

# Fondements théoriques du deep learning

## Expressivité des réseaux de neurones

**Sébastien Gerchinovitz**<sup>1,2</sup>, François Malgouyres<sup>2</sup>, Edouard Pauwels<sup>3</sup> et Nicolas Thome<sup>4</sup>

<sup>1</sup>IRT Saint Exupéry (projet DEEL), Toulouse

<sup>2</sup>Institut de Mathématiques de Toulouse, Université Paul Sabatier

<sup>3</sup>Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier

<sup>4</sup>Centre D'Étude et de Recherche en Informatique et Communications, Conservatoire National des Arts et Métiers

# Plan

- 1 Introduction
  - Compromis approximation-estimation-optimisation
  - Expressivité des réseaux de neurones ?
  - Autres exemples d'approximation
- 2 Universalité des réseaux feedforward à 1 couche cachée
  - Le théorème d'approximation de Cybenko (1989)
  - Extensions et raffinements
- 3 L'effet de la profondeur sur l'expressivité des réseaux
  - Transition de 1 à 2 couches cachées
  - Vitesses d'approximation en fonction de  $W$  et  $L$

# Plan

- 1 Introduction
  - Compromis approximation-estimation-optimisation
  - Expressivité des réseaux de neurones ?
  - Autres exemples d'approximation
- 2 Universalité des réseaux feedforward à 1 couche cachée
  - Le théorème d'approximation de Cybenko (1989)
  - Extensions et raffinements
- 3 L'effet de la profondeur sur l'expressivité des réseaux
  - Transition de 1 à 2 couches cachées
  - Vitesses d'approximation en fonction de  $W$  et  $L$

# Compromis approximation-estimation-optimisation

- Pour un échantillon  $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{i.i.d.}{\sim} P_{X,Y}$  et une fonction  $g$  mesurable, le **risque** et le **risque empirique** de  $g$  relativement à une perte positive  $L$  sont :

$$R(g) = \mathbb{E}(L(g(X), Y)) \quad \text{et} \quad \widehat{R}(g) = \frac{1}{n} \sum_{i=1}^n L(g(X_i), Y_i)$$

- $R^* = \inf_g R(g)$  le risque de Bayes sur l'ensemble des fonctions mesurables
- Paramètres quasi-optimaux **du réseau** :  $\mathbf{w}^*$  pour le risque  $R$ ,  $\widehat{\mathbf{w}}$  pour le risque empirique  $\widehat{R}$  :

$$R(\mathbf{f}_{\mathbf{w}^*}) \leq \inf_{\mathbf{w}} R(\mathbf{f}_{\mathbf{w}}) + \varepsilon \quad \text{et} \quad \widehat{R}(\mathbf{f}_{\widehat{\mathbf{w}}}) \leq \inf_{\mathbf{w}} \widehat{R}(\mathbf{f}_{\mathbf{w}}) + \varepsilon$$

Si notre algorithme d'optimisation retourne  $\mathbf{w}$ , l'excès de risque se décompose en (ce n'est pas la seule approche pour étudier le risque) :

$$\begin{aligned} R(\mathbf{f}_{\mathbf{w}}) - R^* &= R(\mathbf{f}_{\mathbf{w}^*}) - R^* && \text{(erreur d'approximation)} \\ &+ \widehat{R}(\mathbf{f}_{\mathbf{w}^*}) - R(\mathbf{f}_{\mathbf{w}^*}) && \text{(erreur de généralisation 1)} \\ &+ \widehat{R}(\mathbf{f}_{\widehat{\mathbf{w}}}) - \widehat{R}(\mathbf{f}_{\mathbf{w}^*}) && \leq \varepsilon \\ &+ \widehat{R}(\mathbf{f}_{\mathbf{w}}) - \widehat{R}(\mathbf{f}_{\widehat{\mathbf{w}}}) && \text{(erreur d'optimisation)} \\ &+ R(\mathbf{f}_{\mathbf{w}}) - \widehat{R}(\mathbf{f}_{\mathbf{w}}) && \text{(erreur de généralisation 2)} \end{aligned}$$

# Focus sur l'erreur d'approximation

$$\begin{aligned} R(\mathbf{f}_{\mathbf{w}}) - R^* &= R(\mathbf{f}_{\mathbf{w}^*}) - R^* && \text{(erreur d'approximation)} \\ &+ \widehat{R}(\mathbf{f}_{\mathbf{w}^*}) - R(\mathbf{f}_{\mathbf{w}^*}) && \text{(erreur de généralisation 1)} \\ &+ \widehat{R}(\mathbf{f}_{\widehat{\mathbf{w}}}) - \widehat{R}(\mathbf{f}_{\mathbf{w}^*}) && \leq \varepsilon \\ &+ \widehat{R}(\mathbf{f}_{\mathbf{w}}) - \widehat{R}(\mathbf{f}_{\widehat{\mathbf{w}}}) && \text{(erreur d'optimisation)} \\ &+ R(\mathbf{f}_{\mathbf{w}}) - \widehat{R}(\mathbf{f}_{\mathbf{w}}) && \text{(erreur de généralisation 2)} \end{aligned}$$

- **Erreur d'approximation** : Quelles fonctions peut-on approximer avec quels réseaux ? ("expressive power", "expressivité", etc)
- **Erreur de généralisation/estimation** : Quelles conditions sur le réseau pour que  $\widehat{R}(\mathbf{f}_{\mathbf{u}}) \simeq R(\mathbf{f}_{\mathbf{u}})$  pour  $\mathbf{u}$  (ou, du moins, pour tout  $\mathbf{u} \in \{\mathbf{w}, \mathbf{w}^*\}$ ) ? Comment contrôler les fluctuations ? (cf. dimension de Vapnik-Chervonenkis, complexité de Rademacher, etc)
- **Erreur d'optimisation** : Quand est-ce que l'optimisation fonctionne ? (optimisation non-convexe, paysage de la fonction objectif, etc)

# Expressivité des réseaux de neurones ?

Nous allons étudier les capacités d'approximation des réseaux de neurones feedforward.

Questions naturelles :

- Quelles fonctions peut-on approcher avec un réseau à  $k$  couches cachées ?
- Une seule couche cachée est-elle suffisante ?
- L'expressivité augmente-t-elle avec la profondeur ?
- Pour un nombre de neurones donné, vaut-il mieux un réseau peu profond et large ou un réseau profond et étroit ?

Nous allons répondre partiellement à ces questions.

## Warm-up : approximation par des fonctions en escalier

Soit  $f : [a, b]^d \rightarrow \mathbb{R}$  continue et  $\varepsilon > 0$ . Par uniforme continuité de  $f$ , il existe  $\delta > 0$  tel que  $\|x - y\|_\infty \leq \delta \Rightarrow |f(x) - f(y)| \leq \varepsilon$ .

On découpe  $[a, b]^d$  en  $N^d$  cubes  $A_i$  de largeur  $(b - a)/N \approx 2\delta$  avec  $N = \lceil (b - a)/(2\delta) \rceil$ . Sur chaque cube  $A_i$ , on approche  $f$  par la valeur  $f(c_i)$  en son centre  $c_i$ . On a alors :

$$\sup_{x \in [a, b]^d} \left| f(x) - \sum_{i=1}^{N^d} f(c_i) \mathbb{1}_{A_i}(x) \right| \leq \varepsilon.$$

### Théorème

*Toute fonction réelle continue sur  $[a, b]^d$  peut-être arbitrairement bien approchée (au sens de la norme infinie) par une fonction constante par morceaux.*

# Décomposition en série de Fourier

Soit  $f : [0, 2\pi] \rightarrow \mathbb{R}$  de carré intégrable. Considérons ses coefficients de Fourier

$$\widehat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx \quad (n \in \mathbb{Z})$$

et les sommes partielles  $S_N : [0, 2\pi] \rightarrow \mathbb{R}$  de sa série de Fourier

$$S_N(x) = \sum_{n=-N}^N \widehat{f}(n) e^{inx}$$

D'après le théorème de Riesz-Fischer,  $S_N \rightarrow f$  dans  $\mathbb{L}^2([0, 2\pi], \mathbb{R})$  quand  $N \rightarrow +\infty$ .

## Théorème

*L'ensemble des fonctions de la forme*

$$x \mapsto \sum_{n=-N}^N c_n e^{inx} \quad (\text{avec } c_n \in \mathbb{C} \text{ et } c_{-n} = c_n^* \text{ pour tout } 0 \leq n \leq N)$$

*est dense dans  $\mathbb{L}^2([0, 2\pi], \mathbb{R})$ .*

# Théorème de représentation de Kolmogorov-Arnold

Voici un exemple de représentation **exacte** avec un nombre **fini** de termes.

## Théorème (Kolmogorov 1957; Arnold 1957)

Toute fonction continue  $f : [0, 1]^d \rightarrow \mathbb{R}$  de plusieurs variables peut être représentée comme une superposition finie de fonctions continues d'une variable  $\Phi_i, \psi_{i,j} : \mathbb{R} \rightarrow \mathbb{R}$  et de l'opération somme (les  $\psi_{i,j}$  sont indépendantes de  $f$ ) :

$$f(x_1, \dots, x_d) = \sum_{i=0}^{2d} \Phi_i \left( \sum_{j=1}^d \psi_{i,j}(x_j) \right)$$

- Ressemble un peu à un réseau feedforward à 2 couches cachées.
- Travaux de Lorentz (1962) et Sprecher (1996, 1997); Braun and Griebel (2009) : pas nécessaire de choisir des fonctions d'activation  $\Phi_i$  et  $\psi_{i,j}$  toutes différentes.
- Mais, pour les réseaux feedforward, on impose une **unique** fonction d'activation  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ . On se contentera d'une représentation **approximative**.

## Lien avec le 13-ème problème de Hilbert

Le 13-ème problème de Hilbert, formulé dans une liste de 23 problèmes en 1900 par David Hilbert, cherche à savoir si on peut exprimer une solution  $x(a, b, c)$  de l'équation

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

à l'aide d'une superposition finie de fonctions algébriques/continues de deux variables.

- Il était conjecturé que cela n'était pas possible.
- Kolmogorov et Arnold ont contredit cette conjecture en 1956-57 en montrant qu'en fait, toute fonction continue de plusieurs variables pouvait s'exprimer comme une superposition d'un nombre fini de fonctions continues de 2 variables.

Pour ceux qui souhaitent approfondir : cf. article de synthèse de Morris (2021).

# Plan

- 1 Introduction
  - Compromis approximation-estimation-optimisation
  - Expressivité des réseaux de neurones ?
  - Autres exemples d'approximation
- 2 Universalité des réseaux feedforward à 1 couche cachée
  - Le théorème d'approximation de Cybenko (1989)
  - Extensions et raffinements
- 3 L'effet de la profondeur sur l'expressivité des réseaux
  - Transition de 1 à 2 couches cachées
  - Vitesses d'approximation en fonction de  $W$  et  $L$

# Le théorème d'approximation de Cybenko (1989)

On s'intéresse aux réseaux de neurones feedforward à 1 couche cachée, i.e., aux fonctions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de la forme

$$f(x) = \sum_{i=1}^N v_i \sigma(\langle w_i, x \rangle + b_i) \quad (v_i, b_i \in \mathbb{R} \text{ et } w_i \in \mathbb{R}^d)$$

avec une fonction d'activation  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ .

## Théorème (Cybenko 1989)

*Soit  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  continue et sigmoïdale ( $\lim_{-\infty} \sigma = 0$  et  $\lim_{+\infty} \sigma = 1$ ).*

*Alors, l'ensemble  $\mathcal{N}_1$  des réseaux de neurones feedforward à 1 couche cachée est dense dans  $\mathcal{C}([0, 1]^d, \mathbb{R})$ .*

# Warm-up en dimension $d = 1$

Pour commencer : fonctions  $f$  d'une seule variable ( $d = 1$ ), et fonctions d'activation  $\sigma$  très spécifiques (qui ne vérifient d'ailleurs pas les hypothèses du théorème !).

Soit  $f : [0, 1] \rightarrow \mathbb{R}$  une fonction continue et  $\varepsilon > 0$ . Il est facile de construire à la main un réseau  $g : [0, 1] \rightarrow \mathbb{R}$  à 1 couche cachée tel que  $\|g - f\|_\infty \leq \varepsilon$ .

- Avec  $\sigma(x) = \mathbb{1}_{x \geq 0}$  (Heaviside) :  
Il suffit d'approcher  $f$  par une fonction  $g$  constante par morceaux, puis de remarquer que  $g$  est un réseau à 1 couche cachée avec la fonction de Heaviside.
- Avec  $\sigma(x) = \max\{x, 0\}$  (ReLU) :  
Il suffit d'approcher  $f$  par une fonction continue affine par morceaux (interpolation linéaire), puis de remarquer que  $g$  est un réseau ReLU à 1 couche cachée.

# Cas général : schéma de preuve

Il s'agit de montrer que  $\overline{\mathcal{N}_1} = \mathcal{C}([0, 1]^d, \mathbb{R})$  (au sens de la norme infinie).

Preuve non-constructive, qui procède **par contradiction** (méthode classique pour montrer un résultat de densité) :

- On suppose qu'il existe  $f_0 \in \mathcal{C}([0, 1]^d, \mathbb{R}) \setminus \overline{\mathcal{N}_1}$ .
- On invoque un **théorème de Hahn-Banach** qui fournit une forme linéaire continue  $L$  sur  $\mathcal{C}([0, 1]^d, \mathbb{R})$  telle que  $L(f_0) = 1$  mais  $L = 0$  sur  $\mathcal{N}_1$ .
- On utilise ensuite un **théorème de représentation de Riesz**, qui permet de représenter  $L$  à l'aide d'une mesure de Borel signée sur  $[0, 1]^d$  :

$$\forall f \in \mathcal{C}([0, 1]^d, \mathbb{R}), \quad L(f) = \int f d\mu$$

- En exploitant cette représentation intégrale pour  $f(\cdot) = \sigma(\langle w, \cdot \rangle + b) \in \mathcal{N}_1$ , on montre que la transformée de Fourier de  $\mu$  est nulle, et donc que  $\mu = 0$ .
- Dès lors,  $L(f_0) = \int f_0 d\mu = 0$ , ce qui contredit  $L(f_0) = 1$  et prouve le théorème.

# L'extension de Hornik (1991)

Hornik a prouvé en 1991 le résultat plus général suivant.

## Théorème (Hornik 1991)

Soit  $K \subset \mathbb{R}^d$  un **compact**. Supposons que  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  est continue, **bornée et non-constante**. Alors, l'ensemble des réseaux de neurones feedforward à 1 couche cachée est dense dans  $\mathcal{C}(K, \mathbb{R})$ .

Il a aussi prouvé un résultat analogue en dehors du cas  $K$  compact.

## Théorème (Hornik 1991)

Soit  $\mu$  une mesure de Borel positive sur  $\mathbb{R}^d$ , de **masse finie**. Supposons que  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  est **bornée et non-constante**. Alors, l'ensemble des réseaux de neurones feedforward à 1 couche cachée est dense dans  $L^p(\mathbb{R}^d, \mathbb{R}, \mu)$  pour tout  $1 \leq p < +\infty$ .

# Conséquence en régression

**Rappels.** Soit  $(X, Y)$  un couple aléatoire à valeurs  $\mathcal{X} \times \mathbb{R}$ , avec  $Y$  de carré intégrable. Le *risque*  $L^2$ , pour  $f : \mathcal{X} \rightarrow \mathbb{R}$  mesurable,

$$\mathbb{E}[(Y - f(X))^2]$$

est minimisé en la *fonction de régression*  $f^*(x) = \mathbb{E}[Y|X = x]$ , et l'*excès de risque* vaut

$$\mathbb{E}[(Y - f(X))^2] - \mathbb{E}[(Y - f^*(X))^2] = \mathbb{E}[(f^*(X) - f(X))^2] = \|f^* - f\|_{L^2(P_X)}^2.$$

**Conséquence du théorème d'universalité d'Hornik (version  $L^2$ ) :**

En régression, si  $\sigma$  est bornée non constante, *il existe* un réseau à 1 couche cachée avec des performances prédictives (quasi) optimales.

Par densité, la conclusion reste vraie si  $P_X$  est à support compact et si  $\sigma$  est continue et permet d'approcher toute fonction continue sur un compact (ex : ReLU).

Attention : cela ne résout pas le problème *d'apprendre* un bon réseau.

# Conséquence en classification

**Rappels.** Soit  $(X, Y)$  un couple aléatoire à valeurs  $\mathcal{X} \times \{0, 1\}$ . Le *risque de mauvaise classification*, pour  $f : \mathcal{X} \rightarrow \{0, 1\}$  mesurable,

$$\mathbb{P}(Y \neq f(X))$$

est minimisé en  $f^*(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}}$ , avec  $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ , et l'*excès de risque* vaut

$$\mathbb{P}(Y \neq f(X)) - \mathbb{P}(Y \neq f^*(X)) = 2\mathbb{E} \left[ \left| \eta(X) - \frac{1}{2} \right| \mathbb{1}_{f^*(X) \neq f(X)} \right].$$

Si on tente d'estimer  $\eta(x)$  par un réseau  $g(x)$ , et qu'on prédit  $f(x) = \mathbb{1}_{g(x) \geq \frac{1}{2}}$ ,

$$\mathbb{P}(Y \neq f(X)) - \mathbb{P}(Y \neq f^*(X)) \leq 2\mathbb{E} \left[ \left| \eta(X) - g(X) \right| \right] = 2\|\eta - g\|_{L^1(P_X)}.$$

## Conséquence du théorème d'universalité d'Hornik (version $L^1$ ) :

En classification, si  $\sigma$  bornée non constante (autres conditions possibles), *il existe* un réseau à 1 couche cachée associé à des performances prédictives (quasi) optimales.

Attention : cela ne résout pas le problème *d'apprendre* un bon réseau.

## Une borne quantitative de Barron (1993)

Barron a majoré le nombre de neurones  $N$  d'un réseau à 1 couche cachée pour approcher une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  admettant une représentation de Fourier de la forme

$$f(x) = \int_{\mathbb{R}^d} \widehat{f}(w) e^{i\langle w, x \rangle} dw$$

avec  $\widehat{f} : \mathbb{R}^d \rightarrow \mathbb{C}$  telle que  $C_f := \int_{\mathbb{R}^d} \|w\|_2 |\widehat{f}(w)| dw$  est finie.

### Théorème (Barron 1993)

Soit  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  une fonction sigmoïdale. Pour toute  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  comme ci-dessus, tout  $r > 0$  et toute mesure de probabilité  $\mu$  sur  $B_r = \{x : \|x\|_2 \leq r\}$ , il existe un réseau  $g_N$  à 1

couche cachée  $g_N(x) = \sum_{i=1}^N v_i \sigma(\langle w_i, x \rangle + b_i) + v_0$  tel que

$$\int_{B_r} (f - g_N)^2 d\mu \leq \frac{(2rC_f)^2}{N}.$$

Attention : pour certaines fonctions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , la constante  $C_f$  est exponentiellement grande en  $d$ , voire infinie (fonction triangle par ex).

# Autres résultats

Plusieurs résultats similaires ou extensions ont été prouvés depuis. Par exemple :

- Leshno et al. (1993) ont montré que sous des hypothèses faibles sur  $\sigma$ , l'ensemble  $\mathcal{N}_1$  permet d'approcher toute fonction continue sur tout compact (en norme  $L^\infty$ ) si et seulement si  $\sigma$  est non-polynomiale.
- Kidger and Lyons (2020) ont étudié le problème de l'approximation par des réseaux de profondeur arbitraire mais de largeur petite. Ils ont montré que si  $\sigma$  est continue, non-affine, et de classe  $C^1$  au voisinage d'un point  $t_0$  avec  $\sigma'(t_0) \neq 0$ , alors, pour tout compact  $K \subset \mathbb{R}^d$ , l'ensemble des réseaux feedforward de largeur  $d + 3$  et de profondeur arbitraire est dense dans  $(\mathcal{C}(K; \mathbb{R}), \|\cdot\|_\infty)$ .

# Plan

- 1 Introduction
  - Compromis approximation-estimation-optimisation
  - Expressivité des réseaux de neurones ?
  - Autres exemples d'approximation
- 2 Universalité des réseaux feedforward à 1 couche cachée
  - Le théorème d'approximation de Cybenko (1989)
  - Extensions et raffinements
- 3 L'effet de la profondeur sur l'expressivité des réseaux
  - Transition de 1 à 2 couches cachées
  - Vitesses d'approximation en fonction de  $W$  et  $L$

# Effet de la profondeur sur l'expressivité des réseaux

Informellement, l'**expressivité** d'une architecture de réseau feedforward désigne sa capacité à approcher correctement certains espaces de fonctions après calibration des paramètres du réseau.

Questions naturelles :

- Quelles fonctions peut-on approcher avec un réseau à  $k$  couches cachées ?
- Une seule couche cachée est-elle suffisante ?
- L'expressivité augmente-t-elle avec la profondeur ?
- Pour un nombre de neurones donné, vaut-il mieux un réseau peu profond et large ou un réseau profond et étroit ?

Plusieurs articles ont étudié l'effet de la profondeur sur l'expressivité des réseaux.

Nous allons voir deux exemples de résultats : le phénomène de **depth separation**, et la **super-approximation** par des réseaux profonds.

# Ex de depth separation: transition de 1 à 2 couches cachées

Travaux d'Eldan and Shamir (2016), cf. également Daniely (2017). Comparaison entre réseaux feedforward à 1 ou 2 couches cachées :

$$x \mapsto \sum_{i=1}^N v_i \sigma(\langle w_i, x \rangle + b_i) \quad x \mapsto \sum_{i=1}^{N_2} u_i \sigma \left( \sum_{j=1}^{N_1} v_{i,j} \sigma(\langle w_{i,j}, x \rangle + b_{i,j}) + c_i \right)$$

## Theorem (Eldan and Shamir 2016)

Si  $\sigma$  est "universelle" et au plus polynomiale, il existe  $c, C > 0$  telles que : pour toute dimension  $d > C$ , il existe une probabilité  $\mu$  sur  $\mathbb{R}^d$  et  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  vérifiant :

- 1  $g$  est à valeurs dans  $[-2, 2]$ , supportée sur  $\{x : \|x\| \leq C\sqrt{d}\}$  et implémentable par un réseau à 2 couches cachées de largeur polynomiale en  $d$  ;
- 2 toute fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  implémentable par un réseau à 1 couche cachée de largeur au plus  $ce^{cd}$  vérifie

$$\mathbb{E}_{x \sim \mu} \left[ (f(x) - g(x))^2 \right] \geq c.$$

Preuve par transformée de Fourier (fonction  $g$  radiale et oscillante).

En clair : il existe des fonctions représentables aisément par un réseau à 2 couches cachées, mais qu'on ne peut pas approcher par un réseau à 1 couche cachée, sauf à considérer une couche cachée de taille exponentielle en la dimension d'entrée  $d$ .

# Vitesses d'approximation en fonction de $W$ et $L$

On considère :

- Un espace de fonctions  $F$  à approcher ; par ex,  $F = \text{Lip}_1([0, 1]^d)$ .
- Un réseau feedforward (avec connexions résiduelles) à  $W$  paramètres (poids et biais), profondeur  $L$ , fonction d'activation  $\sigma$ , sortie scalaire linéaire.

La fonction représentée par le réseau paramétré par  $\mathbf{w} \in \mathbb{R}^W$  est notée  $g_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ .

On note  $G = \{g_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R} : \mathbf{w} \in \mathbb{R}^W\}$  l'ensemble des fonctions obtenues.

Une façon de quantifier l'expressivité du réseau est d'évaluer la distance de  $F$  à  $G$  définie par

$$\sup_{f \in F} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty}$$

A la fois pessimiste en  $f \in F$  et optimiste en  $\mathbf{w} \in \mathbb{R}^W$ .

La norme infinie peut être remplacée par  $L^p(\mu)$  (e.g., Achour et al. 2022).

**Question** : comment la distance de  $F$  à  $G$  dépend-elle de  $\sigma$ ,  $W$  et  $L$  ?

# Une borne inférieure d'approximation

Par simplicité, considérons  $F = \text{Lip}_1([0, 1]^d) = \{f : [0, 1]^d \rightarrow \mathbb{R} : \forall x, y, |f(x) - f(y)| \leq \|x - y\|\}$ .

Le théorème suivant formalise l'intuition selon laquelle un réseau expressif doit être complexe et donc que l'ensemble  $\text{sgn}(G)$  des classifieurs associés doit être de VC-dimension élevée.

## Theorem (inspiré de Yarotsky 2017)

Soit  $F = \text{Lip}_1([0, 1]^d)$ ,  $G \subset \mathbb{R}^{[0,1]^d}$  et  $\varepsilon = \sup_{f \in F} \inf_{g \in G} \|f - g\|_\infty$  la distance de  $F$  à  $G$ . On a :

$$\text{VCdim}(\text{sgn}(G)) \gtrsim \left(\frac{1}{\varepsilon}\right)^d$$

De façon équivalente, un réseau de petite VC-dimension ne peut pas bien approcher uniformément  $F = \text{Lip}_1([0, 1]^d)$  :

$$\sup_{f \in F} \inf_{g \in G} \|f - g\|_\infty \gtrsim \text{VCdim}(\text{sgn}(G))^{-1/d}.$$

La conclusion demeure (essentiellement) vraie avec une norme  $L^p(\mu)$  ; cf. Achour et al. (2022).

# Vitesse d'approximation avec $\sigma$ constante par morceaux

Si  $\sigma$  est constante par morceaux, le théorème suivant montre qu'on peut approcher  $\text{Lip}_1([0, 1]^d)$  avec une erreur de l'ordre de  $W^{-1/d}$ , et qu'il suffit de deux couches cachées.

En particulier, la profondeur n'est pas utile (au sens de l'approximation pire cas) pour un tel  $\sigma$ .

## Theorem (Yarotsky 2018)

Soit  $\sigma$  une fonction constante par morceaux et une architecture feedforward à  $W$  paramètres fixée. Alors,

$$\sup_{f \in \text{Lip}_1([0, 1]^d)} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty} \gtrsim (W \ln W)^{-1/d}.$$

Réciproquement, il existe une architecture feedforward à 2 couches cachées, au plus  $W$  paramètres, et  $\sigma(x) = \mathbb{1}_{x \geq 0}$  telle que

$$\sup_{f \in \text{Lip}_1([0, 1]^d)} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty} \lesssim W^{-1/d}.$$

Preuve borne inférieure : on utilise la borne  $\text{VCdim} \lesssim W \ln W$  prouvée par Bartlett et al. (2019).

Preuve borne supérieure : on approche  $f$  par une fonction constante par morceaux.

# Effet de la profondeur pour $\sigma$ polynomiale par morceaux

Si  $\sigma$  est affine par morceaux, une vitesse d'approximation en  $W^{-2/d}$  est théoriquement possible (on parle de **super-approximation**), à condition de considérer un réseau **très profond**.

Considérer  $\sigma$  polynomiale par morceaux de degré max  $\geq 2$  n'améliore pas la vitesse pire cas.

## Theorem (Yarotsky 2018)

Soit  $\sigma$  une fonction polynomiale par morceaux et une architecture feedforward à  $W$  paramètres fixée. Alors,

$$\sup_{f \in Lip_1([0,1]^d)} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty} \gtrsim W^{-2/d}.$$

Réciproquement, il existe une architecture feedforward ReLU à  $\leq W$  paramètres telle que

$$\sup_{f \in Lip_1([0,1]^d)} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty} \lesssim W^{-2/d}.$$

De plus, pour  $k \in [1, 2]$ , toute architecture feedforward ReLU à  $W$  paramètres telle que

$\sup_{f \in Lip_1([0,1]^d)} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_{\infty} \lesssim W^{-k/d}$  est nécessairement de profondeur

$$L \gtrsim \frac{W^{k-1}}{\ln W}.$$

# Sur l'instabilité d'une super-approximation

Mauvaise nouvelle : la vitesse (plus rapide) en  $W^{-2/d}$  théoriquement possible via un réseau ReLU profond est nécessairement associée à un **encodage instable (discontinu)**.

## Theorem ( DeVore et al. 1989 )

Soit

- $\Phi_{enc} : Lip_1([0, 1]^d) \rightarrow \mathbb{R}^W$  une "application d'encodage" **continue (pour la norme infinie)**
- $\Phi_{dec} : \mathbb{R}^W \rightarrow \mathcal{C}([0, 1]^d)$  une "application de décodage" quelconque

Alors,

$$\sup_{f \in Lip_1([0, 1]^d)} \|f - \Phi_{dec}(\Phi_{enc}(f))\|_{\infty} \gtrsim W^{-1/d}.$$

Ainsi, pour un réseau ReLU  $\Phi_{dec} : \mathbf{w} \in \mathbb{R}^W \mapsto f_{\mathbf{w}} \in \mathcal{C}([0, 1]^d)$ , la seule façon d'approcher  $Lip_1([0, 1]^d)$  à une vitesse meilleure que  $W^{-1/d}$  est d'encoder chaque  $f \in Lip_1([0, 1]^d)$  à l'aide de paramètres  $\Phi_{enc}(f)$  qui dépendent de façon **discontinue** de  $f$  (cf. "bit extraction technique").

La "super-vitesse"  $W^{-2/d}$  est donc d'intérêt pratique modéré.

# Le cas de fonctions plus régulières

Soit  $\beta = q + \alpha$  un paramètre de régularité avec  $q \in \mathbb{N}$  et  $0 < \alpha \leq 1$ , et  $L > 0$  une constante. On considère l'ensemble  $\mathcal{F}_{\beta,L}$  ("boule de Hölder") de toutes les fonctions  $f : [0, 1]^d \rightarrow \mathbb{R}$  telles que

$$\begin{aligned} |D^n f(\mathbf{x})| &\leq 1 && \text{pour tout } |\mathbf{n}| \leq q \text{ et } \mathbf{x} \in [0, 1]^d \\ |D^n f(\mathbf{x}) - D^n f(\mathbf{y})| &\leq L \|\mathbf{x} - \mathbf{y}\|^\alpha && \text{pour tout } |\mathbf{n}| = q \text{ et } \mathbf{x}, \mathbf{y} \in [0, 1]^d \end{aligned}$$

Pour ces fonctions plus régulières, une meilleure approximation que  $W^{-1/d}$  ou  $W^{-2/d}$  est possible, avec un effet similaire de la profondeur.

## Theorem (Yarotsky and Zhevnerchuk 2020)

Soit  $\sigma$  une fonction polynomiale par morceaux et une architecture feedforward à  $W$  paramètres fixée. Alors,

$$\sup_{f \in \mathcal{F}_{\beta,L}} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_\infty \gtrsim W^{-2\beta/d}.$$

Réciproquement, pour tout  $k \in [1, 2]$ , il existe une architecture feedforward ReLU à  $\leq W$  paramètres telle que

$$\sup_{f \in \mathcal{F}_{\beta,L}} \inf_{\mathbf{w} \in \mathbb{R}^W} \|f - g_{\mathbf{w}}\|_\infty \lesssim W^{-k\beta/d}.$$

Dans le cas  $k = 1$ , un réseau peu profond avec encodage continu suffit. Dans le cas  $k \in (1, 2]$ , un réseau de profondeur  $L \gtrsim W^{k-1} / \ln W$  avec encodage discontinu est nécessaire.

# Expressivité : exemples d'autres résultats

Il existe de nombreux autres résultats sur l'expressivité des réseaux de neurones, et la recherche est loin d'être close. Par exemple :

- Petersen and Voigtlaender (2018); Gühring and Raslan (2020) (et d'autres) étudient l'approximation par des **réseaux quantifiés**.
- Yarotsky and Zhevnerchuk (2020) prouvent une vitesse d'approximation exponentielle (mais avec encodage instable) avec des réseaux feedforward combinant les fonctions d'activation ReLU/sin ("Deep Fourier").
- Vershynin (2020) étudie la **capacité mémoire** de réseaux (la taille maximale  $n$  d'un jeu de données générique que le réseau est capable de mémoriser).
- Gribonval et al. (2022) décrivent les **espaces d'approximation**  $F$  que sont capables d'approcher des réseaux à une vitesse préspecifiée.
- Grohs and Voigtlaender (2023) quantifient à quel point une fonction qui est proche d'un réseau ReLU sera bien reconstruite **si on ne l'observe qu'en un nombre  $n$  de points**, même avec un algo de reconstruction bien choisi. Réponse : assez mal dans le pire cas.
- etc

Conseils de lecture : le survey "Neural Network Approximation" ( DeVore et al., 2021)

# References I

- 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. In *Advances in Neural Information Processing Systems*, volume 35, pages 22396–22408, 2022.
- Vladimir I. Arnold. On functions of three variables. *Dokl. Akad. Nauk SSSR (Russian)*, 114:679–681, 1957.
- Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.
- Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.
- Jürgen Braun and Michael Griebel. On a constructive proof of kolmogorov’s superposition theorem. *Constr. Approx.*, 30: 653–675, 2009.
- G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2:303–314, 1989.
- Amit Daniely. Depth separation for neural networks. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 690–696, 2017.
- Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.
- Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63 (4):469–478, 1989.
- Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, 2016.
- Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55(1):259–367, 2022. ISSN 1432-0940.
- Philipp Grohs and Felix Voigtlaender. Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces. *Foundations of Computational Mathematics*, 2023.
- Ingo Gühring and Mones Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 11 2020.

# References II

- Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080.
- Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327, 2020.
- Andreï N. Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR (Russian)*, 114:953–956, 1957.
- Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.
- G. G. Lorentz. Metric entropy, widths, and superpositions of functions. *Amer. Math. Monthly*, 69(6):469–485, 1962.
- Sidney A. Morris. Hilbert 13: Are there any genuine continuous multivariate real-valued functions? *Bull. Amer. Math. Soc.*, 58:107–118, 2021.
- Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018. ISSN 0893-6080.
- David A. Sprecher. A numerical implementation of kolmogorov's superpositions. *Neural Networks*, 9(5):765–772, 1996.
- David A. Sprecher. A numerical implementation of kolmogorov's superpositions ii. *Neural Networks*, 10(3):447–457, 1997.
- Roman Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM Journal on Mathematics of Data Science*, 2(4):1004–1033, 2020.
- Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.
- Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649, 2018.
- Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015, 2020.