UAB
Universitat Autònoma
de Barcelona

Departament de Matemàtiques

La profunditat en l'aprenentatge automàtic: una perspectiva matemàtica

Treball de Final de Grau en Matemàtiques

Gerard Ginés Pérez

Supervisat per Dr. Roberto Rubio Núñez

25/06/2024

Abstract

En aquest treball, hem explorat les xarxes neuronals des d'un punt de vista matemàtic usant l'anàlisi funcional, preguntant-nos quin paper juga la profunditat a l'hora d'aproximar funcions. Hem analitzat els espais de Sobolev i Sobolev-Slobodeckij on hem comprovat que, des del punt de vista computacional (nombre de neurones), és més eficient utilitzar xarxes neuronals profundes. A més, hem presentat una família de funcions definides en $S^{d-1} \times S^{d-1}$*, on hem demostrat que una xarxa neuronal de profunditat* 2 *no pot aproximar amb un error* $\epsilon > 0$ *fixat ni tan sols amb un nombre de neurones exponencial en la dimensió d'entrada* d*, mentre que una xarxa de profunditat* 3 *és capaç d'aproximar-les amb un nombre de neurones polinomial en* d*. Aquests resultats posen de manifest la rellevància de la profunditat en les xarxes neuronals a l'hora d'aproximar de funcions complexes.*

Agraïments

Vull expressar la meva més sincera gratitud al tutor del treball Dr. Roberto Rubio Núñez per haver sigut un pilar en aquest treball, gràcies per tot el temps que has invertit i per la paciència mostrada al llarg de tot el procés.

Índex

1 Evolució del Deep Learning: Un Viatge a través del Temps

El camí cap a l'aprenentatge profund es remunta als anys 40, quan McCulloch i Pitts [10] van desenvolupar un model de computadora basat en les xarxes neuronals del cervell humà. Les seves neurones, conegudes com a neurones de McCulloch-Pitts, van inspirar la creació de dispositius de xarxes formats per diverses d'aquestes neurones. El perceptró, inventat posteriorment en els anys 50 per Frank Rosenblatt [11], seria la primera d'aquestes xarxes amb capacitat d'aprenentatge mitjançant la modificació progressiva dels seus paràmetres. No obstant això, la seva limitació per resoldre problemes no lineals va ser destacada per Marvin i Seymour [9] en el seu llibre de 1969, «Perceptrons».

No va ser fins a la dècada de 1980, amb la redescoberta de l'algoritme de retropropagació de l'error per part de Geoffrey Hinton, David Rumelhart i Ronald Williams [12], que les xarxes neuronals van tornar a guanyar protagonisme. La retropropagació permetia entrenar xarxes neuronals amb múltiples capes ocultes, conegudes com a xarxes neuronals profundes, millorant significativament la seva capacitat per resoldre problemes complexos. Aquest progrés va ser complementat el 1989 per Yann LeCun [8], qui va combinar xarxes neuronals convolucionals amb retropropagació per llegir dígits escrits a mà, demostrant una aplicació pràctica de xarxes profundes.

En les darreres dècades s'ha plantejat la qüestió de si la profunditat és realment necessària a l'hora d'aproximar funcions. En Yarotsky [13] va analitzar l'eficàcia de les xarxes neuronals profundes en l'aproximació de funcions de l'espai de Sobolev, subratllant que les xarxes profundes poden superar significativament les xarxes poc profundes en termes d'eficiència computacional. Finalment, en Daniely [4] va demostrar amb un exemple concret que una xarxa neuronal de profunditat 3 és capaç d'aproximar amb un error $\epsilon > 0$ fixat mentre que una xarxa de profunditat 2 no ho pot fer, ni tan sols amb un nombre exponencial de neurones.

En aquest treball, ens centrarem en aquests dos últims articles, la nostra contribució és proporcionar exemples i petites demostracions per tal de comprendre i donar una visió més detallada i des d'un punt de vista estrictament matemàtic, dels seus resultats amb l'anàlisi funcional.

2 Preliminars

Per tal de comprendre com funcionen les xarxes neuronals, primerament explorarem la creació del perceptró.

2.1 Perceptró

El perceptró, com a sistema inicial de xarxa neuronal, funciona mitjançant una analogia simplificada del procés neuronal biològic. Aquest model elemental consta d'una capa d'entrada i una de sortida, sense capes ocultes, i és capaç de realitzar tasques de classificació binària. Com es mostra a la Figura 1, les senyals d'entrada x_i són rebudes pel sistema, cadascuna multiplicada per un pes sinàptic w_i, que representa la força o la importància de la connexió respectiva, similar a com una sinapsi regula la força d'un senyal entre neurones en el cervell.

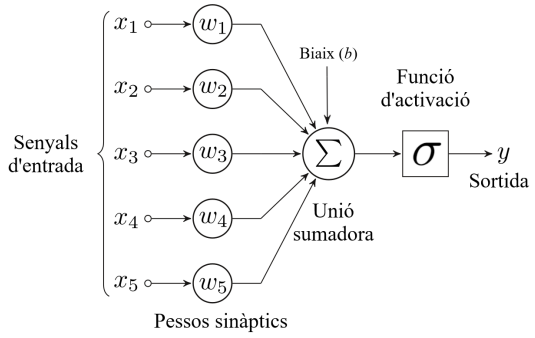

Figura 1: Representació d'un perceptró de 5 entrades.

La informació processada per la unió sumadora és la suma ponderada de les

entrades, a la qual se li suma un terme de biaix b. Aquesta suma ponderada $z = \sum_i (w_i x_i) + b$ constitueix l'entrada neta a la funció d'activació. En el cas del perceptró, la funció d'activació és una funció esglaó, que proporciona una sortida binària; és a dir, si z supera un cert llindar, la sortida y és 1, i si no, és 0.

$$y = \begin{cases} 1 & \text{si } z \geq \text{llindar} \\ 0 & \text{si } z < \text{llindar} \end{cases} \tag{1}$$

Aquest mecanisme de llindar permet al perceptró distingir entre dues classes linealment separables. Durant la fase d'aprenentatge, si la sortida del perceptró no coincideix amb la sortida esperada, els pesos sinàptics s'ajusten en direcció contrària al gradient de l'error, que es calcula com la diferència entre la sortida esperada i la sortida actual. Aquest procés d'ajust, conegut com a regla d'aprenentatge del perceptró, permet al model aprendre i adaptar-se progressivament al conjunt de dades sobre el qual s'entrena.

Tot i la seva simplicitat i les seves limitacions per modelar només funcions linealment separables, el perceptró va establir els fonaments sobre els quals s'han construït xarxes neuronals més complexes i profundes. Aquestes xarxes modernes utilitzen múltiples capes ocultes i una varietat de funcions d'activació, com veurem a continuació, per capturar relacions no lineals i patrons més complexos dins de grans volums de dades.

2.2 Definició Matemàtica de Xarxes Neuronals

La xarxa neuronal artificial presenta una estructura en capes amb un flux d'informació seqüencial i transformacions no lineals. Aquestes capes són interconnectades de tal manera que la sortida de cada una és la base per al processament de la següent, fins arribar a la predicció final.

Per tal de definir matemàticament una xarxa neuronal, primer considerarem les

següents transformacions lineals $f_s : \mathbb{R}^d \to \mathbb{R}^s$ tal que

$$f_s(x) = W^{[s]}x + b^{[s]},$$

on $d, s \in \mathbb{N}$, $x \in \mathbb{R}^d$, $W^{[s]} \in \mathbb{R}^{s\times d}$ direm que és una matriu de pesos, i $b^{[s]} \in \mathbb{R}^s$ un vector de biaixos. Les podem visualitzar com

$$\begin{pmatrix} f_s(x_1) \\ f_s(x_2) \\ \vdots \\ f_s(x_d) \end{pmatrix} = \begin{pmatrix} W_{1,1} & W_{1,2} & \cdots & W_{1,d} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ W_{s,1} & W_{s,2} & \cdots & W_{s,d} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{pmatrix}$$

Ara definirem, el concepte de funció d'activació que ja havia sigut prèviament esmentat en la definició del perceptró.

Definició 2.1. Una funció d'activació $\sigma : \mathbb{R} \to \mathbb{R}$ és una funció no lineal i que s'aplica element a element. És a dir, per un vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, tenim:

$$\sigma(x) = (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)),$$

Un exemple comú és la funció ReLU (Rectified Linear Unit), definida per:

$$\sigma(x) = \max(0, x).$$

Aquesta funció d'activació introdueix no linealitat, el que permetrà a la xarxa neuronal modelar relacions més complexes entre les dades. Vist això, ja podem donar la definició de xarxa neuronal.

Definició 2.2. *(Xarxa neuronal)* Definim una xarxa neuronal com una classe de funció $\Phi : \mathbb{R}^d \to \mathbb{R}^s$ que es pot expressar com una composició de transformacions

lineals f_n i funció d'activació σ de manera que

$$\Phi : x \in \mathbb{R}^d \xrightarrow{f_1} \mathbb{R}^{d_1} \xrightarrow{\sigma} \mathbb{R}^{d_1} \xrightarrow{f_2} \cdots \xrightarrow{f_N} \mathbb{R}^{d_N} \xrightarrow{\sigma} \hat{y} \in \mathbb{R}^s, \tag{2}$$

on $N \in \mathbb{N}$ és el nombre de capes de la xarxa neuronal, $d = d_0, d_1, \ldots, d_{N-1}, s = d_N \in \mathbb{N}$ són les dimensions de cada capa.

Per una xarxa neuronal Φ amb dimensió d'entrada d, dimensió de sortida s i N capes, la seva arquitectura seria

$$\begin{aligned}
\text{Capa d'entrada} \quad & x \in \mathbb{R}^d, \\
& f_{d_1} = W^{[d_1]}x + b^{[d_1]} \\
\text{Primera capa oculta} \quad & \sigma(f_{d_1}) \in \mathbb{R}^{d_1} \\
& \vdots \\
& f_{d_n} = W^{[d_n]}\sigma(f_{d_{n-1}}) + b^{[d_n]}, \quad \text{per } n = 1, \ldots, N, \\
n\text{-èssima capa oculta} \quad & \sigma(f_n) \in \mathbb{R}^{d_n} \\
& \vdots \\
\text{Capa de sortida} \quad & \Phi(x) = \sigma(f_{d_N}) = W^{[d_N]}\sigma(f_{d_{N-1}}) + b^{[d_N]} = \hat{y}, \quad \text{on } \hat{y} \in \mathbb{R}^s.
\end{aligned}$$

Definició 2.3. (Altres notacions) Denotem $M(\Phi) := \sum_{j=1}^{N-1} d_j$ com el total de neurones de les capes ocultes, $M'(\Phi) := d + \sum_{j=1}^{N} d_j$ com el nombre de neurones totals, $P(\Phi) := \sum_{n=1}^{N} \|W^{[n]}\|_0 + \|b^{[n]}\|_0$ com el nombre de pesos, on $\|\cdot\|_0$ indica el nombre d'entrades no nul·les i $A(\Phi) := \max\{d_1, \ldots, d_{N-1}\}$ com l'amplada de la xarxa neuronal Φ.

La qüestió central d'aquest treball és explorar la relació entre la precisió d'aproximació i la complexitat d'una xarxa neuronal Φ, la qual es mesura segons el nombre de neurones $M(\Phi)$, la quantitat de pesos i biaixos no nuls $P(\Phi)$, i el nombre de capes $C(\Phi) = N$.

Definició 2.4. *(Conjunt de xarxes neuronals)* Sigui $d, d_1, \ldots, d_{N-1}, s \in \mathbb{N}$ per

algun $N \in \mathbb{N}$, i $\sigma : \mathbb{R} \to \mathbb{R}$. Denotem

$$\Phi_{d,d_1,\ldots,d_{N-1},s} := \left\{ \Phi : \begin{array}{l} \text{N capes i } d_n \text{ neurones a la capa } n \\ \text{per totes les possibles } \sigma,\ W^{[d_n]} \text{ i } b^{[d_n]} \end{array} \right\}.$$

Aquest conjunt inclou totes les funcions possibles que es poden aprendre amb una xarxa neuronal amb dimensió d'entrada d, dimensió de sortida s, N capes i d_n neurones a la capa n, per qualsevol funció d'activació σ, pesos $W^{[d_n]}$ i biaixos $b^{[d_n]}$.

2.3 Teorema d'Aproximació Universal

Ara, dirigim la nostra atenció cap a un resultat fonamental en la teoria de xarxes neuronals: el teorema d'aproximació universal. Aquest teorema il·lustra la capacitat inherent de les xarxes neuronals per modelar qualsevol funció contínua dins d'un marc ben definit. Començarem recordant algunes nocions bàsiques de topologia i anàlisi matemàtica que són clau per a l'expressió formal del teorema.

Recordem que els conjunts compactes a $\mathbb{R}^d$ són exactament els tancats i fitats.

Definició 2.5. *(Espai de Funcions Contínues)* L'espai $C(K)$ denota el conjunt de totes les funcions $f : K \to \mathbb{R}^d$ contínues definides en un conjunt compacte $K \subseteq \mathbb{R}^d$. Aquestes funcions tenen el tret que per cada $\varepsilon > 0$, existeix un $\delta > 0$ tal que per tots $x, y \in K$, si $\|x - y\| < \delta$ llavors $|f(x) - f(y)| < \varepsilon$.

Definició 2.6. *(Norma Infinita)* La norma infinita d'una funció $f \in C(K)$, denotada com $\|f\|_\infty$, és el valor màxim absolut que f pren sobre el conjunt K. Matemàticament, es defineix com $\|f\|_\infty = \max_{x \in K} |f(x)|$.

Ara, estem llestos per enunciar el teorema. El Teorema d'Aproximació Universal va ser demostrat per George Cybenko el 1989 [3]. La prova de Cybenko abordava específicament les xarxes amb funcions d'activació sigmoïdals, no obstant en el següent enunciat tenim en compte altres resultats, com per exemple [7], on s'ha

demostrat per una classe més amplia de funcions d'activació, incloent-hi la popular ReLU.

Teorema 2.7 (Teorema d'Aproximació Universal [6]). *Per cada funció* $f : \mathbb{R}^d \to \mathbb{R}$ *amb* $f \in C(K)$ *i per a tot* $\varepsilon > 0$*, existeix un* $M \in \mathbb{N}$ *i una xarxa neuronal* $\Phi \in \Phi_{d,M,1}$ *tal que*

$$\|f - \Phi\|_\infty \leq \varepsilon.$$

Aquest resultat demostra que les xarxes neuronals poc profundes tenen la capacitat d'aproximar qualsevol funció contínua en un conjunt compacte fins a un grau d'error arbitràriament petit, donat un nombre de neurones suficient. Aquest últim fet, ens fa qüestionar-nos si existeix cap diferència en l'ordre d'aproximació quan considerem més capes. Realment la profunditat és necessària per al Deep Learning?

3 La profunditat en les xarxes neuronals

En l'aprenentatge profund, la profunditat de les xarxes neuronals és un factor crític que impacta directament en la seva capacitat per capturar la complexitat dels patrons dins de les dades. La relació entre el nombre de neurones i la capacitat d'aproximació d'una funció objectiu no només és intrigant des d'un punt de vista teòric, sinó que també és vital per la construcció eficient de models en pràctiques d'aprenentatge automàtic.

3.1 Teoria de la mesura

Abans de submergir-nos en aquesta discussió, necessitem establir alguns conceptes clau de teoria de la mesura, com són els espais de Lebesgue i Sobolev, així com les restriccions associades a aquests espais. Presumirem la mesura de Lebesgue a $\mathbb{R}^d$.

Definició 3.1. *(Espai de Lebesgue* L_p*)* L'espai de Lebesgue $L_p(K)$, per a un número real $p \geq 1$ i un conjunt mesurable $K \subseteq \mathbb{R}^d$, és l'espai de totes les funcions

mesurables $f : K \to \mathbb{R}$ tal que la potència p-èsima del valor absolut de f és integrable respecte a la mesura de Lebesgue sobre K. Formalment, es defineix com

$$L^p(K) = \{f : K \to \mathbb{R} \mid \|f\|_{L^p} < \infty\}$$

on la norma $\|f\|_{L^p}$ és

$$\|f\|_{L_p} = \left(\int_K |f(x)|^p \, dx\right)^{\frac{1}{p}} < \infty,$$

Per a $p = \infty$, l'espai $L_\infty(K)$ conté funcions que són essencialment fitades, és a dir, funcions que són fitades fora d'un conjunt de mesura nul·la, amb la norma definida com el suprem essencial del valor absolut de f:

$$\|f\|_{L_\infty} = \text{ess sup}_{x \in K} |f(x)|.$$

Definició 3.2. *(Derivada Feble)* Una funció f té una derivada feble d'ordre n si existeix una funció $g \in L_p(K)$ tal que per a totes les funcions test ϕ és a dir, funcions infinitament diferenciables amb suport compacte en K, es compleix

$$\int_K f \cdot D^n \phi \, dx = (-1)^n \int_K g \cdot \phi \, dx,$$

on $D^n \phi$ representa la derivada d'ordre n de ϕ. En aquest cas, diem que g és una derivada feble d'orde n de f.

És important destacar que la funció f pot no ser diferenciable en el sentit clàssic en tots els punts de K. No obstant això, l'equació anterior implica que f és diferenciable en el sentit feble, amb g actuant com la seva derivada feble. Això vol dir que l'equació es compleix gairebé a tot arreu, és a dir, excepte en un conjunt de mesura nul·la. Per visualitzar aquest concepte, observem el següent exemple.

Exemple 3.3. Observem els següents casos

1. Sigui $f(x)$ una funció diferenciable clàssicament amb derivada clàssica $f'(x)$.

Verifiquem que la derivada feble coincideix amb la derivada clàssica.

Considerem la integral

$$\int_K f(x)\varphi'(x)\,dx,$$

on φ és una funció test, és a dir, una funció infinitament diferenciable amb suport compacte en K. Aplicant la integració per parts:

$$\int_K f(x)\varphi'(x)\,dx = [f(x)\varphi(x)]_K - \int_K f'(x)\varphi(x)\,dx.$$

Com que φ té suport compacte dins de K, el primer terme $[f(x)\varphi(x)]_K$ s'anul·la als límits del conjunt K. Per tant,

$$\int_K f(x)\varphi'(x)\,dx = -\int_K f'(x)\varphi(x)\,dx.$$

Això demostra que $g(x) = f'(x)$ és la derivada feble de $f(x)$.

2. Considerem la funció característica dels nombres racionals en l'interval $[-1, 1]$, denotada per $\chi_{\mathbb{Q}}$. Aquesta funció es defineix com:

$$\chi_{\mathbb{Q}}(x) = \begin{cases} 1, & \text{si } x \in \mathbb{Q}; \\ 0, & \text{si } x \notin \mathbb{Q}. \end{cases}$$

La funció $\chi_{\mathbb{Q}}$ no és diferenciable en cap punt de $[-1, 1]$, però podem calcular la seva derivada feble. Com que la mesura de Lebesgue dels nombres racionals és zero, tenim:

$$\int_{-1}^{1} \chi_{\mathbb{Q}}(x)\varphi(x)\,dx = 0$$

per a qualsevol funció test φ.

Així, $g(x) = 0$ és la derivada feble de $\chi_{\mathbb{Q}}$. Això està d'acord amb la intuïció ja que, la funció característica dels nombres racionals és zero gairebé a tot arreu (excepte en un conjunt de mesura nul·la).

Un cop definit l'espai de Lebesgue L_p i el concepte de derivada feble, podem definir un subespai de L_p anomenat Espai de Sobolev que ens permet considerar un tipus de funcions amb certa "suavitat".

Definició 3.4. *(Espai de Sobolev $W^{n,p}(K)$)* Sigui $n \in \mathbb{N}_0$, $1 \leq p \leq \infty$. Aleshores, per un conjunt obert $K \subseteq \mathbb{R}^d$, definim l'espai de Sobolev com

$$W^{n,p}(K) := \{f \in L^p(K) : D^\alpha f \in L^p(K) \text{ per tot } \alpha \text{ amb } |\alpha| \leq n\}.$$

A més a més, per a $f \in W^{n,p}(K)$ i $1 \leq p < \infty$, definim la norma com:

$$\|f\|_{W^{n,p}(K)} := \left(\sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|^p_{L^p(K)} \right)^{1/p},$$

i per $p = \infty$, definim la norma com:

$$\|f\|_{W^{n,\infty}(K)} := \max_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^\infty(K)}.$$

L'espai de Sobolev $W^{n,p}(K)$ és un espai de funcions que no només són integrables en el sentit de L^p, sinó que també les seves derivades febles fins a un cert ordre n també ho són. Això significa que aquestes funcions tenen un cert grau de suavitat, depenent de l'ordre n i de l'exponent p. Notem que si $n = 0$, aleshores l'espai de Sobolev és simplement un espai de Lebesgue, és a dir, $W^{0,p}(K) = L^p(K)$.

Exemple 3.5. Considerem els següents tipus de funcions en un conjunt mesurable $K \subseteq \mathbb{R}$:

1. Polinomis: Qualsevol polinomi $P(x)$ pertany a $W^{n,p}(K)$ per a qualsevol n

i p. Això es deu al fet que els polinomis són infinitament diferenciables i les seves derivades són també polinomis, que són integrables en qualsevol interval finit.

2. Funcions contínues a trossos amb derivades a trossos: Una funció f que és contínua a trossos i té derivades contínues a trossos fins a l'ordre n pertany a $W^{n,p}(K)$. Un exemple típic és la funció $f(x) = |x|$, que pertany a $W^{1,p}(K)$.

3. Funcions L^p amb derivades febles en L^p: Qualsevol funció $f \in L^p(K)$ que té derivades febles fins a ordre n en $L^p(K)$ pertany a $W^{n,p}(K)$. Un exemple és la funció característica dels nombres racionals $\chi_{\mathbb{Q}}$ que hem vist que té derivada feble zero en 3.3.

Definim una restricció de l'espai de Sobolev que ens serà de gran utilitat.

Definició 3.6. *(Espai de funcions $F^p_{n,d}$)* L'espai $F^p_{n,d}$ és la restricció de l'espai de Sobolev $W^{n,p}(K)$ a dins de la bola unitat en $W^{n,p}([0,1]^d)$:

$$F^p_{n,d} = \{f \in W^{n,p}([0,1]^d) : \|f\|_{W^{n,p}([0,1]^d)} \leq 1\}$$

També, caldrà definir l'espai de Sobolev fraccionat:

Definició 3.7. *(Espai de Sobolev-Slobodeckij $W^{n,p}_r(K)$)* Definim l'espai de Sobolev-Slobodeckij $W^{n,p}_r(K)$ per a un conjunt obert $K \subseteq \mathbb{R}^d$, $0 < r < 1$, i $1 \leq p < \infty$, com el conjunt de totes les funcions $f \in L_p(K)$ tals que

$$\|f\|_{W^{n,p}_r(K)} = \left(\|f\|^p_{L_p(K)} + \sum_{|\alpha| \leq n} \int_K \int_K \frac{|D^\alpha f(x) - D^\alpha f(y)|^p}{|x-y|^{d+rp}} \, dx \, dy \right)^{1/p} < \infty,$$

on α és un multiíndex i $D^\alpha f$ representa la derivada feble d'ordre α. Per $p = \infty$, definim

$$\|f\|_{W^{n,\infty}_r(K)} = \max \left\{ \|f\|_{L_\infty(K)}, \text{ ess sup}_{x,y \in K, x \neq y} \frac{|D^n f(x) - D^n f(y)|}{|x-y|^r} \right\}.$$

L'espai de Sobolev-Slobodeckij $W_r^{n,p}(K)$ introdueix una idea fraccionada de la suavitat. En comptes de requerir que les derivades fins a l'ordre n siguin L^p-integrables, també es considera un quocient de diferències d'aquestes derivades.

Aquesta definició captura la idea de funcions que són "suaus en mitjana" fins a un cert grau. Això és especialment útil en l'anàlisi de funcions que no són totalment derivables però encara tenen una certa regularitat.

En el treball [5, Prop. 2.2] es demostra que tenim $F_{n,d}^p \subset W^{n,p}(K) \subseteq W_r^{n,p}(K)$.

Finalment, recordarem el concepte d'ordre d'una funció, ja que ens resultarà útil per la següent secció.

Definició 3.8. Siguin f i g dues funcions de variables reals o enteres. Es defineixen les següents notacions asimptòtiques:

1. Es diu que $f(x)$ és de l'ordre de $g(x)$, denotat $f(x) = O(g(x))$, quan $x \to \infty$, si existeixen constants positives C i x_0 tals que per a tot $x \geq x_0$, es compleix

$$|f(x)| \leq C|g(x)|.$$

2. Es diu que $f(x)$ és de l'ordre inferior de $g(x)$, denotat $f(x) = \Omega(g(x))$, quan $x \to \infty$, si existeixen constants positives c i x_0 tals que per a tot $x \geq x_0$, es compleix

$$|f(x)| \geq c|g(x)|.$$

3.2 Primeres intuïcions sobre la millora amb profunditat

En aquesta secció, comprovarem que la profunditat és una millora en la eficiència (nombre de neurones) a l'hora d'aproximar.

El 2017, Yarotsky va demostrar un resultat fonamental sobre l'aproximació de funcions utilitzant xarxes neuronals amb funció d'activació ReLU en el següent

teorema.

Teorema 3.9. *[13, Thm. 1] Per a qualsevol $d, n \in \mathbb{N}$ i $\varepsilon \in (0,1)$, hi ha una xarxa neuronal, amb funció d'activació ReLU que:*

1. *és capaç d'expressar qualsevol funció de $F_{n,d}^{\infty}$ amb un error ε;*
2. *té una profunditat com a màxim $c(\ln(1/\varepsilon)+1)$ i com a màxim $c\varepsilon^{-d/n}(\ln(1/\varepsilon)+1)$ neurones, amb una constant $c = c(d,n)$.*

En el següent teorema, es presenta una generalització d'aquest resultat que es va obtenir per a funcions de $F_{n,d}^{p}$ amb l'error mesurat en la norma $W_r^{n,p}$ per $0 \leq r \leq 1$.

Teorema 3.10. *[6, Thm. 4.1] Sigui $d \in \mathbb{N}$, $n \in \mathbb{N}_{\geq 2}$, $1 \leq p \leq \infty$ i $0 \leq r \leq 1$. Aleshores, existeix una constant $c = c(d,n,p,r) > 0$ amb les següents propietats:*

Per a tot $f \in F_{n,d}^{p}$, existeix una xarxa neuronal $\Phi_{f,\varepsilon}$ la qual és capaç d'aproximar f amb un error menor que ε en la norma $W_p^{n,r}$ (on $0 \leq r \leq 1$) tal que

1. $M(\Phi_{f,\varepsilon}) \leq c \cdot \varepsilon^{-d/(n-r)} \cdot \log_2(\varepsilon^{-n/(n-r)})$,
2. $C(\Phi_{f,\varepsilon}) \leq c \cdot \log_2(\varepsilon^{-n/(n-r)})$,

Observem que si considerem que $r = 0$ i $p = \infty$, és a dir $W_{\infty}^{n,0}([0,1]^d) := F_{n,d}^{\infty}$, aleshores ens trobem en el mateix cas que el teorema 3.9.

I ara ve una peça clau dins del nostre anàlisi de profunditat de les xarxes neuronals. En els teoremes anteriors no hem fixat el nombre de capes (profunditat) que la xarxa neuronal ha de tenir per tal d'aproximar amb un error determinat. Llavors, què passaria si fixem el nombre de capes? Quin serà el nombre de neurones necessari?

En Yarotsky, en el mateix treball, va respondre a les preguntes anteriors amb el següent teorema.

Teorema 3.11. *[13, Thm. 6] Sigui $f \in C^2([0,1]^d)$ una funció no lineal (és a dir, no de la forma $f(x_1,\dots,x_d) = a_0 + \sum_{k=1}^{d} a_k x_k$ en tot $[0,1]^d$). Llavors, per a qualsevol N fix, una xarxa neuronal amb profunditat N i funció d'activació ReLU aproximant f amb un error $\varepsilon \in (0,1)$ ha de tenir almenys un ordre $O(\epsilon^{-1/(2(N-1))})$ de neurones.*

És important remarcar que si $f \in W^{n,\infty}([0,1]^d)$ amb $n > 2$, llavors f té derivades parcials fins a l'ordre 2 que són acotades i contínues gairebé per tot arreu, i per tant, $f \in C^2([0,1]^d)$. D'on es dedueix la inclusió, $W^{n,\infty}([0,1]^d) \subseteq C^2([0,1]^d)$.

Amb aquesta última clarificació, estem preparats per posar en comparació els dos darrers teoremes, considerant $n > 2$ i una funció $f \in W^{n,\infty}([0,1]^d) \subseteq C^2([0,1]^d)$. Comprovem matemàticament que les fites estan en concordança amb que fixar el nombre de capes serà més restrictiu que no fixar-lo, i que, de fet, és estrictament més restrictiu.

Recordem que en el teorema 3.10, considerant $r = 0$, tenim el nombre de neurones de l'ordre $O(\varepsilon^{-d/n}\ln(1/\varepsilon))$ i, pel teorema 3.11 és de l'ordre $O(\varepsilon^{-1/(2(N-1))})$. Per fer la comparativa, considerem $x = 1/\varepsilon$ i fem tendir x a infinit en el següent limit

$$\lim_{x\to\infty} \frac{x^{\frac{1}{2(N-1)}}}{x^{d/n}\ln(x)} = \lim_{x\to\infty} \frac{x^{\alpha}}{\ln(x)}$$

per algun $\alpha > 0$. Ens queda una indeterminació favorable per aplicar la regla de L'Hopital d'on

$$\lim_{x\to\infty} \frac{\alpha x^{\alpha-1}}{1/x} = \lim_{x\to\infty} \alpha x^{\alpha} = +\infty$$

Com el límit és infinit, significa que el numerador creix més ràpidament a infinit que el denominador. És a dir, que

$$O(\varepsilon^{-d/n}\ln(1/\varepsilon)) < O(\varepsilon^{-1/(2(N-1))}).$$

almenys quan es compleix que $\frac{d}{n} < \frac{1}{2(N-1)}$.

4 Limitacions de la profunditat dos

Fins ara, hem tractat l'expressivitat de les xarxes neuronals i la seva capacitat d'aproximació sota certes condicions amb un nombre exponencial de neurones. Hem començat a plantejar la qüestió entre la grandària i la profunditat: hauríem d'utilitzar xarxes que siguin estretes i profundes (moltes capes, amb un petit nombre de neurones per capa), o poc profundes i amples?

Ara, el nostre objectiu es definir una família de funcions que poden ser aproximades per una xarxa de profunditat 3 de grandària polinomial però, en canvi, no ho poden ser per una xarxa de profunditat 2 de grandària exponencial.

Treballarem amb funcions definides sobre l'esfera unitat $\mathbb{S}^{d-1}$. Notem que l'espai $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ està inclòs dins l'espai euclidià $\mathbb{R}^{2d}$. De fet,

$$\mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \subseteq \mathbb{R}^d \times \mathbb{R}^d \cong \mathbb{R}^{2d}$$

Això ens permetrà no haver de redefinir el concepte de xarxa neuronal que prèviament hem definit en la secció 2.2.

Tot i no caler redefinir el concepte de xarxa neuronal amb aquest nou espai d'entrada, ens serà útil tenir una visualització de com son les xarxes neuronals en $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$.

Sigui $\Phi : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ una xarxa neuronal de profunditat 2 amb amplada r i pesos limitats per B, aleshores

$$\Phi(\mathbf{x}, \mathbf{x}') = w_2^T \sigma(W_1 \mathbf{x} + W_1' \mathbf{x}' + b_1) + b_2, \tag{3}$$

on $W_1, W_1' \in [-B, B]^{r \times d}, w_2 \in [-B, B]^r, b_1 \in [-B, B]^r,\ b_2 \in [-B, B]$ i $(x, x') \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$.

Si fos de profunditat 3, aleshores seria

$$\Phi(\mathbf{x}, \mathbf{x}') = w_3^T \sigma\left(W_2 \sigma\left(W_1 \mathbf{x} + W_1' \mathbf{x}' + b_1\right) + b_2\right) + b_3$$

per $W_1, W_1' \in [-B, B]^{r \times d}, W_2 \in [-B, B]^{r \times r}, w_3 \in [-B, B]^r, b_1, b_2 \in [-B, B]^r$, $b_3 \in [-B, B]$ i $(x, x') \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$

4.1 Teorema principal

Introduïm la mesura μ_d, que s'utilitza freqüentment en l'anàlisi de funcions sobre esferes unitat S^{d-1} en l'espai euclidià $\mathbb{R}^d$. Definim la mesura de probabilitat μ_d sobre l'interval $[-1, 1]$ com:

$$d\mu_d(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1 - x^2)^{\frac{d-3}{2}} dx,$$

on Γ representa la funció Gamma. En el següent exemple, podem interpretar millor aquesta mesura.

Exemple 4.1. Considerem el cas $d = 3$. Utilitzant les propietats de la funció Gamma, tenim que

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\sqrt{\pi} \quad \text{i} \quad \Gamma(1) = 1.$$

aleshores queda

$$d\mu_3(x) = \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{\pi}\Gamma(1)}(1 - x^2)^0 \, dx = \frac{\frac{1}{2}\sqrt{\pi}}{\sqrt{\pi} \cdot 1} \, dx = \frac{1}{2} \, dx.$$

Ara, calculem la mesura de l'interval $[a, b] \subset [-1, 1]$ sota aquesta mesura:

$$\mu_3([a, b]) = \int_a^b \frac{1}{2} \, dx = \frac{1}{2} \int_a^b dx = \frac{1}{2}(b - a).$$

Per més context sobre d'on prové la mesura, consultar [1, Ch. 1, Ch. 2].

Definim $\mathcal{A}_{n,d}(f)$ com la norma $L^2(\mu_d)$ de la diferència entre la funció f i el millor polinomi p de grau fins a $n-1$ que s'acosta a f. És a dir

$$\mathcal{A}_{n,d}(f) = \min_{p \text{ polinomi de grau } n-1} \|f - p\|_{L^2(\mu_d)}$$

Definim també $K_{d,n}$ com una constant que depèn de la dimensió d'entrada $d \geq 2$ i el grau $n \geq 1$

$$K_{d,n} = \binom{d+n-1}{d-1} - \binom{d+n-3}{d-1} = \frac{(2n+d-2)(n+d-3)!}{n!(d-2)!}.$$

Així, ja estem llestos per enunciar el teorema principal.

Teorema 4.2. *[4, Thm. 1] Sigui $F : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ la funció objectiu que desitgem aproximar, tal que $F(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle)$, on f és una funció contínua en l'interval $[-1, 1]$ i $\langle \mathbf{x}, \mathbf{x}' \rangle$ denota el producte interior. I sigui $\Phi : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ una xarxa neuronal amb funció d'activació σ, profunditat 2 i amplada r, amb pesos limitats per una constant B. Llavors, per a qualsevol $n \geq 1$, l'error d'aproximació entre Φ i F en norma L^2 compleix:*

$$\|\Phi - F\|_{L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})} \geq \mathcal{A}_{n,d}(f) \left(\mathcal{A}_{n,d}(f) - \frac{2rB \max_{|x| \leq 2\sqrt{d}B+B} |\sigma(x)| + 2B}{\sqrt{K_{d,n}}} \right)$$

Per tal de comprendre la cota inferior que acabem d'enunciar, cal destacar el paper crucial que juga la n:

- La constant $K_{d,n}$ augmenta a mesura que n creix i disminueix si n decreix. El valor mínim que pot prendre és $K_{2,1} = 2$ i no està fitat superiorment.

- No obstant, $\mathcal{A}_{n,d}(f)$ actua contràriament. Quan n creix, $A_{n,d}(f)$ disminueix i viceversa.

Per tant, si n és gran, el valor de la fracció $\frac{2rB \max_{|x| \leq \sqrt{4d}B+B} |\sigma(x)|+2B}{\sqrt{K_{d,n}}}$ tendeix cap a zero perquè el denominador $\sqrt{K_{d,n}}$ creix. Al mateix temps, $\mathcal{A}_{n,d}(f)$ disminueix.

Això fa que la cota inferior també decreixi, permetent una millor aproximació amb un error menor.

Per contra, si n és petita, $K_{d,n}$ com a mínim serà 2, fent que la fracció sigui un terme constant. En aquest cas, $\mathcal{A}_{n,d}(f)$ creixerà molt i llavors l'error d'aproximació serà gran.

Aquest resultat subratlla les limitacions intrínseques en l'ús de xarxes neuronals amb una capacitat finita (limitada per r i B) per aprendre funcions amb un cert grau de complexitat. No podem esperar que la nostra xarxa neuronal faci una feina millor que això.

4.2 Demostració del Teorema Principal

Abans de donar la demostració, necessitem definir certs conceptes.

Definició 4.3. *(Funcions Separables de* $(\mathbf{v}, \mathbf{v}')$*)* Una funció f és separable respecte a parelles de vectors $\mathbf{v}$ i $\mathbf{v}'$ si es pot expressar com $f(\mathbf{x}, \mathbf{x}') = \psi(\langle \mathbf{v}, \mathbf{x} \rangle, \langle \mathbf{v}', \mathbf{x}' \rangle)$, per alguna $\psi : [-1, 1]^2 \to \mathbb{R}$.

Definició 4.4. *(Polinomis de Legendre)* Els polinomis Legendre de dimensió d són la seqüència de polinomis sobre $[-1, 1]$ definits per la fórmula de recursió

$$P_{n,d}(x) = \frac{2n + d - 4}{n + d - 3} x P_{n-1,d}(x) - \frac{n - 1}{n + d - 3} P_{n-2,d}(x)$$
$$P_{0,d} \equiv 1, P_{1,d}(x) = x$$

Definim un cert tipus de funcions que seran útils per a la demostració del teorema.

Definició 4.5. Sigui $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$. Definim $h_n : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ per

$$h_n(\mathbf{x}, \mathbf{x}') = \sqrt{K_{d,n}} P_{n,d}(\langle \mathbf{x}, \mathbf{x}' \rangle)$$

Per comoditat denotarem $L_n^{\mathbf{x}}(\mathbf{x}') = h_n(\mathbf{x}, \mathbf{x}')$ i, d'ara en endavant, denotarem els polinomis de Legendre $P_{n,d}$ simplement amb P_n.

En el treball [1, Prop. 2.26], van demostrar que

$$\int_{S^{d-1}} P_n^2(\langle \mathbf{x}, \mathbf{x}' \rangle)\, d\mu_d(\mathbf{x}) = \frac{1}{K_{d,n}} \tag{4}$$

Al llarg de la demostració del teorema farem ús de les següents propietats dels polinomis Legendre, tenint en compte el resultat anterior equació 4.

Lema 4.6. *(Propietats dels polinomis de Legendre)*

1. *Per a cada $d \geq 2$, la seqüència $\{\sqrt{K_{d,n}} P_n\}$ és una base ortonormal de l'espai de Hilbert $L^2(\mu_d)$.*
2. *Per a cada n, $||P_n||_\infty = 1$ i $P_n(1) = 1$.*
3. *$\langle L_i^x, L_j^{x'} \rangle = P_i(\langle x, x' \rangle)\delta_{ij}$.*

Definició 4.7. *(Operador de Projecció Ortogonal $\mathcal{O}_n$)* Donat un espai de funcions $L^2(\mu_d)$ sobre l'interval $[-1, 1]$ amb la mesura de probabilitat μ_d, definim l'operador de projecció $\mathcal{O}_n : L^2(\mu_d) \to L^2(\mu_d)$ com la projecció a l'espai complementari dels polinomis de grau $\leq n - 1$.

Aleshores, si f es descompon com $f = \sum_{k=0}^{\infty} \alpha_k P_k$, llavors la projecció ortogonal $\mathcal{O}_n f$ és

$$\mathcal{O}_n f = \sum_{k=n}^{\infty} \alpha_k P_k.$$

Notem doncs que llavors $\mathcal{A}_{n,d}(f) = \|\mathcal{O}_n f\|_{L^2(\mu_d)}$.

El següent lemma serà fonamental en la prova del Teorema principal 4.2.

Lema 4.8. *(Lemma principal) Sigui $f : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ una funció de producte interior i sigui $g_1, \ldots, g_r : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ funcions separables. Llavors*

$$\left\| f - \sum_{j=1}^{r} g_j \right\|^2 \geq \|\mathcal{O}_n f\| \left(\|\mathcal{O}_n f\| - \frac{2 \sum_{j=1}^{r} \|g_j\|}{\sqrt{K_{d,n}}} \right)$$

Demostració. Com hem vist abans, els polinomis de Legendre formen una base ortonormal, fet que ens permet descompondre les funcions f i g_j en sumes infinites d'aquests polinomis.

Considerem f com una suma infinita dels polinomis ortonormals

$$f = \sum_{i=0}^{\infty} \alpha_i h_i \tag{5}$$

on h_j estan definits en la secció 4.5 i α_i són coeficients reals que representen les contribucions de cada component ortonormal h_i.

Sigui $\mathbf{v}_j, \mathbf{v}'_j \in S^{d-1}$, escrivim

$$g_j(\mathbf{x}, \mathbf{x}') = \sum_{k,l=0}^{\infty} \beta_{k,l}^j L_k^{\mathbf{v}_j}(\mathbf{x}) L_l^{\mathbf{v}'_j}(\mathbf{x}').$$

on $L_k^{\mathbf{v}_j}, L_l^{\mathbf{v}'_j}$ definits en la secció 4.5 i els coeficients reals $\beta_{k,l}^j$ representen les contribucions de cada component ortonormal $L_k^{\mathbf{v}_j}$ i $L_l^{\mathbf{v}'_j}$ a la funció g_j.

Ara bé, quan $k \neq l$, els productes $L_k^{\mathbf{v}_j}(\mathbf{x}) L_l^{\mathbf{v}'_j}(\mathbf{x}')$ són ortogonals a f. Això significa que només els termes amb $k = l$ contribueixen a la norma de la diferència $\|f - \sum_{j=1}^{r} g_j\|^2$. Per tant, podem assumir sense pèrdua de generalitat que cada g_j és de la forma:

$$g_j(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} \beta_i^j L_i^{\mathbf{v}_j}(\mathbf{x}) L_i^{\mathbf{v}'_j}(\mathbf{x}'). \tag{6}$$

Un cop vist això ja podem procedir a demostrar la desigualtat del teorema. Primer, substituïm la part de la esquerra per les seves expressions en termes de les

equacions 5 i 6

$$\|f-\sum_{j=1}^{r} g_j\|^2 = \left\|\sum_{i=0}^{\infty}\alpha_i h_i - \sum_{i=0}^{\infty}\sum_{j=1}^{r}\beta_i^j L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\|^2$$
$$= \left\|\sum_{i=0}^{\infty}\left(\alpha_i h_i - \sum_{j=1}^{r}\beta_i^j L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right)\right\|^2 = \sum_{i=0}^{\infty}\left\|\alpha_i h_i - \sum_{j=1}^{r}\beta_i^j L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\|^2$$

on en la última igualtat apliquem el conegut Teorema de Pitàgores, [2, Thm 2.2].

Realment només cal considerar els termes a partir de $i=n$ perquè aquests termes representen la part de f que no pot ser aproximada pels polinomis de grau menor o igual a $n-1$. Aquesta part és exactament el que quantifica la projecció ortogonal $\mathcal{O}_n f$. Per tant, podem escriure la desigualtat següent

$$\|f-\sum_{j=1}^{r} g_j\|^2 \geq \sum_{i=n}^{\infty}\left\|\alpha_i h_i - \sum_{j=1}^{r}\beta_i^j L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\|^2$$

Utilitzem la propietat ortogonal per separar els termes quadràtics

$$\|f-\sum_{j=1}^{r} g_j\|^2 \geq \sum_{i=n}^{\infty}\alpha_i^2 - 2\sum_{i=n}^{\infty}\sum_{j=1}^{r}\left\langle \alpha_i h_i, \beta_i^j L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\rangle \tag{7}$$

El porducte intern el podem calcular com

$$\left\langle h_i, L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\rangle = \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} h_i(\mathbf{x},\mathbf{x}')L_i^{\mathbf{v}_j}(\mathbf{x})L_i^{\mathbf{v}'_j}(\mathbf{x}')d\mu_d(\mathbf{x})d\mu_d(\mathbf{x}')$$

$$= \int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \sqrt{K_{d,i}}P_i(\langle\mathbf{x},\mathbf{x}'\rangle)\sqrt{K_{d,i}}P_i(\langle\mathbf{v}_j,\mathbf{x}\rangle)\sqrt{K_{d,i}}P_i(\langle\mathbf{v}'_j,\mathbf{x}'\rangle)d\mu_d(\mathbf{x})d\mu_d(\mathbf{x}')$$

d'on usant la condició d'ortogonalitat $\int_{\mathbb{S}^{d-1}} P_i(\langle\mathbf{v},\mathbf{x}'\rangle)P_i(\langle\mathbf{v}',\mathbf{x}'\rangle)d\mu_d(\mathbf{x}') = \delta_{i,j}P_i(\langle\mathbf{v},\mathbf{v}'\rangle)$ obtenim que

$$\left\langle h_i, L_i^{\mathbf{v}_j}\otimes L_i^{\mathbf{v}'_j}\right\rangle = K_{d,i}^{3/2}\delta_{i,j}\delta_{i,j}P_i(\langle\mathbf{v}_j,\mathbf{v}'_j\rangle) = K_{d,i}^{3/2}\frac{P_i(\langle\mathbf{v}_j,\mathbf{v}'_j\rangle)}{K_{d,i}} = \frac{P_i\left(\langle\mathbf{v}_j,\mathbf{v}'_j\rangle\right)}{\sqrt{K_{d,i}}}$$

Substituint aquest últim resultat en l'equació 7, obtenim

$$\|f - \sum_{j=1}^{r} g_j\|^2 \geq \sum_{i=n}^{\infty} \alpha_i^2 - 2\sum_{i=n}^{\infty}\sum_{j=1}^{r} \alpha_i \beta_i^j \frac{P_i\left(\langle \mathbf{v}_j, \mathbf{v}'_j \rangle\right)}{\sqrt{K_{d,i}}}$$

Però recordant que $\mathcal{O}_n f$ és la projecció ortogonal de f a l'espai complementari dels polinomis de grau menor o igual a $n-1$. Tenim, $\|\mathcal{O}_n f\|^2 = \sum_{i=n}^{\infty} \alpha_i^2$. Finalment quedaria

$$\|f - \sum_{j=1}^{r} g_j\|^2 \geq \|\mathcal{O}_n f\|^2 - 2\sum_{i=n}^{\infty}\sum_{j=1}^{r} \frac{\beta_i^j \alpha_i P_i\left(\langle \mathbf{v}_j, \mathbf{v}'_j \rangle\right)}{\sqrt{K_{d,i}}}$$

Aplicant de la desigualtat de Cauchy-Schwarz per separar les sumes de termes, tenim

$$\begin{aligned}\|f - \sum_{j=1}^{r} g_j\|^2 &\geq \|\mathcal{O}_n f\|^2 - 2\sum_{j=1}^{r}\sum_{i=n}^{\infty} \frac{\left|\beta_i^j \|\alpha_i\right|}{\sqrt{K_{d,n}}} \\ &\geq \|\mathcal{O}_n f\|^2 - 2\sum_{j=1}^{r} \frac{1}{\sqrt{K_{d,n}}} \sqrt{\sum_{i=n}^{\infty} \left|\beta_i^j\right|^2} \sqrt{\sum_{i=n}^{\infty} |\alpha_i|^2}\end{aligned}$$

Finalment, tenint en compte altre cop que $\|\mathcal{O}_n f\|^2 = \sum_{i=n}^{\infty} \alpha_i^2$ i la definició prèvia de g_j, arribem finalment a la desigualtat

$$\|f - \sum_{j=1}^{r} g_j\|^2 \geq \|\mathcal{O}_n f\|^2 - \frac{2\,\|\mathcal{O}_n f\| \sum_{j=1}^{r} \|g_j\|}{\sqrt{K_{d,n}}}$$

□

Ara ens interessa veure la relació del lemma anterior amb el Teorema Principal 4.2. En el context d'una xarxa neuronal de profunditat 2 i amplada r, cada neurona de la capa oculta pren una combinació lineal de les entrades, $(\mathbf{x}, \mathbf{x}') \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, aplica una funció d'activació, σ, i produeix una sortida

$$g_j(\mathbf{x}, \mathbf{x}') = \sigma(\mathbf{w}_j^T \mathbf{x} + \mathbf{w}_j'^T \mathbf{x}' + b_j)$$

on $\mathbf{w}_j$ i $\mathbf{w}'_j$ són les files de les matrius $\mathbf{W}_1$ i $\mathbf{W}'_1$, respectivament, és a dir $\mathbf{w}'_j, \mathbf{w}_j \in$

$[-B, B]^{1\times d}$ i b_j és el biaix de la j-èsima neurona de la capa oculta. Aquestes sortides, les podem entendre com a funcions separables.

Per tant, la funció total implementada per la xarxa neuronal Φ es pot veure com una suma de funcions separables g_j proporcionades per cada neurona

$$\Phi(x, x') = \sum_{i=1}^{r} g_j(x, x') \tag{8}$$

Només queda veure que $\sum_{j=0}^{r} \|g_j\| = rB \max_{|x|\leq 2\sqrt{d}B+B} |\sigma(x)| + B$.

Notem que

$$|\mathbf{w}_j^T \mathbf{x}| \leq \|\mathbf{w}_j\| \|\mathbf{x}\| \leq \sqrt{d}B,$$

on $\|\mathbf{x}\| \leq 1$ perquè $\mathbf{x} \in S^{d-1}$.

De manera similar

$$|\mathbf{w'}_j^T \mathbf{x'}| \leq \sqrt{d}B.$$

Combinem aquestes contribucions i obtenim

$$|\mathbf{w}_j^T \mathbf{x} + \mathbf{w'}_j^T \mathbf{x'} + b_j| \leq |\mathbf{w}_j^T \mathbf{x}| + |\mathbf{w'}_j^T \mathbf{x'}| + |b_j| \leq \sqrt{d}B + \sqrt{d}B + B = 2\sqrt{d}B + B.$$

Aplicant l'activació σ a aquesta entrada màxima, obtenim:

$$|\sigma(\mathbf{w}_j^T \mathbf{x} + \mathbf{w'}_j^T \mathbf{x'} + b_j)| \leq \max_{|x|\leq 2\sqrt{d}B+B} |\sigma(x)|.$$

Per tant, la norma de la funció separable implementada per la neurona oculta està acotada per:

$$\|g_j\| \leq B \max_{|x|\leq 2\sqrt{d}B+B} |\sigma(x)|$$

Aleshores, sumant la contribució de les r neurones i sumant el biaix màxim de la

neurona primera neurona, obtenim

$$\sum_{j=0}^{r} \|g_j\| = rB \max_{|x|\leq 2\sqrt{d}B+B} |\sigma(x)| + B$$

5 La millora amb profunditat tres

Finalment, presentem i demostrem el corol·lari 5.3 el qual serà crucial per respondre, en la secció 6.2, a la qüestió que portem tractant al llarg de tot el treball, són realment les xarxes neuronals profundes millors a l'hora d'aproximar que les poc profundes?

Abans, però, recordarem el concepte de funció Lipschitz i un Lemma que ens serà de gran utilitat.

Definició 5.1. Una funció $f : \mathbb{R} \to \mathbb{R}$ és L-Lipschitz si existeix una constant $L \geq 0$ tal que, per a tots els $x, y \in \mathbb{R}$, es compleix:

$$|f(x) - f(y)| \leq L|x - y|$$

Aquesta condició implica que la funció f no varia més ràpidament que una recta de pendent L, és a dir, f té una variació controlada i és contínua.

Lema 5.2. *([4, Lem. 5]) Sigui $\sigma(x) = \max(x, 0)$ l'activació ReLU, f: $[-R, R] \to \mathbb{R}$ una funció L-Lipschitz, i $\epsilon > 0$. Existeixen $|\beta_i| \leq R, |\alpha_i| \leq 2L, \gamma_i \in \{-1, 1\}$ i*

$$g(x) = f(0) + \sum_{i=1}^{m} \alpha_i \sigma(\gamma_i x - \beta_i)$$

L-Lipschitz en tot $\mathbb{R}$, tals que $\|g - f\|_\infty \leq \epsilon$. A més a més, $m \leq \frac{2RL}{\epsilon}$

Amb aquests conceptes ben definits, tenim tots els ingredients per enunciar el corol·lari.

Corol·lari 5.3. *Sigui* $f : [-1,1] \to [-1,1]$ *una funció L-Lipschitz i sigui* $\epsilon > 0$. *Definim* F*:* $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [-1,1]$ *per* $F(\mathbf{x},\mathbf{x}') = f(\langle \mathbf{x},\mathbf{x}'\rangle)$. *Hi ha una funció* $G : \mathbb{S}^{d-1}\times\mathbb{S}^{d-1} \to [-1,1]$ *que satisfà* $\|F-G\|_\infty \leq \epsilon$ *i a més* G *pot ser implementada per una xarxa neuronal de profunditat 3, de funció d'activació ReLU amb amplada de limit superior* $\leq \frac{16d^2L}{\epsilon}$ *i pesos limitats per* $\max(4, 2L)$.

Demostració. Inicialment, considerem $f(x) = \frac{x^2}{2}$ en $[-2,2]$. Al ser L-Lipschitz en $[-R,R]$ amb $R = 2$ i $L = 2$, pel Lema 5.2, podem aproximar f amb una funció Φ_{square} de la forma:

$$\Phi_{\text{square}}(x) = f(0) + \sum_{i=1}^{m} \alpha_i \sigma(\gamma_i x - \beta_i)$$

tal que, per construcció, $\|\Phi_{\text{square}} - f\|_\infty \leq \frac{\epsilon}{2dL}$ i amb una amplada màxima $m \leq \frac{16dL}{\epsilon}$, biaixos limitats per $|\beta_i| \leq R = 2$ i pesos de la capa de predicció limitats per $|\alpha_i| \leq 2L = 4$.

Per cada $i \in [d]$ podem compondre la funció lineal $(\mathbf{x},\mathbf{x}') \mapsto x_i + x_i'$ amb Φ_{square} per obtenir una xarxa de profunditat 2, Φ_i, que calcula $f(\mathbf{x},\mathbf{x}') = \frac{(x_i+x_i')^2}{2}$ amb un error de $\frac{\epsilon}{2dL}$ i amb la mateixa amplada i límit de pes que Φ_{square}.

Ara ens interessa obtenir una xarxa neuronal que calculi

$$\langle \mathbf{x},\mathbf{x}'\rangle = \sum_{i=1}^{d} x_i x_i'$$

amb un error de $\frac{\epsilon}{2L}$. Per fer-ho, ens adonem que, sumant les xarxes neuronals descrites en el pas anterior tenim que

$$\sum_{i=1}^{d} \Phi_i = \frac{1}{2}\sum_{i=1}^{d}(x_i + x_i')^2 = \frac{1}{2}\sum_{i=1}^{d} x_i^2 + \frac{1}{2}\sum_{i=1}^{d} x_i'^2 + \sum_{i=1}^{d} x_i x_i'$$

i, com $\mathbf{x},\mathbf{x}' \in \mathbb{S}^{d-1}$, la seva norma euclidiana és igual a 1 (p.e. $\|\mathbf{x}\|^2 = \sum_{i=1}^{d} x_i^2 =$

1). Per tant, si considerem

$$\Phi_{\text{inner}} = \sum_{i=1}^{d} \Phi_i - 1 = \sum_{i=1}^{d} x_i x_i'$$

tenim una xarxa neuronal que calcula $\langle \mathbf{x}, \mathbf{x}' \rangle$ amb un error de $d\frac{\epsilon}{2dL} = \frac{\epsilon}{2L}$ i amb una amplada de $d\frac{16dL}{\epsilon} = \frac{16d^2L}{\epsilon}$, biaixos limitats per 2, i pesos de la capa de predicció limitats per 4.

Ara, de nou pel Lema 5.2, hi ha una xarxa de profunditat 2, Φ_f, que calcula f en $[-1, 1]$, amb un error de $\frac{\epsilon}{2}$, té una amplada màxima de $\frac{4L}{\epsilon}$, els biaixos $|\beta_i| \leq R = 1$ i els pesos estan limitats per $|\alpha_i| \leq 2L$, i és L-Lipschitz.

Finalment, considerem la xarxa de profunditat 3, com la composició següent

$$\Phi_F(\mathbf{x}, \mathbf{x}') = \Phi_f(\Phi_{\text{inner}}(\mathbf{x}, \mathbf{x}'))$$

Quan combinem Φ_{inner} i Φ_f, l'amplada màxima de la xarxa composta serà el màxim d'aquestes dues amplades:

$$\max\left(\frac{16d^2L}{\epsilon}, \frac{4L}{\epsilon}\right) = \frac{16d^2L}{\epsilon}$$

Per tant, Φ_F té una amplada màxima de $\frac{16d^2L}{\epsilon}$.

Els límits dels pesos es mantenen perquè, en la composició de xarxes, no afegim més capes que les ja existents en les xarxes originals, per tant, els pesos de les neurones en aquestes capes un límit de pes de $\max(4, 2L)$.

$$\begin{aligned} |\Phi_F(\mathbf{x}, \mathbf{x}') - F(\mathbf{x}, \mathbf{x}')| &= |\Phi_f(\Phi_{\text{inner}}(\mathbf{x}, \mathbf{x}')) - f(\langle \mathbf{x}, \mathbf{x}' \rangle)| \\ &\leq |\Phi_f(\Phi_{\text{inner}}(\mathbf{x}, \mathbf{x}')) - \Phi_f(\langle \mathbf{x}, \mathbf{x}' \rangle)| + |\Phi_f(\langle \mathbf{x}, \mathbf{x}' \rangle) - f(\langle \mathbf{x}, \mathbf{x}' \rangle)| \end{aligned}$$

al ser Φ_f L-Lipschitz i tenint un error d'aproximació de f de $\frac{\epsilon}{2}$, ens queda que

$$\begin{aligned} &\leq L\left|\Phi_{\text{inner}}\ (\mathbf{x},\mathbf{x}') - \langle \mathbf{x},\mathbf{x}'\rangle\right| + \frac{\epsilon}{2} \\ &\leq L\frac{\epsilon}{2L} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

Com volíem veure, hem trobat una xarxa neuronal de profunditat 3 que es capaç de ϵ-aproximar F. □

6 Comparativa de profunditat dos i tres

En aquesta última secció, finalment presentem un exemple on es veu clarament que, sota certes condicions, la profunditat tres supera a la profunditat dos. Abans però, considerarem el següent Lemma.

Lema 6.1. *[4, Lem. 4] Definim $g_{d,m}(x) = \sin(\pi\sqrt{d}mx)$. Llavors, per a qualsevol $d \geq d_0$, per una constant universal $d_0 > 0$, i per a qualsevol polinomi p de grau k tenim que*

$$\int_{-1}^{1} (g_{d,m}(x) - p(x))^2 d\mu_d(x) \geq \frac{m-k}{4e\pi m}$$

Ara, ja estem llestos per comprovar el potencial dels resultats vists al llarg del treball.

Exemple 6.2. Considerem el cas en què la funció d'activació és la funció ReLU $\sigma(x) = \max(0, x)$. Sigui la funció objectiu $f(x) = \sin(\pi d^3 x)$, $n = d^2$ i el limit dels pesos com $B = 2^d$. En aquest cas, pel Lema 6.1 agafant $m = d^{2.5}$ i $k = 0$ tenim

$$\int_{-1}^{1} (\sin(\pi d^3 x) - p(x))^2 d\mu_d(x) \geq \frac{d^{2.5} - 0^2}{4e\pi d^{2.5}} = \frac{1}{4e\pi}$$

Per tant, podem considerar que $\mathcal{A}_{n,d}(f) \geq \frac{1}{5e\pi}$.

Per tenir una aproximació de $\frac{1}{50e^2\pi^2}$ de F, amb una xarxa neuronal de profunditat 2, aplicant el Teorema 4.2 tenim que

$$\frac{1}{50e^2\pi^2} \geq \frac{1}{5e\pi}\left(\frac{1}{5e\pi} - \frac{2r2^d2^d(1+\sqrt{4d})+2\cdot 2^d}{\sqrt{K_{d,d^2}}}\right)$$

Simplificant i reordenant termes amb l'objectiu d'aïllar el nombre de neurones de la capa oculta r, arribem a

$$r \geq \frac{\sqrt{K_{d,d^2}}}{20e\pi 2^{2d}(1+\sqrt{4d})+2^{d+1}}$$

D'on, sabent que $\sqrt{K_{d,d^2}} = O(d^{d-2})$, demostrat en [2, Ch 2. Eq. 2.12].

Per d molt grans, tindrem que

$$r \geq \frac{d^{d-2}}{20e\pi 2^{2d}(1+\sqrt{4d})+2^{d+1}} = \frac{2^{d\log_2(d-2)}}{20e\pi 2^{2d}(1+\sqrt{4d})+2^{d+1}} \approx 2^{d(\log_2(d-2)-2)}$$

Podem dir que el valor de r ha de ser

$$r = 2^{\Omega(d\log_2(d))}$$

on Ω és la contrapartida de O, recordem la definició 3.8.

D'altra banda, $f(x) = \sin(\pi d^3 x)$ és L-Lipschitz amb $L = \pi d^3$ perquè:

$$|f(x_1) - f(x_2)| = |\sin(\pi d^3 x_1) - \sin(\pi d^3 x_2)| \leq \pi d^3 |x_1 - x_2|$$

per tant, el Corol·lari 5.3 implica que F pot ser aproximada amb un error $\epsilon > 0$ fix per una xarxa neuronal de profunditat 3 de funció d'activació ReLU, amplada $\frac{16\pi d^5}{\epsilon}$ i pesos limitats per $2\pi d^3$.

Aquesta comparació il·lustra clarament com una xarxa de profunditat 3 pot proporcionar una aproximació d'error $\epsilon > 0$ amb un nombre de neurones d'ordre

polinòmic en funció de d, comparat amb una xarxa de profunditat 2 que mai arriba a ϵ-aproximar ni tan sols amb un ordre exponencial de neurones, demostrant així la superioritat de les xarxes més profundes en termes d'aproximació.

Referències

[1] Kendall Atkinson i Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*. Vol. 2044. Springer Science & Business Media, 2012.

[2] J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 2019. URL: https://books.google.es/books?id=wWS-DwAAQBAJ.

[3] George V. Cybenko. "Approximation by superpositions of a sigmoidal function". A: *Mathematics of Control, Signals and Systems* 2 (1989), pàg. 303 - 314. URL: https://api.semanticscholar.org/CorpusID:3958369.

[4] Amit Daniely. "Depth separation for neural networks". A: *Conference on Learning Theory*. PMLR. 2017, pàg. 690 - 696.

[5] Eleonora Di Nezza, Giampiero Palatucci i Enrico Valdinoci. "Hitchhiker's guide to the fractional Sobolev spaces". A: *Bulletin des Sciences Mathématiques* 136.5 (2012), pàg. 521 - 573. URL: https://www.sciencedirect.com/science/article/pii/S0007449711001254.

[6] Ingo Gühring, Mones Raslan i Gitta Kutyniok. "Expressivity of deep neural networks". A: *arXiv preprint arXiv:2007.04759* 34 (2020).

[7] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". A: *Neural Networks* 4.2 (1991), pàg. 251 - 257. URL: https://www.sciencedirect.com/science/article/pii/089360809190009T.

[8] Yann LeCun et al. "Backpropagation applied to handwritten zip code recognition". A: *Neural computation* 1.4 (1989), pàg. 541 - 551.

[9] Minsky Marvin i A Papert Seymour. "Perceptrons". A: *Cambridge, MA: MIT Press* 6 (1969), pàg. 318 - 362.

[10] Warren S McCulloch i Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". A: *The bulletin of mathematical biophysics* 5 (1943), pàg. 115 - 133.

[11] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." A: *Psychological review* 65.6 (1958), pàg. 386.

[12] David E Rumelhart, Geoffrey E Hinton i Ronald J Williams. "Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986". A: *Biometrika* 71 (1986), pàg. 599 - 607.

[13] Dmitry Yarotsky. "Error bounds for approximations with deep ReLU networks". A: *Neural Networks* 94 (2017), pàg. 103 - 114. URL: `https://www.sciencedirect.com/science/article/pii/S0893608017301545`.