Réseaux de neurones : expressivité et espaces implicites

Introduction

De très nombreux problèmes d'apprentissage automatique peuvent être formulés ainsi :

Soient X, E des ensembles et $f : X \to E$ une fonction.
Si un échantillon $(x_i, f(x_i))_{i\in I}$ est fourni, comment estimer $f(x)$ pour un x quelconque ?

Les réponses apportées à cette question permettent à un algorithme de proposer de bons coups au jeu de go, d'effectuer des prédictions météorologiques, ou encore de distinguer une route et une autruche.

En imposant certaines hypothèses sur la structure des ensembles et la régularité des fonction étudiées, il est possible de trouver des stratégies intéressantes pour résoudre ce problème.

Les réseaux de neurones formels, introduits dès les années 50 pour modéliser des systèmes neuronaux biologiques, peuvent être utilisés à cette fin. L'augmentation de la taille de nos bases de données et des puissances de calcul en font des solutions algorithmiques de plus en plus adéquates.

Dans ce TIPE, après une rapide présentation des réseaux de neurones formels, on se propose d'utiliser des méthodes d'analyse et de topologie pour établir quelques résultats théoriques à leur sujet.

La deuxième partie de ce rapport sera consacrée à la démonstration d'un résultat classique : sous certaines hypothèses, les réseaux de neurones sont des approximateurs universels.

Enfin, dans la troisième partie, on présentera le théorème de Moore-Aronszajn, qui suggère l'utilisation de nouvelles classes de neurones.

1 Réseaux de neurones

1.1 Définitions

Définition 1 (Fonction d'activation)**.** *On appelle* fonction d'activation *toute fonction continue croissante* $\varphi : \mathbb{R} \to \mathbb{R}$ *telle que* $\lim\limits_{x\to-\infty} \varphi(x) = 0$ *et* $\lim\limits_{x\to+\infty} \varphi(x) = 1$.

Exemple 1 (Sigmoïde)**.** $\varphi : x \mapsto \frac{1}{1+e^{-x}}$ *est une fonction d'activation.*

Définition 2 (Neurone)**.** *Soient* $n \in \mathbb{N}^*$, $\langle\cdot,\cdot\rangle$ *le produit scalaire canonique sur* $\mathbb{R}^n$.
On appelle neurone *toute fonction*

$$\begin{array}{rccc} f & : & \mathbb{R}^n \times \mathbb{R}^n & \to & \mathbb{R} \\ & & (x, y) & \mapsto & \varphi(\langle x, y\rangle), \end{array}$$

où φ *est une fonction d'activation.*
Dans l'expression $f(x, y)$, x *(resp.* y*) est appelé* souvenir *(resp.* stimulus*) du neurone* f.

Remarque 1. *Un neurone compare un souvenir et un stimulus, et fournit un coefficient compris entre 0 et 1.*

Dans la suite de ce rapport, pour tout $n \in \mathbb{N}^*$, on identifie $\mathbb{R}^n$ et $\mathcal{M}_{n,1}(\mathbb{R})$.

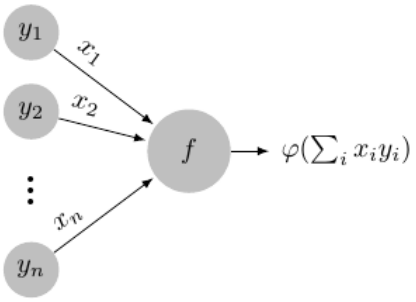

FIGURE 1 – Représentation classique d'un neurone

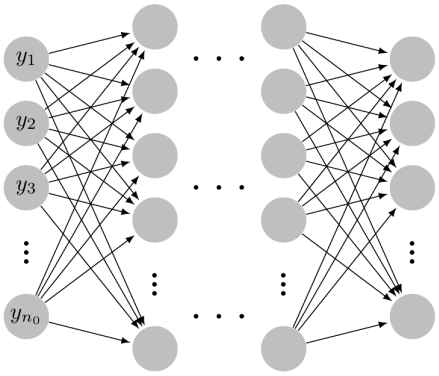

FIGURE 2 – Réseau de neurones

Définition 3 (Couche de neurones)**.** *Soient $n, m \in \mathbb{N}^*$.*

On appelle couche de neurones *de largeur m toute fonction*

$$\begin{array}{rccc} F & : & \mathcal{M}_{m,n}(\mathbb{R}) \times \mathbb{R}^n & \to & \mathbb{R}^m \\ & & (X, y) & \mapsto & (f(x_1, y), \dots, f(x_m, y)), \end{array}$$

où $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est un neurone, et où les x_i sont les lignes de X.

Dans l'expression $F(X, y)$, X (resp. y) est appelé souvenir *(resp.* stimulus*) de la couche F.*

Définition 4 (Couche finale)**.** *Soient $n, m \in \mathbb{N}^*$.*

On appelle couche finale *de largeur m toute couche de neurones sans fonctions d'activations, c'est à dire toute fonction*

$$\begin{array}{rccc} F & : & \mathcal{M}_{m,n}(\mathbb{R}) \times \mathbb{R}^n & \to & \mathbb{R}^m \\ & & (X, y) & \mapsto & (\langle x_1, y\rangle), \dots, \langle x_m, y\rangle) = Xy, \end{array}$$

où X (resp. y) est appelé comme précédemment souvenir *(resp.* stimulus*) de la couche F.*

Remarque 2. *Les couches de neurones sont à valeurs dans $[0, 1]^m$ alors qu'une couche finale peut être à valeurs dans $\mathbb{R}^m$ tout entier.*

Définition 5 (Réseau de neurones)**.** *Soient $p \in \mathbb{N}^*$, $n_0, \dots, n_p \in \mathbb{N}^*$,*
$F_1 : \mathcal{M}_{n_1,n_0}(\mathbb{R}) \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$, …, $F_{p-1} : \mathcal{M}_{n_{p-1},n_{p-2}}(\mathbb{R}) \times \mathbb{R}^{n_{p-2}} \to \mathbb{R}^{n_{p-1}}$ des couches de neurones et $F_p : \mathcal{M}_{n_p,n_{p-1}}(\mathbb{R}) \times \mathbb{R}^{n_{p-1}} \to \mathbb{R}^{n_p}$ une couche finale.

Notons $E = \mathcal{M}_{n_1,n_0}(\mathbb{R}) \times \cdots \times \mathcal{M}_{n_p,n_{p-1}}(\mathbb{R})$.

On appelle alors réseau de neurones *de profondeur p et de couches $F_1, \dots, F_p$ l'application*
$\Phi : E \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_p}$ telle que pour tous $\mathcal{X} = (X_1, \dots, X_p) \in E$, $y \in \mathbb{R}^{n_0}$,

$$\Phi(\mathcal{X}, y) = F_p(X_p, F_{p-1}(X_{p-1}, \dots F_1(X_1, y) \dots)).$$

Dans l'expression $\Phi(\mathcal{X}, y)$, $\mathcal{X}$ est appelé mémoire *du réseau Φ.*

Remarque 3. *Dans les applications, les réseaux de neurones utilisés sont souvent très profonds.*

Nous remarquerons cependant dans la partie 2 que les réseaux de neurones sont des «approximateurs universels» dès $p = 2$.

1.2 Descente de gradient

Le problème est maintenant le suivant : étant donnés $p, n_0, \dots, n_p, E$ et Φ comme dans la définition 5, $\Psi : \mathbb{R}^{n_0} \to \mathbb{R}^{n_p}$, ainsi qu'un échantillon $(e_i, \Psi(e_i))_{i \in I}$, construire $\mathcal{X} \in E$ tel que $\Phi(\mathcal{X}, \cdot)$ soit une «bonne approximation» de Ψ.

La question de la construction de $\mathcal{X}$ ne sera pas étudiée ici en détail, mais présentons tout de même rapidement une méthode canoniquement utilisée dans les implémentations, celle de la descente de gradient, qui suppose les fonctions d'activations φ dérivables.

Pour $t \in \mathbb{N}$, construisons $\mathcal{X}^{(t)}$.

On choisit $\mathcal{X}^{(0)} \in E$ quelconque, ainsi que $\eta > 0$ petit.

Ensuite, par récurrence, pour tout $t \in \mathbb{N}$, on choisit e_{i_t} au hasard dans l'échantillon, on note $Err_t : \mathcal{X} \mapsto \frac{1}{2}\|\Phi(\mathcal{X}, e_{i_t}) - \Psi(e_{i_t})\|^2$, et on fait :

$$\mathcal{X}^{(t+1)} = \mathcal{X}^{(t)} - \eta \overrightarrow{\nabla} Err_t(\mathcal{X}^{(t)}).$$

2 Expressivité

En tâtonnant un peu pour dimensionner le réseau, et en ajoutant des biais aux neurones (cf Figure 4), dès $p \geq 2$, on constate que cet algorithme permet de bien approximer les fonctions à représenter.

Un exemple est donné Figure 3, où l'on cherche à prédire la couleur ($\in [0,1]$) d'un point de $\mathbb{R}^2$, en ne connaissant que sa position et la couleur d'autres points.

On remarque que dans ce cas, les hyperplans affines qui composent les souvenirs du réseau permettent de découper l'espace de façon à coller aux données.

Le théorème 2, dû a Kurt Hornik [1], nous permet en fait d'énoncer un résultat général : quitte à autoriser des biais et des n_1 arbitrairement grands, on peut représenter grâce à un réseau de profondeur $p = 2$ toute fonction continue $\mathbb{R}^{n_0} \to \mathbb{R}^{n_p}$ avec une précision arbitraire sur tout compact.

2.1 Théorème de Stone-Weierstrass

La preuve du théorème 2 s'appuie sur le théorème de Stone-Weierstrass [2], admis ici, qui généralise le théorème de Weierstrass pour les polynômes, et admet pour énoncé :

Théorème 1 (Stone-Weierstrass)**.** *Soient X un espace compact, $\mathcal{C}^0(X, \mathbb{R})$ l'algèbre des applications continues de X dans $\mathbb{R}$.*

Soit A une sous-algèbre de $\mathcal{C}^0(X, \mathbb{R})$, i.e. un sous-espace vectoriel de $\mathcal{C}^0(X, \mathbb{R})$ stable par produit (on n'impose pas $(x \mapsto 1) \in A$), telle que

(i) A sépare les points de X, i.e. pour tous $x, y \in X$, si $x \neq y$, alors il existe $f \in A$ telle que $f(x) \neq f(y)$

(ii) Pour tout $x \in X$ il existe $f \in A$ telle que $f(x) \neq 0$.

Alors, au sens de la convergence uniforme,

$$\overline{A} = \mathcal{C}^0(X, \mathbb{R}).$$

2.2 Approximateurs universels

Soit $d \in \mathbb{N}^*$.

Définition 6. *Soit φ une fonction d'activation.*

On note $\Omega^d(\varphi)$ l'ensemble des $\Phi(\mathcal{X}, \cdot) : \mathbb{R}^d \to \mathbb{R}$ où Φ est un réseau de profondeur 2, dont tous les neurones ont fonction d'activation φ, et où $\mathcal{X}$ est une mémoire de Φ.

Proposition 1. *Soit φ une fonction d'activation.*

$$\Omega^d(\varphi) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \forall x \in \mathbb{R}^d \; f(x) = \sum_{i=1}^{q} \beta_i \varphi(h_i(x)), \text{où } q \in \mathbb{N}^*, \beta_i \in \mathbb{R}, h_i \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}), \text{indépendamment de } x \right\}.$$

Définition 7. *Soit $g \in \mathcal{C}^0(\mathbb{R})$. Soit $\mathbb{A}^d$ l'ensemble des fonctions affines de $\mathbb{R}^d$ dans $\mathbb{R}$. On note*

$$\Sigma^d(g) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \forall x \in \mathbb{R}^d \; f(x) = \sum_{i=1}^{q} \beta_i g(A_i(x)), \text{où } q \in \mathbb{N}^*, \beta_i \in \mathbb{R}, A_i \in \mathbb{A}^r \right\}.$$

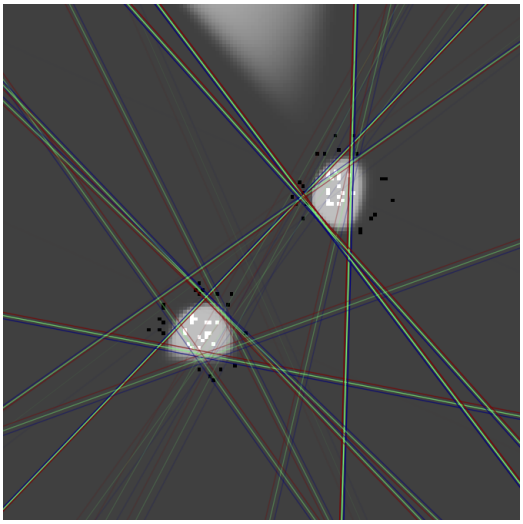

FIGURE 3 – Prédiction du réseau. Selon le réseau, les points blancs sont à trouver dans les zones de couleurs claires, et les points noirs dans les zones foncées. Les échantillons ont une couleur marquée. Les lignes de séparation marquent les hyperplans affines correspondants aux souvenirs de la couche intermédiaire.

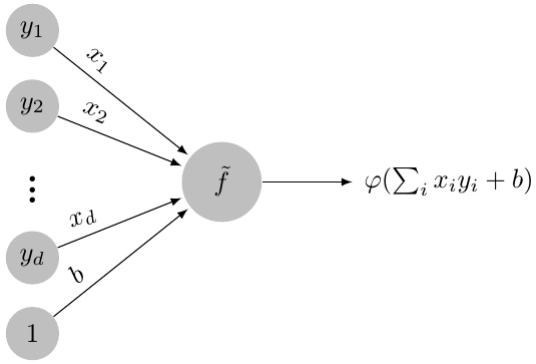

FIGURE 4 – Ajout d'un biais. On impose toujours $y_{d+1} = 1$, mais on permet à $x_{d+1} = b$ d'évoluer (au cours de la descente de gradient).

Remarque 4. *Par l'opération représentée en Figure 4, pour toute fonction d'activation φ, tout élément de $\Sigma^d(\varphi)$ peut être représenté par un réseau de neurone de profondeur 2 utilisant φ pour seule fonction d'activation.*

Le résultat central de cette partie est le suivant :

Théorème 2. *Soit φ une fonction d'activation.*
Alors, $\Sigma^d(\varphi)$ est dense dans $\mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ au sens de la convergence uniforme sur tout compact.

Le reste de cette partie est consacré au schéma d'une preuve du théorème 2, qui s'effectue par une série de réductions.

Définition 8. *Soit $g \in \mathcal{C}^0(\mathbb{R})$. On note*

$$\Sigma\Pi^d(g) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \forall x \in \mathbb{R}^d \; f(x) = \sum_{i=1}^{q} \prod_{k=1}^{l_i} \beta_i g(A_{i,k}(x)), \text{où } q \in \mathbb{N}^*, l_i \in \mathbb{N}^*, \beta_i \in \mathbb{R}, A_{i,k} \in \mathbb{A}^r \right\}.$$

Définition 9. *Soient $K \subset \mathbb{R}^d$ un compact et $g \in \mathcal{C}^0(\mathbb{R})$. On pose $\Sigma^d_{|K}(g) = \left\{ f_{|K} \;\middle|\; f \in \Sigma^d(g) \right\}$ et $\Sigma\Pi^d_{|K}(g) = \left\{ f_{|K} \;\middle|\; f \in \Sigma\Pi^d(g) \right\}$*

Proposition 2. *D'après le théorème de Stone-Weierstrass, pour tous $g \in \mathcal{C}^0(\mathbb{R})$, $K \subset \mathbb{R}^d$ compact, au sens de la convergence uniforme,*

$$\overline{\Sigma\Pi^d_{|K}(g)} = \mathcal{C}^0(K, \mathbb{R}).$$

Proposition 3. *Soient $f, g \in \mathcal{C}^0(\mathbb{R})$. Si $g \in \Sigma^1(f)$, alors $\Sigma^d(g) \subset \Sigma^d(f)$.*

Proposition 4. *Soient $f, g \in \mathcal{C}^0(\mathbb{R})$. Si $\forall M > 0$, $g_{|[-M,M]} \in \overline{\Sigma^1_{|[-M,M]}(f)}$,*
alors pour tout compact $K \subset \mathbb{R}^d$, $\Sigma^d_{|K}(g) \subset \overline{\Sigma^d_{|K}(f)}$.

La démonstration de la proposition suivante requiert une attention particulière.

Proposition 5. *Si ψ, φ sont des des fonctions d'activation alors $\psi \in \overline{\Sigma^1(\varphi)}$.*

Proposition 6. *La fonction suivante est une fonction d'activation :*

$$\psi : x \mapsto \begin{cases} 0 & si\ x < -\pi/2 \\ \frac{1+\cos(x+\frac{3\pi}{2})}{2} & si\ -\pi/2 \leq x < \pi/2 \\ 1 & si\ \pi/2 \leq x. \end{cases}$$

Proposition 7. *Avec les notations de la proposition 6, par translations et combinaisons linéaires, $\forall M > 0$, $\cos_{|[-M,M]} \in \Sigma^1_{|[-M,M]}(\psi)$.*

Démonstration du théorème 2. Soient φ une fonction d'activation et $K \subset \mathbb{R}^d$ un compact. Avec les notations de 6, d'après 7, 3, 5 puis 4,

$$\Sigma^d_{|K}(\cos) \subset \Sigma^d_{|K}(\psi) \subset \overline{\Sigma^d_{|K}(\varphi)}.$$

On peut alors conclure grâce à la proposition 2 et par linéarisation des polynômes en cos, point clé de la démonstration :

$$\mathcal{C}^0(K, \mathbb{R}) = \overline{\Sigma\Pi^d_{|K}(\cos)} = \overline{\Sigma^d_{|K}(\cos)} \subset \overline{\Sigma^d_{|K}(\varphi)} \subset \mathcal{C}^0(K, \mathbb{R}).$$

3 Espaces implicites

Les neurones employés jusqu'à présent utilisent des produits scalaires pour comparer les stimuli avec leurs souvenirs. La Figure 5 présente quelques-uns des souvenirs d'un réseau qui identifie des chiffres écrits à la main avec environ 2% d'erreurs.

La figure 6 montre pourtant qu'il peut être intéressant d'utiliser d'autres fonctions de type positif (cf Définition 13) pour effectuer des séparations «linéaires». Cette technique (appelée «astuce du noyau») pourrait être utilisée sur les réseaux de neurones.

Le théorème 4 fournit une interprétation agréable d'une telle manipulation : elle correspond à se placer implicitement dans un nouvel espace pour effectuer le produit scalaire.

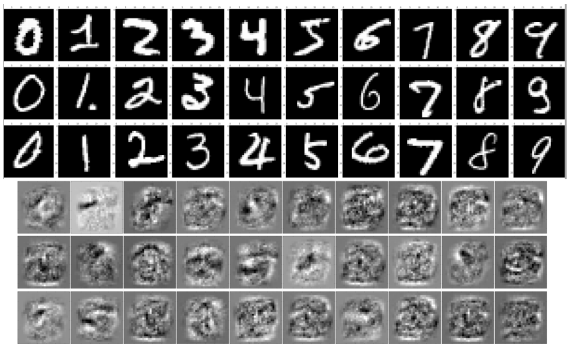

FIGURE 5 – Chiffres écrits à la main (base de données MNIST[3]), et quelques souvenirs des neurones de la couche intermédiaire.

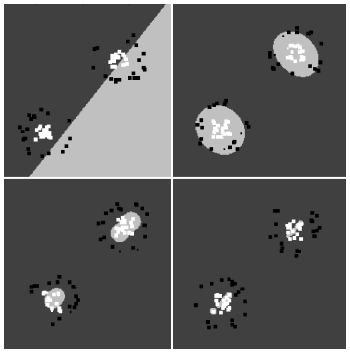

FIGURE 6 – Comparaison entre un classificateur linéaire et trois classificateurs gaussiens, pour $\sigma = 0.3$, $\sigma = 0.1$ puis $\sigma = 0.01$, en utilisant une machine à vecteur de support (de gauche à droite puis de haut en bas).

3.1 Théorème de représentation de Riesz

Définition 10. *On appelle* espace de Hilbert *tout espace $(H, \langle\cdot,\cdot\rangle)$ préhilbertien réel et complet, i.e. tel que toutes les suites de Cauchy de H convergent au sens de la norme induite par le produit scalaire.*

Proposition 8. *Soient $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert et $F \subset H$ un sous-espace vectoriel fermé. Alors, $F^\perp \oplus F = H$.*

Théorème 3 (Fréchet-Riesz, ou Théorème de représentation de Riesz [4])**.**
Soient $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert et $H' = \{f \in \mathcal{L}(H, \mathbb{R}) | f$ est continue$\}$ le dual topologique de H, que l'on munit de la norme d'opérateur.
Pour tout $f \in H'$, il existe un unique $y \in H$ tel que pour tout $x \in H$, $f(x) = \langle y, x\rangle$.
En effet, l'application suivante est un isomorphisme isométrique, et donc une bijection.

$$\begin{array}{rccc} \varphi & : & H & \to & H' \\ & & y & \mapsto & \langle y, \cdot\rangle \end{array}$$

3.2 Théorème de Moore-Aronszajn

Définition 11 (Espace de Hilbert à noyau reproduisant)**.** *Soient X un ensemble et $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert tel que $H \subset \mathbb{R}^X$. Pour tout $x \in X$, on pose*

$$\begin{array}{rccc} \delta_x & : & H & \to & \mathbb{R} \\ & & f & \mapsto & f(x). \end{array}$$

On dit que $(H, \langle\cdot,\cdot\rangle)$ est un espace de Hilbert à noyau reproduisant *sur X si et seulement si pour tout $x \in X$, δ_x est continue.*

Définition 12. *Soient X un ensemble et $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert à noyau reproduisant sur X.*
D'après le théorème 3, pour tout $x \in X$, on dispose d'un unique $K_x \in H$ tel que
$\forall f \in H,\ f(x) = \langle f, K_x\rangle$.
On définit alors le noyau reproduisant K de H :

$$\begin{array}{rccc} K & : & X \times X & \to & \mathbb{R} \\ & & (x, y) & \mapsto & \langle K_x, K_y\rangle. \end{array}$$

Définition 13 (Noyau symétrique de type positif)**.** *Soient X un ensemble et $K : X \times X \to \mathbb{R}$ une application.*
On dit que K est un noyau symétrique de type positif *sur X si et seulement si*

(i) $\forall x, y \in X,\ K(x, y) = K(y, x)$

(ii) $\forall n \in \mathbb{N}^*\ \forall x \in X^n\ \forall a \in \mathcal{M}_{n,1}(\mathbb{R}),\ {}^t aG(x_1, ...x_n)a \geq 0$, *où* $G = (K(x_i, x_j))_{1\leq i,j\leq n}$.

Proposition 9. *La propriété «être un noyau de type positif» est stable par somme, produit, multiplication par un scalaire positif, et passage à la limite simple.*

Exemple 2. *Les produits scalaires sont des noyaux symétriques de type positif - ils sont de plus bilinéaires et définis.*
Ainsi, les $(x, y) \mapsto (1 + m\langle x, y\rangle)^p$ où $m \in \mathbb{R}_+$, $p \in \mathbb{N}^*$ sont également de type positif.
Il est également possible de montrer que, pour tous $d \in \mathbb{N}^*, \sigma > 0$, le noyau gaussien K est de type positif, où $\forall x, y \in \mathbb{R}^d,\ K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

Remarque 5. *Tout noyau reproduisant est de type positif.*

Le théorème suivant présente une réciproque de cette propriété.

Théorème 4 (Théorème de Moore-Aronszajn)**.** *Soient X un ensemble, et $K : X \times X \to \mathbb{R}$ un noyau symétrique de type positif.*

Il existe un unique espace de Hilbert $(H, \langle \cdot, \cdot \rangle)$ où $H \subset \mathbb{R}^X$ dont K est un noyau reproduisant.

De plus, le produit scalaire de H étend K dans le sens suivant : en définissant les K_x comme en 12,

$$\forall x, y \in X,\ K(x, y) = \langle K_x, K_y \rangle.$$

On dit alors que K définit implicitement l'espace *H.*

Présentons la démarche de la preuve de l'existence, développée dans l'article cité ci-dessous.[5]

On pose $H_0 = Vect\{K_x | x \in X\}$ puis, pour tout $f, g \in H_0$, $\langle f, g \rangle_{H_0} = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$ où $f = \sum_i \alpha_i K_{x_i}$, $g = \sum_j \alpha_j K_{y_j}$.

On vérifie alors que

- $\langle \cdot, \cdot \rangle_{H_0}$ est un produit scalaire bien défini.
- Les δ_x sont continues sur H_0.
- Une suite $(f_n) \in H_0^{\mathbb{N}}$ qui converge simplement vers 0 et qui est de Cauchy au sens de $\| \cdot \|_{H_0}$ converge vers 0 au sens de $\| \cdot \|_{H_0}$.

Posons ensuite H l'ensemble des limites simples des suites $(f_n) \in H_0^{\mathbb{N}}$ qui convergent simplement dans $\mathbb{R}^X$ et sont de Cauchy au sens de $\| \cdot \|_{H_0}$.

Pour tous $f, g \in H$, on pose $\langle f, g \rangle_H = \lim_{n \to +\infty} \langle f_n, g_n \rangle_{H_0}$ où $(f_n), (g_n)$ sont des suites de H_0 qui convergent simplement vers f, g et qui sont de Cauchy au sens de $\| \cdot \|_{H_0}$.

On vérifie alors que

- $\langle \cdot, \cdot \rangle_H$ est un produit scalaire bien défini.
- H_0 est dense dans H au sens de $\| \cdot \|_H$, d'où les δ_x continues sur H.
- H est complet.
- H admet bien K pour noyau reproduisant.

Références

[1] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2 : 359–366, 1989.

[2] G. Choquet. *Cours d'Analyse. Tome II : Topologie, ch. VI-4.* Masson & Cie, 1973.

[3] THE MNIST DATABASE of handwritten digits. `http://yann.lecun.com/exdb/mnist/`. Document téléchargé le 25/10/2017.

[4] G. Choquet. *Cours d'Analyse. Tome II : Topologie, ch. VII-4.* Masson & Cie, 1973.

[5] Dino Sejdinovic and Arthur Gretton. What is an RKHS ? 2014. `www.cmap.polytechnique.fr/~zoltan.szabo/teaching/advanced_topics_in_ML/RKHS_note.pdf`.