17. For every $M \in \mathbb{N}^*$, it holds the inclusion $\Sigma(\gamma, h)_M \subset \Sigma_M$ (indeed $\Sigma(\gamma, h)_M$ is a covering of Σ_M) hence $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma(\gamma, h)) \leqslant \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$. This proves the result when $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = -\infty$. From now on we assume $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) > -\infty$. Since $\Sigma(\gamma, h)$ is a (γ, h)-encoding of Σ, there exists a constant $c > 0$ such that for every $M \in \mathbb{N}^*$, the set $\Sigma(\gamma, h)_M$ is a $cM^{-\gamma}$-covering of Σ_M. 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:

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

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

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

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

Proof of Proposition 16. If $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) \leqslant 0$ then the result is trivial since we always have $\gamma^{*\text{encod}}(\mathcal{C}) \geqslant 0$. 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, Σ is γ-encodable so there exists a (γ,h)-encoding of Σ that we denote by $\Sigma(\gamma,h)$. This means that there exists 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 Σ_M of size $|\Sigma(\gamma,h)_M| \leqslant 2^{c_2' M^{1+h}}$. 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:

$$d(f, \Phi_M(f)) \leqslant c_3' M^{-\gamma'}.$$

Since $0 < \gamma' < \gamma$, note that for every $M \in \mathbb{N}^*$, it holds $c_1' M^{-\gamma} + c_3' M^{-\gamma'} \leqslant (c_1' + c_3') M^{-\gamma'}$. 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 $|\Sigma(\gamma,h)_M| \leqslant 2^{c_2 M^{1+h}}$. Now, for every $\varepsilon > 0$, the integer[3] $M_\varepsilon := \left\lceil \left(\frac{c_1}{\varepsilon}\right)^{1/\gamma'} \right\rceil$ satisfies $\varepsilon \geqslant c_1 M_\varepsilon^{-\gamma'}$. By monotonicity of the metric entropy $H(\mathcal{C},d,\cdot)$ we get $H(\mathcal{C},d,\varepsilon) \leqslant H(\mathcal{C},d,c_1 M^{-\gamma'}) \leqslant c_2 M_\varepsilon^{1+h}$. Note that for $0 < \varepsilon < c_1$, denoting by $c = (2c_1^{1/\gamma'})^{1+h}$ it holds $M_\varepsilon^{1+h} \leqslant \left(1 + \left(\frac{c_1}{\varepsilon}\right)^{1/\gamma'}\right)^{1+h} = \left(\frac{c_1}{\varepsilon}\right)^{(1+h)/\gamma'} \left(1 + \left(\frac{\varepsilon}{c_1}\right)^{1/\gamma'}\right)^{1+h} \leqslant c\varepsilon^{-(1+h)/\gamma'}$. Finally for every $0 < \varepsilon < c_1$, it holds

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

As a direct consequence of Definition 9, this implies $\gamma^{*\text{encod}}(\mathcal{C}) \geqslant \frac{\gamma'}{1+h}$ for every $h > 0$ and every $0 < \gamma' < \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma),\gamma)$, hence the desired result. □

Remark 18. — Recall that Σ is γ-encodable in $(\mathcal{F},d)$ if, and only if, for every $h > 0$, there exists a "discrete" version $\Sigma^D(\gamma,h) = (\Sigma_M^D(\gamma,h))_{M\in\mathbb{N}^*}$ of Σ that (γ,h)-encodes Σ. By Lemma 17, this discrete sequence has approximation speeds comparable to those of Σ on every non-empty subset $\mathcal{C} \subset \mathcal{F}$. We will see in Proposition 24 that when each Σ_M is Lipschitz-parameterized by a bounded set in finite dimension, with conditions on the Lipschitz constant, then a discrete version $\Sigma^D(\gamma,h)$ that (γ,h)-encodes Σ can actually be defined in a more structured way and *independently* of $h > 0$: there is a "uniformly quantized" version $\Sigma^Q(\gamma)$ that (γ,h)-encodes Σ for every $h > 0$, and Lemma 17 will guarantee that it has approximation speeds comparable to Σ.

We derive from Proposition 16 a generic lower bound on the encoding speed of the set of functions uniformly approximated at a given speed.

[3]The notation $\lceil\cdot\rceil$ is defined as $\lceil x\rceil := \min\{n \in \mathbb{Z}, n \geqslant x\}$ for every $x \in \mathbb{R}$.

Corollary 19. — 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 γ-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 $\sup_{M \geqslant 1} M^\alpha d(f, \Sigma_M) \leqslant \beta$. This set satisfies

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

Proof. By the very definition of $\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)$, it holds $\gamma^{*\text{approx}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma, \beta)|\Sigma) \geqslant \alpha$. Proposition 16 then gives the result. □

The reader may wonder about the role of β 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 30 suggests this may not be possible without additional assumptions on Σ.

As an immediate corollary of Proposition 16 we also obtain the following result.

Corollary 20. — Consider $\Sigma := (\Sigma_M)_{M \in \mathbb{N}^*}$ an arbitrary sequence of (non-empty) subsets of $\mathcal{F}$ and a (non-empty) set $\mathcal{C} \subset \mathcal{F}$. If Σ is γ-encodable for every $\gamma < \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ then:

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

Proof. For every $\gamma < \gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$, since Σ is γ-encodable, we have $\gamma = \min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leqslant \gamma^{*\text{encod}}(\mathcal{C})$ by Proposition 16. Taking the supremum of such γ, we get the inequality. □

As we will see in section IV, applying Corollary 20 to specific ∞-encodable sequences allows one to unify and generalize different results of the literature [9, Thm. V.3, Thm. VI.4][11, Thm. 5.24]. When Σ is defined with ReLU feed-forward neural networks, we will explicitly study in Proposition 49 how the property of ∞-encodability depends on (bounds on) the neural network sparsity, depth, and weight magnitudes. In particular, we will find a "simple" explicit condition under which Corollary 20 generalizes [9, Thm. VI.4] to other type of constraints.

Remark 21. — The quantity $\gamma^{*\text{encod}}(\mathcal{C})$ is known in several cases, see Example 11. In the next section, we discuss concrete examples of ∞-encodable sequences Σ. For such a sequence Σ and an arbitrary set $\mathcal{C}$, independently of the adequation of Σ and $\mathcal{C}$, Corollary 20 automatically yields an upper bound for the approximation speed of $\mathcal{C}$ by Σ.

IV. The notion of ∞-encodability

We now give several examples of ∞-encodable sequences Σ. According to Corollary 20, these sequences, when used to approximate a function set $\mathcal{C}$, have automatically their (worst-case) approximation speed $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma)$ bounded from above by $\gamma^{*\text{encod}}(\mathcal{C})$. In Lemma 22, we prove that some sequences of balls (in the sense of the metric space $\mathcal{F}$) of increasing radius and dimension are ∞-encodable. Quite naturally, the property of ∞-encodability is preserved under some Lipschitz transformation, as shown in section IV-B in the specific case of the ∞-encodable sequences of balls of Lemma 22 (this can be generalized to other

∞-encodable sequences but this is not useful here). In this specific case, the considered sequences can even be encoded into sequences with approximation speeds comparable to the original ones, using a very simple *uniform quantization scheme*, see Proposition 24. This includes situations where Σ is defined with ReLU neural networks with certain controlled increasing bounds on depth, width and weights' amplitude, as shown in the next section. We conclude this section with examples of ∞-encodable sequences in the context of approximations with dictionaries, see section IV-C, showing that Corollary 20 unifies and generalizes [9, Thm. V.3][11, Thm. 5.24].

A. First example of ∞-encodable sequences

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 $N_M = \mathcal{O}_{M\to\infty}(2^{cM\pi(\log M)})$ balls (with respect to the ambient metric space) centered in Σ_M of radius $\varepsilon_M = \mathcal{O}_{M\to\infty}(M^{-\gamma})$. Since $\mathcal{O}_{M\to\infty}(2^{cM\pi(\log M)}) = \mathcal{O}_{M\to\infty}(2^{M^{1+h}})$ for every $h > 0$, it is clear from the definition that Σ is ∞-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 Σ_M is an exact cover 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 we now describe. Consider $q \in [1, \infty]$ an exponent, $(d_M)_{M\in\mathbb{N}^*}$ a sequence of positive integers, $(r_M)_{M\in\mathbb{N}^*}$ a sequence of real numbers $r_M \geqslant 1$ and define $\Sigma := (\Sigma_M)_{M\in\mathbb{N}^*}$ where $\Sigma_M \subset \ell^q(\mathbb{N})$ is the set of sequences bounded in ℓ^q-norm by r_M and with zero coordinates outside the first d_M ones. Each Σ_M can be identified with the closed ball of radius r_M in dimension d_M with respect to the q-th norm, hence standard bounds on covering numbers [14, Eq. (5.9)] yield for every $0 < \varepsilon \leqslant r_M$:

$$d_M \log_2\left(\frac{r_M}{\varepsilon}\right) \leqslant H(\Sigma_M, \|\cdot\|_q, \varepsilon) \leqslant d_M \log_2\left(\frac{3r_M}{\varepsilon}\right). \tag{5}$$

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

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

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

Lemma 22. — Let $q \in [1, \infty]$ be an exponent, $(d_M)_{M\in\mathbb{N}^*}$ a sequence of positive integers, $(r_M)_{M\in\mathbb{N}^*}$ a sequence of real numbers satisfying $r_M \geqslant 1$ and define $\Sigma := (\Sigma_M)_{M\in\mathbb{N}^*}$, with $\Sigma_M := B_{d_M,\|\cdot\|_q}(0, r)$ being the closed ball of radius r_M in dimension d_M with respect to the q-th norm. Then, Σ as a sequence of subsets

of $\ell^q(\mathbb{N})$, is either ∞-encodable or it is never γ-encodable, whatever $\gamma > 0$ is. Moreover, it is ∞-encodable if, and only if,

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

B. Lipschitz parameterized sequences are ∞-encodable

We've just seen in Lemma 22 that some sequences Σ of balls in finite dimension are ∞-encodable. Quite naturally, this remains true when we consider images of such sequences under Lipschitz maps. Actually, these sequences are not only ∞-encodable: they can be *uniformly quantized*, in the sense that they can be covered with balls centered on a uniform grid, into sequences with comparable approximation speed to the original ones, as we show in Proposition 24 below. In the next section we will prove that this covers the case of ReLU neural networks (Proposition 49). In order to prove Proposition 24, the following lemma will be useful.

Lemma 23. — Consider $q \in [1,\infty]$ and $\gamma > 0$. There exists a constant $c(q,\gamma) > 0$ such that the following holds. Consider arbitrary $n \in \mathbb{N}^*$, $r \geqslant 1$ and consider the set $B_{n,\|\cdot\|_q}(0,r) \subset \ell^q(\mathbb{N})$ that consists of the sequences bounded in ℓ^q-norm by r, and with zero coordinates outside the first n ones. Consider a metric space $(\mathcal{F},d)$ and a Lipschitz-map $\varphi : (B_{n,\|\cdot\|_q}(0,r), \|\cdot\|_q) \to (\mathcal{F},d)$ with Lipschitz constant $\mathrm{Lips}(\varphi) \geqslant 1$:

$$\forall x,y \in B_{n,\|\cdot\|_q}(0,r), d(\varphi(x),\varphi(y)) \leqslant \mathrm{Lips}(\varphi)\|x-y\|_q.$$

For every $M \in \mathbb{N}^*$, define the step size $\eta_M := (M^\gamma n^{1/q}\mathrm{Lips}(\varphi))^{-1}$ and the "quantized" set $Q(B_{n,\|\cdot\|_q}(0,r),\eta_M) := B_{n,\|\cdot\|_q}(0,r) \cap (\eta_M\mathbb{Z})^{\mathbb{N}}$. Then for every integer $M \geqslant 2$, the set $\varphi(Q(B_{n,\|\cdot\|_q}(0,r),\eta_M))$ is an $M^{-\gamma}$-covering of $\varphi(B_{n,\|\cdot\|_q}(0,r))$ of size satisfying:

$$\log_2(|\varphi(Q(B_{n,\|\cdot\|_q}(0,r),\eta_M))|) \leqslant c(q,\gamma)\left(n\left[\log_2(n)+\log_2(r)+\log_2(\mathrm{Lips}(\varphi))+\log_2(M)\right]\right). \tag{6}$$

Proof. When $q=\infty$, it is known [14, Examples 5.2 and 5.6] that $Q(B_{n,\|\cdot\|_q}(0,r),\eta_M)$ is a η_M-covering of $B_{n,\|\cdot\|_q}(0,r)$ of size bounded by $(2r/\eta_M)^n+1$. Since φ is $\mathrm{Lips}(\varphi)$-Lipschitz, we deduce that the set $\varphi(Q(B_{n,\|\cdot\|_q}(0,r),\eta_M))$ is an $M^{-\gamma}$-covering of $\varphi(B_{n,\|\cdot\|_q}(0,r))$ of size satisfying:

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

Since $M \geqslant 2$, it holds $1+\gamma\log_2(M) \leqslant (1+\gamma)\log_2(M)$, hence Equation (6) for $c(q,\gamma) = 1+\gamma \geqslant 1$. This settles the case $q=\infty$.

When $q \in [1,\infty)$, Hölder's inequality yields $\|x\|_q \leqslant n^{1/q}\|x\|_\infty$ 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 φ with respect to $\|\cdot\|_\infty$ is bounded by its Lipschitz constant with respect to $\|\cdot\|_q$, up to a factor $n^{1/q}$. Hence, the case $q \in [1,\infty)$ can be reduced to the case $q=\infty$ by replacing r by $rn^{1/q}$ and $\mathrm{Lips}(\varphi)$ by $n^{1/q}\mathrm{Lips}(\varphi)$. We get:

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

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

Proposition 24. — Let $q \in [1,\infty]$ be an exponent and $(\mathcal{F},d)$ be a metric space. Consider sequences $(d_M)_{M\in\mathbb{N}^*}$ of positive integers, $(r_M)_{M\in\mathbb{N}^*}$ of real numbers satisfying $r_M \geqslant 1$, and define the sequence $\Sigma := (\Sigma_M)_{M\in\mathbb{N}^*}$ of subsets of $\ell^q(\mathbb{N})$ by $\Sigma_M = B_{d_M,\|\cdot\|_q}(0,r_M)$ (with the same notations as in Lemma 22). Consider also a sequence $\phi := (\varphi_M)_{M\in\mathbb{N}^*}$ of maps $\varphi_M : (\Sigma_M, \|\cdot\|_q) \to (\mathcal{F},d)$ that are $\mathrm{Lips}(\varphi_M)$-Lipschitz for some constants $\mathrm{Lips}(\varphi_M) \geqslant 1$. Define $\phi(\Sigma) := (\varphi_M(\Sigma_M))_M$. For $\gamma > 0$, define the *γ-uniformly quantized* version of $\phi(\Sigma)$ as

$$Q(\phi(\Sigma),\gamma) := (\varphi_M(Q(\Sigma_M,\eta_M(\gamma))))_{M\in\mathbb{N}^*},$$

with a step size $\eta_M(\gamma) := (M^\gamma d_M^{1/q}\mathrm{Lips}(\varphi_M))^{-1}$ and an associated quantized set $Q(\Sigma_M,\eta_M(\gamma)) := \Sigma_M \cap (\eta_M(\gamma)\mathbb{Z})^{\mathbb{N}}$ for every $M \in \mathbb{N}^*$. Assume that for every $h > 0$:

$$d_M\left(\log_2(r_M) + \log_2(\mathrm{Lips}(\varphi_M)) + 1\right) = \mathcal{O}_{M\to\infty}(M^{1+h}). \tag{7}$$

Then for every $\gamma > 0$, the *γ-uniform quantization* scheme yields approximation speeds comparable to those of the original nonquantized sequence $\phi(\Sigma)$ for every (non-empty) set $\mathcal{C} \subset \mathcal{F}$:

$$\begin{aligned} &\gamma^{*\mathrm{approx}}\Big(\mathcal{C}|Q(\phi(\Sigma),\gamma)\Big) = \gamma^{*\mathrm{approx}}(\mathcal{C}|\phi(\Sigma)) && \text{if } \gamma \geqslant \gamma^{*\mathrm{approx}}(\mathcal{C}|\phi(\Sigma)),\\ &\gamma^{*\mathrm{approx}}\Big(\mathcal{C}|Q(\phi(\Sigma),\gamma)\Big) \geqslant \gamma && \text{otherwise.}\end{aligned} \tag{8}$$

Moreover, $\phi(\Sigma)$ is ∞-encodable.

Remark 25. — Given the link between approximation speed and encoding speed for ∞-encodable sequences (see Corollary 20), the above results guide the choice of γ to define a concrete γ-quantized sequence in the context of Proposition 24. Indeed, considering $\mathcal{C} \subset \mathcal{F}$ a classical function class for which the quantity $\gamma^{*\mathrm{encod}}(\mathcal{C})$ is known, see Example 11, choosing $\gamma \geqslant \gamma^{*\mathrm{encod}}(\mathcal{C})$ is sufficient to ensure that $\gamma \geqslant \gamma^{*\mathrm{approx}}(\mathcal{C}|\phi(\Sigma))$. Vice-versa, among all such γ, choosing the smallest one $\gamma = \gamma^{*\mathrm{encod}}(\mathcal{C})$ is probably the best choice to yield the largest possible step size η_M and the best concrete compromise.

Proof of Proposition 24. Fix an arbitrary $\gamma > 0$. Lemma 23 guarantees that for every integer $M \geqslant 2$, the set $\varphi_M(Q(\Sigma_M,\eta_M(\gamma)))$ is an $M^{-\gamma}$-covering of $\varphi_M(\Sigma_M)$ of size satisfying:

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

Since $0 \leqslant d_M \leqslant d_M\left(\log_2(r_M) + \log_2(\mathrm{Lips}(\varphi_M)) + 1\right)$, Assumption (7) guarantees that for every $h > 0$, it holds $d_M = \mathcal{O}_{M\to\infty}(M^{1+h})$ so that $d_M\left(\log_2(d_M) + \log_2(M)\right) = \mathcal{O}_{M\to\infty}(M^{1+h}\left(\log(M^{1+h}) + \log_2(M)\right)) = \mathcal{O}_{M\to\infty}(M^{1+h})$. As a consequence, for every $h > 0$, it holds $\log_2(|\varphi_M(Q(\Sigma_M,\eta_M(\gamma)))|) = \mathcal{O}_{M\to\infty}(M^{1+h})$ so that the sequence $Q(\phi(\Sigma),\gamma)$ is a (γ,h)-encoding of $\phi(\Sigma)$. This shows that $\phi(\Sigma)$ is γ-encodable for every $\gamma > 0$, hence ∞-encodable and Lemma 17 proves Equality (8). □

Remark 26. — For arbitrary $\gamma > 0$ and $h > 0$, the discrete sequence $\phi(\Sigma)^D(\gamma,h)$ of Remark 18 can then be defined as the sequence $Q(\phi(\Sigma),\gamma)$, independently of $h > 0$.

C. The case of dictionaries

We now consider sequences Σ defined with dictionaries. As detailed below, results of the literature [11, Thm. 5.24][13, Prop. 11] use arguments that implicitly prove γ-encodability. Let us start with the case of Banach spaces as in [13]. We only explicit the sequence used in [13] which is γ-encodable and we do not delve into more details as results of [13] are out of scope of this paper. A part of the proof of [13, Prop. 11] consists of implicitly showing that some specific sequence Σ^q is s-encodable, for q and s as described below in Proposition 27. In particular, the setup of Proposition 27 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 27. — 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-Telmyakov property [13, Def. 2], *i.e.*, assume that there exists $c > 0$ such that for every finite subset I of $\mathbb{N}^*$:

$$\frac{1}{c}|I|^{1/p} \min_{i\in I} |c_i| \leqslant \| \sum_{i\in I} c_i e_i \|_{\mathcal{F}} \leqslant c|I|^{1/p} \max_{i\in I} |c_i|, \forall (c_i)_{i\in I} \in \mathbb{R}^I. \tag{9}$$

Consider $0 < q < p$. For every $M \in \mathbb{N}^*$, define[4]:

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

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

Proof. The fact that Σ^q is s-encodable is implicitly proven in [13]. It goes as follows. Fix $M \in \mathbb{N}^*$ and $f = \sum_{i=1}^M c_i e_i \in \Sigma^q_M$. Let $0 < \lambda < 1$. Define $Q_\lambda(f) := \sum_{i=1}^M \mathrm{sign}(c_i) \left\lfloor \frac{c_i}{\lambda} \right\rfloor \lambda e_i$. It is proven in [13, Prop. 6] that there exists a constant $c(p,q) > 0$ that only depends on p and q such that:

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

Moreover, it is proven in [13, Lem. 4 and proof of Prop. 11] that $Q_\lambda(f)$ can be encoded with at most $\lambda^{-q}(1 - \log_2(\lambda) + \log_2(M))$ bits. Setting $\varepsilon = \lambda^{1-q/p}$, and observing that $\lambda^{-q} = \varepsilon^{-1/s}$, this proves that every element of Σ^q_M can be encoded within accuracy $\mathcal{O}_{\varepsilon\to 0}(\varepsilon)$ using $\mathcal{O}_{\varepsilon\to 0}(\varepsilon^{-1/s}(\log_2 1/\varepsilon + \log_2 M))$ bits. Setting $\varepsilon = (2M)^{-s}$ (and $0 < \lambda < 1$ accordingly) and translating the result into covering numbers yields that Σ^q is s-encodable. □

In the case of Hilbert spaces, much more generic sequences than Σ^q above are in fact ∞-encodable, as we now discuss. The ∞-encodability can be used to recover [11, Thm. 5.24] (see Corollary 29), and to generalize Corollary 19 (see Corollary 30). Let $\mathcal{F}$ be a Hilbert space and d be the metric associated to the norm on $\mathcal{F}$. A dictionary is, by definition [11, Def. 5.19], a subset $\mathcal{D} = (\phi_i)_{i\in\mathbb{N}^*}$ of $\mathcal{F}$ indexed by a countable set, which

[4] In terms of weak-ℓ^q-space, the set Σ^q_M 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.

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 28. — Let $\mathcal{F}$ be a Hilbert space. Let $\mathcal{D} = (\phi_i)_{i\in\mathbb{N}^*}$ be a dictionary in $\mathcal{F}$, and $\pi : \mathbb{N}^* \to \mathbb{N}^*$ 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 $c > 0$:

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

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

$$\gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi) = \max_{c>0} \gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c}). \tag{10}$$

Proof. Consider $c > 0$. We first prove that $\tilde{\Sigma}^{\pi,c}$ is ∞-encodable. Consider $M \in \mathbb{N}^*, \mathcal{I}_M := \{I \subset \{1,\dots,\pi(M)\}, |I| \leqslant M\}$, 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 $\pi(M)$, it can be encoded using at most $M\lceil \log_2 \pi(M)\rceil$ bits. 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 (6) of Lemma 23 with $n = |I| \leqslant M$, $q = \infty$ and $r = \max(c, 1)$ proves that $\tilde{\Sigma}^{\pi,c}(I)$ can be encoded within accuracy $M^{-\gamma}$ with at most $2c(q,\gamma)M\log_2(rM)$ bits. As a consequence, the elements of the set $\tilde{\Sigma}_M^{\pi,c}$ can be encoded within accuracy $M^{-\gamma}$ using at most $M\lceil\log_2\pi(M)\rceil + 2c(q,\gamma)M\log_2(rM) = \mathcal{O}_{M\to\infty}(M\log M)$ bits, hence (after translation of this result in terms of covering numbers) the ∞-encodability of $\tilde{\Sigma}^{\pi,c}$.

It now remains to prove Equation (10). First, for every $c > 0$ and every $M \in \mathbb{N}^*$, it holds $\tilde{\Sigma}_M^{\pi,c} \subset \Sigma_M^\pi$ so that Σ^π approximates $\mathcal{C}$ at least as quickly as $\tilde{\Sigma}^{\pi,c}$, that is $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi) \geqslant \gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c})$. As we now prove, there is actually equality for $c = \sup_{f\in\mathcal{C}} \sup_{M\in\mathbb{N}^*} \max_{I\in\mathcal{I}_M} \max_{i\in I} |\langle f, \tilde{\phi}_i^I\rangle_\mathcal{F}|$ (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, $c \leqslant \sup_{f\in\mathcal{C}} \|f\|_\mathcal{F}$ which is finite since $\mathcal{C}$ is bounded. If $f \in \mathcal{C}$, then for every $M \in \mathbb{N}^*$, every $I \subset \{1,\dots,\pi(M)\}, |I| \leqslant M$, and every $(c_i)_{i\in I} \in \mathbb{R}^I$, it holds:

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

This implies that $d(f, \tilde{\Sigma}_M^{\pi,c}) \leqslant d(f, \Sigma_M^\pi)$. As a consequence, $\tilde{\Sigma}^{\pi,c}$ approximates $\mathcal{C}$ at least as quickly as Σ^π, that is $\gamma^{*\text{approx}}(\mathcal{C}|\tilde{\Sigma}^{\pi,c}) \geqslant \gamma^{*\text{approx}}(\mathcal{C}|\Sigma^\pi)$. Hence equality (10). □

As a consequence of Theorem 28, one can recover [11, Thm. 5.24]. The proof below is essentially a rewriting in our formalism of the original proof of [11, Thm. 5.24]. We explicitly express it using equality (10) and the ∞-encodability of the sequences $\tilde{\Sigma}^{\pi,c}$ for $c > 0$, which are only implicitly used in the original proof.

Corollary 29 ([11, Thm. 5.24]). — Let $(\mathcal{F}, d)$ be a Hilbert space and $\mathcal{C} \subset \mathcal{F}$. Define $\gamma^*(\mathcal{C}|\Sigma^\pi)$ as in Remark 7. Under the assumptions of Theorem 28, the sequence $\Sigma^\pi = (\Sigma^\pi_M)_{M\in\mathbb{N}^*}$ satisfies for every relatively compact[5] set $\mathcal{C}$:

$$\gamma^*(\mathcal{C}|\Sigma^\pi) = \gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma^\pi) \leqslant \gamma^{*\mathrm{encod}}(\mathcal{C}).$$

Proof. When $\mathcal{C}$ is relatively compact, it holds $\gamma^*(\mathcal{C}|\Sigma^\pi) = \gamma^{*\mathrm{approx}}(\mathcal{C}|\Sigma^\pi)$, see Remark 7. Since $\mathcal{C}$ is relatively compact, it must be bounded so equation (10) of Theorem 28 holds. For every $c > 0$, Proposition 16 applied to $\tilde{\Sigma}^{\pi,c}$ of Theorem 28, which is ∞-encodable, shows that the right hand-side of equation (10) is bounded from above by $\gamma^{*\mathrm{encod}}(\mathcal{C})$. Hence the result. □

We also obtain a generic lower bound on the encoding speed of balls of approximation spaces [6, Sec. 7.9] (also called maxisets [12]) with general dictionaries.

Corollary 30. — Let $(\mathcal{F}, d)$ be a Hilbert space. Under the assumptions of Theorem 28, consider $\alpha, \beta > 0$ and the set[6][7] $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ of all $f \in \mathcal{F}$ such that $\|f\| \leqslant \beta$ and $\sup_{M\geqslant 1} M^\alpha d(f, \Sigma_M) \leqslant \beta$. This set satisfies

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

Corollary 30 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), hence it cannot be encoded at any positive rate.

Remark 31. — In some situations, the converse inequality $\gamma^{*\mathrm{encod}}(\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)) \leqslant \alpha$ 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 [13, Sec. 4] for an example.

Proof of Corollary 30. By the very definition of $\mathcal{C} := \mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$, this is a bounded set so equation (10) of Theorem 28 holds. For every $c > 0$, Proposition 16 applied to $\tilde{\Sigma}^{\pi,c}$ of Theorem 28, which is ∞-encodable, shows that the right hand-side of equation (10) is bounded from above by $\gamma^{*\mathrm{encod}}(\mathcal{C})$, hence

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

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

Remark 32. — Note that if Σ^π was γ-encodable for some $\gamma > 0$ large enough then Corollary 29 would be a special case of Corollary 20 whereas Corollary 30 would be a special case of Corollary 19. But in this situation, Σ^π has no reason to be γ-encodable, whatever $\gamma > 0$ is (since the dictionary is arbitrary and the coefficients of the linear combinations are not bounded). This shows that Corollary 20 and Corollary 19 actually holds more generally for some sequences Σ that are not γ-encodable, whatever $\gamma > 0$ is, as soon as

[5] Recall that a set is relatively compact if its closure is compact. In particular, it must be bounded.

[6] This is the ball of radius β of an approximation space [6, Sec. 7.9]/maxiset[12].

[7] Note that compared to the set in Corollary 19, we additionally require that $\|f\| \leqslant \beta$ so that $\mathcal{A}^\alpha(\mathcal{F}, \Sigma^\pi, \beta)$ is a bounded set and equation (10) of Theorem 28 holds.

Σ can be recovered as a limit of non-decreasing sequences Σ^c, $c > 0$, that are γ-encodable, in the sense that for every $M \in \mathbb{N}^*$, if $0 < c \leqslant c'$ then $\Sigma_M^c \subseteq \Sigma_M^{c'}$ and $\Sigma_M = \cup_{c>0} \Sigma_M^c$.

V. The case of ReLU neural networks

We now consider sequences Σ defined with ReLU neural networks. After specifying in section V-A the details needed about the metric space $(\mathcal{F}, d)$ in which approximation is considered, we formalize in section V-B the fact that the parameterization of sets of functions represented by ReLU networks has the Lipschitz property that makes possible the use of Lemma 23. It has many consequences. First, necessary conditions on the step size $\eta > 0$ used to quantize ReLU networks coordinatewise by $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$ within error $\varepsilon > 0$ are established, see Corollary 39. Second, existing results [7, Thm. 2][9, Lem. VI.8] on approximation properties of quantized ReLU networks are recovered and, sometimes, even improved, see section V-C. Third, bounds on covering numbers of some sets of functions represented by ReLU networks are also established in section V-C, and we compare them to classical bounds [1, Lem. 14.8]. Finally, as a direct consequence of Lemma 23, this leads, in Proposition 49, to a simple explicit condition on the growth with M of the network architectures to guarantee that the sequence $(\Sigma_M)_{M \in \mathbb{N}^*}$ is ∞-encodable, and even encodable with a very *simple uniform quantization scheme* that yields a sequence with comparable approximation speed.

A. Considered functional approximation setting

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 μ be a measure on Ω. Given a norm $\|\cdot\|$ on $\mathbb{R}^{d_{\text{out}}}$, we define for every measurable function $f : \Omega \to \mathbb{R}^{d_{\text{out}}}$:

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

We consider approximation in the space $L^p(\Omega \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$ consisting of all measurable functions f from Ω 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 $\|\cdot\|$, and (for a given p) all norms $\|\cdot\|_{p,\|\cdot\|}$ are equivalent. In light of this fact we will simply denote it $L^p(\Omega \to \mathbb{R}^{d_{\text{out}}}, \mu)$, or even abbreviate it as L^p. We also denote $\|\cdot\|_p := \|\cdot\|_{p,\|\cdot\|_\infty}$. We can establish necessary and sufficient conditions on $\Omega \subset \mathbb{R}^{d_{\text{in}}}$ and μ so that all functions represented by a ReLU neural network with input dimension d_{in} and output dimension d_{out} are in $L^p(\Omega \to \mathbb{R}^{d_{\text{out}}}, \mu)$.

Lemma 33. — Consider an exponent $p \in [1, \infty]$, a dimension d_{in}, a domain $\Omega \subset \mathbb{R}^{d_{\text{in}}}$, and a measure μ on Ω. Define

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

The condition

$$C_p(\Omega, \mu) < \infty \tag{11}$$

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

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

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

Proof. Assume that $C_p(\Omega, \mu) < \infty$ and consider the realization R_θ 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 [2, Thm. 2.1] that R_θ is (continuous and) piecewise linear, hence there is a partition of Ω into finitely many Ω_i, $1 \leqslant i \leqslant n$ 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 $\chi_E f \in L^p(\Omega \to \mathbb{R}^{N_L}, \mu)$ for each set $E \subset \Omega$ and each affine function f. Since $\|\chi_E g\|_p \leqslant \|g\|_p$ 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 Section A) we observe that $\|f(x)\|_\infty \leqslant |||A|||_\infty \|x\|_\infty + \|b\|_\infty \leqslant c(f)(\|x\|_\infty + 1)$ hence $\|f\|_p \leqslant c(f) C_p(\Omega, \mu) < \infty$, showing the result.

Conversely, assume that for every architecture $(L, \mathbf{N})$ and parameter $\theta \in \Theta_{L,\mathbf{N}}$ we have $R_\theta \in L^p(\Omega \to \mathbb{R}^{N_L}, \mu)$. 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 $\int_{x \in \Omega} \|x\|_\infty^p \mathrm{d}\mu(x) = \|R_\theta\|_p^p < \infty$ and $\int_{x \in \Omega} 1 \mathrm{d}\mu(x) = \|R_{\theta'}\|_p^p < \infty$. By the triangle inequality we get $C_p(\Omega, \mu) < \infty$. The case $p = \infty$ is similar. □

Condition (11) holds for every $p \in [1, \infty]$ when the input domain is bounded and μ is the Lebesgue measure. From now on, we fix an input domain $\Omega \subset \mathbb{R}^{d_{\text{in}}}$, an exponent p, and a measure μ on Ω satisfying (11).

B. On the Lipschitz parameterization of ReLU networks

It is known that some sets of functions represented by ReLU networks are Lipschitz-parameterized [5, Rmk. 9.1]. In what follows, we analyze how the Lipschitz constant depends on the depth, the width and the weight's magnitude of the considered networks. Up to our knowledge, this is the first such explicit result. We deduce (see Proposition 49) sufficient conditions on the depth, sparsity and weight's magnitude under which a sequence Σ defined with ReLU networks is ∞-encodable.

Definition 34. — (Parameter set $\Theta^q_{L,\mathbf{N}}(r)$) Given an architecture $(L, \mathbf{N})$ and the set $\Theta_{L,\mathbf{N}}$ (see Equation (2)) we define for each $r \in \mathbb{R}^+$ and $q \in [1, \infty]$ (notation $|||\cdot|||$ refers to the operator norm and is defined in section A):

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

Remark 35. — Instead of constraints on the operator norms, we may encounter constraints on the Frobenius or the max-norm. Let $r \geqslant 0$, and let $(L, \mathbf{N})$ be an architecture. Define by $W := \max_{\ell=1,\dots,L} N_\ell$ the maximal width of the network. Denote by $\|M\|_F = (\sum_{i,j} M_{i,j}^2)^{1/2}$ 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

Section A), and define $\Theta^F_{L,\mathbf{N}}(r)$ (resp. $\Theta^{\max}_{L,\mathbf{N}}(r)$) the set of all $\theta = (W_1, \dots, W_L, b_1, \dots, b_L) \in \Theta_{L,\mathbf{N}}$ such that for every $\ell = 1, \dots, L$:

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

By standard results about equivalence of norms (see e.g. (21) in the appendix) it holds for every $q \in [1, \infty]$:

$$\Theta^F_{L,\mathbf{N}}(r) \subset \Theta^2_{L,\mathbf{N}}(r), \qquad \Theta^{\max}_{L,\mathbf{N}}(r) \subset \Theta^q_{L,\mathbf{N}}(Wr) \subset \Theta^{\max}_{L,\mathbf{N}}(Wr).$$

As we now prove, the set of functions represented by ReLU neural networks with architecture $(L, \mathbf{N})$ and parameters in $\Theta^q_{L,\mathbf{N}}(r)$ is Lipschitz-parameterized.

Proposition 36. — Consider $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}^*, \Omega \subset \mathbb{R}^{d_{\text{in}}}$, μ a measure on Ω satisfying (11), $\|\cdot\|$ a norm on $\mathbb{R}^{d_{\text{in}}}$, $p, q \in [1, \infty]$, and the space $\mathcal{F} := L^p(\Omega \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$. Then there exists a constant $c > 0$ such that for every architecture $(L, \mathbf{N})$ with $N_0 = d_{\text{in}}$ et $N_L = d_{\text{out}}$, and every $r \geqslant 1$, denoting by $W := \max_{\ell=1,\dots,L} N_\ell$ the maximal width of the architecture, the map $\theta \in \Theta^q_{L,\mathbf{N}}(r) \mapsto R_\theta \in L^p$ for ReLU networks satisfies

$$\|R_\theta - R_{\theta'}\|_{p,\|\cdot\|} \leqslant cWL^2 r^{L-1} \|\theta - \theta'\|_\infty \text{ for all } \theta, \theta' \in \Theta^q_{L,\mathbf{N}}(r). \tag{12}$$

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

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

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

$$\|R_\theta - R_{\theta'}\|_{p,\|\cdot\|} \geqslant (1-\varepsilon) c' L r^{L-1} \|\theta - \theta'\|_\infty. \tag{13}$$

This converse result also holds for $1 \leqslant p < \infty$ under the additional assumption that $N_0 = \min_{0 \leqslant \ell \leqslant L} N_\ell$.

Remark 37. — It is open whether the extra factor WL in (12) compared to (13) can be improved, and whether the converse result for $p < \infty$ also holds without the additional assumption. Note that the condition $r \geqslant 1$ in Proposition 36 is reasonable since every parameter $\theta \in \Theta^q_{L,\mathbf{N}}(r)$ represents a function R_θ which is r^L-Lipschitz with respect to the q-norm on the input and output spaces. Constraining $r < 1$ would lead to "very" smooth 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 $r^L \leqslant C$ for some constant, i.e., $1 \leqslant r \leqslant C^{1/L}$.

Proof of Proposition 36. See Section C. □

Here is a list of immediate extensions of Proposition 36:

- *Arbitrary Lipschitz activation:* Proposition 36 can be extended to the case where the ReLU activation function is replaced by any Lipschitz activation function.

- *Pooling-operation:* Proposition 36 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 $\|\cdot\|_\infty \leqslant \|\cdot\|_s$, Proposition 36 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 Proposition 36. In particular, this is the case when $\mathbb{P}$ is supported on a compact set and $\ell(\hat{y}, y)$ is continously differentiable in $\hat{y}$.
- *Skip connections and convolutional layers:* one can also exploit Proposition 36 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 Proposition 36 with improved taylored bounds may be obtained in these settings.

Remark 38 (Related works). — The fact that some sets of functions represented by ReLU neural networks are Lipschitz-parameterized is already known [5, Rmk. 9.1]. To our knowledge, Proposition 36 is however the first result in the literature that explicitly expresses the dependence of the Lipschitz constant on the width, depth and weight's magnitude. Proposition 36 is based on Lemma 54 (Section B), and this lemma is a straightforward generalization of a known inequality for $q = \infty$ (see for instance [3, Eq. (3.12)] or [9, Eq. (37)]) to arbitrary $q \in [1, \infty]$. Moreover, we prove that the inequality established in Lemma 54 is optimal. To our knowledge, even in the case $q = \infty$, the optimality has not been discussed yet in the literature.

Corollary 39. — Consider a dimension $d \in \mathbb{N}^*$, a domain $[-D, D]^d$ for some $D > 0$. Fix an architecture $(L, \mathbf{N}) = (L, (N_0, \ldots, N_L))$ with maximal width $W := \max_{\ell=0,\ldots,L} N_\ell$, a bound $r \geqslant 1$ for the parameters, and an exponent $q \in [1, \infty]$.

- For every $\eta > 0$, let Q_η be a function such that $\|Q_\eta(\theta) - \theta\|_\infty \leqslant \eta$ for every parameter $\theta \in \Theta^q_{L,\mathbf{N}}(r)$. Define $c := Dd^{1/q} + 1$ and consider $0 < \varepsilon < cL^2(2r)^{L-1}$. If $0 < \eta \leqslant \varepsilon \left(cWL^2(2r)^{L-1}\right)^{-1}$, then
$$\max_{\theta\in\Theta^q_{L,\mathbf{N}}(r)} \max_{x\in[-D,D]^d} \|R_\theta(x) - R_{Q_\eta(\theta)}(x)\|_q \leqslant \varepsilon. \tag{14}$$
- Conversely, consider $\eta > 0$ and a function Q_η that acts coordinatewise on vectors and such that[8] for every $x \in \mathbb{R}_+$, $Q_\eta(x) = \lfloor x/\eta \rfloor \eta$. Define $N_{\min} := \min_{0\leqslant\ell\leqslant L} N_\ell$ and $c' = DN_{\min}^{1/q}$. If $\varepsilon, \eta > 0$ are such that (14) holds true then $\min(r, \eta) \leqslant \frac{\varepsilon}{c' r^{L-1}}$. In particular, if $\varepsilon/c' < r^L$ then $\eta \leqslant \frac{\varepsilon}{c' r^{L-1}}$.

Proof. By assumption on η and ε, $0 < \eta \leqslant \varepsilon \left(cWL^2(2r)^{L-1}\right)^{-1} \leqslant 1/W \leqslant r/W$. Note that for a matrix M with input/output dimension bounded by W, it holds $|||M|||_q \leqslant W\|M\|_{\max}$, see (21). This guarantees that

[8] $\lfloor\cdot\rfloor$ is defined as $\lfloor x \rfloor := \max\{n \in \mathbb{Z}, n \leqslant x\}$ for every $x \in \mathbb{R}$.

if $\theta = (W_1, \dots, W_L, b_1, \dots, b_L) \in \Theta^q_{L,\mathbf{N}}(r)$, then for every layer $\ell = 1, \dots, L$, it holds $|||W_\ell - Q_\eta(W_\ell)|||_q \leqslant W\|W_\ell - Q_\eta(W_\ell)\|_{\max} \leqslant W\eta \leqslant r$ and $\|b_\ell - Q_\eta(b_\ell)\|_q \leqslant W^{1/q}\|b_\ell - Q_\eta(b_\ell)\|_\infty \leqslant W\eta \leqslant r$ so that by the triangle inequality $Q_\eta(\theta) \in \Theta^q_{L,\mathbf{N}}(2r)$.

Fix $\theta \in \Theta^q_{L,\mathbf{N}}(r)$. Since $Q_\eta(\theta) \in \Theta^q_{L,\mathbf{N}}(2r)$ we can apply Proposition 36 (replacing r with $2r$), 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 Proposition 36 is actually a maximum. This yields (14) when $0 < \eta \leqslant \varepsilon \left(cWL^2(2r)^{L-1}\right)^{-1}$.

For the converse statement, consider $\varepsilon, \eta > 0$ such that (14) holds true. We must prove that $\min(r, \eta) \leqslant \frac{\varepsilon}{c' r^{L-1}}$. 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 $I_{m\times n} = I^\top_{n\times m}$. Consider $0 < a < \eta$ and define $\theta = (W_1, \dots, W_L, b_1, \dots, b_L)$ with $b_1 = \cdots = b_L = 0$, $W_1 = \lambda I_{N_1\times N_0}$ with $\lambda := \min(r, (\eta - a))$, and for every layer $\ell \geqslant 2$, $W_\ell = rI_{N_\ell \times N_{\ell-1}}$. Since $0 < \lambda \leqslant \eta - a < \eta$, 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 $c' = DN_{\min}^{1/q}$, we get:

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

As $|||W_1|||_q = \lambda = \min(r, (\eta - a)) \leqslant r$, for every $\ell \geqslant 2$, $|||W_\ell|||_q = r$ and for every $\ell \geqslant 1$, $\|b_\ell\|_q = 0 \leqslant r$, we have $\theta \in \Theta^q_{L,\mathbf{N}}(r)$ hence (14) applies. This implies $c'\lambda r^{L-1} \leqslant \varepsilon$, i.e., $\min(r, (\eta - a)) \leqslant \varepsilon/(c' r^{L-1})$. This holds for every $0 < a < \eta$: taking the limit $a \to 0^+$ yields the result. □

Remark 40. — We just saw that for $\eta > 0$ and $\varepsilon/c' < r^{L-1}$, if a function Q_η that acts coordinatewise as $Q_\eta(x) = \lfloor x/\eta \rfloor x$ for $x \in \mathbb{R}_+$ is such that (14) is satisfied, then the number of bits needed to store one coordinate of $Q_\eta(\theta)$, which is proportional to $\ln(1/\eta)$, must at least grow linearly with the network depth L since η is exponential in L. This is essentially due to the fact that parameters in $\Theta^q_{L,\mathbf{N}}(r)$ can represent functions with Lipschitz constant r^L. Less pessimistic bounds can be envisioned under stronger assumptions on the set of parameters or on the network's architecture.

C. How to use the Lipschitz constant?

We now give examples of how Proposition 36 can be used to generalize or recover existing results.

a) Quantization in L^∞: Proposition 36 allows one to establish the following proposition. As a corollary, we recover [9, Lem. VI.8] (of which [3, Lem. 3.7] is a special case). We discuss in Remark 42 how the following proposition, and hence [9, Lem. VI.8], can be generalized to other situations using Proposition 36.

Proposition 41 (extension of [9, Lem. VI.8]). — Consider $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}^*$, $D > 0$, and $(L, \mathbf{N})$ an architecture with input dimension d_{in}, output dimension d_{out} and $L \geqslant 2$ layers. Consider the space $\mathcal{F} = L^\infty([-D, D]^{d_{\text{in}}} \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|_\infty), \mu)$ with μ the Lebesgue measure.

Consider $\varepsilon \in (0, 1/2)$ and $\theta \in \Theta_{L,\mathbf{N}}$. Let $k \geqslant 0$ be the smallest integer such that $\theta \in \Theta^{\max}_{L,\mathbf{N}}(\varepsilon^{-k})$ and $d_{(L,\mathbf{N})} \leqslant \varepsilon^{-k}$, i.e., $k = \lceil \log_2 \max(\|\theta\|_\infty, d_{(L,\mathbf{N})}) / \log_2(1/\varepsilon) \rceil$. For every integer $m \geqslant 2kL + k + 1 + \log_2(\lceil D \rceil)$, the weights of θ can be rounded up to a closest point in $\eta\mathbb{Z} \cap [-\varepsilon^{-k}, \varepsilon^{-k}]$ with $\eta := 2^{-m\lceil \log_2(\varepsilon^{-1}) \rceil} \leqslant \varepsilon^m$ to obtain $\theta' \in \Theta^{\max}_{L,\mathbf{N}}(\varepsilon^{-k}) \cap (\eta\mathbb{Z})^{d_{(L,\mathbf{N})}}$ that satifies:

$$\|R_\theta - R_{\theta'}\|_{L^\infty} \leqslant \varepsilon.$$

Recall that $d_{(L,\mathbf{N})} \geqslant 2$ is the dimension of the ambient space of θ, see Equation (2), so that $k \geqslant 1$. Our result thus implies the result of Elbrächter et al. [9, Lem. VI.8]: for $L \geqslant 2$, since $k \geqslant 1$, we have $k(L-1) \geqslant 1$ hence $3kL \geqslant 2kL + k + 1$ and it is thus sufficient to take $m \geqslant 3kL + \log_2(\lceil D \rceil)$ (which is the sufficient condition given in [9, Lem. VI.8]). Note however the slower growth of m with L in our sufficient condition compared to [9, Lem. VI.8].

Proof. Denote by $W = \max_{\ell=0,\dots,L} N_\ell$ the maximal width of the architecture $(L, \mathbf{N})$. It holds (see Remark 35) $\Theta^{\max}_{L,\mathbf{N}}(\varepsilon^{-k}) \subset \Theta^1_{L,\mathbf{N}}(W\varepsilon^{-k})$ so we can use Proposition 36 with $q = 1$ to get:

$$\|R_\theta - R_{\theta'}\|_{L^\infty} \leqslant cWL^2(W\varepsilon^{-k})^{L-1}\|\theta - \theta'\|_\infty,$$

with $c = 1+D$ in this situation (see Proposition 36). Since $\|\theta-\theta'\|_\infty \leqslant \eta/2 \leqslant \varepsilon^m/2$ and $W, L \leqslant d_{(L,\mathbf{N})} \leqslant \varepsilon^{-k}$, it follows:

$$\|R_\theta - R_{\theta'}\|_{L^\infty} \leqslant ((1+D)/2)\varepsilon^{-k(2L+1)}\varepsilon^m.$$

By assumption $m \geqslant 2kL+k+1+\log_2(\lceil D \rceil)$ so that $-k(2L+1)+m \geqslant 1+\log_2(\lceil D \rceil)$. Hence $\|R_\theta - R_{\theta'}\|_{L^\infty} \leqslant ((1+D)/2)\varepsilon^{1+\log_2(\lceil D \rceil)}$. We are done if $((1+D)/2)\varepsilon^{1+\log_2(\lceil D \rceil)} \leqslant \varepsilon$ *i.e.*, if $((1+D)/2)\varepsilon^{\log_2(\lceil D \rceil)} \leqslant 1$. This is clear when $0 < D \leqslant 1$. While for $D > 1$, $((1+D)/2)\varepsilon^{\log_2(\lceil D \rceil)} \leqslant 1$ holds if and only if $\log_2(\varepsilon) \leqslant -\frac{\log_2((1+D)/2)}{\log_2(\lceil D \rceil)}$. Since $1 < D$, it holds $\frac{1+D}{2} \leqslant \frac{\lceil D \rceil + \lceil D \rceil}{2} = \lceil D \rceil$ so that $-\frac{\log_2((1+D)/2)}{\log_2(\lceil D \rceil)} \geqslant -1$. But since $\varepsilon \in (0, 1/2)$, it holds $-1 \geqslant \log_2(\varepsilon)$, hence $-\frac{\log_2((1+D)/2)}{\log_2(\lceil D \rceil)} \geqslant \log_2(\varepsilon)$ and the result follows. □

Remark 42. — More generally, given bounds on the sparsity and magnitude of network weights, and an *arbitrary* $p \in [1, \infty]$, Proposition 36 can be used to find an appropriate step size that guarantees that a uniform quantization of the considered network is within accuracy $\varepsilon > 0$ in L^p.

b) Quantization in a ball of an L^∞-Sobolev space: Proposition 36 also allows one to recover a special case of [7, Thm. 2] (the other cases can be recovered by combining this special case with [7, Prop. 3]), which gives guarantees on the existence of quantized networks approximating functions in an L^∞-Sobolev space. 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^∞

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:

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

Proposition 43 ([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, d such that for every $\varepsilon \in (0,1)$, there exists $\eta > 0$ satisfying $\ln(1/\eta) \leqslant c \ln^2(1/\varepsilon)$ and a neural 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 a total number of bits to store the network weights bounded by $c\varepsilon^{-d/n}\ln^3(1/\varepsilon)$.

Proof. Using [15, Thm. 1], there exist constants $c(n,d) > 0$ and $r(n,d) > 1$ (for instance, a proof examination of [15, 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 $\|f - R_\theta\|_{L^\infty([0,1]^d)} \leqslant \varepsilon/2$, and such that θ has weight's magnitude bounded by r. Proposition 36 can now be used to quantize the weights of θ, in order to get a quantized ReLU network ε-close to f. Denote by W the maximal width of this network architecture $(L, \mathbf{N})$. Since $\Theta^{\max}_{L,\mathbf{N}}(r) \subset \Theta^1_{L,\mathbf{N}}(Wr)$ (see Remark 35) we can use Proposition 36 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 $\theta \in \Theta^{\max}_{L,\mathbf{N}}(r)$ can be uniformly quantized with a step size $\eta := c'\varepsilon(WL^2(Wr)^{L-1})^{-1}$ to get a quantized network θ' such that $\|R_{\theta'} - R_\theta\|_{L^\infty([0,1]^d)} \leqslant \varepsilon/2$. 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 $\ln(1/\eta) \leqslant c''\ln^2(1/\varepsilon)$ 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 $c'''\ln(1/\eta) \leqslant c'''\ln^2(1/\varepsilon)$ 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. □

Remark 44. — Compared to [7, Thm. 2], Proposition 36 can also be used to establish similar results about quantized networks, not only for a function f in the unit ball of an L^∞-Sobolev space, but for every $f \in L^p$ $(1 \leqslant p \leqslant \infty)$ *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.*

c) Bound on covering numbers: Proposition 36 can be used to derive bounds on covering numbers of classes of functions represented by ReLU neural networks. This is reminiscent of [1, Lem. 14.8] although, as discussed in Remark 46 below, the assumptions and bounds are different.

Proposition 45. — Consider $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}^*, \Omega \subset \mathbb{R}^{d_{\text{in}}}$, μ a measure on Ω satisfying (11), $\|\cdot\|$ a norm on $\mathbb{R}^{d_{\text{in}}}$, $p \in [1,\infty]$, the space $\mathcal{F} := L^p(\Omega \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$, and the corresponding constant $c > 0$ from

Proposition 36. Consider $q \in [1, \infty]$, $r \geqslant 1$, an architecture $(L, \mathbf{N})$ with $N_0 = d_{\text{in}}$ et $N_L = d_{\text{out}}$, and

$$R_{\Theta^q_{L,\mathbf{N}}(r)} := \{R_\theta, \theta \in \Theta^q_{L,\mathbf{N}}(r)\} \subseteq \mathcal{F}$$

the set of realizations of ReLU neural networks with architecture $(L, \mathbf{N})$ and parameters in $\Theta^q_{L,\mathbf{N}}(r)$ (see Definition 34). Denote by $W := \max_{\ell=0,\dots,L} N_\ell$ the maximal width of the architecture and recall that $d_{(\ell,\mathbf{N})}$ denotes the dimension of the ambient space of $\Theta^q_{L,\mathbf{N}}(r)$, cf Equation (2).

Then, for every $0 < \varepsilon \leqslant 2cWL^2r^{L-1}$, the covering numbers (see Definition 8) of $R_{\Theta^q_{L,\mathbf{N}}(r)}$, with respect to the metric $\|\cdot\|_{p,\|\cdot\|}$ of $\mathcal{F}$, are bounded as follows:

$$N(R_{\Theta^q_{L,\mathbf{N}}(r)}, \|\cdot\|_{p,\|\cdot\|}, \varepsilon) \leqslant \left(\frac{6c(WL)^2r^L}{\varepsilon}\right)^{d_{(\ell,\mathbf{N})}}.$$

Proof. Proposition 36 guarantees that the mapping $\varphi : \theta \in (\Theta^q_{L,\mathbf{N}}(r), \|\cdot\|_\infty) \mapsto R_\theta \in (\mathcal{F}, \|\cdot\|_{p,\|\cdot\|})$ is $\text{Lips}(\varphi)$-Lipschitz with $\text{Lips}(\varphi) := cWL^2r^{L-1}$. Hence, for every $\varepsilon > 0$ (see e.g. [1, Lem. 14.13]):

$$N(R_{\Theta^q_{L,\mathbf{N}}(r)}, \|\cdot\|_{p,\|\cdot\|}, \varepsilon) \leqslant N(\Theta^q_{L,\mathbf{N}}(r), \|\cdot\|_\infty, \varepsilon/\text{Lips}(\varphi)).$$

Recall that if $X \subset Y$ are subsets of a metric space $(\mathcal{F}, d)$ we do not generally have $N(X, d, \varepsilon) \leqslant N(Y, d, \varepsilon)$ but only the weaker bound $N(X, d, \varepsilon) \leqslant N(Y, d, \varepsilon/2)$ (see, e.g., [10, Lem. A.1]). The definition of $\Theta^q_{L,\mathbf{N}}(r)$ as a cartesian product of closed balls with respect to operator and vector q-norms, and standard equivalence results on norms (see (21)) yield $\Theta^q_{L,\mathbf{N}}(r) \subset \prod_{\ell=1}^{L} \left(B_{N_\ell N_{\ell-1},\|\cdot\|_\infty}(0, Wr) \times B_{N_\ell,\|\cdot\|_\infty}(0, Wr)\right)$ where $B_{d,\|\cdot\|}(0, r)$ is the closed ball of radius r centered in 0 with respect to the norm $\|\cdot\|$ in dimension d. Hence

$$N(\Theta^q_{L,\mathbf{N}}(r), \|\cdot\|_\infty, \varepsilon/\text{Lips}(\varphi)) \leqslant N(\prod_{\ell=1}^{L} \left(B_{N_\ell N_{\ell-1},\|\cdot\|_\infty}(0, Wr) \times B_{N_\ell,\|\cdot\|_\infty}(0, Wr)\right), \|\cdot\|_\infty, \varepsilon/(2\text{Lips}(\varphi))).$$

Moreover a covering of a product space $X \times Y$ in the uniform norm can be constructed by taking the cartesian product of coverings of X and Y in the uniform norm, meaning that $N(X \times Y, \|\cdot\|_\infty, \varepsilon) \leqslant N(X, \|\cdot\|_\infty, \varepsilon)N(Y, \|\cdot\|_\infty, \varepsilon)$. Hence, for every $0 < \varepsilon \leqslant 2\text{Lips}(\varphi)$, using standard bounds on covering numbers of balls in finite dimension [14, Eq. (5.9)]:

$$\begin{aligned} N(\Theta^q_{L,\mathbf{N}}(r), \|\cdot\|_\infty, \varepsilon/\text{Lips}(\varphi)) &\leqslant \prod_{\ell=1}^{L} N(B_{N_\ell N_{\ell-1},\|\cdot\|_\infty}(0, Wr), \|\cdot\|_\infty, \varepsilon/(2\text{Lips}(\varphi))) \\ &\quad \times N(B_{N_\ell,\|\cdot\|_\infty}(0, Wr), \|\cdot\|_\infty, \varepsilon/(2\text{Lips}(\varphi))) \\ &\leqslant (6Wr\text{Lips}(\varphi)/\varepsilon)^{\sum_{\ell=1}^{L} N_\ell(N_{\ell-1}+1)} = (6Wr\text{Lips}(\varphi)/\varepsilon)^{d_{(L,\mathbf{N})}}. \end{aligned}$$

□

Remark 46. — Proposition 45 gives covering number bounds for neural networks with ℓ^2-norm constraints on the parameters, but the Lipschitz bound of Proposition 36 can be used to get similar covering number bounds with more general constraints. It can be compared to the following covering number bound for ReLU neural networks [1, Lem. 14.8], for $\varepsilon > 0$ small enough:

$$\left(\frac{2d_{(\ell,\mathbf{N})}br^L}{\varepsilon}\right)^{d_{(\ell,\mathbf{N})}}.$$

Here are important differences with the bound of Proposition 45. [1, Lem. 14.8] requires that each neuron output is bounded by some $b \geqslant 0$ while Proposition 45 does not (but if Ω is bounded then each neuron output in Proposition 45 is implicitly bounded by something of order r^L so [1, Lem. 14.8] would yield an extra $r^{Ld_{(\ell,\mathbf{N})}}$ in the situation of Proposition 45). Proposition 45 has an extra $\left(\frac{(WL)^2}{d_{(L,\mathbf{N})}}\right)^{d_{(\ell,\mathbf{N})}}$ while [1, Lem. 14.8] has an extra $b^{d_{(\ell,\mathbf{N})}}$. Notice that in the case of an architecture of constant width ($N_\ell = W$ for every ℓ) we have $d_{(L,\mathbf{N})} = LW(W+1)$ hence $(WL)^2 d^{-1}_{(L,\mathbf{N})}$ is of the order of L. Finally, Proposition 45 holds for arbitrary $L^p(\Omega \to \mathbb{R}^{d_{\text{out}}}, \mu)$ while [1, Lem. 14.8] holds only in the special case where $L^p(\Omega \to \mathbb{R}^{d_{\text{out}}}, \mu)$ is such that $p = \infty$, Ω is bounded and μ is the Lebesgue measure.

D. Application: ∞-encodability of ReLU neural networks

The explicit upper bound for the Lipschitz constant established in Proposition 36, together with Lemma 23, implies that some sequences defined with ReLU neural networks are ∞-encodable, and can even be uniformly quantized to keep comparable approximation speeds, with a step size depending on the growth of the architecture. We first defined the quantization scheme considered, before deriving Proposition 49 that gives explicit conditions on the architectures and weights growth that guarantee that the quantized sequence has an approximation speed comparable to the original one.

Definition 47. — Consider positive integers $d_{\text{in}}, d_{\text{out}}$, a sequence $(L_M)_{M\in\mathbb{N}^*}$ of positive integers, and a sequence $(r_M)_M$ of real numbers such that $r_M \geqslant 1$. For each $M \in \mathbb{N}^*$ define $\mathtt{A}_M$, the set of architectures with input dimension d_{in}, output dimension d_{out}, depth bounded by L_M and widths of the hidden layers bounded by M:

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

For every $M \in \mathbb{N}^*$ and every architecture $(L, \mathbf{N}) \in \mathtt{A}_M$, consider the set $S^M_{(L,\mathbf{N})}$ 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 θ with architecture $(L, \mathbf{N})$. Consider $\Omega \subset \mathbb{R}^{d_{\text{in}}}$, μ a measure on Ω satisfying (11), $\|\cdot\|$ a norm on $\mathbb{R}^{d_{\text{in}}}$, $q \in [1,\infty] \cup \{F, \max\}$ (F and max refers to the Frobenius norm and the max-norm, see Remark 35), and define the sequence $\mathcal{N} := (\mathcal{N}_M)_M$ of sets $\mathcal{N}_M \subset L^p(\Omega \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$ of realizations of ReLU neural networks with an architecture $(L, \mathbf{N}) \in \mathtt{A}_M$ and parameters in $\Theta^q_{L,\mathbf{N}}(r_M)$:

$$\mathcal{N}_M := \bigcup_{(L,\mathbf{N})\in\mathtt{A}_M} \bigcup_{S\in S^M_{(L,\mathbf{N})}} R_{\Theta^q_{L,\mathbf{N}}(r_M),S}$$

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

For any $\gamma > 0$, the γ-*uniformly quantized* version $Q(\mathcal{N}_M, \gamma)$ of $\mathcal{N}_M$ is defined as follows: for every $M \in \mathbb{N}^*$, consider the step size $\eta_M = \eta_M(\gamma, q) := (M^\gamma \text{Lips}(M, q))^{-1}$, with $\text{Lips}(M, q) := ML_M^2 r_M^{L_M - 1} \geqslant 1$ if $q \in [1,\infty] \cup \{F\}$, and $\text{Lips}(M, \max) = \text{Lips}(M, 2) M^{L_M - 1}$, and set

$$Q(\Theta^q_{L,\mathbf{N}}(r_M), \gamma) := \Theta^q_{L,\mathbf{N}}(r_M) \cap (\eta_M \mathbb{Z} \cap [-r_M, r_M])^{d_{(L,\mathbf{N})}}, \forall (L, \mathbf{N}) \in \mathtt{A}_M,$$

$$Q(\mathcal{N}_M, \gamma) := \bigcup_{(L,\mathbf{N})\in\mathtt{A}_M} \bigcup_{S\in S^M_{(L,\mathbf{N})}} R_{Q(\Theta^q_{L,\mathbf{N}}(r_M),\gamma),S}.$$

Remark 48. — The constraint $N_\ell \leqslant M$ in the definition of the architectures in $\mathtt{A}_M$ is written for clarity but is indeed superfluous, given that the realization of a network θ (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 θ' on a "pruned" architecture where $N_\ell \leqslant M$ for every hidden layer.

We can now give explicit conditions on the constraints on the depth, the width and the weight's magnitude, that guarantee that the γ-*uniformly quantized* sequence has an approximation speed comparable to the original one on *every set* $\mathcal{C} \subset L^p$.

Proposition 49. — In the context of Definition 47, assume that for every $h > 0$, it holds:

$$L_M M\,(1 + \log_2(r_M)) = \mathcal{O}_{M\to\infty}(M^{1+h}). \tag{15}$$

Then $\mathcal{N}$ is ∞-encodable in $\mathcal{F} := L^p(\Omega \to (\mathbb{R}^{d_{\text{out}}}, \|\cdot\|), \mu)$. Moreover, for every $\gamma > 0$, the γ-*uniformly quantized* sequence $Q(\mathcal{N}, \gamma) := (Q(\mathcal{N}_M, \gamma))_{M\in\mathbb{N}^*}$ has comparable approximation speeds to $\mathcal{N}$ on every (non-empty) set $\mathcal{C} \subset \mathcal{F}$:

$$\begin{aligned} &\gamma^{*\text{approx}}(\mathcal{C}|Q(\mathcal{N},\gamma)) = \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N}) && \text{if } \gamma \geqslant \gamma^{*\text{approx}}(\mathcal{C}|\mathcal{N}),\\ &\gamma^{*\text{approx}}(\mathcal{C}|Q(\mathcal{N},\gamma)) \geqslant \gamma && \text{otherwise.}\end{aligned} \tag{16}$$

Proof. By Proposition 36 there is a constant $c > 0$ such that for each $M \in \mathbb{N}^*$, each architecture $(L, \mathbf{N}) \in \mathtt{A}_M$ and each support $S \in S^M_{(L,\mathbf{N})}$, the set $R_{\Theta^q_{L,\mathbf{N}}(r_M),S}$ is the image under a Lipschitz map of $(\{\theta \in \Theta^q_{L,\mathbf{N}}(r_M) \text{ supported on } S\}, \|\cdot\|_\infty)$ with a Lipschitz constant bounded by $\text{Lips}_M(q) := cML_M^2 r_M^{L_M-1}$ for $q \in [1,\infty] \cup \{F\}$ and by $\text{Lips}_M(q) := M^{L_M-1}\text{Lips}_M(1)$ for $q = \max$ (for $q \in \{F, \max\}$, this is due to Remark 35). Fix $h > 0$. By assumption (15), we deduce that there exists $c_1 = c_1(h) > 0$ such that for every $M \in \mathbb{N}^*$, and every architecture $(L, \mathbf{N}) \in \mathtt{A}_M$

$$M\,(\log_2(r_M) + \log_2(\text{Lips}_M(q)) + \log_2(M)) \leqslant c_1 M^{1+h}.$$

Fix $\gamma > 0$. Lemma 23 shows that there is a constant $c_2 = c_2(q,\gamma) > 0$ such that for each $M \in \mathbb{N}^*$, each architecture $(L, \mathbf{N}) \in \mathtt{A}_M$ and each support $S \in S^M_{(L,\mathbf{N})}$, using the lemma with $n = |S| \leqslant M$ the cardinality of the support, $r = r_M$, and the same q as here, the quantized set $R_{Q(\Theta^q_{L,\mathbf{N}}(r_M),\gamma),S}$ is a $c_2 M^{-\gamma}$-covering of $R_{\Theta^q_{L,\mathbf{N}}(r_M),S}$ and its number of elements satisfies:

$$\log_2(|R_{Q(\Theta^q_{L,\mathbf{N}}(r_M),\gamma),S}|) \leqslant c_2 M\,(\log_2(M) + \log_2(r_M) + \log_2(\text{Lips}_M(q)) + \log_2(M)) \leqslant 2c_1c_2M^{1+h}.$$

Thus, the quantized set $Q(\mathcal{N}_M, \gamma)$ is a $c_1 M^{-\gamma}$-covering of $\mathcal{N}_M$ and its cardinality satisfies

$$|Q(\mathcal{N}_M,\gamma)| \leqslant \sum_{(L,\mathbf{N})\in\mathtt{A}_M} \sum_{S\in S^M_{(L,\mathbf{N})}} |R_{Q(\Theta^q_{L,\mathbf{N}}(r_M),\gamma),S}| \leqslant |\mathtt{A}_M| \cdot |S^M_{(L,\mathbf{N})}| \cdot 2^{2c_1c_2M^{1+h}}.$$

Note that for every $M \in \mathbb{N}^*$, $|\mathtt{A}_M| \leqslant L_M M^{L_M-1}$ (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_{in} and d_{out}). Similarly, since $S^M_{(L,\mathbf{N})}$ consists at most of all the supports of size M in dimension $d_{(L,\mathbf{N})} \leqslant 2M^2L_M$, its cardinality is bounded by $(2M^2L_M)^M$. Overall, we obtain that

$$\log_2(|Q(\mathcal{N}_M,\gamma)|) \leqslant \log_2(L_M) + L_M\log_2(M) + M\log_2(2M^2L_M) + 2c_1c_2M^{1+h}.$$

Using assumption (15) again, we obtain that there exists $c_3 > 0$ such that $\log_2(|Q(\mathcal{N}_M, \gamma)|) \leqslant c_3 M^{1+h}$ for every $M \in \mathbb{N}^*$. We deduce that for every $\gamma > 0$ and for every $h > 0$, the sequence $Q(\mathcal{N}, \gamma)$ is a (γ, h)-encoding of $\mathcal{N}$ so that $\mathcal{N}$ is ∞-encodable and Lemma 17 gives Equality (16). □

Example 50 (∞-encodable sequences of *sparse* neural networks - [9, Thm. VI.4]). — Let π be a positive polynomial and consider, as in [9, Def. VI.2], $\mathcal{N}_M^\pi$ the set of functions parameterized by a ReLU neural network with weight's amplitude bounded by $\pi(M)$, depth bounded by $\pi(\log M)$ and at most M non-zero parameters. Proposition 49 shows that $\mathcal{N}^\pi := (\mathcal{N}_M^\pi)_{M\in\mathbb{N}^*}$ is ∞-encodable. Indeed, this corresponds to the case where $q = \max$, $L_M \leqslant \pi(\log(M))$, $1 \leqslant r_M \leqslant \max(1, \pi(M))$, and $S^M_{(L,\mathbf{N})}$ consists of all the supports of size M in dimension $d_{(L,\mathbf{N})} \leqslant 2M^2 L_M$. Given Proposition 16, the fact that $\mathcal{N}^\pi$ is ∞-encodable gives the relation between approximation speed and encoding speed stated in [9, Thm. VI.4].

Example 51 (Growth of the step size). — Let $q \in [1, \infty] \cup \{F, \max\}$ be an exponent and π 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 $\Theta^q_{L,\mathbf{N}}(\pi(M))$. As in Example 50, Assumption (15) holds. Consequently, for every $\gamma > 0$, there exists a constant $c(\gamma) > 0$ such that the γ-uniformly quantized sequence $Q(\mathcal{N}, \gamma)$ of $\mathcal{N}$ defined in Proposition 49 is obtained with step size $\eta_M = \mathcal{O}_{M\to\infty}(M^{-c(\gamma)\log M})$ (hence $\mathcal{O}_{M\to\infty}((\log M)^2)$ bits are used to store each parameter), and still has approximation speeds comparable to $\mathcal{N}$. In the same setup, if we assume in addition that L_M is uniformly bounded in M, then for every $\gamma > 0$, a step size $\eta_M = \mathcal{O}_{M\to\infty}(M^{-c(\gamma)})$ (hence $\mathcal{O}_{M\to\infty}(\log M)$ bits per parameter) suffices to get:

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

VI. Conclusion

Notion of γ-encodability. This paper introduced in Definition 13 a new property of approximation families: being γ-encodable. As soon as Σ is γ-encodable in a metric space $(\mathcal{F}, d)$, Proposition 16 shows that there is a simple relation between the approximation speed of every set $\mathcal{C} \subset \mathcal{F}$ and its encoding speed:

$$\min(\gamma^{*\text{approx}}(\mathcal{C}|\Sigma), \gamma) \leqslant \gamma^{*\text{encod}}(\mathcal{C}). \tag{17}$$

As seen in Section IV, several classical approximation families Σ are γ-encodable for some $\gamma > 0$, including classical families defined with dictionaries (section IV-C) or ReLU neural networks (Example 50). As a consequence, γ-encodability lays a generic framework that unifies several situations where Inequality (17) was known, such as when doing approximation with dictionaries [11, Thm. 5.24][13, Prop. 11] or ReLU neural networks [9, Thm. VI.4]. Moreover, some γ-encodable sequences obtained as images of Lipschitz maps can be uniformly quantized while still keeping approximation speeds comparable to the original ones, see Proposition 24.

Lipschitz parameterization of ReLU neural networks. This paper also proves a generic bound on the Lipschitz constant of the mapping that associates the weights of a ReLU architecture to the function

they represent in some L^p space (Proposition 36). As a consequence, our general results on γ-encodability and uniform quantization apply to ReLU networks, see section V-D.

Other consequences of the Lipschitz parameterization. We further used in Section V-C the upper bound on the Lipschitz constant of $\theta \mapsto R_\theta$ to recover and generalize known approximation results on quantized ReLU networks [7, Thm. 2][9, Lem. VI.8]. Moreover, as seen in Corollary 39 and Remark 40, if $\eta > 0$ and $\varepsilon/c' < r^{L-1}$ are such that a function Q_η that acts coordinatewise as $Q_\eta(x) = \lfloor x/\eta \rfloor x$ for $x \in \mathbb{R}_+$ provides ε-accuracy in $L^\infty([-D, D]^d)$ uniformly on a bounded set of parameters $\Theta^q_{L,\mathbf{N}}(r)$ for an architecture $(L, \mathbf{N})$, then the number of bits per coordinates must be linear in the depth L of the considered architecture.

Positioning. From a practical side, our result helps to bound one of the components of the error that is committed when approximating a function f by a neural network in the quantized setting. However to fully deal with this problem one should also bound the distance between the prediction of the quantized network and its numerical evaluation. Indeed, when propagating the input x in the ReLU network with parameters $Q(\theta)$, some state-of-the-art schemes also use very low precision to represent and compute the results of the intermediate layers [16].

Perspectives. In Corollary 39, we saw necessary and sufficient conditions on $\eta > 0$ to guarantee that quantizing coordinatewise by $Q_\eta(x) = \lfloor x/\eta \rfloor x$ provides ε-accuracy in $L^\infty([-D, D]^d)$, uniformly on a bounded set of parameters $\Theta^q_{L,\mathbf{N}}(r)$. 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 accuracy $\varepsilon > 0$ and a "small set" of parameters, such as parameters than 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 Corollary 39, the sufficient value of $\eta > 0$ to ensure a prescribed accuracy $\varepsilon > 0$ 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 [16]). 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 θ of a ReLU neural network (and possibly a finite dataset), we say that θ' is functionally equivalent to θ, 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 θ that can be obtained by rescaling the coordinates of θ (but these are not the only ones since permuting coordinates can also lead to functionally equivalent parameters). When quantizing θ, it would be interesting to take these equivalent parameters into account.

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

[2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding deep neural networks with rectified linear units. *Electron. Colloquium Comput. Complex.*, 24:98, 2017.

[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. L. Donoho. Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Anal.*, 3(4):388–392, 1996.

[9] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei. Deep neural network approximation theory. *IEEE Trans. Inf. Theory*, 67(5):2581–2623, 2021.

[10] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Statistical learning guarantees for compressive clustering and compressive mixture modeling. *Math. Stat. Learn.*, 3(2):165–257, 2020.

[11] P. Grohs. Optimally sparse data representations. In *Harmonic and applied analysis*, Appl. Numer. Harmon. Anal., pages 199–248. Birkhäuser/Springer, Cham, 2015.

[12] G. Kerkyacharian and D. Picard. Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2):283–344, 2000. With comments, and a rejoinder by the authors.

[13] G. Kerkyacharian and D. Picard. Entropy, universal coding, approximation, and bases properties. *Constr. Approx.*, 20(1):1–37, 2004.

[14] 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.

[15] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.

[16] 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 52 (p-th norm). — Let $d \in \mathbb{N}^*$. For an exponent $p \in [1, \infty]$, the p-th 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 53. — ($|||\cdot|||_p$) 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

Optimality of a bound on $\|R_\theta(x) - R_{\theta'}(x)\|_q$

We generalize a known inequality established for $q = \infty$ [9, 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 Section C 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 54. — Let $(L, \mathbf{N})$ be an architecture with any depth $L \geqslant 1$ 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 (2) for the definition of $\Theta_{L,\mathbf{N}}$) be parameters associated to this architecture. For every $\ell = 1, \dots, L-1$, define θ'_ℓ as the parameter deduced from θ', 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 ϱ satisfy:

$$\begin{aligned} \|R_\theta(x) - R_{\theta'}(x)\|_q \leqslant & \sum_{\ell=1}^{L} \left(\prod_{k=\ell+1}^{L} |||W_k|||_q \right) \times |||W_\ell - W'_\ell|||_q \times \|R_{\theta'_{\ell-1}}(x)\|_q \\ & + \sum_{\ell=1}^{L} \left(\prod_{k=\ell+1}^{L} |||W_k|||_q \right) \|b_\ell - b'_\ell\|_q, \end{aligned} \tag{18}$$

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 $\lambda_1, \dots, \lambda_L \geqslant 0$ and $\epsilon \geqslant 0$ 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 (18) 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 $\ell = 1, \dots, L$:

$$W_\ell = \lambda_\ell I_{N_\ell \times N_{\ell-1}}, \quad W'_\ell = (1+\epsilon) W_\ell, \quad b_\ell = b'_\ell = 0. \tag{19}$$

Proof. The proof of Inequality (18) follows by induction on $L \in \mathbb{N}^*$ in a similar way as in the case $q = \infty$ [9, Eq. (37)][3, Eq. (3.12)]. For $L = 1$, this is just saying that

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

Assume that the property holds true for $L \geqslant 1$. Then at rank $L+1$ (using in the last inequality that the activation function ρ is 1-Lipschitz):

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

Using the induction hypothesis gives the desired result.

For the equality case, recall the definition of the parameters θ and θ' in Equation (19). 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 $R_{\theta'}(x) = (1+\varepsilon)^L \lambda (y^\top, 0_{1\times(N_L-s)})^\top$. Hence:

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

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

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

and:

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

Hence the equality case, since:

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

□

Appendix C

Lipschitz-parameterization of ReLU networks (Proof of Proposition 36)

Recall that we fixed a set $L^p(\Omega \to \mathbb{R}^{d_{\text{out}}}, \mu)$ containing all functions represented by ReLU neural networks with input dimension d_{in} and output dimension d_{out}. The parameter set $\Theta^q_{L,\mathbf{N}}(r)$ is defined in Definition 34.

First, Lemma 54 applied to any architecture $(L, \mathbf{N})$, any $\theta \in \Theta_{L,\mathbf{N}}$, and $\theta' = (0, \dots, 0) \in \Theta_{L,\mathbf{N}}$ yields for every $x \in \Omega$:

$$\|R_\theta(x)\|_q \leqslant \prod_{k=1}^{L} |||W_k|||_q \|x\|_q + \sum_{\ell=1}^{L} \left(\prod_{k=\ell+1}^{L} |||W_k|||_q \right) \|b_\ell\|_q, \tag{20}$$

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 $\ell \geqslant 2$ (since $\theta' = 0$).

Let $\theta, \theta' \in \Theta^q_{L,\mathbf{N}}(r)$. We are going to bound $\|R_\theta - R_{\theta'}\|_{p,\|\cdot\|}$ from above using Inequality (18) of Lemma 54. First, we introduce useful notations to write things compactly. Define for every $i, j \in \mathbb{N}$:

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

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

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 (18). Using (20) for the architecture $(\ell - 1, (N_0, \dots, N_{\ell-1}))$ we have:

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

If $\Omega \subseteq \mathbb{R}^{d_{\text{in}}}_+$ and $N_0 = \min_{0 \leqslant \ell \leqslant L} N_\ell$ then for every $x \in \Omega$, the parameters defined in Equation (19) are such that the previous inequality is an equality.

Denote by c_0 a constant such that for every $y \in \mathbb{R}^{d_{\text{out}}}$, $\|y\| \leqslant c_0\|y\|_q$. Note that if $\|\cdot\| = \|\cdot\|_s$ for $s \in [1, \infty]$, then we can take $c_0 = d_{\text{out}}^{\max\left(0, \frac{1}{s} - \frac{1}{q}\right)}$. Now, using the previous inequality and integrating both sides of Inequality (18) of Lemma 54, we get for $1 \leqslant p < \infty$:

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

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

If $\Omega \subseteq \mathbb{R}^{d_{\text{in}}}_+$, $N_0 = \min_{0 \leqslant \ell \leqslant L} N_\ell$, 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 (19).

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 $N_0 = \min_{0\leqslant\ell\leqslant L} N_\ell$), denoting by $N_{\min} := \min_{0\leqslant\ell\leqslant L} N_\ell$, then it holds for the parameters of Equation (19) 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 $W = \max_{\ell=1,\dots L} N_\ell$ is the maximal 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 $1 \leqslant q \leqslant \infty$, it holds $\|b\|_q \leqslant W^{1/q}\|b\|_\infty \leqslant W\|b\|_\infty$ and $\max(\|\!|M|\!\|_1, \|\!|M|\!\|_\infty) \leqslant W\|M\|_{\max}$. The latter, with Riesz-Thorin theorem [6, Chap.2, Thm 4.3], guarantee that for every $1 \leqslant q \leqslant \infty$:

$$\|\!|M|\!\|_q \leqslant W\|M\|_{\max} \text{ and } \|b\|_q \leqslant W\|b\|_\infty. \tag{21}$$

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

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

This time, this is not an equality for the parameters defined in Equation (19). For them it holds instead, assuming that all λ_ℓ are equal:

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

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

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

with a trivial adaptation for $p = \infty$. Since $\theta, \theta' \in \Theta^q_{L,\mathbf{N}}(r)$, it holds $\max(\Pi_{i,j}, \Pi'_{i,j}) \leqslant r^{j-i+1}$ for $i \leqslant j$, and the same also holds for $i = j + 1$ by definition of $\Pi_{i,j}$. Thus:

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

If we define:

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

where we recognize in the second factor the constant $C_p(\Omega,\mu)$ from Lemma 33 when $q=\infty$, then we finally get (12). Let us now explicit c in specific situations where $\Omega=[-D,D]^d$ for some $D>0$, μ is the Lebesgue measure and $\|\cdot\|=\|\cdot\|_q$ so that we can take $c_0=1$. If $q=\infty$ we get $c=C_p(\Omega,\mu)\leqslant(D+1)(2D)^{d/p}$. 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 $\|x\|_q\leqslant d^{1/q}\|x\|_\infty\leqslant d^{1/q}D$ for every $x\in[-D,D]^d$ with equality for $x=(D,\dots,D)^T$.

Let us now discuss the optimality of (12). 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 $N_0=\min_{0\leqslant\ell\leqslant L}N_\ell$, then the parameters θ,θ' defined in Equation (19) with $\lambda_1=\ldots=\lambda_L=\frac{r}{1+\epsilon}\geqslant 0$ are in $\Theta^q_{L,\mathbf{N}}(r)$ and satisfy $\|R_\theta-R_{\theta'}\|_p=c_0\left(\int_{x\in\Omega}\|x\|_q^p\mathrm{d}\mu(x)\right)^{\frac{1}{p}}r^{L-1}\sum_{\ell=1}^L\left(\frac{1}{1+\varepsilon}\right)^{L-\ell}\|\theta-\theta'\|_\infty$.

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 θ,θ' defined in Equation (19) with $\lambda_1=\ldots=\lambda_L=\frac{r}{1+\epsilon}\geqslant 0$ are in $\Theta^q_{L,\mathbf{N}}(r)$ and if we denote by $N_{\min}:=\min_{0\leqslant\ell\leqslant L}N_\ell$ and by $\Omega_{\min}$ the set of $x\in\Omega$ supported on the first $N_{\min}$ coordinates:

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

This yields the conclusion.