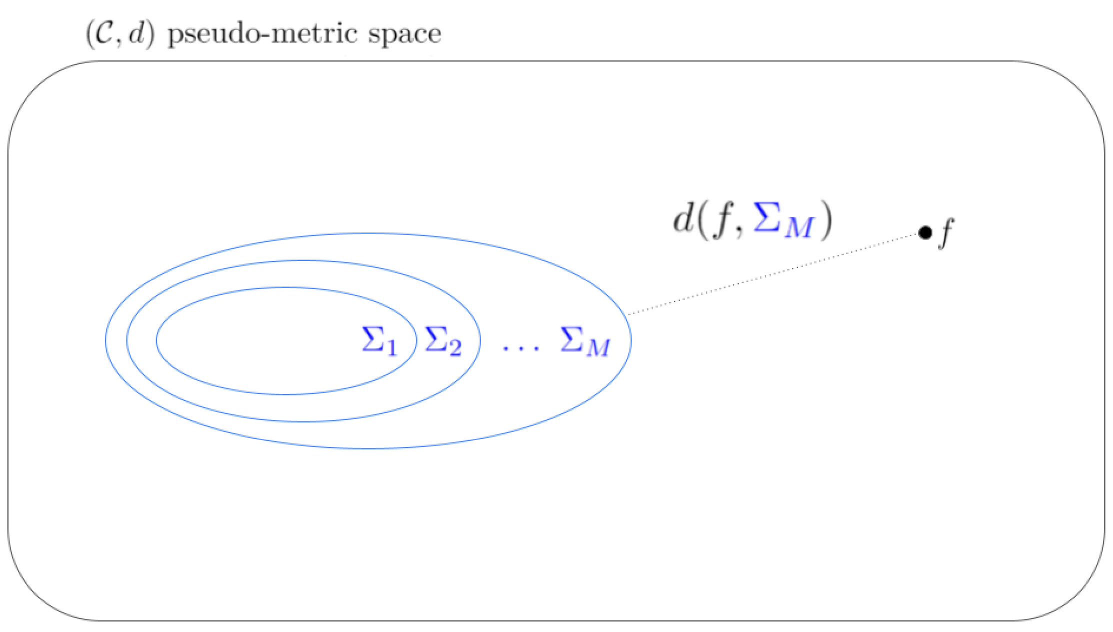
Approximation speed of quantized *vs.* unquantized ReLU neural networks and beyond

Antoine Gonon, Nicolas Brisebarre, Rémi Gribonval, Elisa Riccietti
Univ Lyon, ENS de Lyon, UCBL, CNRS, Inria, LIP, F-69342 Lyon

Problem

Context: Quantized neural networks approximate functions with success in many applications. Does existing theory explain it?

Approximation speed [3]:



- $(\mathcal{C}, d)$ pseudo-metric space
- $\Sigma = (\Sigma_M)_{M \in \mathbb{N}}$ an arbitrary (often nested) sequence of subsets $\Sigma_M \subset \mathcal{C}$

$$\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) := \text{largest } \gamma > 0 \text{ s.t. } \sup_{f \in \mathcal{C}} d(f, \Sigma_M) \underset{M \to \infty}{=} O(M^{-\gamma})$$

Examples of approximation sequences:
Σ_M := M-terms linear combination of a dictionary (polynomials, wavelets etc.)
Σ_M:= functions represented by ReLU networks:

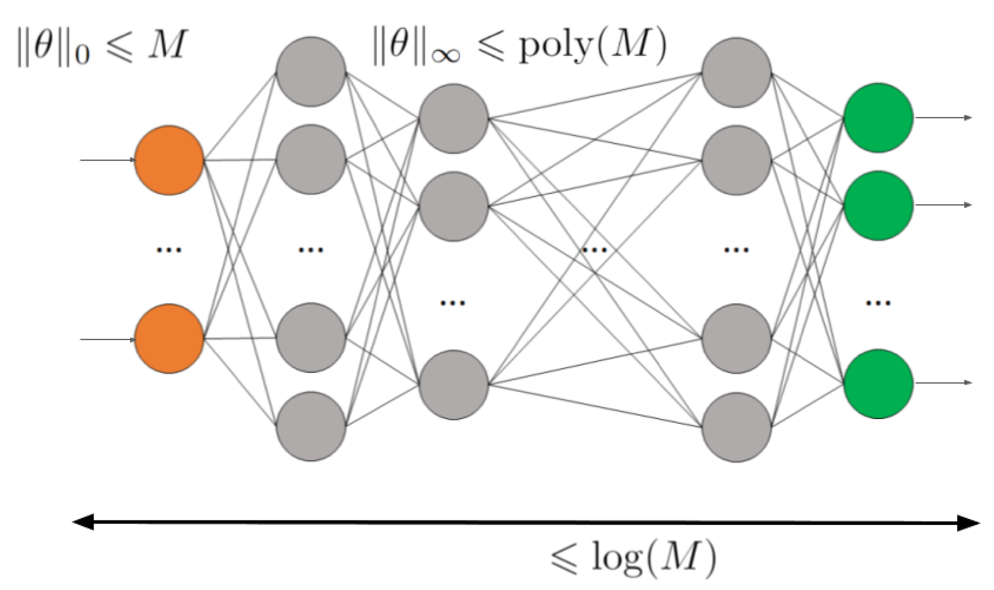


Figure 1

Questions:

- Approximation speed of *quantized versus unquantized* ReLU neural networks?
- Better understand situations where neural networks *cannot* be expected to have higher approximation speed than the best known approximation methods

Contributions:

- Notion of ∞-encodability of Σ
- Analysis of its consequences

References

[1] 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.

[2] 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.

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

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

∞-encodability

Definition:
$(\Sigma_M)_{M \in \mathbb{N}}$ is **∞-encodable** if $\forall \gamma, h > 0$:

$$N(\Sigma_M, M^{-\gamma}) \underset{M \to \infty}{=} O(M^{1+h})$$

Example:
Σ_M := M-terms linear combination of a dictionary, with bounded coefficient growth

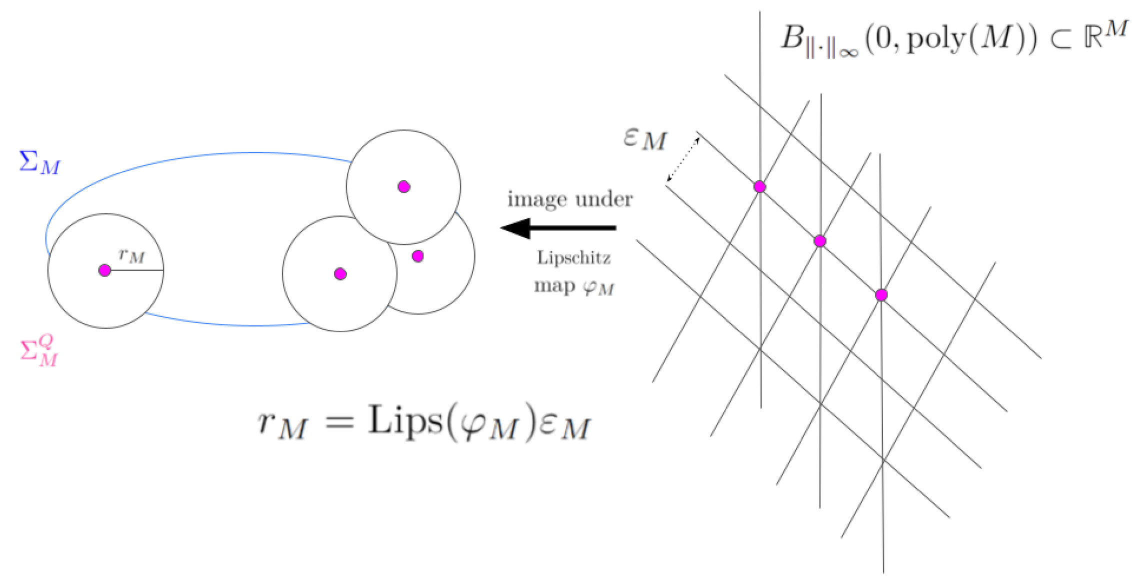
Quantized *vs.* Unquantized

Proposition: If each Σ_M of $\Sigma = (\Sigma_M)_M$ is defined with ReLU networks of Figure 1 then in $L^p([0,1]^d)$:

- it is ∞-encodable,
- it can be *uniformly* quantized into a sequence $(\Sigma_M^Q)_M$ with the same approximation speed as unquantized networks *on every set* $\mathcal{C} \subset L^p$:

$$\boxed{\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = \gamma^{*\text{approx}}(\mathcal{C}|\Sigma^Q)}$$

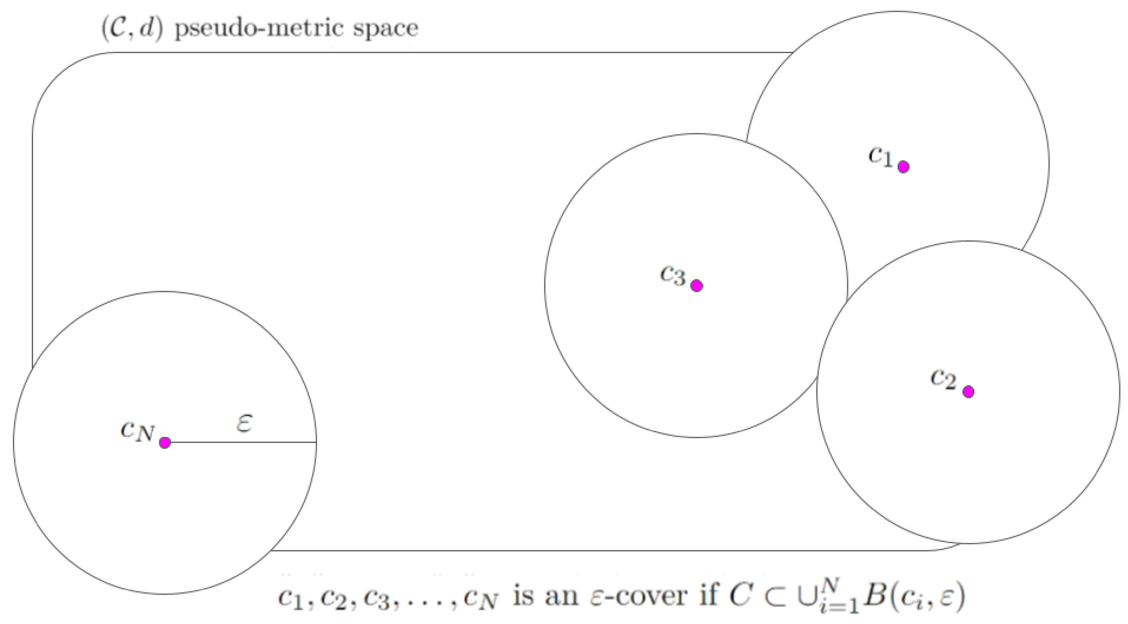
Proof idea: uses Lipschitz-parameterization



Comparison with known results:

- Lipschitz-parameterization proved using a known inequality [1]. What's new is that we proved its optimality.
- [3] also uses [1] to guarantee that on a compact domain, all networks with M weights, weight magnitudes bounded by M, and arbitrary depth, can be uniformly quantized within precision ε in L^∞. Our result generalizes this to other types of constraints.
- [2] constructs ad-hoc quantized networks approximating functions in unit balls of L^p-Sobolev spaces $W_p^m([0,1]^d)$ for $m \in \mathbb{N}^*$, while we quantize *arbitrary* neural networks while controlling the loss in precision, so that *arbitrary* $\mathcal{C} \subset L^p$ can be approximated by *uniformly* quantized networks *as soon as* we know that $\mathcal{C}$ is already approximated by unquantized networks.

Reminder: covering numbers



$$N(C, \varepsilon) := \text{smallest } N \in \mathbb{N} \text{ s.t. } \exists c_1, \ldots, c_N \in \mathcal{C} \text{ an } \varepsilon\text{-cover of } \mathcal{C}$$

Encoding rate

Optimal encoding in terms of *bitrate* [3]:

$$\gamma^{*encod}(C) := \text{largest } \gamma > 0 \text{ s.t. } \log(N(\mathcal{C}, \varepsilon)) \underset{\varepsilon \to 0}{=} O(\varepsilon^{-1/\gamma})$$

Known examples [3]:

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

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

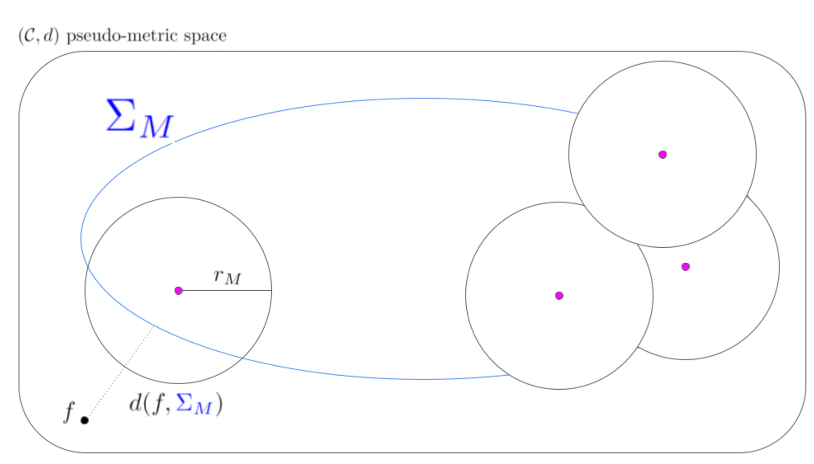
Encoding *vs.* Approximation

If Σ is ∞-encodable then:

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

Comparison with known results: Our concept of ∞-encodability allows us to unify and generalize the proof of this inequality in all the cases we found in the literature: in the case of approximation with dictionaries [4][5] or with ReLU neural networks [3].

Proof idea:



Conclusion

- ∞-encodability guarantees "reasonable" approximation speeds, avoiding degenerate cases such as $\Sigma_1 = \cdots = \Sigma_M = \cdots = \mathcal{C}$
- If an ∞-encodable sequence is known such that $\gamma^{*\text{approx}}(\mathcal{C}|\Sigma) = \gamma^{*\text{encod}}(\mathcal{C})$, then no improved approximation speed can be hoped for using "reasonable" ReLU networks
- Standard growth assumptions on sparsity, depth and weight magnitudes, yield the same approximation speed with uniformly quantized ReLU neural networks as with unquantized ones