Stabilité uniforme de la régression non linéaire par moindres carrés régularisés avec des noyaux à valeurs opérateurs

Julien Audiffren[1] and Hachem Kadri[1]

[1]Équipe QARMA, LIF/CNRS, Université Aix-Marseille

Résumé

Nous étudions dans cet article les propriétés de stabilité et les performances de généralisation de l'algorithme des moindres carrés régularisés dans un espace de Hilbert séparable à noyau reproduisant (EHNR) dans le cas de noyaux à valeurs opérateurs. Ces noyaux, aussi connus sous le nom de noyaux multi-tâches, sont adaptés à des problèmes d'apprentissage à sortie non scalaire, tels que l'apprentissage multi-tâches et la prédiction structurée. Notre résultat principal est une preuve de la stabilité uniforme de l'algorithme de régression à noyau multi-tâches pour des sorties dans un espace de Hilbert. Bien que des similitudes avec le cas scalaire existent, des difficultés inhérentes à la dimension infinie de l'espace de sortie apparaissent. Notre preuve apporte des solutions à ces difficultés, tout en mettant l'accent sur les hypothèses du noyau que nous avons choisies les plus faibles possibles. L'utilisation de la stabilité uniforme ainsi obtenue permet de déduire la consistance de l'algorithme même avec des noyaux opérateurs non Hilbert-Schmidt [1].

Mots-clef : régression par moindres carrés régularisés, EHNR, noyaux à valeurs opérateurs, stabilité, consistance, apprentissage multi-tâches.

1 Introduction

Un des objectifs principaux de la théorie de l'apprentissage est l'étude des propriétés de généralisation des algorithmes proposés dans ce domaine. Depuis les travaux fondateurs de Vapnik et Chervonenkis [VC71], le domaine de l'apprentissage automatique a connu un développement constant dans ces aspects théoriques. Un grand nombre d'outils conçus en ce sens ont été appliqués avec succès pour analyser des algorithmes d'apprentissage de fonctions à valeurs scalaires (comprenant la classification et la régression). Parmi eux, la notion de la stabilité d'un algorithme, c'est-à-dire la manière dont la variation des données d'apprentissage influence le résultat, s'est montrée efficace pour trouver des bornes de généralisation qui dépendent des propriétés algorithmiques de la méthode d'apprentissage ([BE01], [EEP05]).

Depuis quelques années, un intérêt croissant est porté à l'apprentissage de fonctions à valeurs vectorielles [MP05] (ou à valeurs dans un espace de Hilbert plus généralement). Cet intérêt est à l'origine d'un effort considérable consacré au développement d'algorithmes pour des problématiques d'apprentissage à sortie non scalaire, telles que l'apprentissage multi-tâches et la prédiction structurée ([Car97], [BHS+07]). Bien que des avancées considérables aient été faites dans ce sens, l'étude des propriétés de généralisation de ces algorithmes s'est limitée aux cas linéaires et de dimension finie [Mau06]. Le seul travail à notre connaissance qui présente des résultats théoriques sur les bornes de généralisation dans un contexte non linéaire et pour des espaces de sortie de dimension infinie est celui de Caponetto et De Vito [CV06]. Les auteurs ont fourni des bornes de généralisation pour l'algorithme de régression par moindres carrés régularisés dans le cas où l'espace d'hypothèses est un espace de Hilbert à noyau reproduisant (EHNR) séparable. Il est important de noter que, contrairement au cadre d'apprentissage mono-tâche [2], le noyau est une fonction à valeurs opérateurs [3] définie-positive. L'opérateur a l'avantage de donner la possibilité de prendre en compte les dépendances entre les tâches, et ainsi de permettre d'étendre les méthodes d'apprentissage multi-tâches au contexte non-linéaire [EMP05].

1. Pour la définition d'un noyau dit Hilbert-Schmidt, voir la définiton 2.1.

2. c.-à.-d apprentissage de fonctions à valeurs scalaires.

3. Le noyau est à valeurs matricielles dans le cas où l'espace de sortie est de dimension finie.

Les taux de convergence proposés par [CV06], optimaux dans le cas d'un espace de sortie de dimension finie, requièrent des hypothèses sur le noyau qui peuvent être contraignantes dans le cas de la dimension infinie. En effet, leur preuve nécessite que le noyau à valeurs opérateurs soit Hilbert-Schmidt, ce qui restreint l'utilisation de ce résultat. Pour illustrer ce point, soit $K(\cdot,\cdot) = k(\cdot,\cdot)Id$ le noyau à valeurs opérateurs construit à partir de l'opérateur identité, où k est un noyau à valeurs scalaires. Ce noyau K, utilisé dans le contexte de la prédiction structurée [BDBS11] et de l'apprentissage à noyaux de distributions de probabilité [GLB+12], ne vérifie pas l'hypothèse de Hilbert-Schmidt, et par conséquent le résultat de [CV06] ne peut pas être utilisé dans ce cas (voir équation (2.2) pour plus de détails). Il est aussi important de noter que leur analyse est plus centrée sur la complexité de l'espace d'hypothèses que sur les propriétés de l'algorithme.

Le présent article a pour objectif de remédier à ces limitations. Plus précisément, nous nous intéresserons à la stabilité uniforme de l'algorithme des moindres carrés régularisés quand la sortie est un espace de Hilbert séparable, pouvant être de dimension infinie, et nous fournirons un majorant de β, le coefficient de stabilité, dont nous déduirons la consistance de son estimateur grâce à un résultat de [BE01]. Nous nous attacherons à utiliser les hypothèses les plus faibles possibles sur K.

Le reste du papier est organisé comme suit. La Section 2 présente l'algorithme de moindres carrés régularisés dans le cas des EHNR munis de noyaux à valeurs opérateurs, puis détaille notre principale contribution, à savoir la stabilité uniforme de l'algorithme des moindres carrés régularisés avec des noyaux à valeurs opérateurs. La Section 3 contient la preuve des résultats de la Section 2. Pour finir, la Section 4 présente la conclusion et les perspectives du présent article.

2 Résultat principal

Pour commencer, introduisons les notations que nous utiliserons par la suite. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé, $\mathcal{X}$ un espace polonais, $\mathcal{Y}$ un espace de Hilbert séparable, $\mathcal{H}$ un espace de Hilbert à Noyau Reproduisant (EHNR) séparable $\subset \mathcal{Y}^{\mathcal{X}}$ (l'ensemble des fonctions de $\mathcal{X}$ à valeurs dans $\mathcal{Y}$), K son noyau hermitien positif et $L(\mathcal{Y})$ l'espace des endomorphismes de $\mathcal{Y}$ muni de la norme d'opérateur. Soit $\lambda > 0$ un nombre réel, $(X_1, Y_1), ..., (X_m, Y_m)$ m copies i.i.d. du couple de variables aléatoires (X, Y) qui suit la distribution inconnue D.

Du point de vue de l'apprentissage, on notera $Z = \{(x_1, y_1), ..., (x_m, y_m)\}$ une réalisation de m copies i.i.d. de (X, Y) qui constituera les données d'entraînement, $Z^i_{x,y} = Z \cup (x, y) \setminus (x_i, y_i)$ l'ensemble Z où l'on a remplacé le couple (x_i, y_i) par une autre réalisation de (X, Y) indépendante de toutes les précédentes (x, y). Le risque empirique des moindres carrés de la fonction f obtenu à partir des données Z s'écrit

$$R_{emp}(f, Z) = \frac{1}{m}\sum_{k=1}^{m} \|y_k - f(x_k)\|^2_{\mathcal{Y}},$$

et la version régularisée du susdit risque est donnée par

$$R_{reg}(f, Z) = R_{emp}(f, Z) + \lambda\|f\|^2_{\mathcal{H}}.$$

On appellera

$$f_Z = \arg\min_{f\in\mathcal{H}} R_{reg}(f, Z), \tag{2.1}$$

la fonction minimisant le risque régularisé dans l'espace $\mathcal{H}$.

Rappelons la définition du noyau d'un EHNR $\mathcal{H}$ quand $\mathcal{Y}$ est de dimension infinie.

Définition 2.1 *L'application $K : \mathcal{X} \times \mathcal{X} \to L(\mathcal{Y})$ est le noyau hermitien défini-positif du EHNR $\mathcal{H}$ si et seulement si :*

(i) $\forall x \in \mathcal{X}$, $K(., x) \in \mathcal{H}$,

(ii) $\forall f \in \mathcal{H}$, $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$,

$$\langle f(x), y\rangle_{\mathcal{Y}} = \langle f, K(., x)y\rangle_{\mathcal{H}},$$

(iii) $\forall x_1, x_2 \in \mathcal{X}$, $K(x_1, x_2) = K(x_2, x_1)^* \in L(\mathcal{Y})$,

(iv) $\forall n \geq 1$, $\forall (x_i, i \in \{1..n\}), (x'_i, i \in \{1..n\}) \in \mathcal{X}^n$, $\forall (y_i, i \in \{1..n\}), (y'_i, i \in \{1..n\}) \in \mathcal{Y}^n$,

$$\sum_{k,\ell=0}^{n} \langle K(., x_k)y_k, K(., x'_\ell)y'_\ell\rangle_{\mathcal{H}} \geq 0.$$

(i) et (ii) définissent un noyau reproduisant, (iii) hermitien, (iv) défini-positif.

De plus, on dira qu'un tel noyau est Hilbert-Schmidt si et seulement si $\forall x \in \mathcal{X}$, $\exists (y_i)_{i\in\mathbb{N}}$ une base de $\mathcal{Y}$ telle que $Tr(K(x, x)) = \sum_{i\in\mathbb{N}} \langle K(x, x)y_i, y_i\rangle_{\mathcal{H}} < \infty$.

Dans la suite, pour parer aux problèmes de mesurabilité, on supposera que, $\forall y_1, y_2 \in \mathcal{Y}$, l'application :

$$\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$$
$$(x_1, x_2) \mapsto \langle K(x_1, x_2) y_1, y_1 \rangle_{\mathcal{Y}},$$

est mesurable. Avec le fait que $\mathcal{H}$ est séparable, cette hypothèse permet de montrer que toutes les fonctions utilisées dans cet article sont mesurables (pour plus de détails, voir [CV06]).

On appellera algorithme des moindres carrés régularisés (MCR) l'application :

$$\bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{H}$$
$$Z \mapsto f_Z,$$

où l'on rappelle que f_Z réalise le minimum du risque empirique des moindres carrés régularisés (voir équation (2.1)). Cette définition a un sens car dans le cas du MCR f_Z existe. Pour la preuve de cette existence, ainsi que l'étude d'autres propriétés du MCR dans le cas des noyaux à valeurs opérateurs, nous suggérons la lecture de [EMP05], [KDP+10] et [KRP+11].

Rappelons maintenant la définition de la stabilité uniforme.

Définition 2.2 *Un algorithme $Z \mapsto g_Z$ est dit β uniformément stable si et seulement si $\forall m \leq 1$, $\forall 1 \leq i \leq m$, pour presque tout Z une réalisation de m copies i.i.d. de (X, Y), pour presque tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$ une réalisation de (X, Y) indépendante de Z,*

$$|\|y - g_Z(x)\|_{\mathcal{Y}}^2 - \|y - g_{Z^i_{x,y}}(x)\|_{\mathcal{Y}}^2| \leq \beta,$$

où $Z^i_{x,y} = Z \cup (x, y) \setminus (x_i, y_i)$.

Comme il n'est question que de stabilité uniforme dans cet article, on utilisera l'appellation β stable comme abréviation de β uniformément stable dans la suite.

Remarquons que la β-stabilité est en fait une inégalité presque sure des variations de $\|Y - f_Z(X)\|_{\mathcal{Y}}^2$. La première hypothèse dont nous avons besoin pour notre résultat concerne cette norme.

Hypothèse 2.1 *$\exists C > 0$ tel que $\|Y - f_Z(X)\|_{\mathcal{Y}} < C$ p.s.*

Deux remarques découlent de cette hypothèse. D'abord, on peut déduire immédiatement de cette hypothèse que le MCR est β-stable avec $\beta = 2C$. Mais ce β ne dépend d'aucun paramètre du modèle, a fortiori de la taille de l'espace d'entraînement. Le but ici est d'obtenir l'expression explicite des dépendances de β avec les autres paramètres, et tel que $\lim_{m \to \infty} \beta(m) = 0$, ce qui correspond au fait que le MCR devient de plus stable avec l'accroissement de l'espace d'entraînement. Cette propriété permettra de prouver la consistance du MCR au travers d'un résultat de [BE01].

La seconde remarque est que la vérification dans un cas concret de cette hypothèse peut être difficile. En fait, dans le cadre de la β stabilité, on utilise plus souvent l'hypothèse suivante.

Hypothèse 2.2 *$\exists C_y > 0$ tel que $\|Y\|_{\mathcal{Y}} < C_y$ p.s.*

Même si le lien entre ces deux hypothèses n'est pas évident, nous allons montrer que, à l'aide de l'hypothèse 2.3 faite sur le noyau, l'hypothèse 2.2 implique l'hypothèse 2.1. Pour la preuve, nous renvoyons le lecteur au Lemme 3.1 et son Corollaire 3.2. Remarquons qu'une hypothèse similaire à 2.2 est utilisée dans le cas scalaire pour montrer la β stabilité (voir [BE01], [Pre07]).

Passons maintenant à l'hypothèse concernant le noyau K.

Hypothèse 2.3 *$\exists \kappa > 0$ tel que $\forall x \in \mathcal{X}$,*

$$\|K(x, x)\|_{op} \leq \kappa^2,$$

où $\|K(x, x)\|_{op} = \sup_{y \in \mathcal{Y}} \frac{\|K(x, x) y\|_{\mathcal{Y}}}{\|y\|_{\mathcal{Y}}}$ est la norme d'opérateur de $K(x, x)$ sur $L(\mathcal{Y})$.

Une telle hypothèse est nécessaire car si l'espace $\mathcal{H}$ est trop général, on ne peut déduire de propriété sur f_Z qui réalise le minimum du risque sur $\mathcal{H}$. Il est important de noter que cette hypothèse est plus faible que celle utilisée dans [CV06], où l'on suppose que le noyau K est Hilbert-Schmidt et que $\sup_{x \in \mathcal{X}} Tr(K(x, x)) < +\infty$. Alors que ces hypothèses sont équivalentes si $\dim \mathcal{Y} < +\infty$, ce n'est plus le cas en dimension infinie. Ainsi leurs bornes de généralisation ne s'appliquent pas aux noyaux à valeurs opérateurs non Hilbert-Schmidt en dimension infinie des exemples ci-après, bien que ceux-ci apparaissent dans la littérature.

Exemple 1 *Opérateur identité. Soit $c > 0$, $d \in \mathbb{N}^*$. En prenant $\mathcal{X} = \mathbb{R}^d$, I le morphisme identitaire dans $L(\mathcal{Y})$, et $K(x, t) = \exp(-c\|x - t\|_2^2) \times I$, on a*

$$\begin{aligned} \|K(x, x)\|_{op} &= \|I\|_{op} = 1 \\ Tr(K(x, x)) &= Tr(I) = +\infty. \end{aligned} \tag{2.2}$$

De façon générale, c'est également le cas pour les noyaux du type $K(x, t) = k(x, t) I$, où k est un noyau à valeurs scalaires défini positif. De tels noyaux sont utilisés dans le cadre de la prédiction structurée [BDBS11], et de l'apprentissage à noyau de distribution de probabilité [GLB+12].

Exemple 2 *Opérateur multiplication. Soit k un noyau à valeurs scalaires défini positif, $\mathcal{I}$ un intervalle de $\mathbb{R}$ et $C > 0$.*

En prenant $f \in \{g \in L^\infty(\mathcal{I}, \mathbb{R}), \|g\|_\infty < C\}$, $\mathcal{Y} = L^2(\mathcal{I}, \mathbb{R})$, et soit K le noyau, construit à partir de l'opérateur de multiplication, défini de la manière suivante : $K(x,z)y(.) = k(x,z)f^2(.)y(.) \in \mathcal{Y}$. Ce type de noyau est utilisé dans le cas de la régression fonctionnelle non linéaire [KDP+10]. Mais bien que K vérifie toujours l'hypothèse 2.3, il est difficile de vérifier que K est Hilbert-Schmidt, propriété qui dépend du choix de f. Ainsi, pour $f(t) = \frac{C}{2}(exp(-t^2) + 1)$,

$$\begin{aligned} \|K(x,x)\|_{op} &\leq C^2 k(x,x) \\ Tr(K(x,x)) &= \sum_{i\in\mathbb{N}} \langle K(x,x)y_i, y_i\rangle \\ &\geq k(x,x)\frac{C}{2}\sum_{i\in\mathbb{N}} \|y_i\|_2^2 = \infty. \end{aligned} \tag{2.3}$$

Voici maintenant le résultat principal de cet article.

Théorème 1 *Sous les hypothèses 2.3 et 2.1, le MCR avec noyau à valeurs opérateurs est β stable avec* $\beta = \dfrac{8C^2\kappa^2}{m\lambda}$.

Comme annoncé précédemment, $\beta = O(\frac{1}{m})$. Cette propriété permet de déduire le corollaire suivant, qui est principalement une application d'un résultat de [BE01] au théorème précédent.

Corollaire 2.3 *Sous les hypothèses 2.3 et 2.1, L'estimateur obtenu par le MCR est consistant. Plus précisément, on a $\forall m \geq 1$:*

$$\mathbb{P}\left(|R_{emp}(f_Z, Z) - \mathbb{E}(\|Y - f_Z(X)\|_{\mathcal{Y}}^2)| > \varepsilon + \beta\right) \leq 2\exp\left(-\frac{m}{2}\frac{\varepsilon^2\lambda}{8C^2\kappa^2 + C\lambda}\right).$$

3 Preuve

Dans cette section, nous allons montrer les résultats énoncés dans la section 2. Pour commencer nous allons prouver que les hypothèses 2.2 et 2.3 impliquent l'hypothèse 2.1.

Lemme 3.1 *Sous l'hypothèse 2.2, $\exists C_f = \frac{C_y}{\sqrt{\lambda}} > 0$ tel que $\forall Z$, $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}$ une réalisation de (X,Y) indépendante de Z, $\|f_Z\|_{\mathcal{H}} < C_f$ et $\|f_{Z^i_{x,y}}\|_{\mathcal{H}} < C_f$.*

Preuve : Pour prouver cette inégalité, l'astuce est de remarquer que puisque $\mathcal{H}$ est un espace vectoriel, $0 \in \mathcal{H}$, et

$$\begin{aligned} \lambda\|f_Z\|_{\mathcal{H}}^2 &\leq R_{reg}(f_Z, Z) \leq R_{reg}(0, Z) \\ &\leq \frac{1}{m}\sum_{k=1}^{m}\|y_k\|_{\mathcal{Y}}^2 \\ &\leq C_y^2, \end{aligned}$$

où dans la première ligne on a utilisé le fait que f_Z minimise le risque régularisé, puis en deuxième ligne on a explicité $R_{reg}(0,Z)$, et on conclut en utilisant l'hypothèse 2.2. L'inégalité obtenue $\|f_Z\|_{\mathcal{H}} \leq \frac{C_y}{\sqrt{\lambda}}$ est bien uniforme en Z. La preuve est la même pour $f_{Z^i_{x,y}}$ en considérant cette fois $R_{reg}(f_{Z^i_{x,y}}, Z^i_{x,y})$. □

Le corollaire suivant utilise la propriété reproduisante du noyau K et l'hypothèse 2.3 pour obtenir le résultat.

Corollaire 3.2 $\exists C = \left(\frac{\kappa C_y}{\sqrt{\lambda}} + C_y\right) > 0$ *tel que* $\|Y - f_Z(X)\|_{\mathcal{Y}} < C$ *p.s.*

Preuve : $\forall x \in \mathcal{X}$,

$$\begin{aligned} \|f_Z(x)\|_{\mathcal{Y}}^2 &= \langle f_Z(x), f_Z(x)\rangle_{\mathcal{Y}} \\ &= \langle K(x,x)f_Z, f_Z\rangle_{\mathcal{H}} \leq \|K(x,x)\|_{op}\|f_Z\|_{\mathcal{H}}^2 \\ &\leq \kappa^2 C_f^2. \end{aligned}$$

Puis on conclut en utilisant le fait que,

$$\begin{aligned} \|Y - f_Z(X)\|_{\mathcal{Y}} &\leq \|Y\|_{\mathcal{Y}} + \|f_Z(X)\|_{\mathcal{Y}} \\ &\leq C_y + \kappa C_f, \end{aligned}$$

i.e. l'hypothèse 2.1. □

Maintenant nous allons prouver le théorème 1, en utilisant une extension au cas de la dimension infinie de la preuve du théorème 12.3 de [SS02]. Notons que de nombreuses différences avec le cas réel sont présentes. Ainsi, le Lemme 12.7 de [SS02] n'est plus applicable, la fonction coût n'est plus lipschitzienne, et des manipulations techniques sont nécessaire à partir de l'équation (3.5) pour terminer la preuve quand y n'est pas scalaire.

Preuve :

Tout d'abord, remarquons que

$$\begin{aligned} &|\|y - f_Z(x)\|_{\mathcal{Y}}^2 - \|y - f_{Z^i_{x,y}}(x)\|_{\mathcal{Y}}^2| \\ &= |\|y - f_Z(x)\|_{\mathcal{Y}} - \|y - f_{Z^i_{x,y}}(x)\|_{\mathcal{Y}}| \\ &\qquad \times |\|y - f_Z(x)\|_{\|}y + \|y - f_{Z^i_{x,y}}(x)\|_{\mathcal{Y}}| \\ &\leq 2C\kappa\|f_Z - f_{Z^i_{x,y}}\|_{\mathcal{H}} \end{aligned} \tag{3.1}$$

où l'on a utilisé successivement la deuxième inégalité triangulaire et les hypothèses 2.1 et 2.3.

Donc pour obtenir le résultat voulu, il suffit d'obtenir une majoration sur $\|f_Z - f_{Z^i_{x,y}}\|_{\mathcal{H}}$.

Rappelons que dans le cas du MCR :

$$\nabla_f R_{emp}(f, Z) = \frac{2}{m}\sum_{k=1}^{m} K(.,x_k)(f_Z(x_k) - y_k),$$

où ∇_f représente le gradient par rapport à la variable f (voir par exemple [Cia07] page 172 Théorème 8.1-2 et suite pour une définition du gradient dans ce cas).

Puisque f_Z minimise le risque R_{reg} pour Z, et $f_{Z^i_{x,y}}$ pour $Z^i_{x,y}$, on a par définition

$$\begin{aligned} 0 = \nabla_f R_{reg}(f_Z, Z) &= \nabla_f R_{emp}(f_Z, Z) + 2\lambda f_Z \\ &= \nabla_f R_{emp}(f_{Z^i_{x,y}}, Z^i_{x,y}) + 2\lambda f_{Z^i_{x,y}}. \end{aligned} \tag{3.2}$$

On pose maintenant

$$\begin{aligned} R(f) = \Big\langle \nabla_f R_{emp}(f_Z, Z) - \nabla_f R_{emp}(f_{Z^i_{x,y}}, Z^i_{x,y}), \\ f - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} + \lambda \|f - f_{Z^i_{x,y}}\|^2_{\mathcal{H}}. \end{aligned} \tag{3.3}$$

On a $R(f_{Z^i_{x,y}}) = 0$ par construction, et

$$\begin{aligned} \nabla_f R(f) = \nabla_f R_{emp}(f_Z, Z) - \nabla_f R_{emp}(f_{Z^i_{x,y}}, Z^i_{x,y}) \\ + 2\lambda(f - f_{Z^i_{x,y}}), \end{aligned}$$

puis, en utilisant (3.2), on obtient

$$\nabla_f R(f_Z) = 0. \tag{3.4}$$

Considérons l'application $\tilde{R} : \mathbb{R} \mapsto \mathbb{R}$ définie par $\tilde{R}(t) = R(f_{Z^i_{x,y}} + t(f_Z - f_{Z^i_{x,y}}))$. De part sa définition, $\tilde{R(t)} \in \mathbb{R}_2[t]$ et $\lim_{t\to+\infty} \tilde{R}(t) = \lim_{t\to-\infty} \tilde{R}(t) = +\infty$. D'où on déduit que $\tilde{R}$ admet un minimum global au point 1 en utilisant $\tilde{R}(1) = R(f_Z)$ et (3.4). Donc

$$\tilde{R}(1) \leq \tilde{R}(0) = R(f_{Z^i_{x,y}}) = 0.$$

Étudions le produit scalaire qui apparaît dans $\tilde{R}(1)$:

$$\begin{aligned} &\frac{m}{2}\Big\langle \nabla_f R_{emp}(f_Z, Z) - \nabla_f R_{emp}(f_{Z^i_{x,y}}, Z^i_{x,y}) \\ &\qquad , f_Z - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} \\ =&\sum_{j=1, j\neq i}^{m} \Big\langle K(.,x_j)(f_Z(x_j) - y_j) - K(.,x_j)(f_{Z^i_{x,y}}(x_j) - y_j) \\ &\qquad , f_Z - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} \\ &+ \Big\langle K(.,x_i)(f_Z(x_i) - y_i) - K(.,x)(f_{Z^i_{x,y}}(x) - y) \\ &\qquad , f_Z - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} \\ =&\sum_{j=1, j\neq i}^{m} \Big\langle K(.,x_j)(f_Z(x_j) - f_{Z^i_{x,y}}(x_j)), f_Z - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} \\ &+ \Big\langle K(.,x_i)(f_Z(x_i) - y_i) - K(.,x)(f_{Z^i_{x,y}}(x) - y) \\ &\qquad , f_Z - f_{Z^i_{x,y}} \Big\rangle_{\mathcal{H}} \\ =&\sum_{j=1, j\neq i}^{m} \Big\langle f_Z(x_j) - f_{Z^i_{x,y}}(x_j), f_Z(x_j) - f_{Z^i_{x,y}}(x_j) \Big\rangle_{\mathcal{Y}} \\ &+ \Big\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \Big\rangle_{\mathcal{Y}} \\ &- \Big\langle f_{Z^i_{x,y}}(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \Big\rangle_{\mathcal{Y}} \\ \geq&\Big\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \Big\rangle_{\mathcal{Y}} \\ &- \Big\langle f_{Z^i_{x,y}}(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \Big\rangle_{\mathcal{Y}}. \end{aligned}$$

D'où, puisque $\tilde{R}(1) \leq 0$,

$$\begin{aligned} 0 \geq &\Big\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \Big\rangle_{\mathcal{Y}} \\ &- \Big\langle f_{Z^i_{x,y}}(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \Big\rangle_{\mathcal{Y}} \\ &+ \frac{m\lambda}{2}\|f_Z - f_{Z^i_{x,y}}\|^2_{\mathcal{H}}, \end{aligned}$$

et

$$\begin{aligned} \frac{m\lambda}{2}\|f_Z - f_{Z^i_{x,y}}\|^2_{\mathcal{H}} \leq &\Big\langle f_{Z^i_{x,y}}(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \Big\rangle_{\mathcal{Y}} \\ &- \Big\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \Big\rangle_{\mathcal{Y}} \end{aligned} \tag{3.5}$$

Le membre de droite de la précédente inégalité peut être réécrit de la manière suivante :

$$\begin{aligned}
&\left\langle f_{Z^i_{x,y}}(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \right\rangle_{\mathcal{Y}} \\
&\quad - \left\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \right\rangle_{\mathcal{Y}} \\
&= \left\langle f_Z(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \right\rangle_{\mathcal{Y}} \\
&\quad - \left\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \right\rangle_{\mathcal{Y}} \\
&\quad - \|f_Z(x) - f_{Z^i_{x,y}}(x)\|^2_{\mathcal{Y}}.
\end{aligned} \tag{3.6}$$

En combinant (3.5) avec (3.6), on obtient que

$$\begin{aligned}
&\frac{m\lambda}{2}\|f_Z - f_{Z^i_{x,y}}\|^2_{\mathcal{H}} \\
&\leq \left\langle f_Z(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \right\rangle_{\mathcal{Y}} \\
&\quad - \left\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \right\rangle_{\mathcal{Y}}.
\end{aligned} \tag{3.7}$$

D'autre part, $\forall x \in \mathcal{X}$,

$$\begin{aligned}
&\|f_Z(x) - f_{Z^i_{x,y}}(x)\|^2_{\mathcal{Y}} \\
&= \left\langle f_Z(x) - f_{Z^i_{x,y}}(x), f_Z(x) - f_{Z^i_{x,y}}(x) \right\rangle_{\mathcal{Y}} \\
&= \left\langle K(.,x)(f_Z(x) - f_{Z^i_{x,y}}(x)), f_Z - f_{Z^i_{x,y}} \right\rangle_{\mathcal{H}} \\
&= \left\langle K(.,x)K(.,x)^*(f_Z - f_{Z^i_{x,y}}), f_Z - f_{Z^i_{x,y}} \right\rangle_{\mathcal{H}} \\
&\leq \kappa^2 \|f_Z - f_{Z^i_{x,y}}\|^2_{\mathcal{H}},
\end{aligned} \tag{3.8}$$

où l'on a utilisé l'hypothèse 2.3 à la dernière ligne.

On obtient alors que

$$\begin{aligned}
&\frac{m\lambda}{2}\|f_Z - f_{Z^i_{x,y}}\|^2_{\mathcal{H}} \\
&\leq \left\langle f_Z(x) - y, f_Z(x) - f_{Z^i_{x,y}}(x) \right\rangle_{\mathcal{Y}} \\
&\quad - \left\langle f_Z(x_i) - y_i, f_Z(x_i) - f_{Z^i_{x,y}}(x_i) \right\rangle_{\mathcal{Y}} \\
&\leq 2C\kappa \|f_Z - f_{Z^i_{x,y}}\|_{\mathcal{H}},
\end{aligned}$$

où l'on a utilisé (3.7) pour la première inégalité, et l'inégalité de Cauchy Schwartz combinée avec (3.8) pour la deuxième inégalité.

Au final, on obtient que

$$\|f_Z - f_{Z^i_{x,y}}\|_{\mathcal{H}} \leq \frac{4C\kappa}{m\lambda}.$$

D'où, en utilisant (3.1), le MCR est β-stable avec

$$\beta = \frac{8C^2\kappa^2}{m\lambda}.$$

□

Le corollaire 2.3 est une conséquence immédiate du théorème suivant de [BE01].

Théorème 2 *Soit $Z \mapsto f_Z$ un algorithme β-stable, qui vérifie l'hypothèse 2.1. Alors, $\forall m \geq 1$,*

$$\begin{aligned}
&\mathbb{P}\left(|R_{emp}(f_Z, Z) - \mathbb{E}(\|Y - f_Z(X)\|^2_{\mathcal{Y}})| > \varepsilon + \beta\right) \\
&\qquad \leq 2\exp\left(-\frac{m}{2}\frac{\varepsilon^2}{m\beta + C}\right).
\end{aligned}$$

Comme le terme de droite de l'inégalité tend bien vers 0 quand m tend vers $+\infty$, on en déduit la consistance de f_Z.

4 Conclusion

Nous avons prouvé que le MCR est β-stable et que son estimateur est consistant, même dans le cas d'une sortie de dimension infinie. L'utilisation d'hypothèse plus faible sur le noyau K, et notamment le fait qu'il ne soit plus obligatoirement Hilbert-Schmidt, permet d'utiliser d'autres noyaux que ceux étudiés par [CV06] en dimension infinie.

Cependant, la question de l'optimalité de la borne de généralisation en dimension infinie est toujours ouverte, et n'a pas été prouvée. Il est aussi intéressant d'étudier la question de la β-stabilité d'autres algorithmes, tels que la régression à vecteurs de support.

Références

[BDBS11] Céline Brouard, Florence D'Alche-Buc, and Marie Szafranski. Semi-supervised penalized output kernel regression for link prediction. ICML '11, June 2011.

[BE01] Olivier Bousquet and Andre Elisseeff. Algorithm stability and generalisation performance. *Advances in Neural Information Processing Systems*, 13, 2001.

[BHS+07] Gökhan Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S.V.N. Vishwanathan. *Predicting Structured Data*. The MIT Press, 2007.

[Car97] Rich Caruana. *Multitask Learning*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1997.

[Cia07] Philippe G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Dunod, 2007.

[CV06] Andrea Caponetto and Ernesto De Vito. Optimal rates for the regularized least square algorithm. *Fundations of the computational mathematics*, 7 :361–368, 2006.

[EEP05] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6 :55–79, 2005.

[EMP05] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6 :615–637, 2005.

[GLB+12] Steffen Grunewalder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massi Pontil. Conditional mean embeddings as regressors. ICML '12, July 2012.

[KDP+10] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, and Manuel Davy. Nonlinear functional regression : a functional rkhs approach. AISTATS '10, May 2010.

[KRP+11] Hachem Kadri, Asma Rabaoui, Philippe Preux, Emmanuel Duflos, and Alain Rakotomamonjy. Functional regularized least squares classiffication with operator-valued kernels. ICML '11, June 2011.

[Mau06] Andreas Maurer. Bounds for linear multitask learning. *J. of Machine Learning Research*, 7 :117–139, 2006.

[MP05] Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17 :177–204, 2005.

[Pre07] Cristian Preda. Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of statistical planning and inference*, 137 :829–840, 2007.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels*. The MIT Press, 2002.

[VC71] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16(2) :264–280, 1971.