

Cours: Topics in Machine Learning

- Plan
- ① Éléments sur les réseaux de neurones (NN) E.R.
 - ② Théorie de l'approximation pour les NN E.A.
 - ③ Complexité des NN E.R.
 - ④ Régression non paramétrique avec les NN IC
 - ⑤ Generative Adversarial Network E.A.
 - ⑥ Interpolation vs Surapprentissage C.B.
 - ⑦ Apprentissage et confidentialité

Polyycopié disponible (bientôt!) Evaluation sur projets

Orientation sur les aspects plutôt théoriques

Chapitre 1 : Élément sur les réseaux de neurones

Plan

- ① Définitions NN, fonction activation, profondeur, réalisation
- ② Approximations non constructives
Théorème approx universel, approximation rapide
- ③ Opérations de base sur les NN
Concaténation, parallelisation
- ④ Approximation constructive et rôle de la profondeur
Fonction en dents de scie, approx fonction carré
et fonctions C^3 non affines

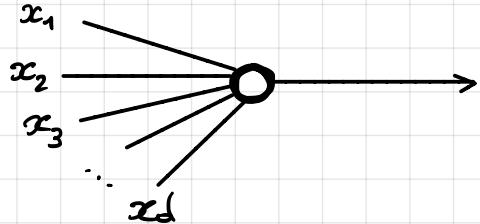
Préférences

Historiques: McCulloch and Pitts (1943)
Rosenblatt (1958)

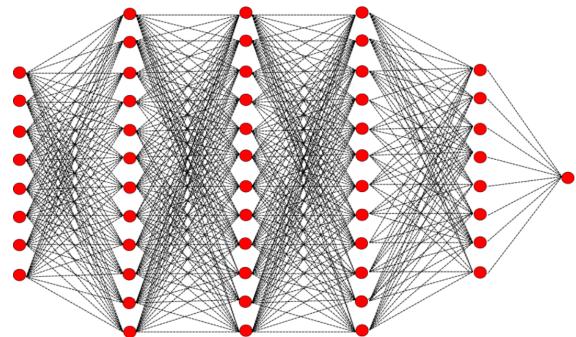
modélisation du fonctionnement d'un neurone du cerveau

Perceptron simple

$$x \mapsto 1 \} \sum_{i=1}^d w_i x_i - \theta \geq 0 \}$$



Multi-couches



Pour ce chapitre :

Petersen (2022) Neural Network Theory

Autres NN
≈ 3 millions sur google scholar

Gribonval et al (2019) Approximation spaces of deep neural networks

① Définitions

Définition: Soit $d, R, L \geq 1$

Un réseau de neurone (NN) de deux entrée d , deux sortie R
de profondeur L

est un couple $\Phi = (W, p)$ avec

- $p : \mathbb{R} \rightarrow \mathbb{R}$ une fonction (fonction d'activation)
 - $W = ((A_1, b_1), \dots, (A_L, b_L))$ suite de couples
où b_ℓ est $N_\ell \times 1$, A_ℓ est $N_\ell \times N_{\ell-1}$ $1 \leq \ell \leq L$
- ($N_0 = d, N_L = R$)
Les éléments des b_ℓ / A_ℓ sont réels (vois des NN)

Topologie / Architecture d'un NN Φ

- $L(\Phi) = L$ profondeur
- $N_p(\Phi) = N_p$ nombre de neurones de la couche p
largeur

$$N_{\max}(\Phi) = \max_{1 \leq p \leq L-1} N_p \quad \text{largeur maximum}$$

$$N(\Phi) = \sum_{0 \leq p \leq L} N_p \quad \text{nombre total de neurones ou largeur totale}$$

$$\bullet \text{sparsité : } \|\Phi\|_0 = \sum_{p=1}^L \|A_p\|_0 + \|b_p\|_0 \quad \text{nb de poids non nuls}$$

nb coeff ≠ 0 dans A_p

support : de Φ donné par la suite de supports

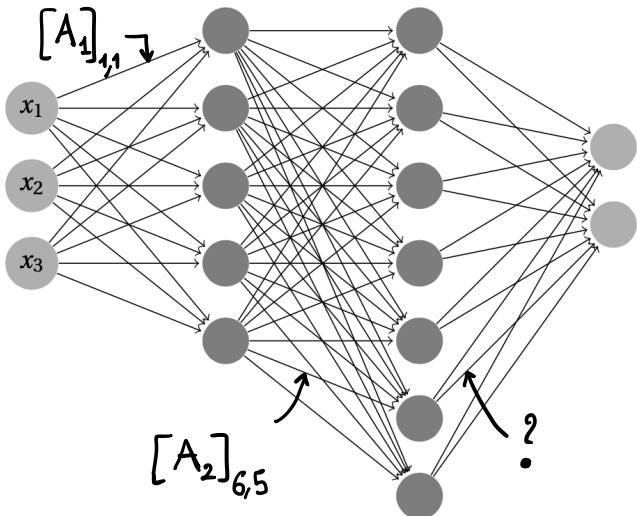
$$\{ (i,j) \in \{1, \dots, N_{p-1}\} \times \{1, \dots, N_p\} : [A_p]_{ij} \neq 0 \} , \quad 1 \leq p \leq L$$

Dans le cas $L=2$, le NN est dit peu profond (shallow network)

Représentation d'un NN

(en fait de la suite des A_i surtout)

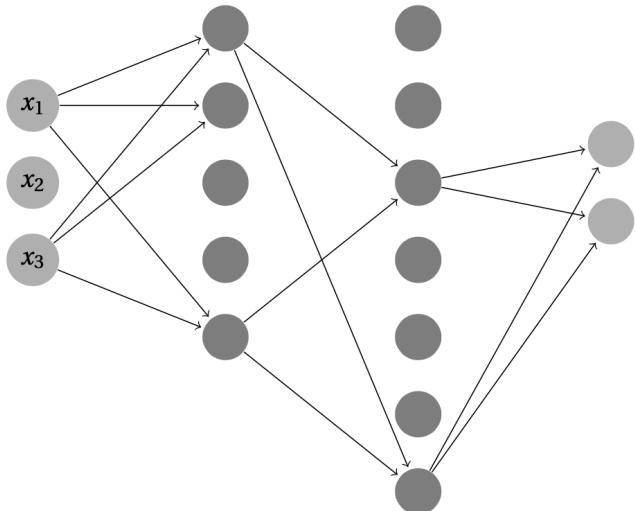
Exemple 1



$$\begin{cases} L = 3 \\ \delta = 3, k = 2, N_1 = 5, N_2 = 7, N_{\max} = 7 \end{cases}$$

entièrement connecté ($[A^l]_{ij} \neq 0$)
 $\forall e, i, j$

Exemple 2



idem mais sparse

$$\text{sparse } \|\Phi\|_0 = 14$$

(si $b^e = 0 \quad 1 \leq e \leq L$)

Réalisation d'un NN

Définition : pour un NN $\Phi = (\mathbf{W}, \rho)$ avec $\mathbf{W} = ((A_1, b_1), \dots, (A_L, b_L))$
 la réalisation de Φ est donnée par la fonction

$$R(\Phi) : x \in \mathbb{R}^d \mapsto R(\Phi)(x) = x^{(L)} \in \mathbb{R}^R$$

où $x^{(0)} = x$

$$x^{(l)} = \rho(A_l x^{(l-1)} + b_l)$$

$$x^{(L)} = A_L x^{(L-1)} + b_{L-1}$$

$$1 \leq l \leq L-1$$

(la fonction $\rho : \mathbb{R} \rightarrow \mathbb{R}$ est composante par composante)

Autrement dit : $R(\Phi) = T_L \circ \rho \circ T_{L-1} \circ \dots \circ T_2 \circ \rho \circ T_1$

où $T_l : x \in \mathbb{R}^{N_{l-1}} \mapsto A_l x + b_l \in \mathbb{R}^{N_l}$

(compositions de fonctions affines et fonction activation)

⚠ Deux NN \neq peuvent avoir la même réalisation ⚡

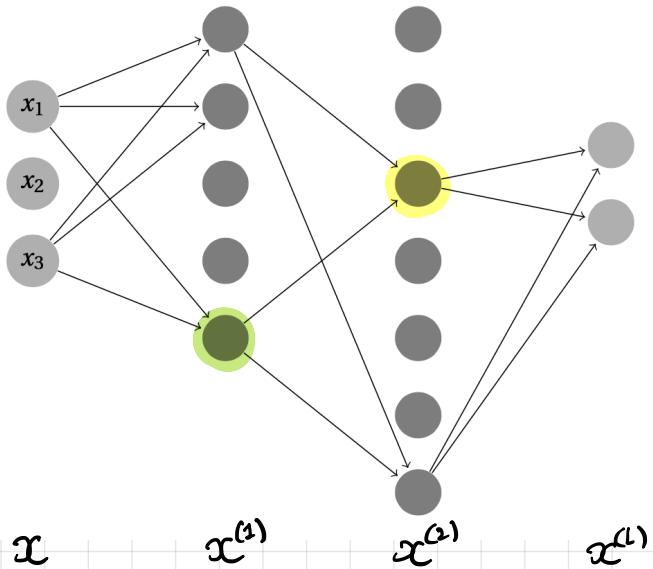
Exemple: réalisation d'un NN peu profond ($L=2$) avec $\begin{cases} d=1 \\ k=1 \end{cases}$

$$x \in \mathbb{R} \mapsto A_2 \rho(A_1 x + b_2) + b_2$$

$$= \sum_{i=1}^{N_1} [A_2]_{1,i} \rho([A_1 x + b_1]_i) + b_2$$

donc de la forme $x \in \mathbb{R} \mapsto b' + \sum_{i=1}^N a'_i \rho(a_i x + b_i)$

Autre exemple: $L=3$ + sparse pour des réels a_i, b_i, a'_i, b'



$$x^{(0)} = x$$

$$x^{(1)} = \begin{pmatrix} \rho([A_1]_{1,1} x_1 + [A_1]_{1,3} x_3) \\ \rho([A_1]_{2,1} x_1 + [A_1]_{2,3} x_3) \\ 0 \\ 0 \\ \vdots \\ \rho([A_1]_{5,1} x_1 + [A_1]_{5,3} x_3) \end{pmatrix}$$

$$\boxed{x^{(2)}_3 = \rho([A_2]_{3,1} x^{(1)}_1 + [A_2]_{3,5} x^{(1)}_5)}$$

Expressivité d'un réseau de neurone

Expressivité = capacité d'approximation de $\mathbb{R}(\Phi)$

Liée à la "taille" de

$$\text{NN}^{k,d}(L, \rho) = \{ \mathbb{R}(\Phi) \text{ , pour } \Phi \text{ NN}$$

dim entrée d
dim sortie k
profondeur L
fonction activation p

Exemples

* si $p(x) = mx + v$ alors $\text{NN}^{k,d}(L, \rho) = 2$ fonctions affines $\mathbb{R}^d \rightarrow \mathbb{R}^k \}$

* si $p(x) = x_+ = \max(0, x)$ ReLU (hyper classique !)

alors $\text{NN}^{k,d}(L, \rho) \subset 2$ fonctions continues $\mathbb{R}^d \rightarrow \mathbb{R}^k$
affines par morceaux }

* si p est polynomiale de degré $r \geq 1$

$\text{NN}^{k,d}(L, \rho) \subset 2$ fonctions polynomiales de degré $\leq r(L-1)$

Réalisation de l'identité avec un NN ReLU

* $R(\Phi)(x) = x$ pour $x \in \mathbb{R}$ pour Φ ReLU ?

En effet, $x = x_+ - (-x)_+ = [1 \ -1] \begin{pmatrix} p(1 \cdot x + 0) \\ p(-1 \cdot x + 0) \end{pmatrix} + 0$

donc ok avec

$$A_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad A_2 = [1 \ -1], \quad b_2 = [0 \ 0]$$

* $R(\Phi)$: $x \in \mathbb{R}^d \mapsto x$ avec Φ RéL profondeur $L \geq 2$?

$$\Phi = \underbrace{\left(\left(\begin{bmatrix} I_d \\ -I_d \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \right), \left(\begin{bmatrix} I_{2d} \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \right), \dots, \left(\begin{bmatrix} I_{2d} \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \right), \left(\begin{bmatrix} I_d & -I_d \end{bmatrix}, \begin{bmatrix} 0 \dots 0 \end{bmatrix} \right) \right)}_{L-2 \text{ fois}}$$

Note: $R(\Phi)$ peut être plus régulière que p

pour p différentiable on peut approcher l'identité à ϵ près
 $(\epsilon \neq 0)$ (cf poly)

2) Approximations non constructives

Théorème (approximation universelle des NN, [Cybernetics (1989)])

Soit p une fonction d'activation continue, $\lim_{-\infty} p = 0$, $\lim_{+\infty} p = 1$
K compact $\subset \mathbb{R}^d$
(sigmoidale)

Alors $NN^{d,1}(2, p)$ est dense dans $C(K, \mathbb{R})$ pour $\|\cdot\|_{\infty, K}$

↓
NN peu profond

↓
fonction continue sur tous les intervalles réels

i.e. $\forall f \in C(K, \mathbb{R})$, $\forall \varepsilon > 0$, $\exists \underline{\Phi} = \underline{\Phi}^{f, \varepsilon, d}$ avec $L(\underline{\Phi}) = 2$ et $\|f - \underline{\Phi}\|_{\infty, K} \leq \varepsilon$

Preuve : théorème Hahn-Banach non constructive !

cf [Petersen]

pas de contrôle sur $N(\underline{\Phi})$!
 $\|\underline{\Phi}\|_0$!

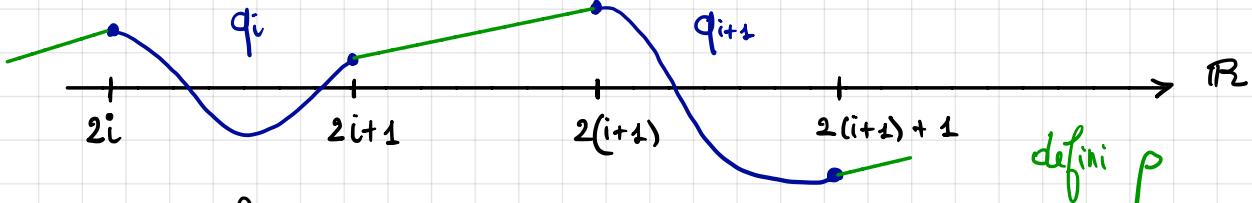
Théorème (approximation rapide, Proposition 2.21 Petersen)

\exists une fonction d'activation p | polynomiale par morceaux telle que continue

$$\forall f \in C([0,1], \mathbb{R}), \forall \varepsilon > 0, \exists \Phi = \Phi^{f,\varepsilon} \text{ NN avec } \begin{cases} L(\Phi) = 2 \\ N(\Phi) \leq 3, \|\Phi\|_6 \leq 3 \\ \text{fonction d'activation } p \end{cases}$$

tel que $\| \Phi(\Phi) - f \|_{\infty, [0,1]} \leq \varepsilon$

Preuve: polynôme avec coeff \mathbb{Q} $\{q_i, i \in \mathbb{Z}\}$ dense dans $C([0,1], \mathbb{R})$



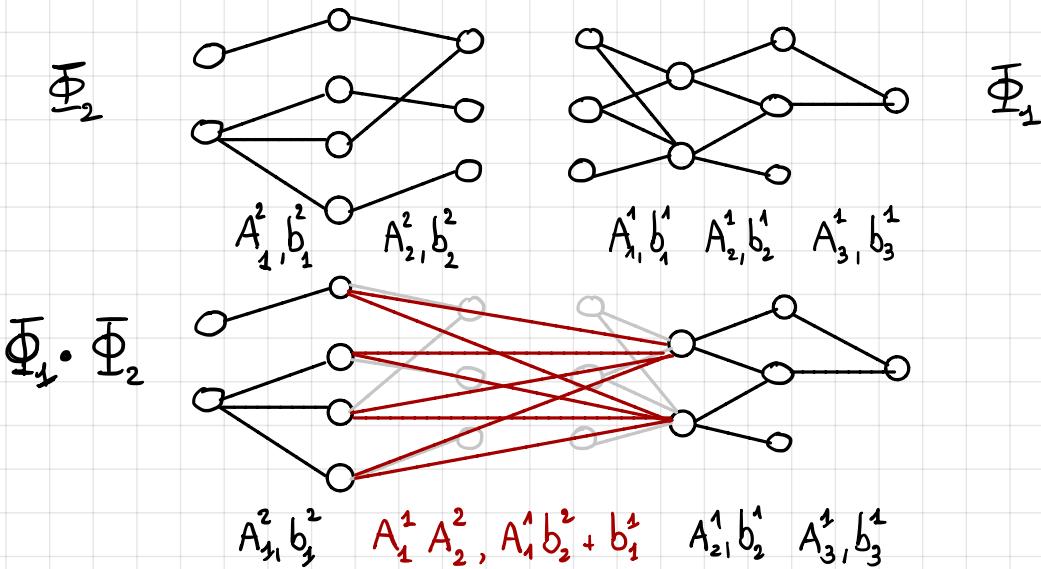
on a bien f arbitrairement proche d'un $q_i(x) = p(x+2i)$

⚠ Fonction p inutile en pratique

$$A_1 = 1, b_1 = 2i, A_2 = 1, b_2 = 0 \quad L=2$$

③ Opérations de base sur les NN

Concaténation Pour deux NN Φ_1 et Φ_2 avec $d_1 = k_2$



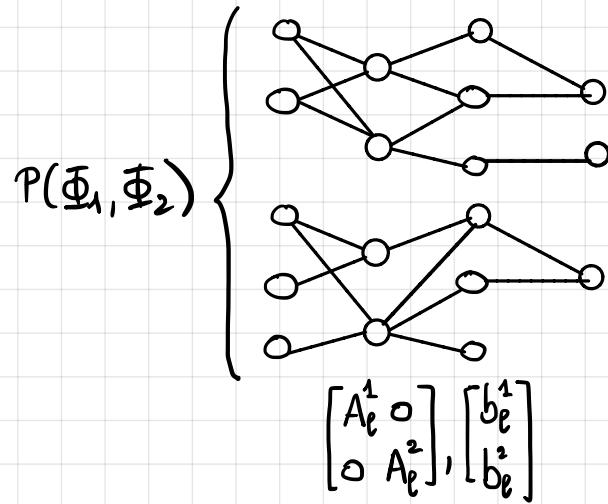
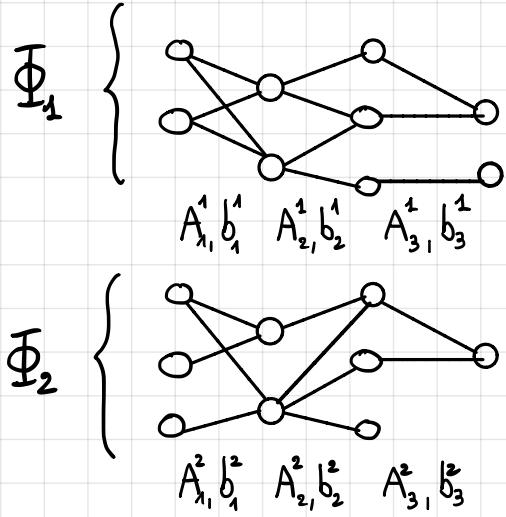
$$\text{Alors } R(\Phi_1) \circ R(\Phi_2) = R(\Phi_1 \cdot \Phi_2)$$

$$L(\Phi_1 \cdot \Phi_2) = L(\Phi_1) + L(\Phi_2) - 1, \quad N(\Phi_1 \cdot \Phi_2) = N(\Phi_1) + N(\Phi_2) - 2k_2$$

$$N_{\max}(\Phi_1 \cdot \Phi_2) = \max(N_{\max}(\Phi_1), N_{\max}(\Phi_2)), \quad \|\Phi_1 \cdot \Phi_2\|_0 \leq \begin{cases} \|\Phi_1\|_0 + \|\Phi_2\|_0 + (N_{k_2-1}^2 + 1)N_2^1 \\ - \|A_{N_{k_2}}^2\|_0 - \|b_{k_2}^2\|_0 - \|A_1^1\|_0 - \|b_1^1\|_0 \end{cases}$$

Mise en parallèle

Pour deux NN Φ_1 et Φ_2 de même profondeur $L_1 = L_2$



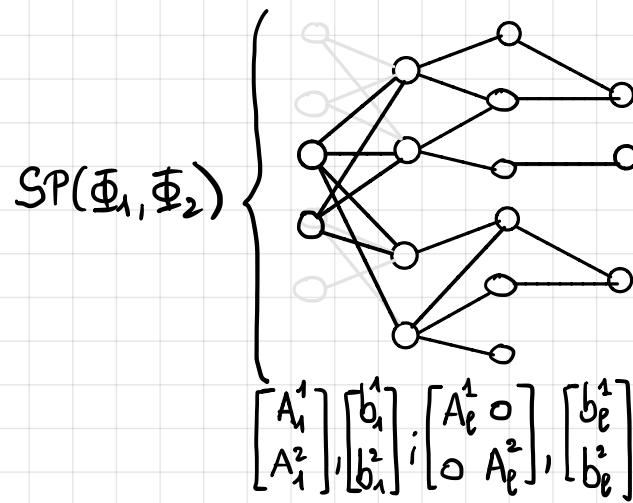
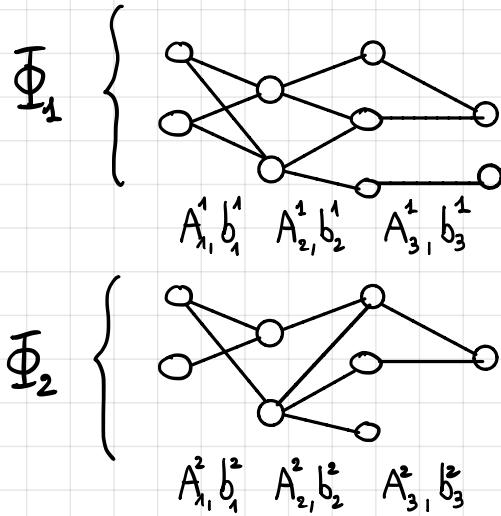
Alors $\forall x = (x_1, x_2) \in \mathbb{R}^{d_1+d_2}$

$$R(P(\Phi_1, \Phi_2))(x) = (R(\Phi_1)(x_1), R(\Phi_2)(x_2)) \in \mathbb{R}^{k_1+k_2}$$

Utile pour augmenter la dimension ! Préserve la complexité .

Mise en parallèle avec entrée partagée

Pour deux NN Φ_1 et Φ_2 avec $L_1 = L_2$ et $d_1 = d_2$



Alors $\forall x \in \mathbb{R}^{d_1} = \mathbb{R}^{d_2}$

$$R(SP(\Phi_1, \Phi_2))(x) = (R(\Phi_1)(x), R(\Phi_2)(x)) \in \mathbb{R}^{k_1+k_2}$$

Présure la complexité

Utile pour augmenter la dimension d'arrivée (puis faire la scission par ex)

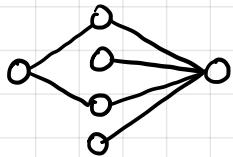
4) Approximations constructives et rôle de la profondeur

- But : • Construire des bonnes approximations avec ReLU NN
• Mettre en évidence le rôle de la profondeur L

Rappel : pour ReLU NN, $R(\Phi)$ continue et affine par morceaux
bonne approximation \rightarrow beaucoup de morceaux possibles

NN ReLU peu profonds : réalisation de la forme

$$\Phi : x \in \mathbb{R} \mapsto b' + \sum_{i=1}^N a_i' \rho(a_i x + b_i), \quad a_i, b_i, a_i', b'$$



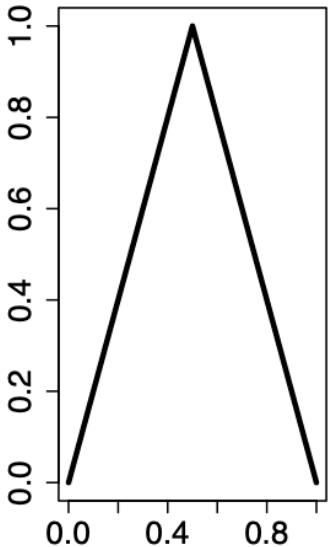
$$\begin{aligned} \text{nb de morceaux} &\leq \#\{i : a_i \neq 0\} + 1 \\ &\leq \|\Phi\|_0 =: M \end{aligned}$$

Question peut-on mieux "dépenser" ces M morceaux ?
en ajoutant de la profondeur ?

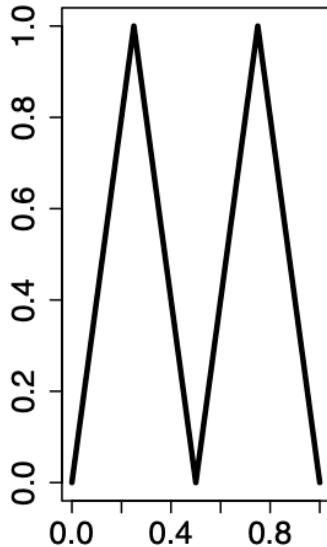
La fonction en "dent de scie"

une fonction avec beaucoup de morceaux !

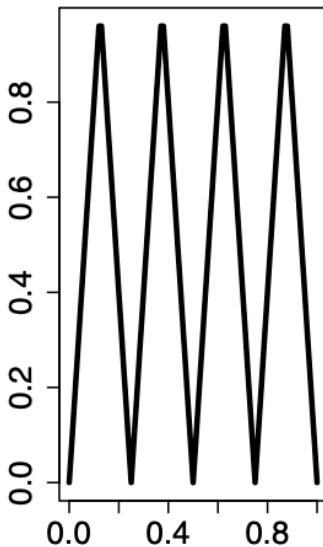
$$T_1(x) = \begin{cases} 2x & \text{si } x \in [0, \frac{1}{2}] \\ 2-2x & \text{sinon} \end{cases}$$



$$T_2 = T_1 \circ T_1$$



$$T_3 = T_1 \circ T_1 \circ T_1$$



$$T_n = \underbrace{T_1 \circ \dots \circ T_1}_{n \text{ fois}}$$

pour $n \geq 1$ qui a 2^{n-1} chapeaux
donc 2^n morceaux sur $[0, 1]$

Proposition (réalisation de T_n avec un NN ReLU)

Pour tout $n \geq 1$, $T_n = R(\Phi)$ pour Φ NN ReLU avec

$$L(\Phi) = n+1, N_{\max}(\Phi) = 3, \|\Phi\|_0 \leq 12n-2$$

Preuve: par récurrence

$$\text{si } n=1 \quad T_1(x) = (2x)_+ - 2(2x-1)_+ + (2x-2)_+$$

$$\text{on a } L(\Phi^{(1)}) = 2, N_{\max}(\Phi^{(1)}) = 3, \|\Phi^{(1)}\|_0 = 8$$

si $\Phi^{(n-1)}$ construit avec $T_{n-1} = R(\Phi^{(n-1)})$

et $L(\Phi^{(n-1)}) = n, N_{\max}(\Phi^{(n-1)}) = 3, \|\Phi^{(n-1)}\|_0 \leq 12(n-1)-2, \Phi^{(n-1)}$ finit avec $[1 \cdot 2 \cdot 1], 0$

On concatène $T_n = T_1 \circ T_{n-1} = R(\Phi^{(1)}) \circ R(\Phi^{(n-1)}) = R(\underbrace{\Phi^{(1)} \cdot \Phi^{(n-1)}}_{:= \Phi^{(n)}})$

et $L(\Phi^{(n)}) = 2+n-1 = n+1, N_{\max}(\Phi^{(n)}) = 3$

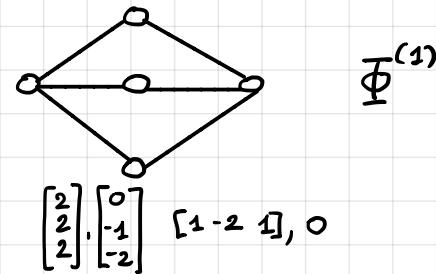
$$\|\Phi^{(n)}\|_0 \leq \|\Phi^{(1)}\|_0 + \|\Phi^{(n-1)}\|_0 + (3+1) \times 3 - (3+2) - 3 = 12(n-1) - 2 + 12 \quad \square$$

nb neurones de mièvre couche
couche cachée de $\Phi^{(n-1)}$

nb neur./ premiere
couche cachée de $\Phi^{(1)}$

sparsité première
couche cachée de $\Phi^{(n-1)}$

sparsité dernière couche
couche cachée de $\Phi^{(1)}$



Première mise en évidence du rôle de la profondeur :

Combien de morceaux peut-on générer avec un NN ReLU $\underline{\Phi}$ sous la contrainte $\|\underline{\Phi}\|_0 \leq M$?

On a

- pour $\underline{\Phi}$ peu profond ($L(\underline{\Phi}) = 2$) , nb morceaux $\leq M$
- pour $\underline{\Phi} = \underline{\Phi}^{(n)}$ avec $n = \lceil (M+2)/12 \rceil$

on a bien $\|\underline{\Phi}\|_0 \leq 12n - 2 \leq M$ (proposition)

et $L(\underline{\Phi}) = n + 1$

et $R(\underline{\Phi}) = T_n$ qui a nb morceaux $= 2^n = 2^{\lceil (M+2)/12 \rceil}$

$\gg M$

Prendre un réseau plus profond permet de réaliser des fonctions plus complexes au prix de la même complexité !

Quasi-optimalité de la fonction en dents de scie

Proposition Soit $L \geq 2$

Tout Φ de L avec $d = 1, R = 1$, $N_{\max}(\Phi) \leq N$
est tel que $R(\Phi)$ a au plus $(2N)^{L-1}$ morceaux

Preuve: par récurrence sur L cf [Petersen]

Du coup:

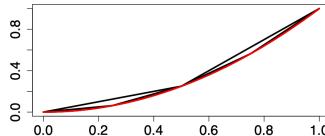
- $\forall \Phi$ de L avec $L(\Phi) = n+1, N_{\max}(\Phi) = 3$
a au plus 6^n morceaux
- $\Phi^{(n)}$ réalise 2^n morceaux

Qualitativement, on peut pas trop faire mieux que $\Phi^{(n)}$

Aproximation de fonctions régulières

explicite et avec quantification de la complexité

1^{er} résultat : borne supérieure



On considère le cas de la fonction $x \in \mathbb{R} \mapsto x^2$

fonction de scie

Lemme: $\forall N \geq 0, \sup_{x \in [0, 1]} \left| x^2 - x + \sum_{n=1}^N 2^{-2n} T_n(x) \right| \leq 2^{-2N-2}$

Proposition: pour tout $\varepsilon \in (0, \frac{1}{2})$, il existe un RLLW NN Φ^ε tel que

$$\sup_{x \in [0, 1]} |R(\Phi^\varepsilon)(x) - x^2| \leq \varepsilon \text{ et}$$

$$\begin{cases} L(\Phi^\varepsilon) \leq 3 \log_2(1/\varepsilon) \\ N_{\max}(\Phi^\varepsilon) \leq 6 \log_2(1/\varepsilon) \\ \|\Phi^\varepsilon\|_\infty \leq 26 \log_2(1/\varepsilon) \end{cases}$$

Preuve

On peut approcher une fonction régulière à l'aide d'un NN de profondeur pas trop grande

2^e résultat : borne inférieure

Proposition: pour toute fonction $f \in C^3([0,1], \mathbb{R})$ non affine

il existe $c = c(f)$ tel que :

pour tout ReLU NN Φ avec profondeur L et $N_{\max}(\Phi) \leq N_{\max}$

on a

$$\sup_{x \in [0,1]} |f(x) - \Phi(\Phi)(x)| \geq c(2N_{\max})^{-2(L-1)}$$

Preuve

Consequence:

Pour approcher une fonction régulière (non-affine) avec précision ε par un ReLU NN
on doit prendre une profondeur L : $c(2N_{\max})^{-2(L-1)} \leq \varepsilon$ ie

$$L \geq 1 + 0.5 \log_2(c/\varepsilon) / \log_2(2N_{\max})$$

profondeur nécessaire !

Pour le corré cette borne inf est de l'ordre de $\log(\frac{1}{\varepsilon}) / \log \log_2(\frac{1}{\varepsilon})$
donc presque la borne $\sup \log_2(\frac{1}{\varepsilon})$ trouvée au dessus

Conclusion

pour approcher des fonctions régulières par des NN ReLU

- Sans contrôle de complexité $L=2$ suffit
- Avec précision ϵ on peut avoir des résultats où $L(\Phi), \|\Phi\|_0, N(\Phi)$ sont contrôlés en fonction de ϵ
plein de résultats à venir dans cette veine !
- Si la fonction est non affine
on doit avoir $L(\Phi)$ assez grand

Question: pourquoi approcher des fonctions régulières par des NN?

↳ intéressant notamment pour réduire la dimension
(cf chapitre 5 de [Petersen])