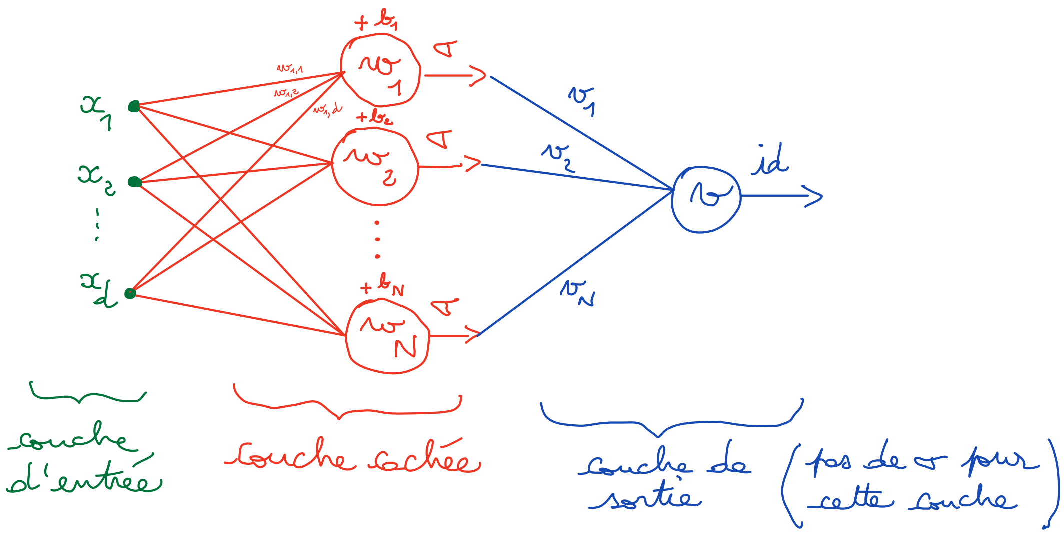
Approximation par des réseaux à 1 couche cachée

Sébastien GERCHINOVITZ

Plusieurs théorèmes s'intéressent à l'universalité des réseaux de neurones feedforward à 1 couche cachée, de la forme :

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

où $v_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$ et $b_i \in \mathbb{R}$ et $\sigma : \mathbb{R} \to \mathbb{R}$ "fonction d'activation"



Nous allons étudier un des premiers théorèmes, dû à Cybenko (1989).

Théorème (Cybenko '89).
Soit $\sigma : \mathbb{R} \to \mathbb{R}$ une fonction continue et sigmoïdale, i.e., $\begin{cases} \lim_{-\infty} \sigma = 0 \\ \lim_{+\infty} \sigma = 1 \end{cases}$.
Alors l'ensemble $\mathcal{N}_1$ des fonctions de la forme

$$x \in [0,1]^d \longmapsto \sum_{i=1}^{N} v_i \, \sigma(\langle w_i, x \rangle + b_i) \; ; \; v_i, b_i \in \mathbb{R}, \; w_i \in \mathbb{R}^d$$

est dense dans l'ensemble $\mathcal{C}([0,1]^d, \mathbb{R})$ des fonctions réelles continues sur $[0,1]^d$ (muni de $\|f\|_\infty = \sup_{x \in [0,1]^d} |f(x)|$).

Le théorème signifie que :

- $\mathcal{N}_1 \subset \mathcal{C}([0,1]^d, \mathbb{R})$
- $\forall f \in \mathcal{C}([0,1]^d, \mathbb{R})$, $\forall \varepsilon > 0$, $\exists N \geq 1$, $v_i, b_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$:

$$\sup_{x \in [0,1]^d} \left| f(x) - \sum_{i=1}^{N} v_i \, \sigma(\langle w_i, x \rangle + b_i) \right| \leq \varepsilon$$

La preuve procède par contradiction (c'est une preuve non constructive) :

- On suppose qu'il existe $f_0 \in \mathcal{C}([0,1]^d, \mathbb{R}) \setminus \overline{\mathcal{N}_1}$
- On invoque alors le <u>théorème de Hahn-Banach</u> pour construire une forme linéaire continue $L : \mathcal{C}([0,1]^d, \mathbb{R}) \to \mathbb{R}$ telle que $L(f_0) = 1$ et $L(f) = 0$ $\forall f \in \mathcal{N}_1$.
- On utilise ensuite le <u>théorème de représentation de Riesz-Markov</u> : $\exists \mu$ mesure de Borel signée sur $[0,1]^d$:

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

- On montre ensuite que $\mu \equiv 0$ nécessairement (via la transformée de Fourier de μ), d'où $L(f_0) = 0$ et donc une contradiction.

Rappelons d'abord les 2 thms d'analyse fonctionnelle utilisés. Toutes les définitions et propriétés énoncées dans les sections 1 et 2 sont valables lorsque le corps des scalaires est $\mathbb{R}$ ou $\mathbb{C}$.

1/ Théorème(s) de Hahn-Banach

Cf. Walter Rudin, Functional Analysis (Chapter 3), 1973.
Il existe plusieurs théorèmes de Hahn-Banach. En voici 2 :

Théorème de Hahn-Banach (propriété de séparation) :

Soit X un espace vectoriel topologique localement convexe*.
M un sous-espace de X.
$x_0 \in X \setminus \overline{M}$

Alors, il existe une forme linéaire continue $L : X \longrightarrow \underbrace{\mathbb{R} \text{ ou } \mathbb{C}}_{\text{cf } X}$
telle que
- $L(x_0) = 1$
- $\forall x \in M, \; L(x) = 0$.

* Tout espace vectoriel normé est un espace vectoriel topologique localement convexe.

Théorème de Hahn-Banach (extension d'une forme lin. cont.)

Soit X un espace vectoriel topologique localement convexe*.
M un sous-espace de X
f une forme linéaire continue sur M (à valeurs $\mathbb{R}$ ou $\mathbb{C}$ selon X)

Alors, il existe une forme linéaire continue L sur X telle que $L(x) = f(x) \;\; \forall x \in M$.

2/ Théorème de représentation de Riesz-Markov

Cf. Walter Rudin, Analyse réelle et complexe (chap 3 et 6), 1998.

a) Quelques brefs rappels : $\mathcal{C}_0(X)$ et mesures complexes.

Déf : Soit X un espace topologique séparé localement compact*.
$f : X \to \mathbb{C}$ ou $\mathbb{R}$
On dit que f s'annule à l'infini ssi : $\forall \varepsilon > 0$,
$\exists K$ compact de X tel que $|f(x)| \leq \varepsilon$ pour tout $x \in X \setminus K$.

* Exemple : partie compacte d'un espace vectoriel normé (e.g., $[0,1]^d$).

On note $\mathcal{C}_0(X)$ l'ensemble des fonctions continues sur X qui s'annulent à l'infini.

Lorsque X est séparé compact, $\mathcal{C}_0(X) = \mathcal{C}(X)$.

Déf : Soit $(X, \mathcal{M})$ un espace mesurable.

- Une mesure complexe μ sur $\mathcal{M}$ est une fonction $\mu : \mathcal{M} \to \mathbb{C}$ telle que : $\mu(E) = \sum_{i=1}^{+\infty} \mu(E_i)$ (convergence abs. dans $\mathbb{C}$) pour tout $E \in \mathcal{M}$ et toute partition mesurable $(E_i)_{i \geq 1}$ de E.
- mesure signée = idem mais avec $\mu : \mathcal{M} \to \mathbb{R}$

- La mesure de variation totale $|\mu|$ est définie par
$$|\mu|(E) := \sup \sum_{i=1}^{+\infty} |\mu(E_i)| \geq |\mu(E)| , \quad E\in\mathcal{M},$$
où le sup est pris sur toutes les partitions mesurables de E.

<u>Propriétés</u> (valables pour μ mesure complexe ou signée)

- $|\mu|$ est une mesure positive sur $\mathcal{M}$ et est toujours de masse finie : $|\mu|(X) < +\infty$.
- $\exists\, h : X \to \mathbb{C}$ ou $\mathbb{R}$ mesurable telle que $|h(x)| = 1 \ \forall x\in X$ et $d\mu = h\, d|\mu|$, i.e., $\forall E\in\mathcal{M}$, $\mu(E) = \int_E h\, d|\mu|$.

On définit alors :
$$\forall f\in L^1(|\mu|), \int f\, d\mu := \int f\, h\, d|\mu|$$

<u>NB</u> : $h(x) = \pm 1$ si μ signée.

b) Théorème de représentation de Riesz-Markov

Plusieurs théorèmes de représentation de Riesz. Le suivant est parfois appelé "théorème de Riesz-Markov".

<u>Théorème de représentation de Riesz (version pour $\mathcal{C}_0(X)$)</u> :

Soit X un espace topologique séparé localement compact.

L une forme linéaire continue sur $(\mathcal{C}_0(X), \|\cdot\|_\infty)$ ← scalaires $= \mathbb{R}$ ou $\mathbb{C}$

Alors, il existe une unique mesure de Borel complexe/signée et régulière μ qui représente L, au sens où :
$$\forall f\in\mathcal{C}_0(X), \quad L(f) = \int f\, d\mu.$$

De plus, $\|L\| := \sup_{\|f\|_\infty\leq 1} |L(f)|$ vérifie $\|L\| = \underbrace{|\mu|(X)}_{\text{"variation totale de } \mu\text{"}}$.

3/ Preuve du théorème d'approximation de Cybenko '89.

On a $\mathcal{N}_1 \subset \mathcal{C}([0,1]^d, \mathbb{R})$ car σ est continue.

- Supposons un instant que $\overline{\mathcal{N}_1} \neq \mathcal{C}([0,1]^d, \mathbb{R})$ et fixons $f_0 \in \mathcal{C}([0,1]^d, \mathbb{R}) \setminus \overline{\mathcal{N}_1}$.

- Appliquons le théorème de Hahn-Banach (version "séparation") à l'espace vectoriel normé $X = \mathcal{C}([0,1]^d, \mathbb{R})$ muni de $\|\cdot\|_\infty$ et au sous-espace $\overline{\mathcal{N}_1}$: on peut fixer une forme linéaire continue $L : \mathcal{C}([0,1]^d, \mathbb{R}) \to \mathbb{R}$ tq $L(f_0) = 1$ mais $L \equiv 0$ sur $\overline{\mathcal{N}_1}$.

- Or, d'après le théorème de représentation de Riesz pour $\mathcal{C}_0(\underbrace{[0,1]^d}_{\text{séparé et compact}}, \mathbb{R}) = \mathcal{C}([0,1]^d, \mathbb{R})$, il existe une mesure de Borel signée μ sur $[0,1]^d$ telle que :

$$\forall f \in \mathcal{C}([0,1]^d, \mathbb{R}), \quad L(f) = \int f \, d\mu \qquad (2)$$

- Montrons que $\mu \equiv 0$, et donc $L(f_0) = 0$, d'où une contradiction.

Puisque L est nulle sur $\mathcal{N}_1$ et $x \mapsto \sigma(\langle w, x \rangle + b) \in \mathcal{N}_1$, on a :

$$\forall w\in\mathbb{R}^d,\ \forall b\in\mathbb{R},\quad \int_{[0,1]^d} \sigma\big(\langle w,x\rangle + b\big)\, d\mu(x) = 0 \qquad (3)$$

car continue et sigmoïdale

Nous allons montrer que puisque σ est bornée et sigmoïdale, la seule mesure de Borel signée vérifiant (3) est la mesure nulle. (On dit dans ce cas que σ est "discriminante".)

Soit $w\in\mathbb{R}^d$ et $b,\lambda,\varphi\in\mathbb{R}$. Pour tout $x\in\mathbb{R}^d$,

$$\sigma_\lambda(x) := \sigma\Big(\lambda\big(\langle w,x\rangle + b\big) + \varphi\Big) \xrightarrow[\lambda\to+\infty]{} \gamma(x) := \begin{cases} 1 \text{ si } \langle w,x\rangle + b > 0 \\ 0 \text{ si } \langle w,x\rangle + b < 0 \\ \sigma(\varphi) \text{ si } \langle w,x\rangle + b = 0 \end{cases}$$

D'où, par convergence dominée (car σ est bornée, $d\mu = h\,d|\mu|$ avec $|h| = 1$, et $|\mu|$ est de masse finie) :

$$0 \overset{\sigma_\lambda\in\mathcal{L}_1}{=} \int_{[0,1]^d} \sigma_\lambda(x)\, d\mu(x) \xrightarrow[\lambda\to+\infty]{} \int_{[0,1]^d} \gamma(x)\, d\mu(x) = \sigma(\varphi)\,\mu\big(\Pi_{w,b}\big) + \mu\big(H_{w,b}\big)$$

où $\Pi_{w,b} := \{x\in[0,1]^d : \langle w,x\rangle + b = 0\}$

$H_{w,b} := \{x\in[0,1]^d : \langle w,x\rangle + b > 0\}$

On a donc montré :

$$\forall w\in\mathbb{R}^d,\ \forall b\in\mathbb{R},\ \forall\varphi\in\mathbb{R},\quad \sigma(\varphi)\,\mu\big(\Pi_{w,b}\big) + \mu\big(H_{w,b}\big) = 0$$

Puisque σ est non-constante (car sigmoïdale), on a donc :

$$\forall w\in\mathbb{R}^d,\ \forall b\in\mathbb{R},\quad \mu\big(\Pi_{w,b}\big) = \mu\big(H_{w,b}\big) = 0.$$

Le fait que μ soit une mesure signée (pas nécessairement positive) ne permet pas tout de suite de conclure que $\mu \equiv 0$ (du moins, ce n'est pas évident à ce stade). [facile en dimension $d=1$ via un argument de classe monotone]

Soit $w \in \mathbb{R}^d$. Définissons, pour toute $h : [-\|w\|_1, \|w\|_1] \to \mathbb{C}$ bornée,

$$\psi(h) = \int_{[0,1]^d} h(\langle w, x \rangle) \, d\mu(x)$$

On remarque que ψ est linéaire et continue car

$$|\psi(h)| \leq \|h\|_\infty \cdot \underbrace{|\mu|([0,1]^d)}_{<+\infty}$$

Or, pour $h = \mathbb{1}_{[\theta, +\infty[}$, $\psi(h) = \mu(\Pi_{w,-\theta}) + \mu(H_{w,-\theta}) = 0$

De même, $\psi(\mathbb{1}_{]\theta,+\infty[}) = \mu(H_{w,-\theta}) = 0$

D'où : $\psi(\mathbb{1}_{[\theta_1,\theta_2]}) = \psi(\mathbb{1}_{[\theta_1,\theta_2[}) = \psi(\mathbb{1}_{]\theta_1,\theta_2]}) = \psi(\mathbb{1}_{]\theta_1,\theta_2[}) = 0$

Par linéarité de ψ, on a donc $\psi(h) = 0$ pour toute fonction h en escalier sur $[-\|w\|_1, \|w\|_1]$.

En approchant $h(t) = e^{it} = \cos(t) + i\sin(t)$ par des fonctions en escalier sur $[-\|w\|_1, \|w\|_1]$, on en déduit, par continuité de ψ :

$$\int_{[0,1]^d} e^{i\langle w, x\rangle} \, d\mu(x) = 0 \qquad \forall w \in \mathbb{R}^d.$$

On reconnaît la transformée de Fourier de μ !
Déduisons-en que μ est nulle.

On admettra pour cela qu'il suffit de montrer que $\int_{[0,1]^d} f \, d\mu = 0$ pour tout $f = g_{|[0,1]^d}$ avec $g : \mathbb{R}^d \to \mathbb{R}$ de classe C^∞ et à support compact.

Pour tout $x \in [0,1]^d$, on a :

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

(formule d'inversion valable car, par ex., g est dans l'espace de Schwartz)

où $\hat{g}(w) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} g(x) \, e^{-i\langle w, x\rangle} \, dx$ est la transformée de Fourier de g.

Dès lors :

$$\int_{[0,1]^d} f(x) \, d\mu(x) = \int_{[0,1]^d} \left(\int_{\mathbb{R}^d} \hat{g}(w) \, e^{i\langle w, x\rangle} \, dw \right) d\mu(x)$$

$$\overset{\text{Fubini}}{=} \int_{\mathbb{R}^d} \hat{g}(w) \underbrace{\left(\int_{[0,1]^d} e^{i\langle w, x\rangle} \, d\mu(x) \right)}_{=0} dw$$

$$= 0$$

- D'où $\mu \equiv 0$ et donc $L(f_0) = \int f_0 \, d\mu = 0$, ce qui contredit $L(f_0) \neq 1$. Donc $\mathcal{N}_1$ est dense dans $C([0,1]^d, \mathbb{R})$. ∎

Take-home messages :

- le résultat de densité $\overline{\mathcal{N}_1} = \mathcal{C}([0,1]^d, \mathbb{R})$

- la technique de preuve, non constructive, par contradiction. L'utilisation du théorème de Hahn-Banach pour prouver un résultat de densité est classique.

4/ Annexe : nullité de μ à partir de fonctions test.

On a utilisé le lemme suivant avec la mesure signée μ étendue à $\mathbb{R}^d$ ($\nu(A) := \mu(A \cap [0,1]^d)$ pour tout borélien $A \subset \mathbb{R}^d$).

Lemme :

plus généralement : complexe

Soit ν une mesure de Borel signée sur $\mathbb{R}^d$ telle que
$$\int g \, d\nu = 0$$
pour toute fonction $g : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^\infty$ à support compact (on écrit : $g \in \mathcal{C}_c^\infty(\mathbb{R}^d)$).
Alors ν est la mesure nulle sur $\mathbb{R}^d$.

Preuve : soit A un borélien de $\mathbb{R}^d$. Par densité de $\mathcal{C}_c^\infty(\mathbb{R}^d)$ dans $L^1(\mathbb{R}^d, |\nu|)$, il existe $g_n \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ telles que

$$\int_{\mathbb{R}^d} |g_n - \mathbb{1}_A| \, d|\nu| \xrightarrow[n \to +\infty]{} 0$$

D'où :
$$\Big| \underbrace{\int g_n \, d\nu}_{=0} - \nu(A) \Big| = \Big| \int (g_n - \mathbb{1}_A) \, d\nu \Big|$$
$$\leq \int |g_n - \mathbb{1}_A| \, d|\nu| \xrightarrow[n \to +\infty]{} 0$$

et donc $\nu(A) = 0$ pour tout borélien A. $\blacksquare$