Introduction au domaine de recherche : Optimisation stochastique pour l'apprentissage

Aymeric Dieuleveut

Supervisé par Francis BACH

8 octobre 2013

Introduction

L'apprentissage statistique est l'étude de la prédiction de phénomènes à partir d'observations. Dans de nombreuses situations le problème peut se ramener à résoudre un problème de minimisation, pour lequel on ne dispose que d'informations incomplètes sur la fonction que l'on souhaite minimiser. Deux problèmes se mêlent alors : quel algorithme utiliser pour approximer efficacement la fonction en conservant un temps de calcul raisonnable, et quel est l'impact de l'imprécision sur la connaissance des fonctions ? Le premier problème est un problème d'optimisation, le second un problème de stochasticité. Les méthodes de descente de gradient ([3]), méthodes récursives qui consistent à actualiser l'estimée du minimiseur en se déplaçant le long de la ligne de plus grande pente, s'avèrent être relativement robustes lorsque l'on rentre dans un cadre stochastique. C'est l'idée que développent Robbins et Monro en 1951. Un point essentiel pour de telles méthodes est le choix de la séquence des pas dans les itérations successives. Les premières propositions utilisent des suites de pas de somme divergente mais de carré sommable, afin d'effectuer un compromis approprié entre biais et variance. Cependant, en utilisant une idée de Polyak et Ruppert ([7], [6]), on peut utiliser des pas beaucoup plus grands, qui ne respectent pas la seconde condition. L'algorithme proposé récemment ([2]) atteint un taux optimal avec une suite de pas constants, dans un espace euclidien. La majorité de mon travail est consacrée à l'étude des généralisations à des espaces de dimension infinie.

Table des matières

1 Apprentissage et optimisation stochastique

1.1 Premières définitions

On rappelle quelques notions fondamentales d'apprentissage statistique, en particulier le cadre de la prédiction : on cherche à prédire une **variable d'intérêt** $Y \in \mathcal{Y}$ à partir d'une **variable explicative** $X \in \mathcal{X}$, en disposant d'un échantillon de n observations $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$, observations qui sont indépendantes et identiquement distribuées, de loi $\mathbb{P}$.

Définition 1. *On appelle un* ***prédicteur*** *toute application mesurable g de $\mathcal{X}$ dans $\mathcal{Y}$. L'ensemble de ces applications est noté* $\mathbb{S}$

On s'attend à ce que $g(X_{n+1})$ soit un "bon prédicteur" de Y_{n+1}. Pour définir une telle notion de bon, il nous faut définir un contraste :

Définition 2. *On appelle contraste toute fonction*

$$\begin{aligned} \ell : \mathbb{S} \times (\mathcal{X} \times \mathcal{Y}) &\to \mathbb{R} \\ (g, (x, y)) &\mapsto \ell(g, (x, y)). \end{aligned}$$

On définit également une fonction de perte :

Définition 3. *La fonction de perte associée à un contraste ℓ est l'espérance du contraste :*

$$\begin{aligned} P_\ell : \mathbb{S} &\to \mathbb{R} \\ g &\mapsto \mathbb{E}\left[\ell(g, (X, Y))\right]. \end{aligned}$$

On appelle prédicteur de Bayes le meilleur des prédicteurs au regard de la fonction de perte : $s^* = \arg\min_{s \in \mathbb{S}} P_\ell(s)$. Notre but est donc de déterminer un prédicteur dont la performance est aussi proche que possible de celle du prédicteur de Bayes.
On peut citer brièvement quelques exemples fondamentaux :

- Régression : dans ce cas $\mathcal{Y} = \mathbb{R}$ et $Y = \eta(X) + \varepsilon$ avec $\eta(X) = \mathbb{E}[Y|X]$ la fonction de régression. On peut alors considérer le contraste des moindres carrés : $\ell(g, (x, y)) = (g(x) - y)^2$. Dans ce cadre le prédicteur de Bayes est la fonction de régression.
- La classification binaire : $\mathcal{Y} = \{0, 1\}$, avec le contraste 0-1 : $\ell(g, (x, y)) = \mathbb{1}_{g(x) \neq y}$. Le prédicteur de Bayes est alors $s^*(X) = \mathbb{1}_{\eta(X) \geq \frac{1}{2}}$.

Comment déterminer effectivement, à partir de nos observations, un prédicteur dont la performance soit aussi proche que possible de celle du prédicteur de Bayes ?

1.2 Deux problèmes d'optimisation

On cherche donc à résoudre le problème de minimisation suivant : $\min_{g \in \mathbb{S}} P_\ell(g)$, à partir de nos observations. Une première approche consiste à minimiser le **risque empirique** : $P_{n,\ell}(g) := \frac{1}{n} \sum_{i=1}^{n} \ell(g, (x_i, y_i))$. Cependant, on ne peut souhaiter minimiser un tel objectif sur l'ensemble des modèles, car on se heurte à un problème de sur-apprentissage : on va choisir un prédicteur trop complexe qui ne permettra pas une bonne généralisation. C'est pourquoi on s'intéresse plutot au critère pénalisé :

$$\arg\min_f P_{n,\ell}(g) + \mathrm{pen}(g).$$

A condition que ℓ soit convexe une pénalité fortement convexe permettra d'aboutir à un problème qui sera, dans sa globalité, fortement convexe, ce qui est important d'un point de vue algorithmique. Cependant, il faudra combiner le résultat obtenu avec une borne uniforme sur $\sup_g |P_{n,\ell} - P_\ell|(g)$, car on ne cherche pas à minimiser la vraie fonction P_ℓ.
Une seconde approche consiste à trouver une technique pour obtenir à partir des observations un bon minimiseur P_ℓ, sans passer par le risque empirique. C'est cette méthode qu'on appelle "approximation stochastique".

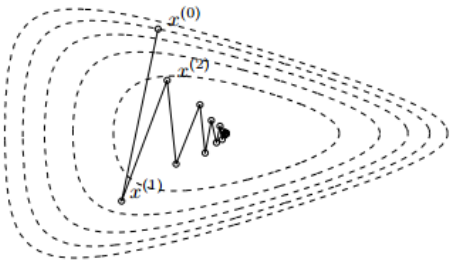

FIGURE 1 – Premières itérées d'une descente de gradient ([3])

1.3 Convexité

Un point fondamental en optimisation est le caractère convexe de la fonction que l'on cherche à optimiser. La fonction P_ℓ est généralement convexe (elle l'est si le contraste est convexe en sa première variable). Ce n'est pas systématique : le contraste 0-1 n'est pas convexe ! Il est néanmoins possible de convexifier le risque, typiquement en utilisant un contraste convexe, vérifiant de bonnes conditions ([1]). Pour cette raison, par la suite, on supposera toujours la fonction de contraste convexe.

1.4 Les algorithmes de descente de gradient

Les algorithmes de descente de gradient, introduits initialement par Cauchy en 1847, sont des algorithmes itératifs qui procèdent par améliorations successives pour s'approcher d'un minimiseur d'une fonction différentiable ou sous différentiable définie sur un espace euclidien E ou un Hilbert $\mathcal{H}$. L'idée fondamentale est de suivre, à chaque étape, la direction de la plus forte pente qui est exactement l'opposé du gradient. (Figure 1).
Pour minimiser une fonction f différentiable sur $\mathcal{H}$, l'algorithme s'exprime donc de façon générale sous la forme :
– Initialisation : choisir un point de départ $\theta_0 \in \mathcal{H}$
– Itérer : étant obtenu θ_k, déterminer le gradient $\nabla f(\theta_k)$ et renvoyer $\theta_{k+1} := \theta_k - \gamma_k \nabla f(\theta_k)$.
Ces algorithmes présentent des vitesses de convergence dépendant généralement des hypothèses sur la forte convexité ou non de la fonction à minimiser, et du choix de la suite $(\gamma_k)_k$ des pas.

2 Optimisation stochastique

2.1 Descente de gradient stochastique

Par la suite, on s'intéressera parfois à des prédicteurs linéaires : $g_\theta(x) = \langle \theta, x \rangle$. On notera donc indifféremment g_θ ou θ.
Dans le cadre stochastique, on n'a pas directement accès au gradient de la fonction que l'on cherche à minimiser, puisque l'on ne connaît pas la loi de distribution des (X, Y), donc on ne connaît pas P_ℓ. On va donc mettre en place l'algorithme suivant, dit "algorithme de gradient stochastique" :

- Initialisation : choisir un point de départ $\theta_0 \in \mathcal{H}$
- Itérer : étant obtenu θ_k, déterminer un estimateur ψ_k sans biais du gradient $\nabla f(\theta_k)$ et renvoyer $\theta_{k+1} := \theta_k - \gamma_k \psi_k$.

Remarque : Il est important de noter que cet algorithme peut s'appliquer aux deux approches évoquées plus haut : à la minimisation du risque empirique pénalisé, comme à l'approximation stochastique. Le point crucial est d'être capable d'exhiber un estimateur sans biais du gradient : dans la proposition suivante, la fonction f ci dessus est tantôt $P_{n,\ell} + \text{pen}$, tantôt P_ℓ.

Proposition 1.

- ***Minimisation du risque empirique pénalisé (MRE) :*** *Soit i_k de loi uniforme sur $\{1, ..., n\}$. Alors $\psi_k := \nabla \left(\ell(g_{k-1}, (x_{i_k}, y_{i_k})) + pen(g_{k-1}) \right)$ est un estimateur sans biais de $\nabla \left(P_{n,\ell}(g_{k-1}) + pen(g_{k-1}) \right)$.*
- ***Approximation stochastique (AS) :*** *Soit (x, y) indépendants de g_{k-1}. Alors $\psi_k := \nabla(\ell(g_{k-1}, (x, y)))$ est un estimateur sans biais de $\nabla P_\ell(g_{k-1})$.*

On constate que dans le cadre de la MRE on peut utiliser plusieurs fois chaque observation de D_n, alors que dans le cadre SA, on doit utiliser un exemple indépendant à chaque itération. En fait on évite le sur-apprentissage en effectuant un seul passage dans les données.

2.2 Un résultat fondamental

La forme de base de la récurrence est donc la suivante : $\theta_{k+1} = \theta_k - \gamma \psi_k$. Un premier résultat découle en quelques lignes de calcul des propriétés de convexité de f.

Théorème 1. *Si on considère l'algorithme ci dessus pour une suite de pas constants, en supposant que pour tout $k \in \{1, .., n\}, ||\psi_k|| \leqslant R$:*

$$\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta_*) \right] \leqslant \frac{1}{2n\gamma} ||\theta_0 - \theta_*||^2 + \frac{\gamma}{2} R^2.$$

Où $\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k$.

Remarques :

– On déduit de ce théorème qu'avec une suite de pas constants, proportionnels à $\frac{1}{\sqrt{n}}$ on obtient un taux de l'ordre de $O\left(\frac{1}{\sqrt{n}}\right)$.

– Avec une hypothèse de forte convexité, on peut obtenir un taux $O\left(\frac{1}{n}\right)$, voir par exemple [5].

2.3 Une vitesse en O(1/n) sans forte convexité

On se place dans le cadre de la régression des moindres carrés sur un espace euclidien. Bach et Moulines ([2]) ont montré récemment qu'on pouvait obtenir le taux optimal $O(1/n)$ sans hypothèse de forte convexité en effectuant une simple descente de gradient à pas constant.

Théorème 2. *Supposons que :*

H1 *: $\mathcal{H}$ est un espace de dimension finie d.*

H2 *: Les observations (x_n, z_n) sont i.i.d. ($z_n = y_n x_n$).*

H3 *: $\mathbb{E}||x_n||^2$ et $\mathbb{E}||z_n||^2$sont finies. Soit $\Sigma = \mathbb{E}(x_n \otimes x_n)$ l'opérateur de covariance de $\mathcal{H}$ dans $\mathcal{H}$.*

H4 *: Le minimum global de $f(\theta) = \frac{1}{2}\mathbb{E}\left[\langle\theta, x_n\rangle^2 - 2\langle\theta, z_n\rangle\right]$ est atteint en θ_*. On note $\xi_n = z_n - \langle\theta_*, x_n\rangle x_n$ le terme résiduel.*

H5 *: On étudie la descente de gradient stochastique définie par*

$$\theta_n = \theta_{n-1} - \gamma(\langle\theta_{n-1}, x_n\rangle x_n - z_n) = (I - \gamma x_n \otimes x_n)\theta_{n-1} + \gamma z_n,$$

avec θ_0 dans $\mathcal{H}$. On note $\bar{\theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\theta_k$.

H6 *: Il existe $R > 0$ et $\sigma > 0$ tels que $\mathbb{E}\left[\xi_n \otimes \xi_n\right] \preccurlyeq \sigma^2\Sigma$, and $\mathbb{E}(||x_n||^2 x_n \otimes x_n) \preccurlyeq R^2\Sigma$ ($\preccurlyeq$ désigne l'ordre naturel entre les opérateurs auto-adjoints).*

Alors pour tout pas constant $\gamma < \frac{1}{R^2}$, on a :

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant \frac{1}{2n}\left[\frac{\sigma\sqrt{d}}{1-\sqrt{\gamma R^2}} + R||\theta_0 - \theta_*||\frac{1}{\sqrt{\gamma R^2}}\right]^2.$$

Par exemple avec $\gamma = \frac{1}{4R^2}$ on obtient $\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant \frac{2}{n}\left[\sigma\sqrt{d} + R||\theta_0 - \theta_*||\right]^2$.

Ce théorème ne couvre que le cadre de la dimension finie. Le terme de gauche de la somme doit être interprété comme un terme de "variance" qui augmente avec la variance du bruit et avec la taille des pas, tandis que le terme de droite est un terme qui représente la difficulté à s'extraire de la condition initiale : il dépend de $||\theta_0 - \theta_*||$ et diminue si on augmente la taille du pas γ. Dans la suite, on généralise ce théorème au cadre de la dimension infinie.

2.4 Cas hilbertien

On va reprendre la majorité des hypothèses du cas euclidien, et en ajouter deux afin de s'adapter à la difficulté de la dimension infinie. On suppose en fait :

H1' : $\mathcal{H}$ est un espace de Hilbert.

On note toujours $\Sigma = \mathbb{E}[x_n \otimes x_n]$, qui n'est plus une matrice mais un opérateur compact symétrique de $\mathcal{H}$. Bien que non inversible en général, il n'y a pas de difficulté à se restreindre à un supplémentaire du noyau.
Pour être capable d'obtenir des résultats, il nous faut effectuer des hypothèses sur l'opérateur de covariance Σ et sur la distance initiale $||\theta_0 - \theta_*||$.

H7 : On note $(\lambda_j)_{j\in\mathbb{N}}$ la suite des valeurs propres de l'opérateur Σ en ordre décroissant. On suppose que $\frac{u^2}{j^\alpha} \leqslant \lambda_j \leqslant \frac{s^2}{j^\alpha}$ pour un certain $\alpha > 1$ (de sorte que $\mathrm{tr}(\Sigma) < \infty$) .

H8 : On suppose que les coordonnées $(\nu_j)_j$ de $\theta_* - \theta_0$ dans la base orthonormale des vecteurs propres de Σ sont telles que $\nu_j \leqslant \left(\frac{1}{T j^{\frac{\beta}{2}}}\right)_{j\in\mathbb{N}}$, pour un certain $\beta > 1$.

Sous ces hypothèses, on peut obtenir le théorème suivant :

Théorème 3. *Supposons **H1', H2-8** :*

1. *Si* $\alpha + 1 > \beta$

$$\left(2\ \mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right]\right)^{1/2} \leqslant \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\ s^{\frac{2}{\alpha}}\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}\right)^{1/2} + \left(\frac{s^{2\frac{-\beta+\alpha+1}{\alpha}}}{T^2u^2} K(\alpha,\beta) \frac{1}{(n\gamma)^{\frac{\alpha+\beta-1}{\alpha}}}\right)^{1/2} + \left(\frac{1}{1-\gamma R^2}\frac{||\eta_0||R^2}{n}\right)^{1/2}.$$

2. *Si* $\alpha + 1 < \beta$

$$\left(2\ \mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right]\right)^{1/2} \leqslant \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\ s^{\frac{2}{\alpha}}\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}\right)^{1/2} + \left(\frac{1}{T^2u^2}\left(\frac{1}{\beta-\alpha-1} + \frac{2}{\alpha+\beta-1}\right)\frac{1}{(\gamma n)^2}\right)^{1/2} + \left(\frac{1}{1-\gamma R^2}\frac{||\eta_0||R^2}{n}\right)^{1/2}.$$

Ce qui se réécrit asymptotiquement :

1. Si $\alpha + 1 > \beta$

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] = O\left(\frac{1}{(n\gamma)^{1-\frac{1}{\alpha}+\frac{\beta}{\alpha}}}\right) + O\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right).$$

2. Si $\alpha + 1 < \beta$

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] = O\left(\frac{1}{(n\gamma)^2}\right) + O\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right).$$

En optimisant le choix de γ, on obtient :

Corollaire 1. *Supposons* ***H1', H2-8*** *alors on a :*

1. *Si* $\alpha + 1 > \beta$

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant K_1\ n^{-1+\frac{1}{\alpha+\beta}}, \qquad \text{avec } \gamma = \frac{\gamma_0}{n^{\frac{\beta}{\alpha+\beta}}}.$$

2. *Si* $\alpha + 1 < \beta$

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant K_2\ n^{-1+\frac{1}{2\alpha+1}}, \qquad \text{avec } \gamma = \frac{\gamma_0}{n^{\frac{\alpha+1}{2\alpha+1}}}.$$

Avec γ_0 une constante t.q. $\gamma_0 \leqslant \frac{1}{R^2}$, et K_1, K_2 des constantes affreuses.

Remarques :

– On doit utiliser un pas constant dans la descente de gradient : une partie de la preuve repose sur ce fait. Cependant, si on travail à horizon fixé, on peut choisir γ qui dépend de n. Dans ce cas l'algorithme n'est pas a priori "en ligne" : si on augmente n, il faudrait recalculer toutes les itérations avec un pas légéremment modifié. Prouver que l'on peut utiliser une séquence de pas décroissante est un enjeu de la suite de mon travail.

– On constate un phénomène connu sous le nom de saturation : avoir un plus grand α ou β constitue une hypothèse plus forte. Il est donc naturel que la vitesse de convergence s'améliore quand α ou β augmente. Cependant, au delà d'une certaine valeur, une augmentation de β n'apporte plus de progrès.

– On retrouve une décomposition condition initiale-bruit identique à celle du cas euclidien.

– Plus encore, ce théorème est cohérent avec le résultat du cas euclidien : en effet, si on est en dimension finie, on peut avoir l'hypothèse **H7** pour α arbitrairement grand, et avec $\alpha \to \infty$, on retrouve exactement les asymptotiques du théorème 2.

3 Régression dans un RKHS

Dans le cadre de la dimension infinie, un cas particulier mérite d'être abordé car la majorité des calculs peuvent alors être menés sans excès de complexité majeur. C'est le cas des espaces à noyau reproduisant, espaces de Hilbert particuliers dans lesquels le produit scalaire peut être calculé efficacement.

3.1 RKHS

On appelle espace à noyau reproduisant (reproducing kernel Hilbert space) un espace de fonctions qui est caractérisé par les propriétés suivantes.
Soit $\mathcal{X}$ un espace quelconque.

Définition 4. *On appelle noyau de Mercer une application continue symétrique $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ semi définie positive dans le sens où $\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geq 0$ pour tout $m \in \mathbb{N}$, tout $(x_i)_{i=1..m} \in \mathcal{X}^m$ et tout $(c_i)_{i=1..m} \in \mathbb{R}^m$.*

Pour tout $x \in \mathcal{X}$ on peut définir une fonction $K_x : x' \mapsto K(x, x')$.

Définition 5. *On appelle espace à noyau reproduisant la complétion de l'espace vectoriel engendré par les fonctions $K_x, x \in \mathcal{X}$, muni du produit scalaire qui étend la forme bilinéaire telle que $\langle K_x, K_{x'} \rangle = K(x, x')$. On note cet espace $\mathcal{H}_K$.*

Proposition 2. *C'est un espace de Hilbert. De plus pour toute fonction f de $\mathcal{H}_K$ et x de $\mathcal{X}$ on a la propriété dite "reproduisante" : $f(x) = \langle f, K_x \rangle$.*

Remarque : On peut aussi définir un RKHS par un autre point de vue : tout espace de Hilbert $\mathcal{H}$ de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ tel que pour tout x de $\mathcal{X}$ la forme linéaire $g \mapsto g(x)$ est continue est un RKHS. En effet par le théorème de Riesz on peut définir une application $\Phi : \mathcal{X} \to \mathcal{H}$ qui à x associe l'unique vecteur K_x tel que $g(x) = \langle g, K_x \rangle$. On peut alors définir le noyau K par $K(x, x') = \langle K_{x'}, K_x \rangle$.
Le théorème d'Aronszjan lie ces deux approches :

Théorème 4 (Aronsjan, 1950)**.** *K est un noyau semi défini positif si et seulement si il existe un espace de Hilbert $\mathcal{H}_K$ et une application $\Phi : \mathcal{X} \to \mathcal{H}_K$, telle que $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$.*

Dans un tel espace, le problème de la régression des moindres carrés se réécrit de façon linéaire :

$$\min_{f \in \mathcal{H}_K} \mathbb{E}[(f(X) - Y)^2] = \min_{f \in \mathcal{H}_K} \mathbb{E}[(\langle f, K_X \rangle - Y)^2].$$

On se retrouve donc dans le même cadre que précédemment et on peut utiliser notre algorithme de descente de gradient stochastique pour trouver un bon prédicteur f_n.

3.2 Algorithme

La descente de gradient décrite précédemment s'écrit :
– Choisir g_0 dans $\mathcal{H}_K$

– $g_n = \sum_{i=1}^{n} a_i K_{x_i}$, avec une suite $(a_n)_n$ définie récursivement par $a_0 = 0$ et

$$a_n := -\gamma(g_{n-1}(x_n) - y_n) = -\gamma\left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_i\right).$$

– On note $\bar{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} \bar{g}_k$.

Remarque : Cet algorithme est particulièrement simple à mettre en place, et de complexité $O(n^2)$.

3.3 Théorème de convergence

Toute l'analyse a été menée dans un espace de Hilbert général, puis transposée au cadre du RKHS. On peut dresser un bref tableau récapitulatif des différences entre les notations de la première et de la seconde partie :

	Espace de Hilbert	RKHS
Espace :	$\mathcal{H}$	$\mathcal{H}_K$
Observations :	$y_n = \langle \theta_*, x_n \rangle + \varepsilon_n$	$y_n = g_\rho(x_n) + \varepsilon_n$
Objectif :	θ_*	g_ρ
Vecteur :	x	K_x
Gradient :	$\gamma(\langle \theta_{n-1}, x_n \rangle x_n - z_n)$	$-(g_{n-1}(x_n) - y_n)K_{x_n}$
Opérateur de covariance :	$\Sigma = \mathbb{E}[x_n \otimes x_n]$	L_K

Il n'y a donc aucune difficulté supplémentaire à obtenir le théorème suivant, avec des hypothèses qui sont l'exacte transposition des hypothèses du cas "Hilbert" au cas "RKHS". On obtient ainsi exactement les vitesses de convergence détaillées ci dessus.
Il est intéressant de donner un exemple d'une situation dans laquelle les hypothèses **H7,8** sont réalisées.

3.4 L'exemple des splines

Cet exemple est tiré de [8] : on considère l'ensemble $W^{m,2}]0;1[$ des fonctions de $]0;1[$ dans $\mathbb{R}$ dont la dérivée m$^{\text{e}}$ est intégrable. C'est un espace de Sobolev qui est aussi un espace de Hibert. En décomposant les fonctions sur une base de cosinus et sinus, on peut obtenir une forme simple pour le noyau, ainsi qu'une expression des éléments propres de l'opérateur de covariance.

Théorème 5. *L'espace $W^{m,2}]0;1[$ est un RKHS de noyau $R(s,t) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}(\{s-t\})$, avec B_m le m^e polynome de Bernoulli.*
*Dans cet espace, les valeurs propres de l'opérateur de covariance sont de multiplicité deux, de valeurs $\left(\frac{1}{(2i\pi)^{2m}}\right)_{i\geq 1}$. On a donc notre hypothèse **H7** avec $\alpha = 2m$.*

Ce théorème donne une interprétation simple de la condition "α plus grand" qui correspond ici à travailler sur un plus petit RKHS, ce qui rend bien naturel le fait d'obtenir une meilleure performance.
Dans ces conditions, l'hypothèse sur les coordonnées de θ_* correspond à une hypothèse de régularité du signal.

3.5 Optimalité du résultat

Le résultat démontré est optimal. En effet on trouve dans [4] la borne minimax suivante :

Théorème 6. *Soit $\mathcal{P}(\alpha, r)$ $(\alpha > 1, r \in [1/2, 1])$ l'ensemble de toutes les mesures de probabiltés ρ sur $\mathcal{X} \times \mathcal{Y}$, telles que :*
– *ps, $|y| \leqslant M_\rho$,*
– *$L_K^{-r} g_\rho \in \mathcal{L}^2_{\rho(X)}$,*
– *les valeurs propres $(\mu_j)_{j\in\mathbb{N}}$ en ordre décroissant, vérifient $\mu_j = O(n^{-j})$.*
Alors :

$$\liminf_{n\to\infty} \inf_{f_n} \sup_{\rho\in\mathcal{P}(b,r)} \mathbb{P}\left\{ f(g_n) - f(g_\rho) > C n^{-2r/(2r+1)} \right\} = 1,$$

pour une constante $C > 0$. L'infimum du milieu est pris sur tous les algorithmes vus comme des applications $((x_i, y_i)_{1\leqslant i\leqslant n}) \mapsto f_n \in \mathcal{H}_K$.

Des expériences sur des jeux de données aléatoires simulées illustrent ces résultats.

3.6 Problèmes ouverts

Un certain nombre d'enjeux doivent encore être abordés :
– Peut-on démontrer le résultat pour une suite de pas décroissants, afin d'obtenir un algorithme "en ligne" ?
– Peut-on améliorer la complexité algorithmique sans trop perdre sur le résultat, pour obtenir une complexité sub-quadratique ?
– Aborder les problèmes d'adaptativité du choix du noyau et de la suite de pas.

Références

[1] F. Bach. *Notes du cours d'apprentissage statistique*. 2009.

[2] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *ArXiv e-prints*, June 2013.

[3] S. Boyd and L. Vanderberghe. *Convex optimisation*. 2004.

[4] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3) :331–368, 2007.

[5] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O(1/t) rate for the stochastic projected subgradient method. *ArXiv e-prints*, December 2012.

[6] B. T. Polyak and A. B. Juvisy. Acceleration of stochastic approximation by averaging. 1992.

[7] D Ruppert . *Efficient estimation from a slowly convergent Robbins-Monro process*. 1988.

[8] G. Wahba. *Spline Models for observationnal data*. 1990.