# Introduction au domaine de recherche

## Apprentissage statistique : méthodes RKHS

Pierre Wolinski

29 octobre 2015

Diplôme de l'ENS

ENS Paris

# Table des matières



# 1 Introduction

Parmi les problèmes étudiés en apprentissage statistique, la régression fait toujours l'objet de nombreuses publications. Le cadre est le suivant : on dispose d'une suite de couples d'observations $\mathcal{D}_n = ((X_1, Y_1), \cdots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ identiquement distibués, copies du couple de variables aléatoires $(X, Y)$ de loi P, et on se donne une fonction de perte $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ (généralement, $\forall y \in \mathcal{Y}, L(y, y) = 0$ et $L(x, y) = l(y - x)$, où $l$ est convexe), ainsi qu'un ensemble $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$. On cherche alors une fonction $f \in \mathcal{G}$ minimisant le risque :

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X)).$$

En pratique, la formulation explicite de $\mathcal{R}_{L,\mathrm{P}}$ nécessite de connaître P, ce qui n'est jamais le cas, donc on se contente de trouver un estimateur $\hat{f}_{\mathcal{D}_n}$ de $f$ minimisant le risque empirique :

$$\hat{\mathcal{R}}_{L,\mathrm{P}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

Malheureusement, ce risque peut être facilement minimisé par une fonction bien adaptée aux observations, mais très sensible à l'ajout d'autres observations. On appelle ce phénomène le *surapprentissage*. Pour éviter cela, on cherche plutôt à minimiser le risque régularisé :

$$\mathcal{R}^{\mathrm{reg}}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X)) + \mathrm{pen}(f),$$

où la pénalité pen : $\mathcal{G} \to \mathbb{R}$ est souvent proportionnelle à une norme sur $\mathcal{G}$, avec un coefficient de proportionnalité noté $\lambda$.

Les méthodes RKHS (*Reproducing Kernel Hilbert Space*) s'inscrivent notamment dans ce cadre. À la différence d'autres méthodes utilisées en régression, celles-ci sont non paramétriques et nécessitent peu d'hypothèses sur les espaces $\mathcal{X}$ et $\mathcal{Y}$. Le premier cadre de régression dans lequel les RKHS ont montré leur intérêt est le suivant : on cherche à estimer une fonction d'un ensemble $\mathcal{X}$ mesuré dans $\mathbb{R}$ à partir d'une suite d'observations i.i.d. $\mathcal{D}_n = ((X_1, Y_1), \cdots, (X_n, Y_n)) \in (\mathcal{X} \times \mathbb{R})^n$, minimisant un risque $\mathcal{R}$. L'absence de structure hilbertienne sur $\mathcal{X}$ rend impossible les modélisations habituelles, telles que le modèle linéaire gaussien. En revanche, il est plus aisé de manipuler un espace fonctionnel $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ déjà muni d'une structure d'espace vectoriel. Si de plus $\mathcal{H}$ est hilbertien et vérifie une propriété détaillée plus loin, on le qualifie de RKHS. On remplace ainsi la difficulté liée à l'absence de structure hilbertienne sur $\mathcal{X}$ par l'usage d'un espace fonctionnel. La jonction entre $\mathcal{X}$ et $\mathcal{H}$ est assurée par le noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. L'estimateur a alors la forme suivante :

$$\forall x \in \mathcal{X}, \quad \hat{f}_n(x) = \sum_{i=1}^{n} c_i(\mathcal{D}_n) k(x, X_i),$$

où les $c_i(\mathcal{D}_n)$ sont des coefficients réels dépendant uniquement de l'ensemble des observations $\mathcal{D}_n$. Il est souvent important de préciser si $\mathcal{H}$ est dense dans $\mathcal{C}(\mathcal{X}, \mathbb{R})$ pour savoir si l'on peut approcher autant que l'on veut une fonction cible continue. Si c'est le cas, on dit que le noyau est *universel*.

On peut se référer au livre de Wahba ([14], 1990) pour l'utilisation de RKHS dans l'étude des *spline models*, qui a inspiré leur usage en apprentissage statistique, et plus particulièrement en régression non paramétrique.

L'approche théorique des RKHS qui nous intéressera ici est celle qui fut développée par De Vito et al. ([3], 2004). Purement fonctionnelle, elle permit de trouver une formulation de $f = \arg\min_g \mathcal{R}(g)$ ne faisant pas intervenir la perte, et de démontrer la consistance des méthodes RKHS en régression, pour une perte *convexe* et pour des fonctions cibles à valeur scalaire (Christmann et al. [12], 2007).

En parallèle, l'estimation de fonctions cibles à valeur vectorielle a conduit à définir les RKHS plus largement, pour aboutir aux RKHS à noyau opérateur (Evgeniou et al. [5], 2005 ; Carmelli et al. [1], 2006). On nomme $\mathcal{Y}$ l'espace d'arrivée des fonctions cibles. Les RKHS à noyau scalaire pourraient suffire si l'on veut effectuer une régression coordonnée par coordonnée, mais, dans la mesure où les coordonnées d'un vecteur de $\mathcal{Y}$ sont susceptibles d'être corrélées, il est légitime de construire un estimateur à valeur dans $\mathcal{Y}$. Sur le plan théorique, les efforts se sont concentrés sur la notion d'universalité du noyau (Carmelli et al. [2], 2010), et uniquement dans le cas d'un risque quadratique avec pénalité quadratique.

Dans un premier temps, nous rappellerons quelques résultats concernant les RKHS à noyau scalaire et les RKHS à noyau opérateur [7] [1].

Dans un deuxième temps, nous donnerons le schéma de la preuve de la consistance dans un cadre i.i.d. pour une perte convexe, avec un RKHS à noyau scalaire [12].

Dans un troisième temps, nous donnerons une extension de ce résultat un cas où les observations sont non i.i.d.

# 2 Définitions et propriétés des RKHS

## 2.1 Les RKHS à noyau scalaire

Pour comprendre comment fonctionnent les méthodes RKHS, on commence par étudier un cas simple. On cherche à estimer une fonction $f : \mathcal{X} \to \mathbb{R}$, où $\mathcal{X}$ est un ensemble. Si $\mathbb{R}$ est muni d'une structure euclidienne, $\mathcal{X}$ n'est pourvu d'aucune structure. On tente donc de créer artificiellement une structure pour $\mathcal{X}$, en exploitant l'espace vectoriel réel $\mathcal{F}(\mathcal{X}, \mathbb{R})$, qui émerge naturellement du problème initial.

L'idée est de construire une sorte de produit scalaire sur $\mathcal{X}$, que l'on appellera *noyau*. Si l'on choisit un sous-espace vectoriel $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ hilbertien réel, alors on dispose d'un véritable produit scalaire à partir duquel on peut définir le noyau. Sur le plan formel, tout se passe comme si l'on effectuait les calculs dans un « espace de travail » $\mathcal{H}$, avec un noyau servant de pont entre $\mathcal{H}$ et $\mathcal{X}$.

**Cadre général.**

- $\mathcal{X}$ est un ensemble ;
- $\mathcal{H}$ est un sous-espace vectoriel inclus dans $\mathcal{F}(\mathcal{X}, \mathbb{R})$, l'espace des applications de $\mathcal{X}$ dans $\mathbb{R}$,
  on suppose que $\mathcal{H}$ est un espace de Hilbert muni du produit scalaire $\langle \cdot, \cdot \rangle$.

**Définition 1.** *On dit que $\mathcal{H}$ est un* RKHS *si, et seulement si :*

$$\forall x \in \mathcal{X}, \quad f \mapsto f(x) \quad \textit{est continue sur } \mathcal{H}.$$

Dans le cas où $\mathcal{H}$ est effectivement un RKHS, le théorème de Riesz assure que :

$$\forall y \in \mathcal{X}, \exists k_y \in \mathcal{H} : \forall f \in \mathcal{H}, \quad f(y) = \langle k_y, f \rangle .$$

On note alors $k$ le *noyau* défini par :

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad k : (x, y) \mapsto k_y(x).$$

On peut ainsi formuler la propriété reproduisante :

$$\forall f \in \mathcal{H}, x \in \mathcal{X}, \quad f(x) = \langle k(\cdot, x), f \rangle .$$

**Théorème 1.** *On suppose que $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une fonction telle que :*

$$\forall n \geq 2, \forall (x_1, \cdots, x_n) \in \mathcal{X}^n, \quad (k(x_i, y_i))_{1 \leq i,j \leq n} \in \mathcal{S}_n^+(\mathbb{R}),$$

*où $\mathcal{S}_n^+(\mathbb{R})$ est l'ensemble des matrices symétriques réelles de rang $n$ dont les valeurs propres sont dans $\mathbb{R}^+$.*
*Alors il existe un unique RKHS $\mathcal{H}$ (à une isométrie près) admettant $k$ comme noyau reproduisant.*

Ce théorème signifie que le choix initial du noyau détermine entièrement l'espace de travail $\mathcal{H}$. On peut donc se contenter de choisir le noyau sans se préoccuper du RKHS lui-même.

Par ailleurs, $\mathcal{X}$ peut être lui-même un espace hilbertien réel, mais cela n'invalide pas la démarche proposée ici : la structure hilbertienne de $\mathcal{X}$ n'est pas nécessairement adaptée au traitement de données que l'on souhaite effectuer. Dans ce cas, on constate que le choix du noyau est primordial pour passer d'une structure « inadaptée » à une structure « adaptée ». Il faut donc dans tous les cas sélectionner soigneusement le noyau en fonction du jeu de données à traiter.

**Exemple 1.** *Noyau gaussien :*

$$k_G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad k_G : (x, x') \mapsto \exp\left(-\gamma \|x - x'\|\right),$$

*où* $\gamma > 0$.

Il est tout à fait possible d'étendre ce cadre à l'estimation de fonctions $f : \mathcal{X} \to \mathcal{Y}$, où $\mathcal{Y}$ est un espace hilbertien réel. Dans ce cas, on a $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ et $k : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$. Les résultats sont plus techniques, mais le fond est pratiquement identique.

**Exemple 2.** *Noyau gaussien opérateur :*

$$K_G : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{L}(\mathbb{R}^D) = \mathcal{M}_D(\mathbb{R}), \quad K_G : (x, x') \mapsto (k_G(x_i, x'_j))_{1 \leq i,j \leq D},$$

## 2.2 Utilisation des RKHS

On détaille ici quelques outils spécifiques aux méthodes RKHS. Par exemple, l'*universalité* est une notion permettant de juger la qualité de noyaux.

**Définition 2.** *On suppose que $\mathcal{X} \subset \mathbb{R}^d$. On dit que le noyau continu $k$ est* universel *si, et seulement si, son RKHS $\mathcal{H}$ est dense dans $(\mathcal{C}(\mathcal{X}, \mathbb{R}), \|\cdot\|_\infty)$, l'espace des fonctions continues de $\mathcal{X}$ dans $\mathbb{R}$ muni de la norme $\|\cdot\|_\infty$.*

On suppose dorénavant que $\mathcal{X}$ est un compact de $\mathbb{R}^d$. Cela réduit le champ d'étude, mais, au vu des démonstrations de consistance existantes, la compacité de $\mathcal{X}$ est difficilement négociable. En revanche, on pourrait se contenter de choisir $\mathcal{X}$ mesuré, métrique et compact.

On se place dans le cas où l'on dispose d'une suite d'observations de $\mathcal{X} \times \mathbb{R}$ i.i.d. de loi P : $\mathcal{D}_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$. On note $\mathrm{D}_n$ la distribution empirique associée et $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ une fonction de perte convexe. Pour définir la consistance, on rappelle les définitions suivantes :

- Risque non régularisé :

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L\left(Y, f(X)\right);$$

- Risque non régularisé optimal :

$$\mathcal{R}^\star_{L,\mathrm{P}} = \min_{f \in \tilde{\mathcal{F}}} \mathcal{R}_{L,\mathrm{P}}(f),$$

où $\tilde{\mathcal{F}} = \{f \in \mathcal{F}(\mathcal{X}, \mathbb{R}) : f \text{ mesurable}\}$ ;

- Risque régularisé :

$$\mathcal{R}_{L,\mathrm{P},\lambda}(f) = \mathbb{E}_\mathrm{P} L\left(Y, f(X)\right) + \lambda \left\|f\right\|^2$$

et son minimiseur :

$$f_{\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P},\lambda}(f).$$

Celui-ci est bien défini et unique, car $L$ est convexe, donc $\mathcal{R}_{L,\mathrm{P},\lambda}$ est fortement convexe ;

- Estimateur :

$$\hat{f}_{n,\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{D}_n,\lambda}(f).$$

Comme $\mathrm{D}_n$ est la distrubution empirique, $\hat{f}_{n,\lambda}$ est bien un minimiseur du risque empirique appartenant à $\mathcal{H}$.

**Définition 3.** *On dit que l'estimateur $\hat{f}_{n,\lambda_n}$ est* $L$-consistant *si, et seulement si :*

$$\mathcal{R}_{L,\mathrm{P}}(\hat{f}_{n,\lambda_n}) \xrightarrow{\mathrm{P}} \mathcal{R}^\star_{L,\mathrm{P}}.$$

Dans la pratique, les méthodes RKHS demandent de nombreux réglages. On a vu précédemment que le noyau devait être sélectionné en fonction du problème posé (d'Alché-Buc et al. [6]), mais la suite des coefficients de régularisation $(\lambda_n)_n$ doit vérifier certaines propriétés détaillées plus loin pour que l'estimateur soit consistant. En ce qui concerne les coefficients $c_i(\mathcal{D}_n)$ [6], il faudra la plupart du temps utiliser un algorithme pour les trouver, les méthodes analytiques ne fonctionnant que dans des cas bien particuliers.

# 3 Consistance dans le cas scalaire i.i.d. avec une perte convexe

## 3.1 Schéma de la preuve

La preuve de la consistance présentée ici est inspirée de celle de Christmann et al. [12].

**Notations.** On note :

- $|h|_1$ la constante de Lipschitz de $h$, si $h$ est lipschitzienne ;
- $|\mu|_a = \int a(y)\,\mathrm{d}\mu(y)$, où $\mu$ est une mesure sur $\mathbb{R}$, et $a : \mathbb{R} \to \mathbb{R}^+$,
  $|\mathrm{P}|_p = \int |y|^p \,\mathrm{dP}(y)$, où $p > 0$.

**Cadre.** On pose :

- $\mathcal{X} \subset \mathbb{R}^d$ compact ;
- P loi sur $\mathcal{X} \times \mathbb{R} = \mathcal{Z}$ ;
- $\mathcal{H}$ un RKHS de noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ;
- $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ une fonction de perte ;
- on cherchera à minimiser les risques suivants sur $\mathcal{H}$ :

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X)),$$
$$\mathcal{R}_{L,\mathrm{P},\lambda}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X)) + \lambda \|f\|^2 .$$

Le premier problème auquel on est confronté est la définition implicite de la fonction $f_{\mathrm{P},\lambda}$ minimisant le risque $\mathcal{R}_{L,\mathrm{P},\lambda}$, avec pour seule information la convexité de la perte $L$. L'idée est de trouver une formulation explicite de $f_{\mathrm{P},\lambda}$ pour ensuite effectuer les calculs nécessaires à la preuve de la consistance. On y arrive en exploitant un résultat fonctionnel démontré par De Vito et al. [3].

Le second problème est la démonstration de la consistance elle-même. Il faudrait pouvoir comparer la fonction $\mathcal{R}_{L,\mathrm{P},\lambda}$ avec son estimateur $\hat{f}_{n,\lambda}$. Cela est rendu possible par l'utilisation du résultat fonctionnel en remplaçant P par $\mathrm{D}_n$, la distribution empirique. On arrive finalement au résultat en utilisant une inégalité de concentration.

Avant toute chose, on donne quelques définitions utiles dans la suite, et on précise quelques hypothèses.

**Définition 4.** *Soient $p \in [1, +\infty[$ et $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$.*
*On dit que $L$ est une* perte de type $(a, p)$ *par rapport à* P *si, et seulement si :*

1. $\forall y \in \mathbb{R}$, $L(y, \cdot)$ *est convexe sur* $\mathbb{R}$ ;
2. $L$ *est mesurable sur* $\mathbb{R} \times \mathbb{R}$ ;
3. $\exists b \in \mathbb{R}^+$ *et* $\exists a : \mathbb{R} \to \mathbb{R}^+$ *tels que :*

$$\forall (y, w) \in \mathbb{R} \times \mathbb{R}, \quad L(y, w) \leq a(y) + b\,|w|^p ,$$
$$|\mathrm{P}|_a = \mathbb{E}_{\mathrm{P}}\left[a(Y)\right] < \infty.$$

**Définition 5.** *On dit que $L$ est* invariante *si, et seulement si, $\exists l : \mathbb{R} \to \mathbb{R}^+$ telle que $\forall (y, t) \in \mathbb{R} \times \mathbb{R}, L(y, t) = l(y - t)$.*
*On note alors $\forall r \in \mathbb{R}^+, V(r) = \left|l_{|\mathcal{B}(r)}\right|_1$, où $\mathcal{B}(r)$ est la boule ouverte centrée en 0 de rayon $r$.*

**Hypothèses.** On suppose :

- $L$ est une perte convexe par rapport à la deuxième variable, invariante et de type $(a, p)$ par rapport à P.
  On en déduit que $L$ est continue ;

- $p'$ tel que $\frac{1}{p} + \frac{1}{p'} = 1$ ;
- $k$ est un noyau universel,
  $k$ est borné, on note :

$$\|k\|_\infty = \sup\left\{\sqrt{\|\|k(x,x)\|\|} : x \in \mathcal{X}\right\} < \infty.$$

Il est clair que la perte $L$ doit vérifier certaines conditions dépendant de P. Si ce n'était pas le cas, il serait impossible de contrôler les expressions faisant intervenir $L$, qui plus est lorsqu'on les intègre par rapport à P. Son invariance n'est pas strictement nécessaire, mais elle simplifie les calculs. De plus, il est légitime de supposer que la perte est indépendante de la position dans $\mathbb{R}$ des vecteurs considérés.

## 3.2 Un résultat fonctionnel

Comme le risque $\mathcal{R}_{L,\mathrm{P},\lambda}$ est fortement convexe, on en déduit que son unique minimum est atteint lorsque son gradient, s'il existe, s'annule. En partant de là, on peut arriver à exprimer son minimiseur. En fait, l'existence du gradient n'est pas nécessaire. Pour une fonction convexe, la sous-différentielle suffit à mener à bien la démonstration.

**Proposition 1.** *On suppose que $L$ est de type $(a,p)$ par rapport à* P *et que $|\mathrm{P}|_a < \infty$. Soit $\lambda > 0$.*
*Alors il existe un unique minimiseur $f_{\mathrm{P},\lambda}$ de $\mathcal{R}_{L,\mathrm{P},\lambda}$ et $\|f_{\mathrm{P},\lambda}\| \leq \sqrt{\mathcal{R}_{L,\mathrm{P}}(0)/\lambda} = \delta_{\mathrm{P},\lambda}$.*
(cf. [12], prop. 8)

**Définition 6.** *On rappelle que la* sous-différentielle *d'une fonction $F : \mathbb{R} \to \mathbb{R}$ convexe est définie par :*

$$\forall w \in \mathbb{R}, (\partial F)(w) = \{w^* \in \mathbb{R} : \forall v \in \mathbb{R}, w^*(v-w) \leq F(v) - F(w)\}.$$

**Théorème 2.** *Soient $\lambda > 0$ et $f^\lambda \in \mathcal{H}$. Alors :*

$$f^\lambda \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \{\mathcal{R}_{L,\mathrm{P},\lambda}(f)\}$$

*si, et seulement si, il existe $h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}, h \in \mathcal{L}^{p'}(\mathcal{Z}, \mathrm{P})$ tel que :*

$$\begin{aligned} & & h(x,y) &\in (\partial_2 L)(y, f^\lambda(x)) \quad p.s. \\ \forall s \in \mathcal{X}, & & f^\lambda(s) &= -\frac{1}{2\lambda}\mathbb{E}_\mathrm{P}\left[k(s,X)h(X,Y)\right], \end{aligned}$$

*où $\partial_2$ est la sous-différentielle par rapport à la deuxième variable.*
*On garde cette définition de $h$ dans la suite.*
(cf. [12], th. 10)

### 3.3 Démonstration de la consistance

En raison de l'expression fonctionnelle du minimiseur $f_{\mathrm{P},\lambda}$ du risque $\mathcal{R}_{L,\mathrm{P},\lambda}$, l'inégalité de concentration dont on a besoin doit pouvoir s'appliquer sur un espace fonctionnel, ici $\mathcal{H}$. L'inégalité démontrée par Christmann et al. [12] est utilisable sur les espaces hilbertiens, ce qui convient parfaitement.

**Lemme 1.** *Soit $g : \mathcal{Z} \to \mathcal{H}$ une fonction mesurable avec $\|g\|_q = (\mathbb{E}_\mathrm{P}\|g\|^q)^{1/q}$, où $q \in ]1, +\infty]$.*
*On note $q^* = \min\{1/2, 1/q'\}$, où $q'$ est tel que : $\frac{1}{q} + \frac{1}{q'} = 1$.*
*Il existe une constante $c_q > 0$ telle que, pour tout $\epsilon > 0, n \geq 1$ :*

$$\mathbb{P}_{\mathrm{P}^{\otimes n}}\left((z_1, \cdots, z_n) \in \mathcal{Z}^n : \left\|\frac{1}{n}\sum_{i=1}^{n} g(z_i) - \mathbb{E}_\mathrm{P} g\right\| \geq \epsilon\right) \leq \left(\frac{\|g\|_q}{\epsilon n^{q^*}}\right)^q.$$

(cf. [12], Lemme 21)

**Théorème 3.** *On suppose que $l$ est d'*ordre inférieur et supérieur *$p$, c'est-à-dire :*

$$\exists (c_-, c_+) \in \mathbb{R}^+ \times \mathbb{R}^+ : \forall y \in \mathbb{R}, \quad c_-(|y|^p - 1) \leq l(y) \leq c_+(|y|^p + 1).$$

*Soit $(\lambda_n)_n$ une suite de $\mathbb{R}^{+*}$ telle que, en posant $p^* = \max\{2p, p^2\}$, on ait :*

$$\lambda_n \to 0, \quad \lambda_n^{p^*} n \to \infty. \tag{1}$$

*Alors la suite d'estimateurs $(\hat{f}_{n,\lambda_n})_n$ est convergente pour les lois* P *telles que $|\mathrm{P}|_p < \infty$.*
(cf. [12], th. 12)

Maintenant la preuve de la consistance terminée, on entrevoit un avantage bien particulier des formulations fonctionnelles. On a réussi à circoncire l'usage de l'indépendance des variables aléatoires à l'inégalité de concentration. Cela signifie que, pour étendre le résultat de consistance à un type de processus donné, il suffit de démontrer une telle inégalité pour celui-ci, le reste de la preuve ne nécessitant aucun changement. La seule contrainte imposée par la preuve sur le processus est sa *stationnarité*. En effet, on a besoin d'une mesure P invariante pour utliser le théorème 2.

## 4 Consistance dans un cas non i.i.d.

### 4.1 Présentation du problème

On considère le système stochastique $(X_n, Y_n)$ à valeurs dans $\mathcal{X} \times \mathbb{R}$ défini par :

$$\begin{aligned} Y_n &= f(X_n, \epsilon_n) \\ X_{n+1} &= T^\star(X_n, \delta_{n+1}) \end{aligned} \tag{2}$$

où $\epsilon_n$ et $\delta_n$ correspondent respectivement à du bruit d'observation et du bruit de dynamique, et $T^\star$ est le noyau de transition. On suppose que $(X_n)_n$ est stationnaire ergodique et que $(\epsilon_n)_n$ et $(\delta_n)_n$ sont i.i.d. On note :

$$\mathrm{P} = \mu \otimes \mathrm{P}_\delta$$

où $\mu$ est la mesure stationnaire de la chaîne de Markov $(X_n)_n$ et $\mathrm{P}_\delta$ est la distribution des $\delta_n$.

On souhaite construire un estimateur de $T^\star$ en fonction de la partie dynamique $(X_n)_n$, on ne s'occupera pas de la partie d'observation $(Y_n)_n$. On recherche donc la fonction $\hat{T}$ minimisant le risque sur un espace hilbertien $\mathcal{H}$ :

$$\mathcal{R}_{L,\mathrm{P},\lambda}(T) = \mathbb{E}_\mathrm{P} L\left[T^\star(X,\delta), T(X)\right] + \lambda \left\|T\right\|^2$$

où $L$ est une fonction de perte convexe et $\|\cdot\|$ est la norme hilbertienne de $\mathcal{H}$.

Pour estimer $T^\star$ en utilisant les RKHS, il faut généraliser les résultats de consistance existants dans un cadre non i.i.d. Nous préciserons les hypothèses que nous faisons sur le système dynamique. La stationnarité et l'ergodicité de la chaîne $(X_n)_n$ sont des hypothèses difficiles à exploiter dans l'adaptation de la preuve de consistance. Nous nous limiterons donc à une classe de systèmes dynamiques : les *causal Bernoulli shifts* [10], que nous définissons dans le paragraphe suivant.

## 4.2 Un premier résultat de consistance

Le choix du cas particulier non i.i.d. s'est porté sur le *causal Bernoulli shift*, une famille de processus stationnaires née d'un cadre théorique très général (Shields, [11], 1973). Ceux-ci sont assez généraux, englobant notamment les processus autorégressifs linéaires, et leur formulation est adaptée à un cadre fonctionnel. En revanche, la contrainte lipschitzienne qu'on leur impose au départ amène rapidement à une limitation du cadre d'étude (hypothèses du type Hölder ou Lipschitz sur la fonction cible, la perte, etc.).

L'objectif de cette partie est principalement d'établir un résultat de consistance pour l'estimation de la fonction $f$ d'observation du système dynamique (2). On verra enfin le cas de l'estimation de la fonction de transition $T^\star$.

**Définition 7.** *On dit que la fonction* $\eta : \mathbb{R}^\mathbb{Z} \to \mathcal{X}$ *est* $(\alpha_s)_s$-lipschitzienne *si, et seulement si :*

$$\forall t \in \mathbb{Z}, \forall (\xi_s)_s, (\tilde{\xi}_s)_s, \quad \left\|\eta((\xi_s)_{s\leq t}) - \eta((\tilde{\xi}_s)_{s\leq t})\right\| \leq \sum_{s\geq 0} \alpha_s \left|\xi_{t-s} - \tilde{\xi}_{t-s}\right|.$$

**Cadre.** Soit $(\xi_t)_t$ une suite de variables aléatoires i.i.d. On pose :

- $X_t = \phi((\xi_s)_{s\leq t}) \in \mathcal{X}$, où $\phi$ est $(\alpha_s)_s$-lipschitzienne ;
- $Y_t = f(X_t) + \epsilon_t \in \mathbb{R}$, où $(\epsilon_t)_t$ sont des variables aléatoires centrées i.i.d. indépendantes de $(\xi_t)_t$ ;
- $Z_t = k(\cdot, X_t)h(X_t, Y_t) - \mathbb{E}_\mathrm{P}(k(\cdot, X)h(X,Y)) = \psi(\epsilon_t, (\xi_s)_{s\leq t}) \in \mathcal{H}$.

D'après les définitions de $(X_t)_t$ et de $(Y_t)_t$, il est clair que le processus $(X_t, Y_t)_t$ est stationnaire. On note P la loi de $(X_t, Y_t)$ et de $((X_t, Y_t), \cdots, (X_{t+n-1}, Y_{t+n-1}))$. Cela permet aussi d'omettre l'indice $t$ ou de le remplacer par 0 dans le calcul des espérances. Si $\sigma_n : (\mathcal{X} \times \mathbb{R})^n \to \mathbb{R}$, on pourra écrire :

$$\mathbb{E}_{\mathrm{P}}\sigma_1(X_t, Y_t) = \mathbb{E}_{\mathrm{P}}\sigma_1(X, Y),$$
$$\mathbb{E}_{\mathrm{P}}\sigma_n((X_t, Y_t), \cdots, (X_{t+n-1}, Y_{t+n-1})) = \mathbb{E}_{\mathrm{P}}\sigma_n((X_0, Y_0), \cdots, (X_{n-1}, Y_{n-1})).$$

Dans la suite, on établit une inégalité de concentration pour le processus $(Z_t)_t$. On note que les variables $Z_t$ sont centrées. Ce processus est directement défini à partir du théorème 2. En effet, si l'on calcule la moyenne empirique de $(Z_t)_{1 \leq t \leq n}$, on obtient :

$$\mathbb{E}_{\mathrm{D}_n}(k(\cdot, X)h(X, Y)) - \mathbb{E}_{\mathrm{P}}(k(\cdot, X)h(X, Y)),$$

ce qui est exactement la quantité que permet de contôler le théorème 3. Par ailleurs, on retrouve dans le premier terme de $Z_t$ l'estimateur de $f_{\mathrm{P},\lambda}$ :

$$\hat{f}_{\mathrm{P},\lambda} = \frac{1}{2\lambda n}\sum_{t=1}^{n} k(\cdot, X_t)h(X_t, Y_t) = \frac{1}{2\lambda}\mathbb{E}_{\mathrm{D}_n}(k(\cdot, X)h(X, Y)).$$

**Hypothèses.** On suppose :

- $\mathcal{X}$ est compact,
  on note $\|\cdot\|_{\mathcal{X}}$ la norme dans $\mathbb{R}^d$ ;
- $f$ est $C_f$-lipschitzienne ;
- le noyau $k$ est $\hat{C}_k$-lipschitzien par rapport à la première et la seconde variable, ce qui signifie :

  $$\forall (x_0, x_1, x_2) \in \mathcal{X}, \qquad |||k(x_0, x_1) - k(x_0, x_2)||| \leq \hat{C}_k \|x_1 - x_2\|_{\mathcal{X}},$$
  $$\forall (x_0, x_1, x_2) \in \mathcal{X}, \qquad |||k(x_1, x_0) - k(x_2, x_0)||| \leq \hat{C}_k \|x_1 - x_2\|_{\mathcal{X}},$$

  on en déduit que $k$ est $C_k$-lipschitzien, où $C_k \geq 0$ ;
- la fonction $h$ définie par :

  $$h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto (\partial_2 L)(y, f(x))$$

  est $\alpha$-hölderienne, avec $\alpha \geq \frac{1}{2}$ ;
- il existe une fonction $m : \mathbb{R}^+ \to \mathbb{R}^+$ croissante telle que :

  $$\left\{ \begin{array}{lll} \forall (x, y) \in \mathcal{X} \times \mathbb{R}, & |h(x, y)| & \leq \quad m(|y|) \\ & \mathbb{E}\left[m(|\epsilon_0|)^2\right] & < \quad \infty \end{array} \right. ; \tag{3}$$

- la suite des coefficients de Lipschitz de $\phi$ vérifie :

  $$\sum_s \alpha_s < \infty,$$
  $$\sum_k R_k^{1/2} < \infty \quad \text{où } R_k = \sum_{s \geq k} \alpha_s.$$

  Cela signifie que, si $\alpha_s = (s+1)^{-\beta}$, alors la condition est vérifiée si, et seulement si, $\beta > 3$.

Les deux hypothèses nouvelles sont le caractère $\alpha$-hölderien de $h$ avec $\alpha \geq \frac{1}{2}$, et la condition (3). Cela signifie que si la perte est de la forme $l : y \mapsto y^r$, alors on doit avoir $r \geq \frac{3}{2}$, ce qui inclut heureusement la perte quadratique. Quant à la condition (3), il était nécessaire d'imposer une contrainte sur les relations entre la loi de $\epsilon_0$ et $h$. Celle-ci est, dans la démonstration, assez naturelle.

Pour la démonstration de la consistance, on s'appuie largement sur la preuve de la partie 3. La seule partie à modifier est celle qui concerne l'inégalité de concentration. On peut notamment chercher à contrôler $\mathbb{E} \left\| \sum_{k=1}^{n} Z_k \right\|^2$.

**Proposition 2.** *Majoration de la covariance entre $Z_k$ et $Z_l$ ($k \leq l$).*

$$|\mathbb{E}_{\mathrm{P}} \langle Z_k, Z_l \rangle| \leq \left[ \hat{C} \mathbb{E}_{\mathrm{P}} \left\| Z_0 \right\|^2 \mathbb{E} \left| \xi_0 \right| R_{l-k} \right]^{1/2} \quad \text{où } R_n = \sum_{s \geq n} \alpha_s,$$

*où $\langle \cdot, \cdot \rangle$ est le produit scalaire dans $\mathcal{H}$.*

**Proposition 3.** *Majoration de $\mathbb{E}_{\mathrm{P}} \left\| \sum_{i=1}^{n} Z_i \right\|^2$ :*

$$\mathbb{E}_{\mathrm{P}} \left\| \sum_{i=1}^{n} Z_i \right\|^2 \leq n \left[ \mathbb{E} \left\| Z_0 \right\|^2 + 2 \left( \hat{C} \mathbb{E}_{\mathrm{P}} \left\| Z_0 \right\|^2 \mathbb{E} |\xi_0| \right)^{1/2} R \right] \quad \text{où } R = \sum_{k \geq 1} R_k^{1/2}.$$

**Corollaire 1.** *L'estimateur est consistant si $n \lambda_n^{p+1} \to \infty$.*

Ce corollaire achève la preuve de la consistance des méthodes RKHS pour l'estimation de la fonction d'observation d'un système dynamique, à partir des données $X_t$ et $Y_t$.

## 4.3 Retour au problème initial

On peut assez facilement adapter la preuve à l'estimation de la fonction de transition $T^\star$. On reprend celle-ci en remplaçant $Y_t$ par $Y'_t = X_{t+1}$.

**Notations.**

- $Y'_t = X_{t+1} = T^\star(X_t, \xi_{t+1})$ ;
- $Z'_t = k(\cdot, X_t) h(X_t, Y'_t) - \mathbb{E}(k(\cdot, X_t) h(X_t, Y'_t)) = \psi'((\xi_s)_{s \leq t+1}) \in \mathcal{H}$.

**Hypothèses.**

- $T$ est $C_T$-lipschizienne et bornée ;
- $\mathcal{X} \subset \mathbb{R}$ compact.

**Proposition 4.** *Majoration de la covariance entre $Z'_k$ et $Z'_l$ ($k \leq l$) :*

$$|\mathbb{E}_{\mathrm{P}} \langle Z'_k, Z'_l \rangle| \leq \left[ C_X \mathbb{E}_{\mathrm{P}} \left\| Z'_0 \right\|^2 \mathbb{E} \left| \xi_0 \right| R_{l-k} \right]^{1/2},$$

*où $R_n = \sum_{s \geq n} \alpha_s$ et $C_X \geq 0$.*

Le reste de la démonstration de la consistance se déroule exactement de la même manière que dans la partie 4.2, d'où le théorème suivant.

**Théorème 4.** *L'estimateur est consistant si* $n\lambda_n^{p+1} \to \infty$.

On peut donc, sous certaines conditions, utiliser des méthodes RKHS pour estimer la fonction de transition d'un système dynamique, ce qui répond à la problématique posée dans la partie 4.1. Par contre, au niveau de l'apprentissage, on a besoin des grandeurs « réelles » $X_t$, alors que l'on aimerait se contenter des grandeurs « observables » $Y_t$.

# 5 Conclusion

Finalement, on peut estimer la fonction de transition d'un système dynamique à l'aide des RKHS à noyau opérateur, ce qui répond à la problématique de départ. Dans les démonstrations, on a pu constater la puissance du théorème 2, qui donne une formulation fonctionnelle très générale d'une fonction minimisant un risque régularisé.

On notera également que les résultats de consistance restent valables pour $\mathcal{Y}$ hilbertien réel, ce qui est plus général que $\mathcal{Y} = \mathbb{R}$. Entre autres, $\mathcal{Y}$ peut être un espace fonctionnel, auquel cas on chercherait à apprendre une fonctionnelle. Cela pose des problèmes d'un autre ordre. En pratique, si l'on apprend une fonctionnelle, on a besoin de réalisations $(Y_1, \cdots, Y_n)$ dans $\mathcal{Y}$. Dans ce cas, comment les obtient-on ? Comment les encode-t-on ?

Toutefois, des généralisations sont encore possibles. Par exemple, la plupart des démonstrations présentées ici sont valides pour $\mathcal{X}$ métrique séparable. On pourrait alors prendre $\mathcal{X} = \mathcal{Y}$ hilbertiens réels dans la partie 4.3, sur l'estimation de $T^\star$.

De plus, le problème initial n'est pas complètement résolu. Il serait intéressant en pratique d'estimer $T^\star$ uniquement à partir des données observables $(Y_t)_t$.

Pour finir, j'aimerais rappeler que, malgré l'intérêt pratique des RKHS à noyau scalaire, les RKHS à noyau opérateur nécessitent de nombreux réglages, dépendant des données pour fournir des résultats probants, comme le choix du noyau $k$. Le problème de l'apprentissage se situe alors sur les nombreux paramètres : si $\mathcal{Y}$ est de dimension infinie, il faudra certainement mettre en œuvre des stratégies complexes pour trouver un bon noyau, ce qui n'est pas le cas pour les RKHS à noyau scalaire, où $\mathcal{Y} = \mathbb{R}$.

Les RKHS à noyau opérateur sont donc théoriquement utilisables en régression non paramétrique, mais, en pratique, leur usage risque d'être nettement plus complexe.

# Références

[1] Claudio Carmelli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(4) :377–408, 2006.

[2] Claudio Carmelli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanità. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(1), 2010.

[3] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5 :1363–1390, 2004.

[4] Ivar Ekeland and Thomas Turnbull. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, 1983.

[5] Theodoros Evgeniou, Charles Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6 :615–637, 2005.

[6] Néhémy Lim, Florence d'Alché Buc, Cédric Auliac, and George Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3) :489–513, 2015.

[7] Vern I. Paulsen. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. 2006.

[8] Robert Phelps. *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Mathematics. Springer, 1986.

[9] Ulrich Rieder. Measurable selection theorems for optimization problems. *manuscripta mathematica*, 24(1) :115–131, 1978.

[10] Andres Sanchez-Perez. Time series prediction via aggregation : An oracle bound including numerical cost. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, volume 217 of *Lecture Notes in Statistics*, pages 243–265. 2015.

[11] Paul Shields. *The Theory of Bernoulli Shifts*. Chicago Lectures in Mathematics. The University of Chicago Press, 1973.

[12] Ingo Steinwart and Andreas Christmann. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3) :799–819, 2007.

[13] Ingo Steinwart, Don Hush, and Clint Scovel. Function classes that approximate the bayes risk. In *Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 79–93. 2006.

[14] Grace Wahba. *Spline Models for Observational Data*. Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990.