TER M1 :
NN and RKHS

Matthieu Denis

16 juin 2021

Table des matières

1 Reproducing Kernel Hilbert space (RKHS)

Nous allons dans cette section s'intéresser aux RKHS, des espaces de Hilbert réels qui satisfont certaines propriétés, et qui ont des applications intéressantes en machine learning. Par exemple, on montrera un théorème qui nous permet de simplifier un problème de minimisation de risque empirique de dimension infini à un problème en dimension fini. Ou encore des applications dans plusieurs algorithmes de ML, les transformant d'algorithmes linéaires à non-linéaires à très faible prix, et d'autres encore.

Soit X un ensemble quelconque, H un espace de Hilbert de fonctions réelles sur X, muni de l'addition point par point ainsi que de la multiplication par scalaire point par point. On introduit aussi une forme linéaire qui à chaque fonction de H l'évalue en un certain point $x \in X$,

$$L_x : f \mapsto f(x) \; \forall f \in H$$

Définition 1 (RKHS)**.** On dit d'un espace de Hilbert qu'il est un RKHS si $\forall x \in X$, L_x est continue sur H, ou encore si L_x est bornée sur H, i.e

$$\forall x \in X, \; \exists M_x > 0, \; \forall f \in H \text{ t.q } |L_x(f)| \coloneqq |f(x)| \leq M_x ||f||_H$$

Dans ce qui suit, H sera un RKHS.

Propriété 1 (Convergence en norme dans un RKHS implique pointwise convergence)**.** Soient f_n, $f \in H$. Si $f_n \overset{H}{\to} f$, alors $\forall x \in X, \; f_n(x) \to f(x)$.

Démonstration.

$$\forall x \in X, \; |f_n(x) - f(x)| = |L_x(f_n) - L_x(f)| = |L_x(f_n - f)| \leq M_x ||f_n - f||_H$$

■

Définition 2 (Noyau / Kernel)**.** Une fonction $k : X \times X \to \mathbb{R}$ est un noyau si

$$\exists \phi : X \to H \text{ t.q } k(x, y) = \langle \phi(x), \phi(y) \rangle_H \; \forall x, y \in X$$

Propriété 2. Tout noyau k est symétrique défini positif.

Démonstration.
- Symétrie : découle de la symétrie du produit scalaire.
- Défini positif :
  $\forall x_1, \cdots, x_n \in X, \forall c_1, \cdots, c_n \in X$

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) = \left\langle \sum_{i=1}^{n} c_i \phi(x_i), \sum_{j=1}^{n} c_j \phi(x_j) \right\rangle_H = \left\| \sum_{i=1}^{n} c_i \phi(x_i) \right\|_H^2 \geq 0$$

■

Définition 3 (Noyau reproduisant / Reproducing kernel)**.** Une fonction $k : X \times X \to \mathbb{R}$ est un noyau reproduisant de H si $\forall x \in X$, $f \in H$:
- $k(\cdot, x) \in H$
- $f(x) = \langle f, k(\cdot, x) \rangle_H$

Propriété 3. Tout noyau reproduisant k est un noyau.

Démonstration. En prenant $f = k(\cdot, y) \in H$ pour un certain $y \in X$ dans la définition, on a en particulier $\forall x, y \in X\ k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_H$. Ici $\phi(u) = k(\cdot, u)\ \forall u \in X$. ■

Théorème 1 (Théorème de représentation de Riesz)**.** Soient :

- H un espace de Hilbert réel, muni de son produit scalaire $\langle \cdot, \cdot \rangle_H$
- $L \in H'$ une forme linéaire continue sur H.

Alors

$$\exists!\, g \in H,\ \forall f \in H,\ L(f) = \langle f, g\rangle_H$$

Propriété 4 (Existence et Unicité)**.** [mettre une hypothèse sur H] Il existe un unique noyau reproduisant de H.

Démonstration. Montrons l'existence puis l'unicité :

- On applique le théorème de Riesz à L_x :

  $$\forall x \in X,\ \exists!\, k_x \in H,\ \forall f \in H,\ f(x) = L_x(f) = \langle f, k_x\rangle_H$$

  Soit $k(y, x) \coloneqq k_x(y)\ \forall x, y \in X$, alors $\forall x \in X,\ f \in H,\ k(\cdot, x) \in H$ et $f(x) = \langle f, k(\cdot, x)\rangle_H$.
  Donc k est un noyau reproduisant de H.
- Soient k_1 et k_2 deux noyau reproduisant de H. Alors $\forall x \in X,\ f \in H$

  $$\langle f, k_1(\cdot, x) - k_2(\cdot, x)\rangle_H = f(x) - f(x) = 0$$

  En prenant $f = k_1(\cdot, x) - k_2(\cdot, x)$, on a :

  $$||k_1(\cdot, x) - k_2(\cdot, x)||_H^2 = 0$$

  C'est-à-dire que $k_1 = k_2$: le noyau reproduisant de H est unique.

■

Théorème 2 (Lien des deux visions)**.** H est un RKHS si et seulement si il existe un unique noyau reproduisant de H.

Démonstration.

- ($\Rightarrow$) Donné par la propriété 2. [ici c'est 4. On ne met jamais les numéros en dur (sinon un petit changement au début oblige à tout changer) : on utilise label / ref. A faire partout]
- ($\Leftarrow$) $\forall x \in X,\ f \in H$

  $$|L_x(f)| = |f(x)| = |\langle f, k(\cdot, x)\rangle_H| \leq ||f||_H ||k(\cdot, x)||_H = \underbrace{\sqrt{k(x, x)}}_{M_x} ||f||_H$$

  M_x ne dépendant pas de f, on a bien l'inégalité $\forall x \in X,\ \exists M_x > 0,\ \forall f \in H$, i.e H est un RKHS.

■

Théorème 3 (Théorème de Moore-Aronszajn)**.** Soit K un kernel. Alors il existe un unique espace de Hilbert H de fonctions sur X pour lequel K est un noyau reproduisant.

Démonstration. $\forall x \in X$, on pose $K_x(\cdot) \coloneqq K(x, \cdot)$. Soit H_0 le sous-espace vectoriel engendré par $\{K_x : x \in X\}$. Sous couvert d'existence de H, la reproducing property de K nous assure que $\forall x, y \in X,\ K(x,y) = \langle K_x, K_y \rangle_H$. On a de même :

$$\left\langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{j=1}^{m} b_j K_{x_j} \right\rangle_H = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(x_i, x_j)$$

Il est alors naturel de définir pour toutes fonctions de H_0 : $f \coloneqq \sum_{i=1}^{n} a_i K_{x_i}$ et $g \coloneqq \sum_{j=1}^{m} b_i K_{y_j}$

$$\langle f, g \rangle_{H_0} \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(x_i, x_j) = \sum_{i=1}^{n} a_i g(x_i) = \sum_{j=1}^{m} b_j f(y_j)$$

[pourquoi les deux dernières égalitées sont-elles importantes ? que disent-elles ?]

Déjà, $K_x \in H_0$, et comme $f(x) = \langle f, K_x \rangle_{H_0} < \infty$ [pas clair pourquoi cité ici], $\langle \cdot, \cdot \rangle_{H_0}$ est bien défini, billinéaire, est symétrique et positif, et on a la reproducing property : [pourquoi peux-tu utiliser Cauchy-Schwartz alors que tu ne sais pas encore que c'est un produit scalaire ?]

$$\forall x \in X,\ f \in H_0,\ \langle f, K_x \rangle_{H_0} = \sum_{i=1}^{n} a_i K_x(x_i) = \sum_{i=1}^{n} a_i K_{x_i}(x) = f(x)$$

Il reste à montrer que $||f||_{H_0} = 0 \Rightarrow f = 0$ pour que ca soit un produit scalaire :

$$|f(x)| = |\langle f, K_x \rangle_{H_0}| \leq ||f||_{H_0} ||K_x||_{H_0} = ||f||_{H_0} \sqrt{K(x,x)}$$

Donc $\langle \cdot, \cdot \rangle_{H_0}$ est un produit scalaire.

On complète H_0 avec le produit scalaire $\langle \cdot, \cdot \rangle_{H_0}$ en H, et toutes les propriétés sont gardées.

Unicité : Supposons G un autre espace de Hilbert pour lequel K est un noyau reproduisant. On a alors

$$\forall x, y \in X,\ \langle K_x, K_y \rangle_H = K(x,y) = \langle K_x, K_y \rangle_G$$

Par linéarité, $\langle \cdot, \cdot \rangle_H = \langle \cdot, \cdot \rangle_G$ sur tout H_0. **FINIR PREUVE UNICITE** ■

Tout compte fait, on a montré que : [pas clair du tout comme statement, mettre quelque chose qui soit rigoureux]

H est un RKHS $\Leftrightarrow \exists!\ k$ noyau reproduisant de $H \Leftrightarrow k$ est un noyau

2 Introduction : Réseau de neurone simple

Commencons par étudier un NN très simple : une fonction $\Phi : (\mathbb{R}^m \times \mathbb{R}^{m \times m} \times \mathbb{R}^m) \times \mathbb{R} \to \mathbb{R}$ combinaison d'applications linéaires, sans non linéarités intermédiaires :

$$\Phi((\beta, A, u), x) \coloneqq \frac{1}{m^\alpha} \beta^T \left(\frac{1}{m^\gamma} A \right) ux$$

On initialise $\theta^0 \coloneqq (\beta^0, A^0, u^0)$ de manière standarde : $\forall i, j \in \{1, \cdots, m\},\ u_i^0, A_{ij}^0, \beta_i^0 \sim_{iid} N(0,1)$

Nous montrerons quelques propriétés asymptotiques en la largeur des couches, et sur l'évolution des paramètres lors du premier pas de la descente de gradient.

2.1 Lois suivant la largeur des couches m

- Loi de $||u^0||_2^2$ pour m grand

  Comme $||u^0||_2^2 = \sum_{i=1}^m (u_i^0)^2 \sim \chi^2(m)$ et $u_i^0 \sim \chi^2(1)$, en appliquant le TCL aux u_i^0, on a :

  $$\frac{||u^0||_2^2 - m}{\sqrt{2m}} \sim_{m\to\infty} N(0,1)$$

  C'est-à-dire que $||u^0||_2^2 \sim N(m, 2m)$ pour m grand.

- Loi de $\left(\frac{1}{m^\gamma} A^0\right) u^0 x$ sachant u^0

  $(A^0 u^0)_i = \sum_{j=1}^m A_{ij}^0 u_j^0$. En sachant u^0, comme $A_{i\cdot}^0$ est un vecteur gaussien, $(A^0u^0)_i \sim N(0, ||u^0||_2^2)$.
  De même, part indépendance des A_{ij}^0, les $(A^0u^0)_i$ sont indépendants et $A^0u^0 \sim N(0_m, ||u^0||_2^2\ Id_m)$.
  Ainsi, $\left(\frac{1}{m^\gamma} A^0\right) u^0 x$ — $u^0 \sim N(0_m, \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\ Id_m)$.

- Loi de $\left(\frac{1}{m^\gamma} A^0\right) u^0 x$

  HISTOIRE DE MEME TRIBU ENGENDREE PAR U0 ET LE RESTE IMPLIQUE QUE CEST LA MEME CHOSE DECONDITIONNEE
  Donc $\left(\frac{1}{m^\gamma} A^0\right) u^0 x \sim N(0_m, \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\ Id_m)$.

- Loi de $|| \left(\frac{1}{m^\gamma} A^0\right) u^0 x ||_2^2$ pour m grand

  On a $\left(\left(\frac{1}{m^\gamma} A^0\right) u^0 x\right)_i^2 \sim \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2 \cdot \chi^2(1)$, d'espérance $\mu \coloneqq \left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2$ et de variance
  $\sigma^2 \coloneqq 2\left(\left(\frac{x}{m^\gamma}\right)^2 ||u^0||_2^2\right)^2$.
  Donc en appliquant le TCL à ceux ci, on a :

  $$\frac{|| \left(\frac{1}{m^\gamma} A^0\right) u^0 x ||_2^2 - m\mu}{\sigma\sqrt{m}} \sim_{m\to\infty} N(0,1)$$

  C'est-à-dire que $|| \left(\frac{1}{m^\gamma} A^0\right) u^0 x ||_2^2 \sim N(m\mu, m\sigma^2)$ pour m grand.

- [•] Loi de $\frac{1}{m^\alpha}(\beta^0)^T x_2$ sachant x_2, avec $x_2 = \left(\frac{1}{m^\gamma}A^0\right)u^0 x$

  On a $\frac{1}{m^\alpha}(\beta^0)^T x_2 | x_2 \sim N(0, \frac{1}{m^{2\alpha}}||x_2||_2^2)$

- [•] Loi de $\frac{1}{m^\alpha}(\beta^0)^T x_2$

  On a $\frac{1}{m^\alpha}(\beta^0)^T x_2 \sim N(0, \frac{1}{m^{2\alpha}}||x_2||_2^2)$

2.2 Choix de γ

On regarde la variance de chaque composante de $\left(\frac{1}{m^\gamma}A^0\right)u^0 x$ pour m grand : $Var = x^2 m^{-2\gamma} m$ comme $||u^0||_2^2$ se comporte en m pour m grand. Or on ne veut pas qu'elle tende vers 0 ou l' infini lorsque m tend vers l'infini car Φ prendrait des valeurs de 0 ou l'infini, ce qui impose le choix $\gamma = 1/2$

Dans la suite, on prendra $\gamma = 1/2$.

2.3 Gradients

Trivialement,

$$\nabla_\beta \Phi = \frac{x}{m^{\alpha+1/2}} Au$$

$$\nabla_u \Phi = \frac{x}{m^{\alpha+1/2}} A^T \beta$$

$$\nabla_A \Phi = \frac{x}{m^{\alpha+1/2}} \beta u^T$$

Etudions les ordres de grandeur des normes correspondantes à l'initialisation pour m grand :
On va simplement utiliser les approximations données par la loi des grands nombres : $||u^0|| \simeq \sqrt{m}$, et comme vu plus haut, $||A^0 u^0|| \simeq \sqrt{m}||u^0|| \simeq m$. Ainsi $||\nabla_\beta \Phi(\theta^0, x)|| \sim m^{-\alpha-1/2} \cdot m = m^{1/2-\alpha}$.

$$||\nabla_\beta \Phi(\theta^0, x)|| \sim m^{1/2-\alpha}$$

Exactement de la même manière, on aboutit à :

$$||\nabla_u \Phi(\theta^0, x)|| \sim m^{1/2-\alpha}$$

Pour A, on prend la norme de Frobenius : la LGN nous dit que $||\beta^0 (u^0)^T|| \simeq m$ et donc :

$$||\nabla_A \Phi(\theta^0, x)||_F \sim m^{1/2-\alpha}$$

2.4 Descente de gradient

On va étudier ici le premier pas de descente de gradient.
Posons une fonction de perte $F : \mathbb{R} \to \mathbb{R}$ t.q $F'(0) \neq 0$ et $\Delta F := F(\Phi(\theta^1, x)) - F(\Phi(\theta^0, x))$, avec $\theta^1 := \theta^0 - \eta \nabla_\theta F(\Phi(\theta^0, x))$
Il semble honnête de prendre η dépendant de m, le produit scalaire final ayant plus de chance d'exploser en grande dimension. Prenons $\eta := m^a$, $a \in \mathbb{R}$

2.4.1 Choix de η

On veut que ΔF ne diverge pas ni ne tende vers 0 lorsque m tend vers l'infini.
Pour cela, on utilise l'approximation $\Delta F \simeq < \Delta\theta, \nabla_\theta F(\Phi(\theta^0, x)) >$.
On a

$$\Delta F \simeq < -\eta \nabla_\theta F(\Phi(\theta^0, x)), \nabla_\theta F(\Phi(\theta^0, x)) >= -\eta ||\nabla_\theta F(\Phi(\theta^0, x))||^2$$

$$\nabla_\theta F(\Phi(\theta^0, x)) = \underbrace{F'(\Phi(\theta^0, x))}_{\text{constante en } m} \cdot \nabla_\theta \Phi(\theta^0, x)$$

Or $||\nabla_\theta \Phi(\theta^0, x)||^2 = ||\nabla_\beta \Phi(\theta^0, x)||^2 + ||\nabla_u \Phi(\theta^0, x)||^2 + ||\nabla_A \Phi(\theta^0, x)||^2$

Donc pour m grand :

$$\begin{aligned}
\Delta F &\simeq -\eta ||\nabla_\theta F(\Phi(\theta^0, x))||^2 \\
&\sim \eta ||\nabla_\theta \Phi(\theta^0, x)||^2 \\
&\simeq \eta (3 \cdot (m^{1/2-\alpha})^2) \\
&\simeq m^a \cdot m^{1-2\alpha}
\end{aligned}$$

Ce qui impose le choix $a = 2\alpha - 1$

2.4.2 Ordres de grandeur des écarts relatifs

Pour cela introduisons $\Delta\theta \coloneqq \theta^1 - \theta^0$. Remarquons que $\Delta\theta = -\eta \nabla_\theta F(\Phi(\theta^0, x))$.

$$\begin{aligned}
||\Delta u|| &\sim \eta ||\nabla_u \Phi(\theta^0, x)|| &(1)\\
&\sim m^{2\alpha-1} \cdot m^{1/2-\alpha} &(2)\\
&\sim m^{\alpha-1/2} &(3)
\end{aligned}$$

Ce qui nous donne un ordre de grandeur de l'écart relatif :

$$\frac{||\Delta u||}{||u^0||} \sim m^{\alpha-1}$$

On a le même résultat pour l'écart relatif de β^0 :

$$\frac{||\Delta \beta||}{||\beta^0||} \sim m^{\alpha-1}$$

Pour A, la LGN nous donne $||A^0||_F \simeq m$ pour m grand, on a alors par les mêmes calculs :

$$\frac{||\Delta A||_F}{||A^0||_F} \sim m^{\alpha-3/2}$$

Concernant l'écart entrywise de A, on a $|\Delta A_{ij}| \sim \eta |(\nabla_A \Phi(\theta^0, x))_{ij}| \sim m^{2\alpha-1} \cdot m^{-1/2-\alpha} \cdot 1 \sim m^{\alpha-3/2}$ car $|\beta^0 (u^0)^T| \sim 1$. $|A_{ij}| \sim 1$, donc :

$$\frac{|\Delta A_{ij}|}{|A_{ij}|} \sim m^{\alpha-3/2}$$

Maintenant avec la norme opétateur : le corollaire 7.9 du cours de MIA2 nous donne une majoration sur $|A^0|_{op}$: $|A^0|_{op} \leq \sqrt{m} + 7\sqrt{m+\xi} \sim \sqrt{m}$, avec $\xi \sim Exp(1)$.

De plus, on peut trouver la la norme opérateur de ΔA comme suit : tout le travail est de trouver $|\beta^0(u^0)^T|_{op} \coloneqq \sup\{||\beta^0(u^0)^T x|| \text{ avec } ||x|| = 1\}$.

$$(\beta^0(u^0)^T x)_{ij} = \beta_i^0 u_j^0 \tag{4}$$

$$(\beta^0(u^0)^T x)_i = \sum_{j=1}^{m} (\beta^0(u^0)^T x)_{ij} \cdot x_j \tag{5}$$

$$= \beta_i^0 < u^0, x > \tag{6}$$

$$||\beta^0(u^0)^T x|| = | < u^0, x > | \cdot ||\beta^0|| \tag{7}$$

Donc le sup est bien atteint en $x = u/||u^0||$ et est égal à $||u^0|| \cdot ||\beta^0||$. En utilisant les approximations précédentes, on a donc $|\beta^0(u^0)^T|_{op} \simeq m$. Ainsi on a $|\Delta A|_{op} \sim m^{2\alpha-1} \cdot m^{-1/2-\alpha} \cdot m \sim m^{\alpha-1/2}$. et :

$$\frac{|\Delta A|_{op}}{|A^0|_{op}} \sim m^{\alpha-1}$$

2.4.3 Choix de α

- $\alpha < 1$

  Dans ce cas là, tous les écarts relatifs d'ordre $m^{\alpha-1} \xrightarrow[m\to\infty]{} 0$. On a donc pour m grand :

  $$||\Delta u|| \ll ||u^0||$$

  $$||\Delta \beta|| \ll ||\beta^0||$$

  $$|\Delta A|_{op} \ll |A^0|_{op}$$

  Regardons maintenant les Δ des gradients en u pour la première itération :

$$\Delta \nabla_u \Phi \coloneqq \nabla_u \Phi(\theta^1, x) - \nabla_u \Phi(\theta^0, x) \tag{8}$$

$$\sim m^{-\alpha-1/2} \underbrace{((\beta^1)^T A^1 - (\beta^0)^T A^0)}_{(\star)} \tag{9}$$

$$(\star) = (\beta^0 - \eta \nabla_\beta F(\Phi(\theta^0, x)))^T (A^0 - \eta \nabla_A F(\Phi(\theta^0, x))) - (\beta^0)^T A^0 \tag{10}$$

$$= \eta^2 [\nabla_\beta F(\Phi)]^T [\nabla_A F(\Phi)] - \eta [\nabla_\beta F(\Phi)] A^0 - \eta \beta^0 [\nabla_A F(\Phi)] \tag{11}$$

$$= (\Delta\beta)^T (\Delta A) - (\Delta\beta) A^0 - \beta^0 (\Delta A) \tag{12}$$

  Donc :

$$
\begin{aligned}
||\Delta\nabla_u\Phi|| &\sim m^{-\alpha-1/2}||(\Delta\beta)^T(\Delta A) - (\Delta\beta)A^0 - \beta^0(\Delta A)|| && (13)\\
&\lesssim m^{-\alpha-1/2}(||\Delta\beta|| \cdot |\Delta A|_{op} + ||\Delta\beta|| \cdot |A^0|_{op} + ||\beta^0|| \cdot |\Delta A|_{op}) && (14)\\
&\lesssim m^{-\alpha-1/2}(m^{\alpha-1/2}m^{\alpha-1/2} + m^{\alpha-1/2}m^{1/2} + m^{1/2}m^{\alpha-1/2}) && (15)\\
&\lesssim m^{-\alpha-1/2}(m^{\alpha-1/2}m^{1/2} + m^{\alpha-1/2}m^{1/2} + m^{1/2}m^{\alpha-1/2}) && (16)\\
&\lesssim m^{-1/2} && (17)
\end{aligned}
$$

C'est-à-dire que $||\Delta\nabla_u\Phi|| = \mathcal{O}(m^{-1/2})$ pour m grand.

Par la même démarche on trouve $||\Delta\nabla_\beta\Phi|| = \mathcal{O}(m^{-1/2})$ et $||\Delta\nabla_A\Phi||_F = \mathcal{O}(m^{-1/2})$ pour m grand.

Ainsi :

$$||\Delta\nabla_{\beta/u/A}\Phi|| = \mathcal{O}(m^{-1/2}) \overset{\alpha<1}{\ll} m^{1/2-\alpha} \sim ||\nabla_{\beta/u/A}\Phi||$$

Cela traduit un comportement linéaire lorsque $\alpha < 1$ pour m grand.

On peut aussi remarquer que

$$\Delta\nabla_\theta F(\Phi) = F'(\Phi) \cdot \Delta\nabla_\theta\Phi$$

On a aussi :

$$||\Delta\nabla_{\beta/u/A}F(\Phi)|| = \mathcal{O}(m^{-1/2}) \overset{\alpha<1}{\ll} m^{1/2-\alpha} \sim ||\nabla_{\beta/u/A}F(\Phi)||$$