A general approximation lower bound in L^p norm, with applications to feed-forward neural networks

El Mehdi Achour – Armand Foucault – Sébastien Gerchinovitz – François Malgouyres

Approximation problem

Motivations

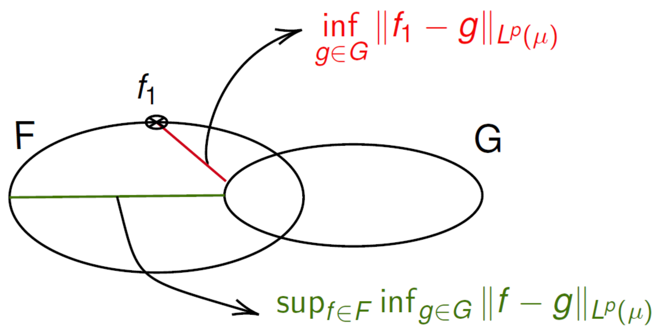
- Feed-forward neural networks are ubiquitous in artificial intelligence applications, that often involve estimating an (unknown) function
- Understanding the fundamental properties of their approximating properties is thus paramount
- Approximating functions in L^p norm can be done even in some cases when the sup norm approximation is intractable, especially in discontinuous settings

Problem statement

Let F be a space of functions with values within a finite range $[a, b]$
Let G be a space of functions implemented by a feed-forward neural network
Question: what is the order of magnitude, and in particular, a lower bound, of the *L^p approximation error*

$$\sup_{f\in F} \inf_{g\in G} \|f-g\|_{L^p(\mu)}$$

in terms of complexity measures of both F and G ?



Earlier works

Sup norm approximation

Several papers establish lower bounds and upper bounds on the *sup norm* approximation error

$$\sup_{f\in F} \inf_{g\in G} \|f-g\|_\infty,$$

for specific function spaces F. Their methods rely on the VC-dimension theory, see for instance Yarotsky [2018].

L^p norm approximation

A few lower bounds in L^p norm, p finite, exist for special classes of feed-forward neural networks (shallow networks, quantized networks...), see for instance Siegel and Xu [2021] or Petersen and Voigtlaender [2018].

Open question

Our main result solves an open question by DeVore et al. [2021]: "What is missing vis-à-vis Problem 8.13 is what the best bounds are and how we prove lower bounds for approximation rates in $L^p(\Omega)$, $p \neq \infty$."

References

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Proceedings of the 31st Conference On Learning Theory*, pages 639–649, 2018.

Jonathan W. Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *arXiv e-prints*, 2021.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, pages 296–330, 2018.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 2002.

A general lower bound

Main result

The following result, involving complexity measures such as the log-packing number and the pseudo-dimension, relies on a key inequality by Mendelson [2002].
In the sequel, we denote respectively $\log M(\varepsilon, F, \|\cdot\|_{L^p(\mu)})$ and $\mathrm{Pdim}(F)$ the log-packing number and the pseudo-dimension of F. We let $1 \leq p < +\infty$.

Theorem 1. *Let F, G be two spaces of functions defined over a set $\mathcal{X}$ endowed with a probability measure μ, with values within a finite range $[a, b]$.*
Assume the following two conditions are satisfied:

(i) $\mathrm{Pdim}(G) < +\infty$

(ii) $\log M(\varepsilon, F, \|\cdot\|_{L^p(\mu)}) \geq c\varepsilon^{-\alpha}$ *for some $c, \alpha, \varepsilon_0 > 0$, and all $\varepsilon \leq \varepsilon_0$.*

Then,

$$\sup_{f\in F} \inf_{g\in G} \|f-g\|_{L^p(\mu)} \geq \min\left\{\varepsilon_1, c_1\, \mathrm{Pdim}(G)^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(\mathrm{Pdim}(G))\right\},$$

for some constants $c_1, \varepsilon_1 > 0$ independent from ε.

Application to feed-forward neural networks

Combining the above theorem with a known bound on the pseudo-dimension of feed-forward neural networks, we obtain the following corollary (see the paper for a more general statement with piecewise-polynomial activation).

Corollary 1. *Let G be a space of functions implemented by a feed-forward neural network with fixed architecture, W variable weights, L layers and the ReLU activation. If $\log M(\varepsilon, F, \|\cdot\|_{L^p(\mu)}) \geq c\varepsilon^{-\alpha}$ for some $c, \alpha, \varepsilon_0 > 0$, and all $\varepsilon \leq \varepsilon_0$, then*

$$\sup_{f\in F} \inf_{g\in G} \|f-g\|_{L^p(\mu)} \geq c_2 (LW)^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(W),$$

for all $W \geq W_{\min}$, where $c_2, W_{\min}$ are positive constants independent from W and L.

Two examples

- If F is the unit Hölder ball over $[0,1]^d$ with smoothness parameter $s > 0$,
$$\sup_{f\in F} \inf_{g\in G} \|f-g\|_{L^p(\mu)} \geq c_3 W^{-\frac{2s}{d}} \log^{-\frac{3s}{d}}(W).$$
This lower bound matches the upper bound, up to the logarithmic factor.
- If F is the set of functions $[0,1]^d \to [0,1]$ that are *non decreasing* (non decreasing along any line parallel to an axis), we show that the L^p approximation of F is feasible , while the sup norm approximation is impossible. This illustrates the qualitative difference between L^p, $p < +\infty$, and sup norm approximation.

Conclusion and future works

Conclusion

We derive a general lower bound on the approximation error in any $L^p(\mu)$ norm, $p < +\infty$, for non-quantized networks of arbitrary depth and general sets F, which was known as a difficult problem (DeVore et al. [2021]).

Future works

Our work focuses on the approximation with neural networks with piecewise polynomial activation. Neural networks with other types of activations is left for future work.