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

El Mehdi Achour[1] Armand Foucault[1] Sébastien Gerchinovitz[2,1]
François Malgouyres[1]

[1]Institut de Mathématiques de Toulouse ; UMR 5219
Université de Toulouse ; CNRS
UPS IMT F-31062 Toulouse Cedex 9, France

[2]Institut de Recherche Technologique Saint Exupéry, Toulouse, France

JSS 2024 - IMT
presented at NeurIPS 2022

A very natural and general question in maths:

How to approximate a function f by g ?

Or, given a function f and a function set G, how well can a function $g \in G$ approximate a function f ?

Typical case:

you want to simulate the output of some $f \in F$, but you only have access to functions in G, which is limited

Typical case:

you want to simulate the output of some $f \in F$, but you only have access to functions in G, which is limited

examples:

G is a set of polynomials, or trigonometrical polynomials, or...

Introduction

- In statistics, very common problem.

Given some function set F and a loss function L,

$$f = argmin_{f \in F} E_{X,y} \left[L(f(X), y) \right]$$

Introduction

- In statistics, very common problem.

Given some function set F and a loss function L,

$$f = argmin_{f \in F} E_{X,y} \left[L(f(X), y) \right]$$

- And you give yourself a model (e.g linear model, neural network) to *approximate* this optimal f

- Thus natural to ask: *what is the approximation error of f by G (wrt $\|.\|$) ? Namely*

$$\inf_{g \in G} \|f - g\|$$

Introduction

- Thus natural to ask: *what is the approximation error of f by G (wrt $\|.\|$) ? Namely*

$$\inf_{g \in G} \|f - g\|$$

- More generally: given a function set F and an *approximation* function set G,

how well can I expect to approximate *any* function in F by *the best* function in G ?

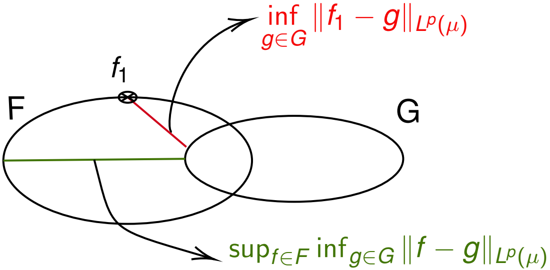
-

What is the *approximation error of* F *by* G *?*

$$\Longrightarrow \quad \sup_{f \in F} \inf_{g \in G} \|f - g\|$$

The problematic

- $F, G \subset [a,b]^{\mathcal{X}}$



$\rightarrow$ **Problematic**: quantify the approximation error (lower bounds) of F by G

$$\sup_{f \in F} \inf_{g \in G} \|f - g\|_{L^p(\mu)} \,, \tag{1}$$

expressed as a function of complexity notions of both F and G

Contributions (and outline)

- A general lower bound
- Lower bounds on the $L^p(\mu)$ approximation error of general sets F by piecewise polynomial feed forward networks

$\Rightarrow$ Improving over known bounds in sup norm

$\Rightarrow$ New proof strategy, suited for the L^p norm (open question by Devore et al. 2021 [2])

Why L^p norm is difficult

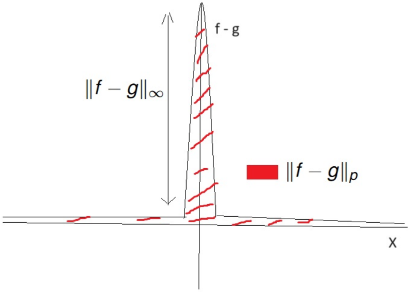
There is a qualitative difference between the L^p norm, $p < \infty$, and the sup norm:

- L^p norm, $p < \infty$: $\|f - g\|_{L^p(\mu)} = \left(\int_{\mathcal{X}} |f(x) - g(x)|^p \mathrm{d}\mu\right)^{1/p}$
- *sup* norm: $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$

Why L^p norm is difficult

"High" distance between f and g at a single point ($|f(x) - g(x)| > \varepsilon$):

- $\implies \|f - g\|_\infty > \varepsilon \implies \sup_{f \in F} \inf_{g \in G} \|f - g\|_\infty > \varepsilon$
- $\not\Longrightarrow \|f - g\|_{L^p} > \varepsilon$



Why L^p norm is difficult

- Existing lower bounds in *sup* norm [8, 7, 9, 6]
- Lower bounds in L^p norm only in very specific cases [3, 4]

Why L^p norm is difficult

- Existing lower bounds in *sup* norm [8, 7, 9, 6]
- Lower bounds in L^p norm only in very specific cases [3, 4]

$\Rightarrow$ Hence our contribution : a lower bound of the approximation error in L^p norm in a general setting

Complexity measures

- Our lower bound on $\sup_{f \in F} \inf_{g \in G} \|f - g\|_{L^p}$ involves complexity measures of F and G

Complexity measures

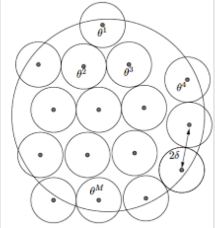
- Our lower bound on $\sup_{f \in F} \inf_{g \in G} \|f - g\|_{L^p}$ involves complexity measures of F and G

Intuition: The more complex / richer is F the harder it is to approximate. Conversely: the more complex / richer is G, the better approximation ability

Complexity measures: the packing number

- An ε-packing (*wrt* norm $\|.\|$) in F is a subset $\{f_1, \dots, f_n\}$ of functions in F that are pairwise at least ε-distant:

$$\|f_i - f_j\| > \varepsilon \qquad \forall\, i, j = 1, \dots, n$$



- The ε-packing number of F (*wrt* $\|.\|$) is the (possibly infinite) maximal cardinality of an ε-packing in F:

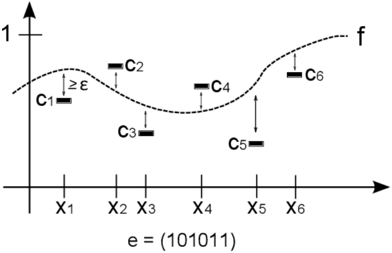
$$M(\varepsilon, F, \|.\|) = \sup\{N \in \mathbb{N}, \text{there exists a packing of size } n \text{ in } F\}$$

Complexity measures: the fat-shattering dimension

- For $\gamma > 0$, a set of points $S = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ is said to be γ-*fat-shattered* by F if

$$\exists r : S \to \mathbb{R},\ \forall E \subset S,\ \exists f \in F\ st \begin{cases} f(x) \geq r(x) + \gamma & \text{if } x \in E \\ f(x) \leq r(x) - \gamma & \text{otherwise.} \end{cases} \tag{2}$$

- The γ-fat-shattering dimension of F $fat_\gamma(F)$ is the maximal cardinality of a subset of $\mathcal{X}$ that is γ-fat-shattered by F

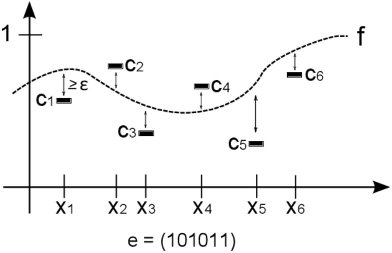


Complexity measures: the fat-shattering dimension

- For $\gamma > 0$, a set of points $S = \{x_1, \dots, x_n\} \subset \mathcal{X}$ is said to be γ-*fat-shattered* by F if

$$\exists r : S \to \mathbb{R},\ \forall E \subset S,\ \exists f \in F\ st \begin{cases} f(x) \geq r(x) + \gamma & \text{if } x \in E \\ f(x) \leq r(x) - \gamma & \text{otherwise.} \end{cases} \tag{2}$$

- The γ-fat-shattering dimension of F $fat_\gamma(F)$ is the maximal cardinality of a subset of $\mathcal{X}$ that is γ-fat-shattered by F



- The *pseudo-dimension* of F, denoted $Pdim(F)$, would be the 0-fat-shattering dimension if we replace the loose inequality by a strict in eq. (2)

Main lower bound

$M(\varepsilon, F, \|\cdot\|_{L^p(\mu)})$ is the ε-packing number of F in the $L^p(\mu)$ norm.

Theorem (informal statement)

- $1 \leq p < +\infty$
- *μ probability measure over $\mathcal{X}$*
- $F, G \subset [a, b]^{\mathcal{X}}$
- $\mathrm{fat}_\gamma(G) < +\infty$

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

$$\inf \left\{ \varepsilon > 0 : \log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \leq c_p \, \mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2\, \mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon/(b-a)} \right) \right\}.$$

Proof: relies on Mendelson 2002 [5].

Main lower bound: corollary

- Assume $\log M(\varepsilon, F, \|.\|_{L^p(\mu)})$ grows at least polynomially with $1/\varepsilon$, i.e, there exists $c_0 > 0$ and $\alpha > 0$ st:

$$\log M(\varepsilon, F, \|.\|_{L^p(\mu)}) \geq c_0 \varepsilon^{-\alpha}$$

- Then solving the equation in theorem 1 for ε yields

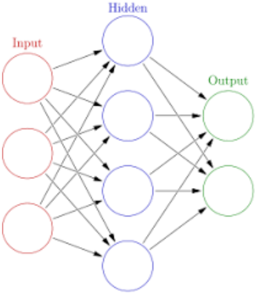
$$\sup_{f \in F} \inf_{g \in G} \|f - g\|_{L^p(\mu)} \geq \min \left\{ \varepsilon_1, Pdim\,(G)^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}} (Pdim\,(G)) \right\}$$

Application to neural networks

What if G is a set of function corresponding to a neural network ?

Informal presentation of neural networks:

- A (feed-forward) neural network is a parametrical model
- It is characterized by a number of parameters W and a depth (number of layers) L
- To a fixed parameter $\theta \in \mathbb{R}^W$, we can associate a function g_θ to the neural network

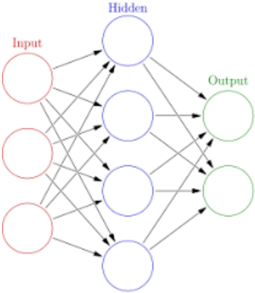


Application to neural networks

What if G is a set of function corresponding to a neural network ?

Informal presentation of neural networks:

- A (feed-forward) neural network is a parametrical model
- It is characterized by a number of parameters W and a depth (number of layers) L
- To a fixed parameter $\theta \in \mathbb{R}^W$, we can associate a function g_θ to the neural network



$$G := \left\{ g_\theta, \theta \in \mathbb{R}^W \right\}$$

Application to neural networks

- G: space of functions implemented by a feed forward neural network with W variable weights, L layers and *ReLU* activation
- Assume $\log M\left(\varepsilon, F, \|\cdot\|_{L^p(\mu)}\right) \geq c\varepsilon^{-\alpha}$ for all $\varepsilon < \varepsilon_0$ for some $\alpha, \varepsilon_0, c > 0$

Corollary

Under the above assumptions:

$$\sup_{f\in F}\inf_{g\in G}\|f-g\|_{L^p(\mu)} \geq c_1(LW)^{-\frac{1}{\alpha}}\log^{-\frac{3}{\alpha}}(W),$$

where the constant c_1 is independent from W and L.

Two examples

F	Holder functions	Monotonic functions
α	$\frac{d}{s}$	$\max(p(d-1), d)$
sup norm	Feasible	Infeasible
L^p norm	same rate as sup norm (does not depend on p)	Feasible (rate depends on p)
Tight bound	for ReLU (upper bound in [9])	for Heaviside (upper bound in this article)

Bibliography I

Thank you! [1]

[1] El Mehdi Achour, Armand Foucault, Sébastien Gerchinovitz, and François Malgouyres. A general approximation lower bound in L^p norm, with applications to feed-forward neural networks, 2022.

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

[3] Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. Manuscripta mathematica, 63(4):469–478, 1989.

[4] Vitaly E. Maiorov, Ron Meir, and Joel Ratsaby. On the approximation of functional classes equipped with a uniform measure using ridge functions. Journal of Approximation Theory, 99:95–111, 1999.

[5] Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. IEEE Transactions on Information Theory, 48, 2002.

[6] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. Journal de Mathématiques Pures et Appliquées, 157:101–135, feb 2022.

[7] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. Neural Networks, 94:103–114, 2017.

Bibliography III

[8] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 639–649, 2018.

[9] Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. Advances in neural information processing systems, 33:13005–13015, 2020.