École normale supérieure
Mémoire de troisième année.

Directeur : Francis Bach

Optimisation stochastique dans les espaces de Hilbert

Aymeric Dieuleveut

Paris, le 6 octobre 2013

Table des matières :

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".

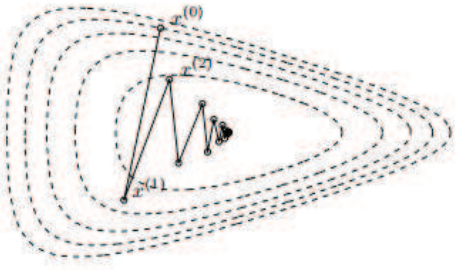


FIGURE 1 – Premières itérées d'une descente de gradient ([3])

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 :

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

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.

Curriculum vitae

Aymeric Dieuleveut
45 rue d'Ulm
75005 Paris

Études

	Né le 9 février 1991.
2010	Entrée à l'École Normale Supérieure de Paris.
2010–2011	Licence de mathématiques, mention AB.
2011–2012	M1 de mathématiques, mention B.
2012–2013	M2 de mathématiques, mention Probabilités et Statistiques à l'univeristé Paris XI, mention TB.

Stages en laboratoire

Avril 2013 -	Optimisation stochastique et apprentissage dans les espaces de Hilbert. Sous la direction de Francis BACH, Inria Paris Rocquencourt, équipe Sierra.

Liste des cours suivis en M2

S1 -	**Probabilités** Grandes déviations et applications (2011, 14/20) Théorèmes limites et processus de Poisson (2012, 16.5/20) Calcul Stochastique (2012, 18/20) **Statistiques** : Concentration et sélection de modèles (2012, 18/20) Méthodes asymptotiques en statistiques (2012, 19/20) Statistiques en grandes dimensions (2012, 17/20)
S2 -	Statistique et théorie de l'information (2013, 14/20) Apprentissage statistique (2013, 18/20) Mémoire M2 (2013, 17/20).

Liste des exposés présentés

juin 2011	Mémoire de première année : **Marches aléatoires controlées** Sous la direction de Gilles Stoltz.
octobre 2011	Exposés de statistiques non paramétrique (groupe de travail) : **Estimateur proximaux** ; **Régression non paramétrique**.
avril 2012	Exposé d'imagerie mathématique (leçons de maths) : **Utilisation de la dérivée topologique pour la détection de petites anomalies**.
mai 2012	Exposé de statistiques non paramétrique (leçon de maths) : **Efficacité asymptotique et adaptation.**
2013	Exposés de validation des cours de M2 : **Reconnaissance d'objets géométriques et sélection de modèle.** Exposé de théorie de l'information : **codage en alphabet infini**. Exposé d'apprentissage : **"SVM optimisation, inverse dependance on training set size".**
sept. 2013	Soutenance de mémoire de M2 : **Optimisation stochastique dans les espaces de Hilbert.** Sous la direction de Francis Bach.

Conférences et séminaires suivis

- Séminaires SMILE (2013)
- École d'été de statistiques mathématiques et applications, Fréjus, 2-6 septembre 2013.
- Séminaire CLARA, 11 septembre 2013.

Divers

Colles en PC au lycée Turgot (Paris 3).
Colles en MP* au lycée Fénelon Sainte Marie.
Anglais courant.
Programmation en Matlab, Scilab.
Connaissance de base du langage C.

Université Paris Sud
Master de mathématiques fondamentales et appliquées
Mémoire de M2

Supervisor : Francis Bach

Stochastic optimisation in Hilbert spaces

Aymeric Dieuleveut

Paris, le September 25, 2013

Contents

Introduction

Given a sequence of independent and identically distributed (i.i.d.) random examples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ *drawn from a probability measure* ρ *on* $\mathcal{X} \times \mathcal{Y}$*, we consider the problem of minimising the mean squared error* $f(g) := \mathbb{E}_\rho \left[(g(X) - Y)^2\right]$*, assuming that our objective function (the regression function, such that* $g_\rho(x) = \int_{\mathcal{X}\times\mathcal{Y}} y d\rho_{\mathcal{Y}|x}$*) lies in a reproducing kernel Hilbert space. Finding an approximation of the regression function has been extensivly studied in [2], [3] by considering stochastic gradient descent algorithms on the problem of regularized mean squares (i.e., adding a penalisation term). To investigate such a problem, we may drive all the analysis in a general Hilbert space, following the ideas proposed in [4], but in a infinite-dimensional space instead of considering the finite-dimensional setting. Traditionnal online stochastic approximation algorithms, as they where introduced by Robbins and Monro [5] are based on stochastic gradient descent methods with steps decreasing with time t which are in most cases proportionnal to* $\frac{1}{t^\rho}$*, with* ρ *between* $1/2$ *and 1. More recently [4] showed that constant steps with averaging led to the best posible convergence rate. This raised the idea that one may consider steps which may be either constant, or decreasing as* $t^{-p},\ 0 < p < 1/2$*. Our analysis, although made in a finite horizon setting, leads to an explicit choice of a constant step to be used in the gradient descent, which depends on the number of training examples we want to use and on the assumptions we make. We show that when tryi ng to minimise the squared loss in the regression setting with the objective function satisfying good assumptions, using regularized least squares doesn't to better than stochastic gradient descent with averaging. This may be compared with [6], [7]. In the following, we first remind of the finite-dimensionnal case (i.e., in an Euclidean space), as it is proposed in [4]. We then study the case of a general Hilbert space and finaly show how these results may be applied to the regression setting in a reproducing kernel Hilbert space (a.k.a., RKHS).*

1 Euclidean spaces

Recently, Bach and Moulines showed in [4] that for least squares regression, averaged stochastic gradient descent achieved a rate of $O\left(\frac{1}{n}\right)$, in a finite-dimensional Hilbert space : if we consider the problem of minimising a convex function $f(\theta) = \mathbb{E}\left[\ell(y, \langle\theta, x\rangle)\right]$, and under the following assumptions, the may get theorem 1 (this content is directly taken from [4]).

H1: $\mathcal{H}$ is a d-dimensional space.

H2: The observations (x_n, z_n) are independently and identically distributed.

H3: $\mathbb{E}||x_n||^2$ and $\mathbb{E}||z_n||^2$ are finite. Let $\Sigma = \mathbb{E}(x_n \otimes x_n)$ be the covariance operator from $\mathcal{H}$ to $\mathcal{H}$.

H4: The global minimum of $f(\theta) = \frac{1}{2}\mathbb{E}\left[\langle\theta, x_n\rangle^2 - 2\langle\theta, z_n\rangle\right]$ is attained at a certain θ_*. We denote by $\xi_n = z_n - \langle\theta_*, x_n\rangle x_n$ the residual. We have $\mathbb{E}\left[\xi_n\right] = 0$ but in general we don't have $\mathbb{E}\left[\xi_n | x_n\right] = 0$.

H5: We study the stochastic gradient recursion defined as

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

with θ_0 in $\mathcal{H}$. We denote $\bar{\theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\theta_k$.

H6: There exist $R > 0$ and $\sigma > 0$ such that $\mathbb{E}\left[\xi_n \otimes \xi_n\right] \preccurlyeq \sigma^2\Sigma$, and $\mathbb{E}(||x_n||^2 x_n \otimes x_n) \preccurlyeq R^2\Sigma$ where $\preccurlyeq$ denotes the order between self-adjoint operators.

For least-squares problems, $z_n = x_n y_n$ where y_n is the response to be predicted as a linear function of x_n. We denote $\eta_n = \theta_n - \theta_*$ (resp. $\bar{\eta}_n = \bar{\theta}_n - \theta_*$).

Theorem 1. *Assume **H1-6**. Then for any constant step-size $\gamma < \frac{1}{R^2}$, we have*

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

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

However this bound is not interesting whenever d is very large, or even infinite. To consider such settings, we are going to make two supplementary assumptions.

2 Hilbert spaces

We are now interested with the case where d is very large or even infinite.
We thus make the hypothesis :

H1': $\mathcal{H}$ is a Hilbert space.

We denote $||.||$ the norm associated with the scalar product.
We still denote $\Sigma = \mathbb{E}[x_n \otimes x_n]$, but it can no longer be considered as a matrix but as a compact operator of $\mathcal{H}$. Moreover it's not assumed to be invertible in general, however we may write "Σ^{-1}" to denote the inverse of the induced operator $\tilde{\Sigma} : F \to \mathrm{Im}(\Sigma)$, where F is a supplementary of $\ker(\Sigma)$. This inverse is only defined on $\mathrm{Im}(\Sigma)$ and may not be continuous, but will always be considered composed with Σ so that there will not be any difficulty.
Without strong convexity, we are only interested in results on $f\left(\bar{\theta}_n\right) - f(\theta_*)$. A calculation, which is exactly lemma 1 gives us :

$$f\left(\bar{\theta}_n\right) - f(\theta_*) = \langle \bar{\theta}_n - \theta_*, \Sigma(\bar{\theta}_n - \theta_*)\rangle = \mathbb{E}\left[\langle x, \bar{\theta}_n - \theta_*\rangle^2\right] =: ||\bar{\theta}_n - \theta_*||^2_{\mathcal{L}^2_{\rho(X)}},$$

a norm which depends on the distribution $\rho(X)$ of X. Using the equalities above, we clearly have $||\Sigma^{1/2}(\bar{\theta}_n - \theta_*)||^2 = ||\bar{\theta}_n - \theta_*||^2_{\mathcal{L}^2_{\rho(X)}}$ so that $\Sigma^{1/2}$ is an isometrical operator from $\left(\mathcal{H}_\rho, ||.||_{\mathcal{L}^2_\rho}\right)$ to $(\mathcal{H}, ||.||)$, where $\mathcal{H}_\rho := \{\theta, ||\theta||_{\mathcal{L}^2_\rho} < \infty\}$. Moreover, for any x in $\ker(\Sigma)$, $||\theta + x||_{\mathcal{L}^2_\rho} = ||\theta||_{\mathcal{L}^2_\rho}$: we may do our study using $\tilde{\Sigma}$ instead of Σ, for simplicity we use the same notation in the following, which is more or less equivalent to considering that Σ is invertible.

In order to be able to establish bounds, we have to make assumptions on the covariance operator Σ and the initial value $||\theta_0 - \theta_*||$.

H7 : We denote $(\lambda_j)_{j\in\mathbb{N}}$ the sequence of eigenvalues of the operator Σ (assumed invertible, as explained above), in decreasing order. We assume $\frac{u^2}{j^\alpha} \leqslant \lambda_j \leqslant \frac{s^2}{j^\alpha}$ for some $\alpha > 1$ (so that $\mathrm{tr}(\Sigma) < \infty$) .

H8 : We assume the coordinates $(\nu_j)_j$ of $\theta_* - \theta_0$ in the eigenbasis of the covariance operator are such that $\nu_j \leqslant \left(\frac{1}{T\, j^{\frac{\beta}{2}}}\right)_{j\in\mathbb{N}}$, for some $\beta > 1$ (so that $||\theta_* - \theta_0|| < \infty$). In fact we assume in the following $\theta_0 = 0$ and $\theta_* \preccurlyeq \left(\frac{1}{Tj^{\frac{\beta}{2}}}\right)_{j\in\mathbb{N}}$ in the eigenbasis. Here $\preccurlyeq$ stands for a comparison of all the coordinates.

However this couple of assumptions may be slightly reduced, even if they make things easier to be understood. It's better to consider :

H7' : We denote $(\lambda_j)_{j\in\mathbb{N}}$ the sequence of eigenvalues of the operator Σ, in decreasing order. We assume $\lambda_j \leqslant \frac{s^2}{j^\alpha}$ for some $\alpha > 1$ (so that $\mathrm{tr}(\Sigma) < \infty$).

H8' : $\Sigma^{-r}\theta_* \in \mathcal{L}^2_{\rho(X)}$ with $r \geq 0$.

In fact when $\beta + \alpha - 2\alpha r > 1$ we can prove that **H7,8** implie **H7',8'** : we have $||\Sigma^{-r}\theta_*||^2_{\mathcal{L}^2_\rho} = ||\Sigma^{-r+1/2}\theta_*||^2$ and in the eigenbasis of Σ we have $\Sigma^{-r+1/2}\theta_* \preccurlyeq \frac{j^{\alpha r-\alpha/2}}{u^{2r-1}Tj^{\beta/2}} = Cj^{\alpha r-\alpha/2-\beta/2}$, so that $||\Sigma^{-r}\theta_*||^2_{\mathcal{L}^2_\rho} \leqslant C^2\sum_{j=1}^\infty j^{2\alpha r-\alpha-\beta} < \infty$. In other words, **H7,8** implie **H7',8'** with any $r < \frac{\alpha+\beta-1}{2\alpha}$: we don't need neither the lower bound on Σ's eigenvalues, nor the particular polynomial decrease for the objective function.

In the following, most of the results are proved and written for **H7,8**, but lemma 6 shows that they remain true under **H7',8'**.

2.1 Assuming the covariance operator Σ is known

We first assume the covariance operator Σ is known.

H5': We study the stochastic gradient recursion defined as

$$\theta_n = (I - \gamma\Sigma)\theta_{n-1} + \gamma z_n,$$

with θ_0 in $\mathcal{H}$.

Following the proof in [4] we have :

Proposition 1. *Assuming **H1', H2-4, H5', H6-8** , and $\mathbb{E}[\xi_n|x_n] = 0$, we have :*

1. *If $\alpha + 1 > \beta$ then :*

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

2. *If $\alpha + 1 < \beta$ then :*

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

With $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$, $K(\alpha,\beta) = \frac{1}{\alpha+\beta-1} + \frac{1}{\alpha+1-\beta}$ and $L(\alpha,\beta) = \frac{1}{\alpha+\beta-1}\ + \frac{2}{\beta-\alpha-1}$.

Note this analysis is tight : we could also get a similar lower bound up to asymptotically negligible terms making natural supplementary assumptions : $\mathbb{E}[\xi_n \otimes \xi_n] \geq \tilde{\sigma}^2\Sigma$ in **H6** and $\nu_j = \Theta\left(j^{-\frac{\beta}{2}}\right)$ in **H8**, we also have a similar minoration, up to asymptoticaly negligeable terms, with constants that could be written exactly.

We have to choose γ as a constant step, as most of the calculations in our proof rely on this fact. However, in the case where we assume that n is a fixed number (i.e., considering a finite horizon), we may choose $\gamma = f(n)$ as a constant step. To get rid of this constraint, one may either consider a doubling trick such as in [8] or [9]. One may also consider a step varrying with time t, leading to a correct online algorithm, but making the analysis much harder (this may be done in a future work).

Corollary 1. *Assuming **H1', H2-4, H5', H6-8** we have :*

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

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

With γ_0 a constant such that $\gamma_0 \leqslant \frac{1}{R^2}$, and :

$$\begin{aligned} K_1 &= C(\alpha)\ s^{2/\alpha}(\sigma^2+||\theta_*||^2R^2)\gamma_0^{\frac{1}{\alpha}}+\sigma^2+2\frac{s^{2\left(\frac{\alpha+1-\beta}{\alpha}\right)}}{T^2u^2}K(\alpha,\beta)\frac{1}{\gamma_0^{\frac{\alpha+\beta-1}{\alpha}}}\ , \\ K_2 &= C(\alpha)\ s^{2/\alpha}(\sigma^2+||\theta_*||^2R^2)\gamma_0^{\frac{1}{\alpha}}+\sigma^2+\frac{1}{T^2u^2}L(\alpha,\beta)\frac{1}{\gamma_0^2}\ . \end{aligned}$$

This speed may be compared with $O\left(n^{-1}\right)$ in the finite-dimensional case. We can see that the bigger α or β is, the better our convergence rate is, which is not surprising as a bigger α or β is a stronger hypothesis on the model. But beyond a certain constant (which is exactly $\alpha+1$) getting a bigger β does not improve our convergence rate anymore. This is a phenomenon known as a saturation problem, see for example [10].

2.2 Assuming the covariance operator Σ is unknown

We are able to get a similar proposition in the case of an unknown covariance operator Σ, using the natural stochastic gradient descent as described in **H5**.

Proposition 2. *Assume **H1', H2-8** :*

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

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

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

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

Asymptotically :

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

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

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

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

Optimising our choice of γ, we have :

Corollary 2. *Assuming* ***H1', H2-8*** *we have :*

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

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant K_3\ n^{-1+\frac{1}{\alpha+\beta}}, \qquad \textit{with } \gamma = \frac{\gamma_0}{n^{\frac{\beta}{\alpha+\beta}}}.$$

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

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f(\theta_*)\right] \leqslant K_4\ n^{-1+\frac{1}{2\alpha+1}}, \qquad \textit{with } \gamma = \frac{\gamma_0}{n^{\frac{\alpha+1}{2\alpha+1}}}.$$

With γ_0 *a constant such that* $\gamma_0 \leqslant \frac{1}{R^2}$*, and :*

$$K_3 = \frac{1}{2}\left(\frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\ s^{\frac{2}{\alpha}}\ \sigma^2\gamma_0^{\frac{1}{\alpha}} + \sigma^2\right)^{1/2} + \left(\frac{s^{2\frac{-\beta+\alpha+1}{\alpha}}}{T^2u^2\gamma_0^{\frac{\alpha+\beta-1}{\alpha}}}K(\alpha,\beta)\right)^{1/2} + \left(\frac{||\eta_0||R^2}{1-\gamma R^2}\right)^{1/2}\right)^2.$$

$$K_4 = \frac{1}{2}\left(\frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\ s^{\frac{2}{\alpha}}\ \sigma^2\gamma_0^{\frac{1}{\alpha}} + \sigma^2\right)^{1/2} + \left(\frac{L(\alpha,\beta)}{T^2u^2\gamma_0^2}\right)^{1/2} + \left(\frac{||\eta_0||R^2}{1-\gamma R^2}\right)^{1/2}\right)^2.$$

with $L(\alpha,\beta) = \left(\frac{1}{\beta-\alpha-1} + \frac{2}{\alpha+\beta-1}\right)$.

2.3 Comments

Several comments should be made on these convergence rates. First, the optimal step size γ is proportionnal to some $1/n^p$, with p varrying from 0 to 1. More precisely it tends to behave just like a constant when $\alpha \to \infty$, i.e., when our assumptions tend to make our model comparable to a finite-dimensional setting, thus being consistant with the results of [4].
We may summarize the different cases in the following tabular :

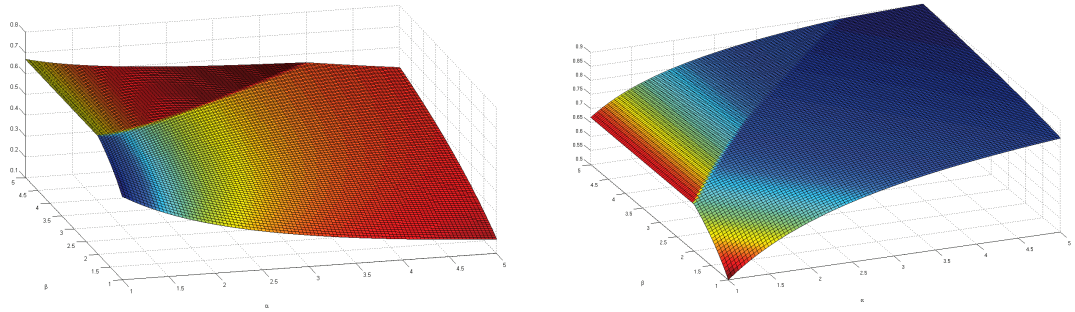
(λ_i)	$\theta_{*\,i}$	Condition	Bias	Variance	Optimal γ	Predicted performance
$i^{-\alpha}$	$i^{\frac{-\beta}{2}}$	$\alpha > \beta - 1$	$\left(\frac{1}{(n\gamma)^{1-\frac{1}{\alpha}+\frac{\beta}{\alpha}}}\right)$	$\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right)$	$\frac{1}{n^{\frac{\beta}{\alpha+\beta}}}$	$\left(n^{-1+\frac{1}{(\alpha+\beta)}}\right)$
$i^{-\alpha}$	$i^{\frac{-\beta}{2}}$	$\alpha < \beta - 1$	$\left(\frac{1}{(n\gamma)^2}\right)$	$\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right)$	$\frac{1}{n^{\frac{\alpha+1}{2\alpha+1}}}$	$\left(n^{-1+\frac{1}{(2\alpha+1)}}\right)$

It appears that whenever $\alpha > \beta$ we have to choose a step proportionnal to n^{-p} with $0 < p < \frac{1}{2}$.

This must be compared with similar tradeoffs and results showed in [1] (using conversions between the constants in this article and the other constants here). We get both the same limit condition and predicted performance. See appendix A for the comparison of both hypothesis and notations. These results are ploted in figure 1.

3 Result in RKHS

We are going to apply the results above to the problem of non parametric estimation, assuming the objective function lies in a reproducing kernel Hilbert space.



(a) $p(\alpha, \beta)$ such that $\gamma = n^{-p}$ is optimal.

(b) $\zeta(\alpha, \beta)$ such that the performance is $n^{-\zeta}$.

Figure 1: optimal choice of step size, saturation, and predicted performance.

3.1 Results on RKHS

We have to introduce some basic background and notations about reproducing kernel Hilbert spaces. Part of this content may be found in [2].

3.1.1 Definition

Let $K : \mathcal{X} \times \mathcal{X}$ be a *Mercer kernel*, i.e., a continuous symmetric real function which is positive semi-definite in the sense that $\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geq 0$ for any $m \in \mathbb{N}$, any $(x_i)_{i=1..m} \in \mathcal{X}^m$ and any $(c_i)_{i=1..m} \in \mathbb{R}^m$. A Mercer kernel induces functions $K_x : \mathcal{X} \to \mathbb{R}$ defined by $K_x(x') = K(x, x')$. We denote $\mathcal{H}_K$ the completion of the span $\{K_x, x \in \mathcal{X}\}$, with respect to the inner product defined as the unique extension of the bilinear form $\langle K_x, K_{x'} \rangle = K(x, x')$. We name $\mathcal{H}_K$ the reproducing kernel Hilbert space (RKHS) associated with the kernel K. This space is of course a Hilbert space, so we may be able to apply the results of the previous part. The norm of $\mathcal{H}_K$ is denoted by $||.||_K$. One of the most important properties of RKHS is the reproducing property : for all $g \in \mathcal{H}_K$ and $x \in \mathcal{X}$, $g(x) = \langle g, K_x \rangle_K$.
Another way to define a RKHS is to consider any Hilbert space $\mathcal{H}_K$ of functions from $\mathcal{X}$ to $\mathbb{R}$ such that for any $x \in \mathcal{X}$ the linear form $g \mapsto g(x)$ is continuous. We may then define by Riesz's representation theorem a map $\Phi : \mathcal{X} \to \mathcal{H}_K, \quad x \mapsto \Phi(x) := K_x$ which is the only vector so that $g(x) = \langle g, K_x \rangle$. Finaly, we define a kernel by $K(x, x') = \langle K_x, K_{x'} \rangle$. Aronszajn's theorem links both approches :

Theorem 2 (Aronszajn, 1950)**.** *K is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}_K$ and a map $\Phi : \mathcal{X} \to \mathcal{H}_K$, so that $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$.*

3.1.2 Sampling Operator

3.1.3 Covariance operator

Let $\rho_{\mathcal{X}}$ be any probability measure on $\mathcal{X}$ and $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ the space of square integrable functions with respect to $\rho_{\mathcal{X}}$.
We define a linear map :

$$\begin{array}{rccc} L_K & : & \mathcal{L}^2_{\rho_{\mathcal{X}}} & \to & \mathcal{L}^2_{\rho_{\mathcal{X}}} \\ & & g & \mapsto & \displaystyle\int_{\mathcal{X}} g(t)\ K_t\ d\rho_{\mathcal{X}}(t). \end{array}$$

So that for any $x \in \mathcal{X}$, $L_K(g)(x) = \int_{\mathcal{X}} g(t)\ K(x, t)\ d\rho_{\mathcal{X}}(t)$. If $X \sim \rho_{\mathcal{X}}$ then $L_K(g) = \mathbb{E}\left[g(X) K_X\right]$.

Theorem 3. *L_K is a compact operator on $\mathcal{L}^2_{\rho_\mathcal{X}}$.*

The proof may be found in [11].

Theorem 4. *L_K is a self-adjoint operator on $\mathcal{L}^2_{\rho_\mathcal{X}}$.*

Proof 1. *For any $g, h \in \mathcal{L}^2_{\rho_\mathcal{X}}$ we have :*

$$\begin{aligned}
\langle h, L_K g\rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} &= \int_\mathcal{X} h(x) L_K(g)(x) d\rho_\mathcal{X}(x) \\
&= \int_\mathcal{X} h(x) \left(\int_\mathcal{X} g(t)\ K(x,t)\, d\rho_\mathcal{X}(t) \right) d\rho_\mathcal{X}(x) \\
&= \int_{\mathcal{X}\times\mathcal{X}} h(x) g(t)\ K(x,t)\ \ d\rho_\mathcal{X}(t) d\rho_\mathcal{X}(x) \\
&= \langle L_K h, g\rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}}.
\end{aligned}$$

The fact that L_K is self-adjoint and compact implies the existence of an ortonormal eigensystem (which is an hilbertian base of $\mathcal{L}^2_{\rho_\mathcal{X}}$) : $(\mu_i, \phi_i)_{i\in\mathbb{N}} \in \mathbb{R} \times \mathcal{L}^2_{\rho_\mathcal{X}}$ such that $L_K\phi_i = \mu_i\phi_i$. Besides as $\mathrm{Im}(L_K) \subset \mathcal{H}_K$, we also have $\phi_i \in \mathcal{H}_K$.
We can thus define (as $\mathrm{span}\{\phi_i\}$ is dense in $\mathcal{L}^2_{\rho_\mathcal{X}}$), for any $r \in \mathbb{R}$:

$$\begin{aligned}
L_K^r\ :\quad \mathcal{L}^2_{\rho_\mathcal{X}} &\to \mathcal{L}^2_{\rho_\mathcal{X}} \\
\sum_i a_i\phi_i &\mapsto \sum_i \mu_i^r a_i \phi_i.
\end{aligned}$$

Under good conditions $ker(L_K) = 0$
In particular, $L_K^{1/2}$ is an isometrical isomorphism between $\mathcal{L}^2_{\rho_\mathcal{X}}$ and $\mathcal{H}_K$. Indeed :

$$\begin{aligned}
||L_K^{1/2} g||_K &= \langle g, L_K g\rangle_K \\
&= \left\langle g, \int_\mathcal{X} g(t)\ K_t\ d\rho_\mathcal{X}(t) \right\rangle_K \\
&= \int_\mathcal{X} g(t)\ \langle g, K_t\rangle_K\ d\rho_\mathcal{X}(t) \\
&= \int_\mathcal{X} g(t)^2 d\rho_\mathcal{X}(t) \\
&= ||g||^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}.
\end{aligned}$$

The restriction of L_K on $\mathcal{H}_K$ induces an operator which is the covariance operator of $\rho_\mathcal{X}$ in $\mathcal{H}_K$ since $L_K|_{\mathcal{H}_K} = \mathbb{E}\left[S_x^* S_x\right]$ by the reproducing property.

3.2 Theorem in RKHS

3.2.1 Hypothesis in RKHS

We make the following assumptions :

A1: $\mathcal{H}_K$ is the RKHS associated with a Mercer kernel K.

A2: The observations (x_n, y_n) are independently and identically distributed according to ρ. We denote $\rho_\mathcal{X}$ the induced marginal probability on X.

A3: $\mathbb{E}\left[K(x_n, x_n)\right]$ and $\mathbb{E}||y_n x_n||^2$ are finite. Let L_K be the covariance operator from $\mathcal{H}_K$ to $\mathcal{H}_K$.

A4: The global minimum of $f(g) := \frac{1}{2}\mathbb{E}_\rho\left[(g(x)-y)^2\right]$ is attained at a certain g_ρ (which is in fact the regression function of ρ, i.e., the function such that $g_\rho(x) = \int_Y y d\rho_{Y|x}$ with $\rho_{Y|x}$ the conditional probability measure on Y with respect to $x \in \mathcal{X}$). We denote by $\xi_n = y_n - g_\rho(x_n)$ the residual. We have $\mathbb{E}[\xi_n] = 0$ but in general we don't have $\mathbb{E}[\xi_n|x_n] = 0$.

A5: We study the stochastic gradient recursion defined by :

$$\begin{aligned} g_0 &\in \mathcal{H}_K \quad \text{(we consider } g_0 = 0\text{)}, \\ g_n &= \sum_{i=1}^{n} a_i K_{x_i}, \end{aligned}$$

with the sequence $(a_n)_n$ such that $a_n := -\gamma_n(g_{n-1}(x_n) - y_n) = -\gamma_n\left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_i\right)$. We denote $\bar{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} \bar{g}_k$.

A6: There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \sigma^2 L_K$, and $\mathbb{E}(||x_n||^2 x_n \otimes x_n) \preccurlyeq R^2 L_K$ where $\preccurlyeq$ denotes the order between self-adjoint operators.

A7: We denote $(\lambda_j)_{j\in\mathbb{N}}$ the sequence of eigenvalues of the operator L_K, in decreasing order. We assume $\frac{u^2}{j^\alpha} \leqslant \lambda_j \leqslant \frac{s^2}{j^\alpha}$ for some $\alpha > 1$.

A8: We assume that the objective function g_ρ lies in the RKHS and that its coordinates (ν_i) in the orthonormal eigenbasis of L_K (with respect to $||.||_K$) are such that $\nu_i \leqslant \left(\frac{1}{T\, i^{\frac{\beta}{2}}}\right)$.

Remark : As it has been noticed in the previous part, the two norms ($||.||_{\mathcal{L}_2}$ and $||.||_K$) are rather different : $||.||_K$ is the norm in the Hilbert space, so in our setting we don't expect to get results like $||g_n - g_\rho||_K \to 0$. However :

$$||g_n - g_\rho||^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} = \int_\mathcal{X} (g - g_\rho)(t)^2 d\rho_\mathcal{X}(t) = \mathbb{E}\left[(g-g_\rho)^2(X)\right] = \mathbb{E}\left[f(g) - f(g_\rho)\right]$$

which can also be seen as a consequence of lemma 1 as (using the isometry proved above) :

$$||g_n - g_\rho||^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} = ||L_K^{1/2}(g_n - g_\rho)||_K = \mathbb{E}\left[f(g) - f(g_\rho)\right].$$

Discussion: For any n, g_n is in $\text{span}\{K_{x_i}, 1 \leqslant i \leqslant t\}$, moreover, at any step n, the cost of calculating (a_n) is about n (thanks to the kernel trick). Thus, the complete cost of calculating $g_n(x)$ for an $x \in \mathcal{X}$ is about n^2. It must be noted that when studying regularized least square problem (i.e., adding a penalisation term), one has to update every coefficient $(a_i)_{1\leqslant i\leqslant t}$ at step t. More content on learning with kernels may be found in [12].
It may be noted that unlike in the Euclidean case, we could not try to take benefit of a knowledge of the covariance operator : the kernel trick can be used only for the classical stochastic gradient descent.

3.2.2 Summing up notations

In the following tabular, we establish a few links between the notations in the first part and in this one, in order to make it easier to understand how we may derive theorem 5 from proposition 2.

	Hilbert space	RKHS
Space :	$\mathcal{H}$	$\mathcal{H}_K$
Observations :	$y_n = \langle \theta_*, x_n\rangle + \varepsilon_n$	$y_n = g_\rho(x_n) + \varepsilon_n$
Objective :	θ_*	g_ρ
Vector :	x	K_x
Gradient :	$\gamma(\langle\theta_{n-1}, x_n\rangle x_n - z_n)$	$-(g_{n-1}(x_n) - y_n)K_{x_n}$
Covariance operator :	$\Sigma = \mathbb{E}[x_n \otimes x_n]$	L_K

3.2.3 Theorem

Theorem 5. *Assuming **A1-8**, we have*

If $\alpha + 1 > \beta$

$$\mathbb{E}\left[f\left(\bar{g}_n\right) - f(g_\rho)\right] = O\left(n^{-1+\frac{1}{\alpha+\beta}}\right), \qquad \textit{with } \gamma = \frac{\gamma_0}{n^{\frac{\beta}{\alpha+\beta}}}.$$

If $\alpha + 1 < \beta$

$$\mathbb{E}\left[f\left(\bar{g}_n\right) - f(g_\rho)\right] = O\left(n^{-1+\frac{1}{2\alpha+1}}\right), \qquad \textit{with } \gamma = \frac{\gamma_0}{n^{\frac{\alpha+1}{2\alpha+1}}}.$$

With γ_0 *a constant such that* $\gamma_0 < \frac{1}{\sup_t R(t,t)}$.

3.3 Example : splines on the circle

The best example to match our assumptions may be found in [13]. We consider $y = g_\rho(x) + \varepsilon$, with $x \sim \mathcal{U}[\,0;1]$, and g_ρ in a particular RKHS : $W_m^0(\text{per})$ the collection of all functions on $[\,0;1]$ of the form

$$f : t \mapsto \sqrt{2}\sum_{i=1}^{\infty} a_i(f)\cos(2\pi i t) + \sqrt{2}\sum_{i=1}^{\infty} b_i(f)\sin(2\pi i t),$$

with

$$\sum_{i=1}^{\infty}(a_i(f)^2 + b_i(f)^2)(2\pi i)^{2m} < \infty.$$

This means that the m-th derivative of f, $f^{(m)}$ is in $\mathcal{L}^2[0\,;1]$. We consider the inner product :

$$\langle f, g\rangle_{W_m^0} = \sum_{i=1}^{\infty}(2\pi i)^{2m}\left(a_i(f)a_i(g) + b_i(f)b_i(g)\right).$$

One can see that the reproducing kernel $R(s,t)$ for $W_m^0(\text{per})$ is

$$\begin{aligned} R(s,t) &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}[\cos(2\pi i s)\cos(2\pi i t) + \sin(2\pi i s)\sin(2\pi i t)] \\ &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}\cos(2\pi i(s-t)). \end{aligned}$$

Indeed, for any $s \in [\,0;1]$ we have :

$$\begin{aligned} \langle f, R_s\rangle_{W_m^0} &= \sum_{i=1}^{\infty}(2\pi i)^{2m}\left(a_i(f)a_i(R_s) + b_i(f)b_i(R_s)\right) \\ &= \sum_{i=1}^{\infty}(2\pi i)^{2m}\left(a_i(f)\frac{\sqrt{2}\cos(2\pi i s)}{(2\pi i)^{2m}} + b_i(f)\frac{\sqrt{2}\sin(2\pi i s)}{(2\pi i)^{2m}}\right) \\ &= \sqrt{2}\sum_{i=1}^{\infty}\left(a_i(f)\cos(2\pi i s) + b_i(f)\sin(2\pi i s)\right). \end{aligned}$$

The eigenvalues of the covariance operator are all of multiplicity 2 and are $\lambda_i = (2\pi i)^{-2m}$, and the eigenfunctions are $\Phi_i : t \mapsto \sqrt{2}\cos(2\pi i t)$ and $\Psi_i : t \mapsto \sqrt{2}\sin(2\pi i t)$. Indeed, for example :

$$\begin{aligned} L_K(\Phi_i)(s) &= \int_0^1 R(s,t)\sqrt{2}\cos(2\pi i t)dt = \left(\int_0^1 \frac{2}{(2i\pi)^{2m}}\sqrt{2}\cos(2\pi i t)^2 dt\right)\cos(2\pi i s) \\ &= \lambda_i\sqrt{2}\cos(2\pi i s) = \lambda_i\Phi_i(s). \end{aligned}$$

It's well known that $(\Phi_i, \Psi_i)_i$ is an orthogonal system in $\mathcal{L}^2[0;1]$ and it is easy to check that $((2i\pi)^{-m}\Phi_i, (2i\pi)^{-m}\Psi_i)_i$ is an orthogonal basis of our RKHS $(W_m^0,\ \langle .,. \rangle_{W_m^0})$. (this may also be seen as a consequence of the fact that $L_K^{1/2}$ is an isometry).
In the context above, we have our assumption **H7** with $\alpha = 2m$. Moreover, we can derive from the study of Bernoulli polynomials a close formula for $R(s,t)$:

$$R(s,t) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}\left(\{s-t\}\right),$$

with B_m denoting the m-th Bernoulli polynomial and $\{s-t\}$ the fractional part of $s-t$.
Moreover, considering $f(x) = \sum_{i=1}^{\infty} \frac{1}{i^{\frac{\beta}{2}}} \frac{\sqrt{2}}{(2i\pi)^m} cos(2i\pi x)$, our hypothesis **H8** stands.
We give in figure 2 the functions used for simulations in a few cases.

α	β	m	f	$K \quad (u := \{s-t\})$
2	2	1	$g_{\rho,2,2}(x) = \frac{1}{\sqrt{2}}\pi\left(x^2 - x + \frac{1}{6}\right)$	$K_2(s,t) = \frac{1}{2}\left(u^2 - u + \frac{1}{6}\right)$
2	6	1	$g_{\rho,2,6}(x) = \frac{-\pi^3}{3\sqrt{2}}\left(x^4 - 2x^3 + x^2 - \frac{1}{30}\right)$	$K_2(s,t) = \frac{1}{2}\left(u^2 - u + \frac{1}{6}\right)$
6	2	3	$g_{\rho,6,2}(x) = \frac{-\pi}{6}\left(x^4 - 2x^3 + x^2 - \frac{1}{30}\right)$	$K_6(s,t) = \frac{1}{6!}\left(u^6 - 3u^5 + \frac{5}{2}u^4 - \frac{1}{2}u^2 + \frac{1}{42}\right)$

Figure 2: A few examples of different kernels and objective functions matching with different α, β our assumptions **H7,8**.

3.4 Optimality of the result

It's interesting to link our results to what has been done in [2] and [3] in the case of regularized least mean square, and with the results presented in [14] which deals with the best we can hope on our purpose. This can be done with little work trying to match the different hypothesis.
In [3] the following result is proved (*Remark 2.8* following *Theorem C*) :

Theorem 6. *Assume that $L_K^{-r} g_\rho \in \mathcal{L}^2_{\rho(X)}$ for some $r \in [1/2, 1]$. Then with probability at least $1-\delta$, for all $t \in \mathbb{N}$,*

$$f(g_n) - f(g_\rho) \leqslant O\left(n^{-2r/(2r+1)}\right).$$

The assumption are not exactly the same as no assumption is made on the covariance matrix, but only on $L_K^{-r} g_\rho$.

In [14] a minimax lower bound was given in *Theorem 2* : let $\mathcal{P}(\alpha, r)$ $(\alpha > 1, r \in [1/2, 1])$ be the set of all probability measures ρ on $\mathcal{X} \times \mathcal{Y}$, such that :

- a.s., $|y| \leqslant M_\rho$,
- $L_K^{-r} g_\rho \in \mathcal{L}^2_{\rho(X)}$,
- the eigenvalues $(\mu_j)_{j \in \mathbb{N}}$ arranged in a non increasing order, are subject to the decay $\mu_j = O(n^{-j})$.

Then the following minimax lower rate stands :

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

for some constant $C > 0$ where the infimum in the middle is taken over all algorithms as a map $((x_i, y_i)_{1 \leqslant i \leqslant n}) \mapsto f_n \in \mathcal{H}_K$.

3.5 What if $g_\rho \notin \mathcal{H}$

In most of what we have done before, we assume that g_ρ is in the RKHS : that is $\beta > 1$. But how do we have to choose our RKHS and what does happen if it is chosen to small ? A first point to notice is that

$$g_\rho \in \mathcal{H}_K \Longleftrightarrow \exists r \geq \frac{1}{2}, \quad L_K^{-r} g_\rho \in \mathcal{L}_\rho^2 \tag{1}$$

However the proof of lemma 6 relies on $r \geq 0$ the inequality on the bias remains true even if $g_\rho \notin \mathcal{H}_K$.

Theorem 7. *blabla*

How to choose the rkhs How to illlustrate it : avec 1/n rac n : rkha peut etre a=2 b=1 ou a =3 (?) b=0. il faudrait avoir une bonne illustration qu'on trouve la decroissance attendue, et pouvoir comparer les deux noyaux.

4 Experiments on artificial data.

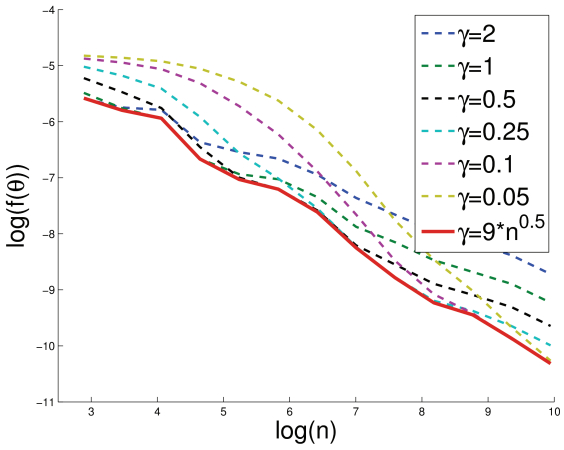
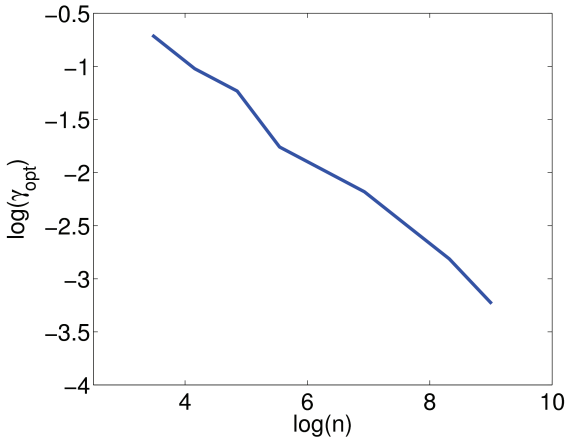
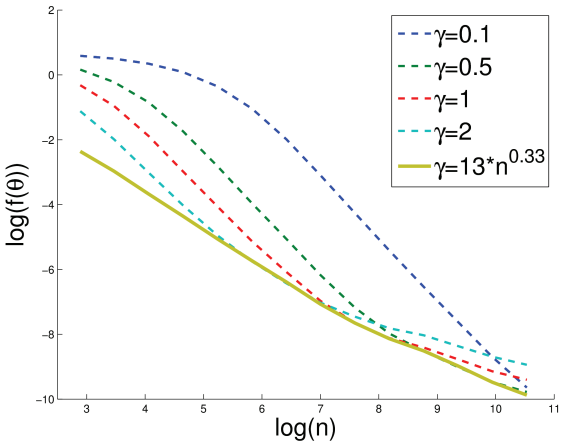
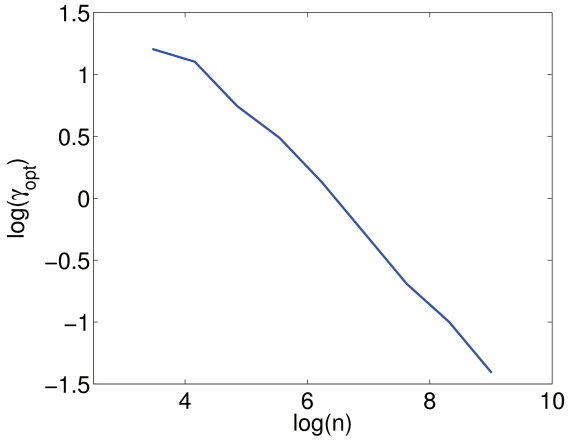
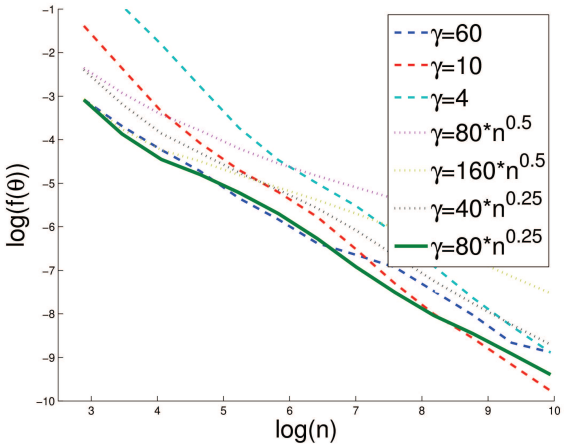
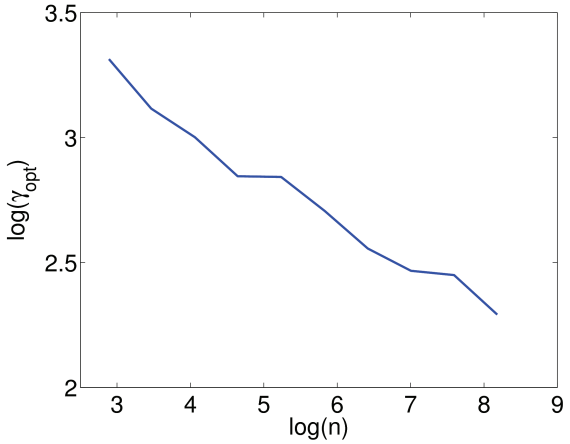
We may illustrate what has been done above in different RKHS as above : W_1^0, W_3^0, with different functions g_ρ, leading to a panel of assumptions **H7,8**.

We are interested in representing :

- the decrease in the generalisation error, as a function of the number of training examples we have, for different constant steps : which may depend on n or not. It must be noticed that when we're using a constant step which is typically a function of n, we do not have an "online" algorithm, as we would have to recompute the entire gradient descent in order to use a bigger number of examples.
- the best choice of γ as a constant step when trying to get the smallest error with a certain number of training examples n.

All the plots will be in logarithmic scale, in order to be able to evaluate the decrease as a power of n.

We use the different functions and kernels as described in figure 2.

α, β	loss for diff.γ slope of the error for γ_{pred} : found / theoretic	γ optimal as a function of n slope of γ_{opt} found / theoretic
2,2	 0.71/0.75	 0.45 / 0.5
2,6	 0.8/0.96	 0.5 / 0.6
6,2	 0.87/0.87	 0.19 / 0.16

When we choose $\gamma_n = \gamma_0 n^{-p}$ we have to take $\gamma_0 \leqslant \frac{1}{R^2}$ with $R^2 = K(0,0)$. When $\alpha = 2$ this leads

to $\gamma_0 \leqslant 12$ and $\alpha = 6$ leads to $\gamma_0 \leqslant 30240$.

5 Proof of Proposition 1

We have to introduce a few lemmas to prove proposition 1. The lemmas are given in section 5.1, the proofs in section 5.2.

5.1 Lemmas

In the following, we will use constantly the following observation :

Lemma 1. *Assume **H3-4** , let $\eta_n = \theta_n - \theta_*$, $\bar{\eta}_n = \bar{\theta}_n - \bar{\theta}_*$:*

$$\begin{aligned} f(\theta_n) - f(\theta_*) &= \frac{1}{2}\langle \eta_n, \Sigma\eta_n\rangle = \frac{1}{2}\langle \theta_n - \theta_*, \Sigma(\theta_n - \theta_*)\rangle = \frac{1}{2}\mathbb{E}\left[\langle x, \theta_n - \theta_*\rangle^2\right] \left(:= \frac{1}{2}||\theta_n - \theta_*||^2_{\mathcal{L}^2_{\rho(X)}}\right), \\ f\left(\bar{\theta}_n\right) - f(\theta_*) &= \frac{1}{2}\langle \bar{\eta}_n, \Sigma\bar{\eta}_n\rangle. \end{aligned}$$

Therefore all the following proofs rely on calculations made on the right side quantity. The two next lemma's proofs may be found in [4].

Lemma 2. *Assume $(x_n, \xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\xi_n|\mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[||\xi_n||^2|\mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[||x_n||^2 x_n \otimes x_n|\mathcal{F}_{n-1}\right] \preccurlyeq R^2\Sigma$, with $\mathbb{E}\left[x_n \otimes x_n|\mathcal{F}_{n-1}\right] = \Sigma$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator Σ. Consider the recursion $\alpha_n = (I - \gamma x_n \otimes x_n)\alpha_{n-1} + \gamma\xi_n$, with $\gamma R^2 \leqslant 1$. Then :*

$$(1 - \gamma R^2)\,\mathbb{E}\left[\left\langle \bar{\alpha}_{n-1}, \Sigma\bar{\alpha}_{n-1}\right\rangle\right] + \frac{1}{2n\gamma}\mathbb{E}||\alpha_n||^2 \leqslant \frac{1}{2n\gamma}||\alpha_0||^2 + \frac{\gamma}{n}\sum_{k=1}^{n}\mathbb{E}||\xi_k||^2.$$

Especially, if $\alpha_0 = 0$, we have

$$\mathbb{E}\left[\left\langle \bar{\alpha}_{n-1}, \Sigma\bar{\alpha}_{n-1}\right\rangle\right] \leqslant \frac{1}{(1 - \gamma R^2)}\frac{\gamma}{n}\sum_{k=1}^{n}\mathbb{E}||\xi_k||^2.$$

Lemma 3. *Assume $(\xi_n) \in \mathcal{H}$ is $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\xi_n|\mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[||\xi_n||^2|\mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\xi_n \otimes \xi_n\right] \preccurlyeq C$. Consider the recursion $\alpha_n = (I - \gamma\Sigma)\alpha_{n-1} + \gamma\xi_n$, with $\gamma\Sigma \preccurlyeq I$ for some invertible Σ. Then :*

$$\mathbb{E}\left[\alpha_n \otimes \alpha_n\right] = (I - \gamma\Sigma)^n\alpha_0 \otimes \alpha_0(I - \gamma\Sigma)^n + \gamma^2\sum_{k=1}^{n}(I - \gamma\Sigma)^{n-k}C(I - \gamma\Sigma)^{n-k},$$

and :

$$\mathbb{E}\left[\left\langle \bar{\alpha}_{n-1}, \Sigma\bar{\alpha}_{n-1}\right\rangle\right] = \frac{1}{n^2}\Big\langle \sum_{k=0}^{n-1}\left(I - \gamma\Sigma\right)^k\alpha_0, \sum_{k=0}^{n-1}\left(I - \gamma\Sigma\right)^k\Sigma\alpha_0\Big\rangle + \frac{1}{n^2}\sum_{j=1}^{n-1}\mathrm{tr}\left(\left[I - (I - \gamma\Sigma)^{n-j}\right]^2\Sigma^{-1}C\right).$$

In **lemma 3**, in the last inequality, the first term shows the impact of the initial setting and the hardness to forget the initial condition, whereas the second one stands for the effect of the noise. Thus the first one tends to decrease when γ is increasing, whereas the second one increases when γ increases. We understand we may have to choose our step γ in order to optimize the tradeoff between these two factors.

In the finite-dimensional case, it results from the last inequallity of lemma 3 that if $C = \sigma^2\Sigma$ then $\mathbb{E}\left[\left\langle \bar{\alpha}_{n-1}, \Sigma\bar{\alpha}_{n-1}\right\rangle\right] \leqslant \frac{1}{n\gamma}||\alpha||_0^2 + \frac{\sigma^2 d}{n}$. But of course this majoration is vacuous when d is either

large or infinite. Making further assumptions, namely **H7** and **H8**, we can derive comparable bounds in the infinite-dimensional setting.

The next two lemmas are exactly these two bounds, and result of a calculation given below.

Lemma 4 (Variance majoration)**.** *Assume **H6-7**, let α be the constant in **H7** and assume $C = \sigma^2\Sigma$ in **H6**. We consider :*

$$\mathrm{Var}(n,\gamma,\alpha,\sigma^2) := \frac{1}{n^2}\sum_{j=1}^{n-1} \mathrm{tr}\left(\left[I-(I-\gamma\Sigma)^{n-j}\right]^2\Sigma^{-1}C\right)$$

We have :

$$(1-e^{-1})^2\ C(\alpha)\ s^{2/\alpha}\ \sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{(n-1)^{1-\frac{1}{\alpha}}} \leqslant \mathrm{Var}(n,\gamma,\alpha,\sigma^2) \leqslant C(\alpha)\ s^{2/\alpha}\ \sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n},$$

with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$. If we only have $C \preccurlyeq \sigma^2\Sigma$, only the majoration stands.

Lemma 5 (Bias Majoration)**.** *Assume **H7-8**, , let α (resp. β) be the constant in **H7** (resp. **H8**): we denote* $\mathrm{Bias}(n,\gamma,\alpha,\beta) := \frac{1}{n^2}\left\langle \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\alpha_0, \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma\alpha_0\right\rangle$.

If** $\alpha+1>\beta$ **:

$$\mathrm{Bias}(n,\gamma,\alpha,\beta) \leqslant \frac{s^{2\left(\frac{-\beta+\alpha+1}{\alpha}\right)}}{T^2u^2}\ 2\ K(\alpha,\beta)\left(\frac{1}{(n\gamma)^{\frac{\alpha+\beta-1}{\alpha}}}\right),$$

with $K(\alpha,\beta) = \frac{1}{\alpha+\beta-1} + \frac{1}{\alpha+1-\beta}$.

If** $\alpha+1<\beta$ **:

$$\mathrm{Bias}(n,\gamma,\alpha,\beta) \leqslant \frac{1}{T^2u^2}\left(\frac{1}{\beta-\alpha-1}+\frac{2}{\alpha+\beta-1}\right)\left(\frac{1}{(n\gamma)^2}\right).$$

Lemma 6 (Bias majoration with **H7', H8'**)**.** *Assume **H7', H8'** and let α (resp. r) be the constant in **H7'** (resp. **H8'**) : we denote* $\mathrm{Bias}(n,\gamma,\alpha,r) := \frac{1}{n^2}\left\langle \sum_{k=0}^{n}(I-\gamma\Sigma)^k\alpha_0, \sum_{k=0}^{n}(I-\gamma\Sigma)^k\ \ \Sigma\alpha_0\right\rangle$.
***If** $r \leqslant 1$:*

$$\mathrm{Bias}(n,\gamma,\alpha,r) \leqslant ||\Sigma^{-r}\theta_*||^2_{\mathcal{L}^2}\left(\frac{1}{(n\gamma)^{2r}}\right).$$

***If** $r \geq 1$:*

$$\mathrm{Bias}(n,\gamma,\alpha,r) \leqslant ||\Sigma^{-r}\theta_*||^2_{\mathcal{L}^2}\left(\frac{1}{(n\gamma)^{2}}\right).$$

5.2 Proofs

*Proof : **Lemma 1**.*

$$\begin{aligned}
f(\theta_n)-f(\theta_*) &= \frac{1}{2}\mathbb{E}\left[\langle\theta_n,x\rangle^2 - 2\langle\theta_n,z\rangle\right] - \frac{1}{2}\mathbb{E}\left[\langle\theta_*,x\rangle^2-2\langle\theta_*,z\rangle\right]\\
&= \frac{1}{2}\left(\mathbb{E}\left[\langle\theta_n,x\rangle^2 - 2\langle\theta_n,\langle\theta_*,x\rangle x+\xi\rangle\right] - \mathbb{E}\left[\langle\theta_*,x\rangle^2-2\langle\theta_*,\langle\theta_*,x\rangle x+\xi\rangle\right]\right)\\
&= \frac{1}{2}\left(\mathbb{E}\left[\langle\theta_n,x\rangle^2 - 2\langle\theta_*,x\rangle\langle\theta_n,x\rangle + \langle\theta_*,x\rangle^2\right]\right)\\
&= \frac{1}{2}\mathbb{E}\left[\langle x,\bar\theta_n-\theta_*\rangle^2\right] = \frac{1}{2}\mathbb{E}\left[\langle x,\eta_n\rangle^2\right]\\
&= \frac{1}{2}\mathbb{E}\left[\langle\eta_n, x_n\otimes x_n\eta_n\rangle\right]\\
&= \frac{1}{2}\langle\eta_n,\Sigma\eta_n\rangle = \frac{1}{2}\langle\bar\theta_n-\theta_*,\Sigma(\bar\theta_n-\theta_*)\rangle = \left(:=\frac{1}{2}||\bar\theta_n-\theta_*||^2_{\mathcal{L}^2_{\rho(X)}}\right)
\end{aligned}$$

Using the facts that $z = \langle \theta_*, x\rangle x + \xi$ and $\mathbb{E}\langle \theta_n, \xi\rangle = 0$. All the expectations in this calculation are expectations on x, y, ξ (the stochasticity of θ_n is not considered). In fact this calculation is true for any θ (with $\eta = \theta - \theta_*$) f.e. $\bar{\theta}_n$.

□

Proposition 1.
To prove proposition 1 we have to understand how and why we can use the lemmas above. We remind that we assume **H1', H2-4, H5' H6-8**. We study $\eta_n = \theta_n - \theta_*$. By **H5'**, η_n satisfies $\eta_0 = \theta_*$ and for any $n \geqslant 1$, $\eta_n = (I - \gamma\Sigma)\eta_{n-1} + \gamma\xi'_n$, with $\xi'_n = z_n - \Sigma\theta_*$. For any n, ξ'_n is $\mathcal{F}_n$ mesurable, with $\mathcal{F}_n = \sigma((x_i, z_i)_{1\leqslant i\leqslant n})$. Moreover, $\mathbb{E}\left[\xi'_n|\mathcal{F}_{n-1}\right] = 0$ (with **H2**, $\mathbb{E}\left[\xi'_n|\mathcal{F}_{n-1}\right] = \mathbb{E}\left[\xi'_n\right]$ and with **H4**, $\mathbb{E}\left[\xi'_n\right] = \mathbb{E}\left[\xi_n\right] = 0$). $\mathbb{E}||\xi'_n||^2$ is finite (**H3**) and finaly $\mathbb{E}\left[\xi'_n \otimes \xi'_n\right] \preccurlyeq C$ using **H6** and assuming $\mathbb{E}\left[\xi_n|x_n\right] = 0$ (which is true for example in the least squares setting with an iid noise) : $\xi'_n = \xi_n + (x_n \otimes x_n - \Sigma)\theta_*$ So :

$$\mathbb{E}\left[\xi'_n \otimes \xi'_n\right] = \mathbb{E}\left[\xi_n \otimes \xi_n\right] + \mathbb{E}\left[(x_n \otimes x_n - \Sigma)\theta_* \otimes \theta_*(x_n \otimes x_n - \Sigma)\right] + \mathbb{E}\left[\xi_n \otimes \theta_*(x_n \otimes x_n - \Sigma)\right] + \mathbb{E}\left[(x_n \otimes x_n - \Sigma)\theta_* \otimes \xi_n\right].$$

And we have :

$$\mathbb{E}\left[(x_n \otimes x_n - \Sigma)\theta_* \otimes \theta_*(x_n \otimes x_n - \Sigma)\right] \leqslant ||\theta_*||^2 R^2 \Sigma.$$

For any γ such as $\gamma\Sigma \preccurlyeq I$ we may apply lemme 3. That is (as $\theta_0 = 0$) :

$$\mathbb{E}\left[\left\langle \bar{\eta}_{n-1}, \Sigma\bar{\eta}_{n-1}\right\rangle\right] = \frac{1}{n\gamma}\left\langle \theta_*, \left[I - (I-\gamma\Sigma)^n\right]^2 (n\gamma\Sigma)^{-1}\theta_*\right\rangle + \frac{1}{n^2}\sum_{j=1}^{n-1} \operatorname{tr}\left(\left[I - (I-\gamma\Sigma)^{n-j}\right]^2 \Sigma^{-1} C\right).$$

Under **H7-8** we can apply lemmas 4 and 5 and we get proposition 1 using lemma 1. □

Corollary 1 .
We just have to choose γ in order to get the best compromise. In order to balance both terms, the best choice for γ is $\alpha + 1 > \beta$ then $\gamma = \frac{1}{n^{\frac{\beta}{\alpha+\beta}}}$, else if $\alpha + 1 < \beta$ then $\gamma = \frac{1}{n^{\frac{\alpha+1}{2\alpha+1}}}$.
These choices lead to the majorations given in Corollary 1. □

Lemma 4 .
In the following proof, we consider $s = 1$. It's easy to get the complete result replacing in the proof below " γ" by " $s^2\gamma$".
We have, for $j \in \mathbb{N}$, still assuming $\gamma\Sigma \preccurlyeq I$, and by a comparison to the integral :

$$\begin{aligned}
\operatorname{tr}\left(I - (I-\gamma\Sigma)^j\right)^2 \Sigma^{-1} C &= \sigma^2 \operatorname{tr}\left(I - (I-\gamma\Sigma)^j\right)^2 \\
&\leqslant 1 + \sigma^2 \int_{u=1}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du \\
&\quad \text{(1 stands for the first term in the sum)} \\
&= 1 + \sigma^2 \int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du + \sigma^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du.
\end{aligned}$$

Note that the first integral may be empty if $\gamma j \leqslant 1$. We also have:

$$\operatorname{tr}\left(I - (I-\gamma\Sigma)^j\right)^2 \Sigma^{-1} C \geqslant \sigma^2 \int_{u=1}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du.$$

Considering that $g_j : u \mapsto \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2$ is a decreasing fonction of u we get :

$$\forall u \in [1; (\gamma j)^{\frac{1}{\alpha}}], \quad (1 - e^{-1})^2 \leqslant g_j(u) \leqslant 1.$$

Where we have used the fact that $\left(1 - \frac{1}{j}\right)^j \leqslant e^{-1}$ for the left hand side inequality. Thus we have proved :

$$(1 - e^{-1})^2 (\gamma j)^{\frac{1}{\alpha}} \leqslant \int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du \leqslant (\gamma j)^{\frac{1}{\alpha}}.$$

For the other part of the sum, we consider $h_j : u \mapsto \left(\frac{1-\left(1-\frac{\gamma}{u^\alpha}\right)^j}{\frac{\gamma}{u^\alpha}}\right)^2$ which is an increasing function of u. So :

$$\forall u \in [(\gamma j)^{\frac{1}{\alpha}}; +\infty], \quad (1 - e^{-1})^2 j^2 \leqslant h_j(u) \leqslant j^2,$$

using the same trick as above. Thus :

$$\begin{aligned}
\int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du &= \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} h_j(u) \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2 \gamma^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{1}{u^\alpha}\right)^2 du \\
&= j^2 \gamma^2 \left[\frac{1}{(1-2\alpha) u^{2\alpha-1}}\right]_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \\
&= j^2 \gamma^2 \frac{1}{(2\alpha-1)((\gamma j)^{\frac{1}{\alpha}})^{2\alpha-1}} \\
&= \frac{1}{(2\alpha-1)} (j\gamma)^{\frac{1}{\alpha}}
\end{aligned}$$

And we could get, by a similar calculation :

$$\int_{u=(\gamma j)^{\frac{1}{\alpha}}+1}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du \geqslant (1 - e^{-1})^2 \frac{1}{(2\alpha-1)} (j\gamma)^{\frac{1}{\alpha}}$$

Finaly, we have shown that :

$$C_1 (j\gamma)^{\frac{1}{\alpha}} \leqslant \operatorname{tr}\left(I - (I - \gamma\Sigma)^j\right)^2 \leqslant C_2 (j\gamma)^{\frac{1}{\alpha}} + 1.$$

Where $C_1 = (1 - e^{-1})^2 \left(1 + \frac{1}{(2\alpha-1)}\right)$ and $C_2 = \left(1 + \frac{1}{(2\alpha-1)}\right)$ are real constants. To get the complete variance term we have to calculate : $\frac{\sigma^2}{n^2} \sum_{j=1}^{n-1} \operatorname{tr}\left(I - (I - \gamma\Sigma)\right)^j$. We have :

$$\begin{aligned}
\frac{\sigma^2}{n^2} \sum_{j=1}^{n-1} \operatorname{tr}\left(I - (I - \gamma\Sigma)^j\right)^2 &\leqslant \frac{\sigma^2}{n^2} \sum_{j=1}^{n-1} \left(C_2 (j\gamma)^{\frac{1}{\alpha}} + 1\right) \\
&\leqslant \frac{\sigma^2}{n^2} C_2 \gamma^{\frac{1}{\alpha}} \int_{u=2}^{n} u^{\frac{1}{\alpha}} du + \frac{\sigma^2}{n} \\
&\leqslant \frac{\sigma^2}{n^2} C_2 \gamma^{\frac{1}{\alpha}} \frac{\alpha}{\alpha+1} n^{\frac{\alpha+1}{\alpha}} + \frac{\sigma^2}{n} \\
&\leqslant \frac{\alpha\ \sigma^2\ C_2}{\alpha+1}\ \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n}.
\end{aligned}$$

That is :

$$(1-e^{-1})^2\ C(\alpha)\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{(n-1)^{1-\frac{1}{\alpha}}} \leqslant \mathrm{Var}(n,\gamma,\alpha,\sigma^2) \leqslant C(\alpha)\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n},$$

with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$.

If we only assume $C \preccurlyeq \sigma^2\Sigma$, we clearly keep the majoration. □

Lemma 5.

The calculations for the bias term will be essentially the same as above, even if the bias is dependent of the initial condition. In the following we denote :

$$\mathrm{Bias}(n,\gamma,\alpha,\beta) = \frac{1}{n^2}\Big\langle \sum_{k=0}^{n-1} (I-\gamma\Sigma)^k\,\alpha_0, \sum_{k=0}^{n-1} (I-\gamma\Sigma)^k\,\Sigma\alpha_0 \Big\rangle$$

Using **H7-8** and calculating in the (orthonormal) eigenbasis, we have :

$$\begin{aligned}
\mathrm{Bias}(n,\gamma,\alpha,\beta) &= \frac{1}{n^2}\sum_{i=1}^{\infty}\left(\sum_{k=0}^{n-1}(1-\gamma\mu_i)^k\,(\alpha_0)_i\right)\left(\sum_{k=0}^{n-1}(1-\gamma\mu_i)^k\,\mu_i\,(\alpha_0)_i\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{\infty}\left(\frac{1-(1-\gamma\mu_i)^n}{\gamma\mu_i}\right)^2\mu_i\,(\alpha_0)_i^2\\
&\leqslant \frac{1}{n^2}\sum_{i=1}^{\infty}\frac{\left(1-\left(1-\frac{\gamma s^2}{i^\alpha}\right)^n\right)^2}{\gamma^2\frac{u^2}{i^\alpha}}\frac{1}{T^2 i^\beta}\\
&\leqslant \frac{1}{\gamma^2 n^2}\frac{1}{T^2u^2}\sum_{i=1}^{\infty}\left(1-\left(1-\frac{\gamma s^2}{i^\alpha}\right)^n\right)^2\frac{i^\alpha}{i^\beta}.
\end{aligned}$$

Let's first try to find an upper bound for

$$S(n,\tilde\gamma,\alpha,\beta) := \sum_{i=1}^{\infty}\left(1-\left(1-\frac{\tilde\gamma}{i^\alpha}\right)^n\right)^2\frac{i^\alpha}{i^\beta}.$$

The problem we have to deal with here is that $u \mapsto \frac{u^\alpha}{u^\beta}\Big(1-\big(1-\frac{\tilde\gamma}{u^\alpha}\big)^n\Big)^2$ is not a non increasing function and so we can't compare it to its integral. However :

$$S(n,\tilde\gamma,\alpha,\beta) = \underbrace{\sum_{k=1}^{\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\left(1-\left(1-\frac{\tilde\gamma}{k^\alpha}\right)^n\right)^2}_{S_1} + \underbrace{\tilde\gamma^2\sum_{k=\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}\left[\frac{\left(1-\left(1-\frac{\tilde\gamma}{k^\alpha}\right)^n\right)}{\frac{\tilde\gamma}{k^\alpha}}\right]^2}_{S_2}$$

First, using the functions defined above g_n and h_n we have :

Behaviour of S_1 :

$$\begin{aligned}
S_1 &= \sum_{k=1}^{\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\left(1-\left(1-\frac{\tilde\gamma}{k^\alpha}\right)^n\right)^2 = \sum_{k=1}^{\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\,g_n(k),\\
\text{so that}\quad S_1 &\leqslant \sum_{k=1}^{\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta},\\
\text{and}\quad S_1 &\geqslant (1-e^{-1})^2\sum_{k=1}^{\lfloor(\tilde\gamma n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}.
\end{aligned}$$

$*$ If $\alpha+1>\beta$:

$$\sum_{k=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\leqslant\int_{u=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor+1}u^{\alpha-\beta}du\leqslant\frac{2}{-\beta+\alpha+1}(n\tilde{\gamma})^{\frac{-\beta+\alpha+1}{\alpha}}.$$

$*$ If $\alpha+1=\beta$: a supplementary logarithmic term would appear, we won't consider this case in the following.
$*$ But if $\alpha+1<\beta$, we have :

$$\sum_{k=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\leqslant\int_{u=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor+1}\frac{1}{u^{\beta-\alpha}}du\leqslant\int_{u=1}^{\infty}\frac{1}{u^{\beta-\alpha}}du\leqslant\frac{1}{\beta-\alpha-1}.$$

And

$$\sum_{k=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}\frac{k^\alpha}{k^\beta}\geqslant\int_{u=1}^{\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}\frac{1}{u^{\beta-\alpha}}du\geqslant\frac{1}{\beta-\alpha-1}\left(1-\frac{1}{(n\tilde{\gamma})^{\beta-\alpha-1}}\right)\geqslant\frac{1}{\beta-\alpha-1}C.$$

As $\beta-\alpha-1>0$ and $\tilde{\gamma}n$ is decreasing with n (i.e., $\tilde{\gamma}$'s dependence on n is greater than n^{-1}). However, one must notice that our analysis is made for α, β fixed : the minoration claimed below contains a constant C depending on α and β and there's no contradiction when $\beta\to\alpha+1$ as the constant $C\to 0$.

Behaviour of S_2 :

$$\begin{aligned}
S_2 &= \tilde{\gamma}^2\sum_{k=\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}\left[\frac{\left(1-\left(1-\frac{\tilde{\gamma}}{k^\alpha}\right)^n\right)}{\frac{\tilde{\gamma}}{k^\alpha}}\right]^2=\tilde{\gamma}^2\sum_{k=\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}\;h(k)\;,\\
\text{so that}\quad S_2 &\leqslant n^2\tilde{\gamma}^2\sum_{k=\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}\;,\\
\text{and}\quad S_2 &\geqslant (1-e^{-1})^2\;n^2\tilde{\gamma}^2\sum_{k=\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}.
\end{aligned}$$

And

$$\sum_{k=\lfloor(\tilde{\gamma}n)^{\frac{1}{\alpha}}\rfloor}^{\infty}\frac{1}{k^{\beta+\alpha}}\leqslant\frac{2}{\alpha+\beta-1}(n\tilde{\gamma})^{\frac{-\alpha-\beta+1}{\alpha}}.$$

Behaviour of $\mathrm{Bias}(n,\gamma,\alpha,\beta)$ **:** We just have to compute the results of the two paragraphs above :

If $\alpha+1>\beta$: We get :

$$S(n,\tilde{\gamma},\alpha,\beta)\leqslant\frac{2}{-\beta+\alpha+1}(n\tilde{\gamma})^{\frac{-\beta+\alpha+1}{\alpha}}+\frac{2}{\alpha+\beta-1}(n\tilde{\gamma})^{2+\frac{-\alpha-\beta+1}{\alpha}}.$$

So that :

$$\begin{aligned}
\mathrm{Bias}(n,\gamma,\alpha,\beta) &\leqslant \frac{1}{\gamma^2n^2}\frac{1}{T^2u^2}S(n,s^2\gamma,\alpha,\beta)\\
&\leqslant \frac{s^{2\frac{-\beta+\alpha+1}{\alpha}}}{T^2u^2}\left(\frac{2}{\alpha+\beta-1}+\frac{2}{\alpha+1-\beta}\right)\left(\frac{1}{(n\gamma)^{\frac{\alpha+\beta-1}{\alpha}}}\right).
\end{aligned}$$

And a similar minoration with the constant $(1-e^{-1})^2\left(\frac{1}{\alpha+\beta-1}+\frac{C}{\alpha+1-\beta}\right)$. Asymptotically :

$$\text{Bias}(n,\gamma,\alpha,\beta)=\Theta\left(\frac{1}{(n\gamma)^{\frac{\alpha+\beta-1}{\alpha}}}\right).$$

If $\alpha+1<\beta$ **:** In this case, $S_2=o\left(S_1\right)$:

$$S(n,\tilde{\gamma},\alpha,\beta)\leqslant\frac{1}{\beta-\alpha-1}+\frac{2}{\alpha+\beta-1}\frac{1}{(n\tilde{\gamma})^{\frac{\beta-\alpha-1}{\alpha}}}.$$

So that :

$$\begin{aligned}\text{Bias}(n,\gamma,\alpha,\beta) &\leqslant \frac{1}{\gamma^2n^2}\frac{1}{T^2u^2}S(n,s^2\gamma,\alpha,\beta)\\ &\leqslant \frac{1}{T^2u^2}\left(\frac{1}{\beta-\alpha-1}+\frac{2}{\alpha+\beta-1}\frac{1}{(n\gamma s^2)^{\frac{\beta-\alpha-1}{\alpha}}}\right)\left(\frac{1}{(n\gamma)^2}\right).\end{aligned}$$

Which implies :

$$\text{Bias}(n,\gamma,\alpha,\beta) \leqslant \frac{1}{T^2u^2}\left(\frac{1}{\beta-\alpha-1}+\frac{2}{\alpha+\beta-1}\right)\left(\frac{1}{(n\gamma)^2}\right).$$

And we could derive a similar minoration under the natural supplementary hypothesis. That is, asymptotically :

$$\text{Bias}(n,\gamma,\alpha,\beta)=\Theta\left(\frac{1}{(n\gamma)^2}\right).$$

□

We see that we'll have to make a distinction between the two cases. In fact these asymptotics show that the bias term can never decrease farther than $\frac{1}{(\gamma n)^2}$. As long as $\beta<\alpha+1$ the decrease will improve as β grows, but there won't be any improvement after $\alpha+1$ whatever β might be.

Lemma 6.

If $0\leqslant r\leqslant 1$:

$$\begin{aligned}\text{Bias}(n,\gamma,\alpha,r) &= \frac{1}{n^2}\Big\langle\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\alpha_0,\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\ \ \Sigma\alpha_0\Big\rangle\\ &= \frac{1}{n^2}\Big\langle\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^{2r}\Sigma^{-r+1/2}\alpha_0,\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\ \ \Sigma^{-r+1/2}\alpha_0\Big\rangle\\ &= \frac{1}{n^2}\left\|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^r(\Sigma^{-r+1/2}\alpha_0)\right\|^2\\ &\leqslant \frac{1}{n^2}\left\|\left\|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\Sigma^r\right\|\right\|^2\ \ \left\|\Sigma^{-r+1/2}\alpha_0\right\|^2\\ &= \frac{1}{n^2}\gamma^{-2r}\left\|\left\|\sum_{k=0}^{n-1}(I-\gamma\Sigma)^k\gamma^r\Sigma^r\right\|\right\|^2\ \ \left\|\Sigma^{-r}\alpha_0\right\|^2_{\mathcal{L}^2_\rho}\\ &\leqslant \frac{1}{n^2}\gamma^{-2r}\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^kx^r\right)^2\left\|\Sigma^{-r}\alpha_0\right\|^2_{\mathcal{L}^2_\rho}\\ &\leqslant \left(\frac{1}{(n\gamma)^{2r}}\right)\left\|\Sigma^{-r}\alpha_0\right\|^2_{\mathcal{L}^2}.\end{aligned}$$

Using the inequality : $\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right)\leqslant n^{1-r}$: indeed

$$\begin{aligned}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) &= \frac{1-(1-x)^n}{x}x^r\\ &= (1-(1-x)^n)x^{r-1}\end{aligned}$$

And we have, for any $n\in\mathbb{N}, r\in[\,0;1], x\in[\,0;1]$: $(1-(1-x)^n)\leqslant(nx)^{1-r}$:

1. if $nx\leqslant 1$ then $(1-(1-x)^n)\leqslant nx\leqslant(nx)^{1-r}$ (the first inequality can be proved by deriving the difference).
2. if $nx\geq 1$ then $(1-(1-x)^n)\leqslant 1\leqslant(nx)^{1-r}$.

If $r\geq 1$, $x\mapsto(1-(1-x)^n)$ is increasing on $[\,0;1]$ so $\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right)=1$: there is no improvement in comparison to $r=1$:

$$\text{Bias}(n,\gamma,\alpha,r) \leqslant \left(\frac{1}{(n\gamma)^2}\right)\left|\left|\Sigma^{-r}\alpha_0\right|\right|^2_{\mathcal{L}^2}.$$

□

6 Proof of Proposition 2

Here we no longer consider the case where the covariance operator Σ is known, so that we've got to use $x_n\otimes x_n$ in the gradient descent. We may note that in the RKHS setting, we have no choice to use Σ as the kernel trick can be used only to calculate with $x_n\otimes x_n$. We're going to use the same proof as in [4], with a few adaptations to keep the improvement due to our stronger hypothesis (the decreases in θ_* and Σ's eigenvalues).

6.1 Proof principle

We remind that $(\eta_n)_n$ is defined by :

$$\eta_0=\theta_0-\theta_*, \text{ and the recusion } \eta_n=(I-\gamma x_n\otimes x_n)\eta_{n-1}+\gamma\xi_n.$$

We consider $(\eta_n^{init})_n$ defined by :

$$\eta_0^{init}=\theta_0-\theta_* \text{ and } \eta_n^{init}=(I-\gamma x_n\otimes x_n)\eta_{n-1}^{init}.$$

η_n^{init} is the part of $(\eta_n)_n$ which is due to the initials conditions.
Respectively, let $(\eta_n^{noise})_n$ be defined by :

$$\eta_0^{noise}=0 \text{ and } \eta_n^{noise}=(I-\gamma x_n\otimes x_n)\eta_{n-1}^{noise}+\gamma\xi_n.$$

η_n^{noise} is the part of $(\eta_n)_n$ which is due to the noise.
A straightforward induction shows that for all n, $\eta_n=\eta_n^{init}+\eta_n^{noise}$ and $\bar\eta_n=\bar\eta_n^{init}+\bar\eta_n^{noise}$. Thus Minkowski's inequality (applied to the scalar product $\langle x,y\rangle_H=\mathbb{E}\left[\langle x,\Sigma y\rangle\right]$), leads to :

$$\left(\mathbb{E}\left[\langle\bar\eta_n,\Sigma\bar\eta_n\rangle\right]\right)^{1/2}\leqslant\left(\mathbb{E}\left[\langle\bar\eta_n^{noise},\Sigma\bar\eta_n^{noise}\rangle\right]\right)^{1/2}+\left(\mathbb{E}\left[\langle\bar\eta_n^{init},\Sigma\bar\eta_n^{init}\rangle\right]\right)^{1/2}.$$

That means we can always consider separatly the effect of the noise and the effect of the initial conditions. We'll first study η_n^{noise} and then η_n^{init}.

6.2 Noise process

Multiple recursion : Still following [4] , we are going to define multiple recursions changing $x_n \otimes x_n$ into its expectation Σ each time :

$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma x_n \otimes x_n)\eta_{n-1}^{noise} + \gamma\xi_n.$$

Let (η_n^0) be the following sequence :

$$\eta_0^0 = 0 \text{ and } \eta_n^0 = (I - \gamma\Sigma)\eta_{n-1}^0 + \gamma\xi_n.$$

We have

$$\eta_n^{noise} - \eta_n^0 = (I - \gamma x_n \otimes x_n)(\eta_{n-1}^{noise} - \eta_{n-1}^0) + \gamma(\Sigma - x_n \otimes x_n)\eta_{n-1}^0.$$

Thus we may consider, denoting $\xi_n^1 = (\Sigma - x_n \otimes x_n)\eta_{n-1}^0$ the following sequence, changing $x_n \otimes x_n$ into its expectation Σ :

$$\eta_0^1 = 0 \text{ and } \eta_n^1 = (I - \gamma\Sigma)\eta_{n-1}^1 + \gamma\xi_n^1.$$

And we have

$$\eta_n^{noise} - \eta_n^0 - \eta_n^1 = (I - \gamma x_n \otimes x_n)(\eta_{n-1}^{noise} - \eta_{n-1}^0 - \eta_n^1) + \gamma(\Sigma - x_n \otimes x_n)\eta_{n-1}^1.$$

And so on... For all $r \geqslant 0$ we define a sequence (η_n^r) by :

$$\eta_0^r = 0 \text{ and } \eta_n^r = (I - \gamma\Sigma)\eta_{n-1}^r + \gamma\xi_n^r, \quad \text{with } \xi_n^r = (\Sigma - x_n \otimes x_n)\eta_{n-1}^{r-1}.$$

We have, for any $r, n \in \mathbb{N}^2$:

$$\eta_n^{noise} - \sum_{i=0}^{r} \eta_n^r = (I - \gamma x_n \otimes x_n)(\eta_{n-1}^{noise} - \sum_{i=0}^{r} \eta_{n-1}^r) + \gamma(\Sigma - x_n \otimes x_n)\eta_{n-1}^r.$$

Minkowski's inequality : Considering this decomposition, we have, for any r, using minkowski's inequality :

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\left\langle \bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^i, \Sigma\left(\bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^i\right)\right\rangle\right]\right)^{1/2} + \sum_{i=0}^{r}\left(\mathbb{E}\left[\langle \bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right]\right)^{1/2}. \tag{2}$$

Moment Bounds For any $i \geqslant 0$, we find that we may apply lemma 3 to the sequence (η_n^i). Indeed :

- For any $r \geqslant 0$, (η_n^r) is defined by :

$$\eta_0^r = 0 \text{ and } \eta_n^r = (I - \gamma\Sigma)\eta_{n-1}^r + \gamma\xi_n^r, \quad \text{with } \xi_n^r = \begin{cases} (\Sigma - x_n \otimes x_n)\eta_{n-1}^{r-1} & \text{if } r \geqslant 1, . \\ \xi_n \quad \text{if } r = 0. \end{cases}$$

- for any $r \geqslant 0$, for all $n \geqslant 0$, ξ_n^r is $\mathcal{F}_n := \sigma\left((x_i, z_i)_{1 \leqslant i \leqslant n}\right)$ measurable. (for $r = 0$ we use the definition of ξ_n (**H4**), and by induction, for any $r \geqslant 0$ if we have $\forall n \in \mathbb{N}$, ξ_n^r is $\mathcal{F}_n$ measurable, then for any $n \in \mathbb{N}$, by induction on n, η_n^r is $\mathcal{F}_n$ measurable, thus for any $n \in \mathbb{N}$, ξ_n^{r+1} is $\mathcal{F}_n$ measurable.)
- for any $r, n \geqslant 0$, $\mathbb{E}\left[\xi_n | \mathcal{F}_{n-1}\right] = 0$: as shown above, η_{n-1}^{r-1} is $\mathcal{F}_{n-1}$ mesurable so $\mathbb{E}\left[\xi_n | \mathcal{F}_{n-1}\right] = \mathbb{E}\left[\Sigma - x_n \otimes x_n | \mathcal{F}_{n-1}\right]\eta_{n-1}^{r-1} = \mathbb{E}\left[\Sigma - x_n \otimes x_n\right]\eta_{n-1}^{r-1} = 0$ (as x_n is independent of $\mathcal{F}_{n-1}$ by **H2** and $\mathbb{E}\left[\Sigma - x_n \otimes x_n\right] = \mathbb{E}\left[\Sigma - x_n \otimes x_n\right]$ by **H4**).

- $\mathbb{E}\left[||\xi_n^r||^2\right]$ is finite (once again, by **H3** if $r = 0$ and by a double recursion to get the result for any $r, n \geqslant 0$).
- We have to find a bound on $\mathbb{E}\left[\xi_n^r \otimes \xi_n^r\right]$. To do that, we are going, once again to use induction on r.

Lemma 7. *For any $r \geqslant 0$ we have $\mathbb{E}\left[\xi_n^r \otimes \xi_n^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2 I$.*

Lemma 7. We make an induction on n.

Initialisation : for $r = 0$ we have by **H6** that $\mathbb{E}\left[\xi_n^0 \otimes \xi_n^0\right] \leqslant \sigma^2\Sigma$. Thus we can apply the first part of lemma 3 to (η_n^0). We get

$$\forall n \geqslant 0, \quad \mathbb{E}\left[\eta_n^0 \otimes \eta_n^0\right] \preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{2n-2-k}\Sigma \preccurlyeq \gamma\sigma^2 I$$

If we assume $\forall n \geqslant 0, \quad \mathbb{E}\left[\xi_n^r \otimes \xi_n^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 I$ then: $\forall n \geqslant 0$,

$$\begin{aligned}
\mathbb{E}\left[\xi_n^{r+1} \otimes \xi_n^{r+1}\right] &\preccurlyeq \mathbb{E}\left[(\Sigma - x_n \otimes x_n)\eta_{n-1}^r \otimes \eta_{n-1}^r(\Sigma - x_n \otimes x_n)\right] \\
&= \mathbb{E}\left[(\Sigma - x_n \otimes x_n)\mathbb{E}\left[\eta_{n-1}^r \otimes \eta_{n-1}^r\right](\Sigma - x_n \otimes x_n)\right] \quad (\text{as } \eta_{n-1} \in \mathcal{F}_{n-1}) \\
&\preccurlyeq \gamma^{r+1} R^{2r}\sigma^2 \mathbb{E}\left[(\Sigma - x_n \otimes x_n)^2\right] \\
&\preccurlyeq \gamma^{r+1} R^{2r+2}\sigma^2\Sigma.
\end{aligned}$$

And applying lemma 3 to (η_n^{r+1}), for any n:

$$\begin{aligned}
\mathbb{E}\left[\eta_n^{r+1} \otimes \eta_n^{r+1}\right] &\preccurlyeq \gamma^2 \mathbb{E}\left[\sum_{k=1}^{n} (I - \gamma\Sigma)^{n-1-k}\xi_n^{r+1} \otimes \xi_n^{r+1}(I - \gamma\Sigma)^{n-1-k}\right] \\
&\preccurlyeq \gamma^{r+3} R^{2r+2}\sigma^2 \sum_{k=1}^{n} (I - \gamma\Sigma)^{2n-2-2k}\Sigma \\
&\preccurlyeq \gamma^{r+2} R^{2r+2}\sigma^2 I.
\end{aligned}$$

□

Using the second part of lemma 3 we conclude that :

$$\mathbb{E}\left[\langle \bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right] \leqslant \gamma^i R^{2i} \operatorname{Var}(n, \gamma, \alpha, \sigma). \tag{3}$$

Moreover, using lemma 2 for $(\bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^i)_n$ (all conditions are satisfied) we have :

$$\begin{aligned}
(1 - \gamma R^2)\ \mathbb{E}\left[\left\langle \bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^i, \Sigma\left(\bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^i\right)\right\rangle\right] &\preccurlyeq \frac{\gamma}{n}\sum_{i=1}^{n} \mathbb{E}||\xi_k^{r+1}||^2 \\
&\preccurlyeq \gamma \operatorname{tr}\left(\mathbb{E}\left[\xi_k^{r+1} \otimes \xi_k^{r+1}\right]\right) \\
&\preccurlyeq \gamma^{r+2} R^{2r+2}\sigma^2 \operatorname{tr}(\Sigma). \qquad (4)
\end{aligned}$$

Conclusion Thus using (2), (3) and (4) :

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \left(\frac{1}{1 - \gamma R^2}\gamma^{r+2}\sigma^2 R^{2r+2} \operatorname{tr}(\Sigma)\right)^{1/2} + \operatorname{Var}(n, \gamma, \alpha, \sigma)^{1/2} \sum_{i=0}^{r} \left(\gamma R^2\right)^{i/2}.$$

And using the fact that $\gamma R < 1$ (we will have $\gamma = 1/n^\rho, \rho > 0$!), when $r \to \infty$:

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma \bar{\eta}_n^{noise} \rangle\right]\right)^{1/2} \leqslant \mathrm{Var}(n, \gamma, \alpha, \sigma)^{1/2} \frac{1}{1 - \sqrt{\gamma R^2}}.$$

Discussion : It must be noticed that unlike in the finite-dimensional setting, where our final choice for γ is a constant, It is not necessary to have $r \to \infty$ to get a satisfying bound. Indeed, in most cases, one (or a few) steps of the recursion are enough to have the left term be a $o(\frac{1}{n})$. But as far as the next steps hardly change the rigth side constant (as $1/(1 - \sqrt{\gamma R^2})$ is close to 1 anyway), there's no reason to bother with the first term.

6.3 Initial conditions

We are now interested in getting such a bound for $\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle\right]$. As this part stands for the initial conditions effect we may keep in mind that we would like to get a majoration comparable to what we found for the Bias term in the proof of proposition 1.
We remember that :

$$\eta_0^{init} = \theta_0 - \theta_* \text{ and } \eta_n^{init} = (I - \gamma x_n \otimes x_n)\eta_{n-1}^{init}.$$

and define $(\eta_n^0)_{n \in \mathbb{N}}$ so that :

$$\eta_0^0 = \theta_0 - \theta_*, \quad \eta_n^0 = (I - \gamma\Sigma)\eta_{n-1}^0.$$

Minkowski's again : As above

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init} - \bar{\eta}_n^0, \Sigma\left(\bar{\eta}_n^{init} - \bar{\eta}_n^0\right) \rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0 \rangle\right]\right)^{1/2}.$$

Applying Lemma 3 and 5 : First, using our bias term majoration, we have

$$\mathbb{E}\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0 \rangle \leqslant \mathrm{Bias}(n, \gamma, \alpha, \beta).$$

Residual term

Either we use lemma 3 Then, using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma\Sigma)(\eta_n^0 - \eta_n^{init}) + \gamma(x_n \otimes x_n - \Sigma)\eta_{n-1}^{init},$$

we may apply **lemma 3** and **4** to the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$ and $\xi_n = (x_n \otimes x_n - \Sigma)\eta_{n-1}^{init}$.
We note that, as $\eta_n^{init} = \prod_{k=n}^1 (I - \gamma x_k \otimes x_k)\ \eta_0^{init} = \prod_{k=n}^1 (I - \gamma x_k \otimes x_k)\ \eta_0$, where $\prod_{k=n}^1 M_i$ stands for $M_n M_{n-1} ... M_2 M_1$ (the operators do not commute).

$$\begin{aligned}
\xi_n \otimes \xi_n &= (\Sigma - x_n \otimes x_n) \prod_{k=n}^{1} (I - \gamma x_k \otimes x_k)\ \eta_0^{init} \otimes \eta_0^{init} \prod_{k=1}^{n} (I - \gamma x_n \otimes x_n)\ (\Sigma - x_n \otimes x_n) \\
&\quad \text{and as } \eta_0 \otimes \eta_0 \preccurlyeq ||\eta_0||^2 I, \\
&\preccurlyeq ||\eta_0||^2 (\Sigma - x_n \otimes x_n) \prod_{k=n}^{1} (I - \gamma x_k \otimes x_k) \prod_{k=1}^{n} (I - \gamma x_n \otimes x_n)\ (\Sigma - x_n \otimes x_n).
\end{aligned}$$

We denote $V_k = (I - \gamma x_k \otimes x_k)$.

$$\begin{aligned}
\mathbb{E}\left[\xi_n \otimes \xi_n\right] &\preccurlyeq \mathbb{E}\left[||\eta_0||^2(\Sigma - x_n \otimes x_n) \prod_{k=n}^{1} V_k \prod_{k=1}^{n} V_k \ (\Sigma - x_n \otimes x_n)\right] \\
&\preccurlyeq \mathbb{E}\left[\mathbb{E}\left[||\eta_0||^2(\Sigma - x_n \otimes x_n) \prod_{k=n}^{1} V_k \prod_{k=1}^{n} V_k \ (\Sigma - x_n \otimes x_n)\middle|(x_2, ..x_n)\right]\right] \\
&\preccurlyeq \mathbb{E}\left[||\eta_0||^2(\Sigma - x_n \otimes x_n) \prod_{k=n}^{2} V_k \ \mathbb{E}\left[(I - \gamma x_1 \otimes x_1)^2\right] \prod_{k=2}^{n} V_k(\Sigma - x_n \otimes x_n)\right]
\end{aligned}$$

But we have :

$$\begin{aligned}
\mathbb{E}\left[(I - \gamma x_1 \otimes x_1)^2\right] &\preccurlyeq \mathbb{E}\left(I - 2\gamma x_1 \otimes x_1 + \gamma^2 ||x_1||^2 x_1 \otimes x_1\right) \\
&\preccurlyeq I - 2\gamma\Sigma + \gamma^2 R^2 \Sigma \qquad \text{as by } \mathbf{H6} \quad \mathbb{E}\left[||x_1||^2 x_1 \otimes x_1\right] \preccurlyeq R^2\Sigma \\
&= I + \gamma(\gamma R^2 - 2)\Sigma \\
&\preccurlyeq I \qquad \text{as } \gamma R^2 \leqslant 1.
\end{aligned}$$

Iterating the process, we get :

$$\begin{aligned}
\mathbb{E}\left[\xi_n \otimes \xi_n\right] &\preccurlyeq \mathbb{E}\left[||\eta_0||^2(\Sigma - x_n \otimes x_n)(\Sigma - x_n \otimes x_n)\right] \\
&\preccurlyeq ||\eta_0||^2 R^2 \Sigma.
\end{aligned}$$

As $\mathbb{E}\left[\xi_n \otimes \xi_n\right] \preccurlyeq \mathbb{E}\left[(\Sigma - x_n \otimes x_n)^2\right] = \mathrm{Var}(x_n \otimes x_n) \preccurlyeq \mathbb{E}(||x_n||^2 x_n \otimes x_n) \preccurlyeq R^2\Sigma$.

Thus we get using our variance term majoration (lemma 4) (with $\sigma = ||\eta_0||R$!) :

$$\left(\mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle\right)^{1/2} \leqslant \mathrm{Var}(n, \gamma, \alpha, ||\eta_0||R) = \Theta\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right).$$

Conclusion Summing both bounds we get :

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \sqrt{\mathrm{Var}(n, \gamma, \alpha, ||\eta_0||R)} + \sqrt{Bias(n, \gamma, \alpha, \beta)}.$$

Or we use lemma 2 : Then, using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma x_n \otimes x_n)(\eta_n^0 - \eta_n^{init}) + \gamma(x_n \otimes x_n - \Sigma)\eta_{n-1}^0,$$

we may apply **lemma 2** to the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$ and $\xi_n = (x_n \otimes x_n - \Sigma)\eta_{n-1}^0$. That is (as $\alpha_0 = 0$):

$$\mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle \leqslant \frac{1}{1 - \gamma R^2}\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^{n} ||\xi_k||^2\right]$$

Now

$$\begin{aligned}
\mathbb{E}||\xi_k||^2 &= \mathbb{E}\left[\left\langle \eta_0, (I - \gamma\Sigma)^n(\Sigma - x_n \otimes x_n)^2(I - \gamma\Sigma)^n\eta_0\right\rangle\right] \\
&\leqslant \left\langle \eta_0, (I - \gamma\Sigma)^n R^2\Sigma(I - \gamma\Sigma)^n\eta_0\right\rangle \\
&\leqslant R^2\left\langle \eta_0, (I - \gamma\Sigma)^{2n}\Sigma\eta_0\right\rangle \qquad \text{and calculating in the orthonormal basis:} \\
&= R^2\sum_{i=1}^{\infty}\left(1 - \frac{\gamma}{i^\alpha}\right)^{2n}\frac{1}{i^{\alpha+\beta}}.
\end{aligned}$$

Thus :

$$\begin{aligned}
\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^{n}||\xi_k||^2\right] &\leqslant \frac{\gamma R^2}{n}\sum_{k=1}^{n}\sum_{i=1}^{\infty}\left(1-\frac{\gamma}{i^\alpha}\right)^{2k}\frac{1}{i^{\alpha+\beta}} \\
&\leqslant \frac{\gamma R^2}{n}\sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\left(1-\frac{\gamma}{i^\alpha}\right)^{2k}\frac{1}{i^{\alpha+\beta}} \\
&\leqslant \frac{\gamma R^2}{n}\sum_{i=1}^{\infty}\frac{1}{1-\left(1-\frac{\gamma}{i^\alpha}\right)^2}\frac{1}{i^{\alpha+\beta}} \\
&\leqslant \frac{\gamma R^2}{n}\sum_{i=1}^{\infty}\frac{i^\alpha}{\gamma}\frac{1}{i^{\alpha+\beta}} \qquad \text{as } \forall x\in[\,0;1],\quad 1-(1-x)^2\geqslant x \\
&\leqslant \frac{||\eta_0||R^2}{n}.
\end{aligned}$$

Which is even better than what we found in the previous way.

Conclusion Summing both bounds we get :

$$\left(\mathbb{E}\left[\langle\bar{\eta}_n^{noise},\Sigma\bar{\eta}_n^{noise}\rangle\right]\right)^{1/2}\leqslant\left(\frac{1}{1-\gamma R^2}\frac{||\eta_0||R^2}{n}\right)^{1/2}+\left(Bias(n,\gamma,\alpha,\beta)\right)^{1/2}.$$

Discussion: It seems the first analysis makes appear the same constant as in the "Σ known" setting ($\sigma^2+||\theta_*||R^2$) which may be rather artificial as far as this constant came from the passage from the "unknown Σ" to "known Σ". The second part of the analysis shows this was indeed an artefact of the proof.

6.4 Conclusion

Using the conclusions of the preceeding sections, we get, using lemma 4 and 5:

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

$$\begin{aligned}
\left(\mathbb{E}\left[\langle\bar{\eta}_n,\Sigma\bar{\eta}_n\rangle\right]\right)^{1/2}\leqslant\frac{1}{1-\sqrt{\gamma R^2}}&\left(C(\alpha)\;\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}\right)^{1/2}+\left(\frac{s^{2\frac{-\beta+\alpha+1}{\alpha}}}{T^2u^2}K(\alpha,\beta)\frac{1}{(n\gamma)^{\frac{\alpha+\beta-1}{\alpha}}}\right)^{1/2} \\
&+\left(\frac{1}{1-\gamma R^2}\frac{||\eta_0||R^2}{n}\right)^{1/2}.
\end{aligned}$$

With $K(\alpha,\beta)=\frac{1}{\alpha+\beta-1}+\frac{1}{\alpha+1-\beta}$ and $C(\alpha)=\frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$. Asymptotically :

$$\mathbb{E}\left[\langle\bar{\eta}_n,\Sigma\bar{\eta}_n\rangle\right]=O\left(\frac{1}{(n\gamma)^{1-\frac{1}{\alpha}+\frac{\beta}{\alpha}}}\right)+O\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right).$$

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

$$\begin{aligned}
\left(\mathbb{E}\left[\langle\bar{\eta}_n,\Sigma\bar{\eta}_n\rangle\right]\right)^{1/2}\leqslant\frac{1}{1-\sqrt{\gamma R^2}}&\left(C(\alpha)\;\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}\right)^{1/2} \\
&+\left(\frac{1}{T^2u^2}\left(\frac{1}{\beta-\alpha-1}+\frac{2}{\alpha+\beta-1}\right)\frac{1}{(\gamma n)^2}\right)^{1/2} \\
&+\left(\frac{1}{1-\gamma R^2}\frac{||\eta_0||R^2}{n}\right)^{1/2}.
\end{aligned}$$

With $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$. Asymptotically :

$$\mathbb{E}\left[\langle \bar{\eta}_n, \Sigma \bar{\eta}_n \rangle\right] = O\left(\frac{1}{(n\gamma)^2}\right) + O\left(\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right).$$

Using **lemma 1** it is **proposition 2**.

Appendices

A Link with [1].

In this paragraph we are intereted in comparing our notations and hypothesis with those used in [1]. For better comprehension, we denote in this paragraph α^0 and β^0 the constants appearing in **H7** and **H8**, and β^1 and δ^1 the constants corresponding to the hypothesis made in [1].

Our hypothesis	[1] hypothesis
H7	"the kernel matrix K has eigenvalues of the form $\Theta(n\mu_i)$ with $\mu_i = i^{-2\beta^1}$"
H8	"the coordinates of z in the eigenbasis of K have the asymptotic behaviour $\Theta(\sqrt{n\nu_i})$ with $\nu_i = i^{-2\delta^1}$"

We clearly get from the first point that $\alpha^0 = 2\beta^1$, then from the second point that $\alpha^0 + \beta^0 = 2\delta^1$. Thus the limit condition $2\delta^1 = 4\beta^1 + 1$ is exactly the same as $\beta^0 = \alpha^0 + 1$, and the predicted performance are both the same.

References

[1] F. Bach. Sharp analysis of low-rank kernel matrix approximations. *Proceedings of the International Conference on Learning Theory (COLT)*, 2012.

[2] Y. Yao. *A dynamic Theory of Learning*. PhD thesis, University of California at Berkeley, 2006.

[3] P. Tarrès and Y. Yao. Online Learning as Stochastic Approximation of Regularization Paths. *ArXiv e-prints*, (1103.5538), 2011.

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[5] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[6] S. Shalev-Schwartz and N. Srebro. SVM optimisation : Inverse dependance on training set size. *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[7] S. Shalev-Schwartz and K. Sridharan. Theorical basis for more data less work. *COST*, 2011.

[8] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. *Proceedings of the International Conference on Learning Theory (COLT)*, 2011.

[9] 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*, (1212.2002), 2012.

[10] H. W. Engl, M. Hanke, and Neubauer A. Regularization of inverse problems. *Klüwer Academic Publishers*, 1996.

[11] F. Paulin. *Topologie, analyse et calcul différentiel*. Notes de cours, École Normale Supérieure, 2009.

[12] J. Kivinen, Smola A.J., and R. C. Williamson. Online learning with kernels. *IEEE transactions on signal processing*, 52(8):2165–2176, 2004.

[13] G. Wahba. *Spline Models for observationnal data*. SIAM, 1990.

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

Exposé de Maîtrise : Marches Aléatoires Contrôlées

Nicolas Coudray Aymeric Dieuleveut

Table des matières

1 Stratégies déterministes

Avant tout, donnons quelques définitions :

Nous considérons ici un jeu entre deux joueurs. On a deux ensembles finis d'actions $\mathcal{A}$ et $\mathcal{B}$. A chaque tour, les joueurs choisissent un élément de $\Delta(\mathcal{A})$ et $\Delta(\mathcal{B})$ (ensembles des lois de probabilité sur $\mathcal{A}$ et $\mathcal{B}$, identifiés aux ensembles des vecteurs de $\mathbb{R}^{\mathcal{A}}$ et $\mathbb{R}^{\mathcal{B}}$ à coordonnées positives et dont la somme des coordonnées vaut 1). La fonction de gain du joueur 1 est $m : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ (vue comme une matrice, $[m(i,j)]_{i,j}$), prolongée linéairement à $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ par la formule $m(p,q) = \sum_{i,j} p_i q_j m(i,j)$.
(Par abus de notation, on notera parfois $m(i,q) := m(\delta_i, q) := \sum_j q_j m(i,j)$ et $m(p,j) := m(p,\delta_j)$.)

Par exemple, si on considère le jeu "pierre - feuille - ciseaux", la matrice des gains sera :

	Pierre	Feuille	Ciseaux
Pierre	0	-1	1
Feuille	1	0	-1
Ciseaux	-1	1	0

On a $\mathcal{A} = \mathcal{B} = \{Pierre, Feuille, Ciseaux\}$ et le gain du joueur 1 s'il joue "Feuille" et que son adversaire 2 joue "Pierre" est 1.
Dans le cas des stratégies déterministes, le joueur 1 peut choisir de jouer le vecteur $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ et dans ce cas il est assuré que son gain sera nul.

Si on autorise au joueur 1 le "puits", qui bat les ciseaux et la pierre , la matrice devient alors :

	Pierre	Feuille	Ciseaux
Pierre	0	-1	1
Feuille	1	0	-1
Ciseaux	-1	1	0
Puits	1	-1	1

Les gains que l'on considère par la suite sont, contrairement à l'exemple, des vecteurs de $\mathbb{R}^d$.

Définition 1. *Dans le cas des stratégies déterministes, les joueurs choisissent $p_t, q_t \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$. Le gain reçu est $m(p_t, q_t)$. A la fin de chaque tour, q_t est révélé au joueur 1.*

On appelle stratégie (pour le joueur 1) une suite de fonctions $f_T : (\Delta(\mathcal{A}) \times \Delta(\mathcal{B}))^{(T-1)} \mapsto \Delta(\mathcal{A})$, qui à partir des choix passés (p_t, q_t pour $t \leqslant T-1$) donne le coup suivant, p_T pour le joueur 1 (ou $q_T \in \Delta(\mathcal{B})$ pour le joueur 2).

On pose alors D_T distance entre le gain moyen à l'instant $T(\in \mathbb{N})$ et un ensemble C :

$$D_T := \inf_{c \in C} \left\| c - \frac{1}{T} \sum_{t=1}^{T} m(p_t, q_t) \right\|_2 .$$

Définition 2. *Avec ces notations, on dira qu'un ensemble $C \subseteq \mathbb{R}^d$ est m-approchable par des stratégies déterministes s'il existe une stratégie pour le joueur 1 telle que pour toute stratégie du joueur 2,*

$$\lim_{T \to \infty} D_T = 0 \ .$$

Remarquons qu'un ensemble est approchable si et seulement si son adhérence l'est ($d(x,C) = d(x,\bar{C})$ car la distance d'un point x à un ensemble C est atteinte sur l'adhérence de ce dernier). Aussi, nous ne nous intéresserons dans la suite qu'à l'approchabilité des ensembles fermés.

Nous allons maintenant démontrer le théorème suivant (dû à Blackwell) :

Théorème 1.
Soit C un convexe fermé. Alors, en reprenant les notations précédentes, C est approchable si et seulement si

$$\forall q \in \Delta(\mathcal{B}), \exists p \in \Delta(\mathcal{A}), \qquad m(p,q) \in C.$$

Une stratégie possible d'approchabilité est la suivante : on choisit p_1 arbitrairement, puis à chaque tour, le joueur 1 joue un coup p_T qui résout le problème suivant :

$$\min_{p\in\Delta(\mathcal{A})} \max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p,q) > ,$$

avec :

- $\overline{m}_T := \frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t)$,
- *π_T la projection orthogonale de $\overline{m}_T$ sur C ($\pi_T = \overline{m}_T$ si $\overline{m}_T \in C$),*
- *$< .,. >$ le produit scalaire euclidien sur $\mathbb{R}^d$.*

De plus, si l'on note $M = \sup_{(i,j)\in\mathcal{A}\times\mathcal{B}} \|m(i,j)\|_2$, alors on a une estimation de la vitesse de convergence en suivant cette stratégie :

$$D_T \leqslant \frac{2M}{\sqrt{T}}.$$

Avant de donner la démonstration, on donne une interprétation simple et des exemples dans le cas ou les gains sont des réels et non des vecteurs.

Le théorème signifie que si $d = 1$, il existe deux valeurs limites $Mm = \max_{q\in\mathcal{B}} \min_{p\in\mathcal{A}} m(p,q)$ et $mM = \min_{q\in\mathcal{B}} \max_{p\in\mathcal{A}} m(p,q)$ telles qu'un convexe $C := [v;V]$ de $\Re$ est appprochable si et seulement si $mM \geqslant v$ et $Mm \leqslant V$.

En effet, on peut vérifier simplement que la condition est nécessaire, même sans le théorème : en reprenant les notations ci-dessus,

1. si on a $Mm > V$ alors $\exists q_0 \in \mathcal{B}, \quad \min_{p\in\mathcal{A}} m(p,q) > V + \varepsilon$ c'est-à-dire $\exists q_0 \in \mathcal{B}, \forall p \in \mathcal{A}, \quad m(p,q) > V + \varepsilon$, donc si le joueur 2 joue la stratégie qui consiste à , quel que soit le passé, jouer q_0, le gain moyen sera strictement supérieur à $V + \varepsilon$.
2. et de même si $mM < V$ alors $\exists q_0 \in \mathcal{B}, \quad \max_{p\in\mathcal{A}} m(p,q) < V - \varepsilon$ c'est-à-dire $\exists q_0 \in \mathcal{B}, \forall p \in \mathcal{A}, \quad m(p,q) < V - \varepsilon$, donc si le joueur 2 joue la stratégie qui consiste à , quel que soit le passé, jouer q_0, le gain moyen sera strictement inférieur à $v - \varepsilon$.

Dans les deux cas, C n'est pas approchable.

Réciproquement, si on a cette condition, on a la condition du théorème : en effet on peut remarquer que dans un cadre tout à fait général (indépendant de d), la convexité de C permet de se contenter de l'hypothèse :

$$\forall q \in \mathcal{B}, \exists p \in \Delta(\mathcal{A}), \qquad m(p,q) \in C,$$

et on a avec notre hypothèse que $\forall q \in \mathcal{B}, \qquad \min_{p\in\mathcal{A}} m(p,q) \leqslant V$ et $\max_{p\in\mathcal{A}} m(p,q) \geqslant v$.
Donc $\forall q \in \mathcal{B}, \exists p \in \Delta(\mathcal{A}), \qquad m(p,q) \in C$.

On peux donner briévement deux exemples à partir de jeux à deux actions :
Cas 1 :

	p_1	p_2
q_1	0	1
q_2	4	3

Ici, $Mm = \max(0,3) = 3$ et $mM = \min(1,4) = 1$ un convexe C est donc approchable si et seulement si : $[1;3] \subset C$.

Cas 2 :

	p_1	p_2
q_1	0	5
q_2	4	3

Ici, $Mm = \max(0,3) = 3$ et $mM = \min(5,4) = 4$ un convexe $C = [v;V]$ est donc approchable si et seulement si : $v \leqslant 4$ et $3 \leqslant V$, c'est à dire si et seulement s'il contient un élément de [3 ;4].

Démonstration. L'idée de la démonstration du sens indirect provient essentielement de l'article de Blackwell [1] (p.6). En remarquant que si

$$\exists q_0 \in \Delta(\mathcal{B}), \forall p, \qquad m(p,q_0) \notin C,$$

alors en considérant

$$R(q_0) := \{m(p,q_0) \ / \ p \in \Delta(\mathcal{A})\} = \left\{ \sum_i p_i m(i,q_0) \ / \ p = (p_i)_{i\in\mathcal{A}} \in \Delta(\mathcal{A}) \right\}$$

l'enveloppe convexe des $m(i,q_0)$, alors pour le jeu associé à la matrice ${}^t m$, on a

$$\forall p, \exists q(=q_0), \qquad {}^t m(q,p) \in R(q_0),$$

$R(q_0)$ étant un convexe fermé d'intersection vide avec C. Si l'on démontre le sens direct, on aura alors $R(q_0)$ approchable par le joueur 2. Or cela signifie par définition qu'il existe une stratégie du joueur 2 telle que pour toute stratégie du joueur 1, $d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), R(q_0)\right) \to 0$. On utilise ensuite le fait que $d\left(R(q_0),C\right) := \inf_{x\in R(q_0)} d\left(x,C\right) = \inf_{x\in R(q_0), y\in C} d\left(x,y\right) > 0$ (car $R(q_0)$ et C sont disjoints, $R(q_0)$ est compact et C fermé) et par inégalité triangulaire, si $a \in C$ et $b \in R(q_0)$, on sait que

$$d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), a\right) + d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), b\right) \geqslant d(a,b),$$

et donc en passant à l'inf sur C et $R(q_0)$,

$$d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), C\right) + d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), R(q_0)\right) \geqslant d\left(C,R(q_0)\right),$$

d'où

$$\liminf_{T\to\infty}\ d\left(\frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t), C\right) \geqslant d(R(q_0),C) > 0,$$

et donc C n'est pas approchable pour le joueur 1.

Il ne reste donc plus qu'à montrer le sens direct. On utilise la stratégie donnée dans l'énoncé du théorème (la demonstration qui suit provient de l'article [4]).

Par le théorème du minimax (démontré en annexe) appliqué à la matrice $A := (< \overline{m}_T - \pi_T, m(i,j) >)_{i,j}$, on sait que le problème "p_T résout $\min_{p\in\Delta(\mathcal{A})} \max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p,q) >$" proposé pour construire une stratégie admet une solution et que :

$$\min_{p\in\Delta(\mathcal{A})}\max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p,q) > = \max_{q\in\Delta(\mathcal{B})}\min_{p\in\Delta(\mathcal{A})} < \overline{m}_T - \pi_T, m(p,q) > .$$

Or on a :

$$\forall q, \exists p, \qquad m(p,q) \in C\ ,$$

et donc, par les propriétés de projection sur les convexes,

$$\forall q, \exists p, \qquad < \overline{m}_T - \pi_T, m(p,q) - \pi_T > \leqslant\ 0\ .$$

(voir par exemple le théorème 6.32 de [7] dans le cas d'un produit scalaire réel).

On a donc prouvé :

$$\min_{p\in\Delta(\mathcal{A})}\max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p,q) - \pi_T > = \max_{q\in\Delta(\mathcal{B})}\min_{p\in\Delta(\mathcal{A})} < \overline{m}_T - \pi_T, m(p,q) - \pi_T > \leqslant\ 0\ .$$

En particulier puisque p_{T+1} est défini comme solution du problème $\min_{p\in\Delta(\mathcal{A})}\max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p,q) >$ on a la relation :

$$\forall q, \quad < \overline{m}_T - \pi_T, m(p_{T+1},q) - \pi_T > \leqslant 0\ ,$$

et donc en particulier :

$$< \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \pi_T > \leqslant 0\ , \tag{1}$$

(qui reste vrai dans le cas $\overline{m}_T \in C$).

Un calcul algébrique donne alors :

$$\begin{aligned}
\overline{m}_{T+1} &= \frac{1}{T+1}\sum_{\tau=1}^{T+1} m(p_\tau,q_\tau) \\
&= \frac{1}{T+1}m(p_{T+1},q_{T+1}) + \frac{1}{T}\sum_{\tau=1}^{T} m(p_\tau,q_\tau) - \frac{1}{T(T+1)}\sum_{\tau=1}^{T} m(p_\tau,q_\tau) \\
&= \overline{m}_T + \frac{1}{T+1}(m(p_{T+1},q_{T+1}) - \overline{m}_T)\ .
\end{aligned}$$

Cela permet de déterminer une inégalité sur D_T :

$$\begin{aligned} D_{T+1}^2 &= \inf_{c\in C} \left\| c - \frac{1}{T+1}\sum_{t=1}^{T+1} m(p_t,q_t) \right\|_2^2 \\ &= \|\pi_{T+1} - \overline{m}_{T+1}\|_2^2 \\ &\leqslant \|\pi_T - \overline{m}_{T+1}\|_2^2 \\ &= \left\| (\pi_T - \overline{m}_T) - \frac{1}{T+1}(m(p_{T+1},q_{T+1}) - \overline{m}_T) \right\|_2^2, \end{aligned}$$

où l'on a utilisé pour l'inégalité le fait que $\pi_T \in C$. Puis en développant :

$$D_{T+1}^2 \leqslant \|\pi_T - \overline{m}_T\|_2^2 + \frac{2}{T+1} < \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \overline{m}_T > + \frac{\|(m(p_{T+1},q_{T+1}) - \overline{m}_T)\|_2^2}{(T+1)^2},$$

et en ajoutant et soustrayant $\frac{2}{T+1} < \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \pi_T >$ on fait apparaître $D_T = \|\pi_T - \overline{m}_T\|_2^2$:

$$D_{T+1}^2 \leqslant \|\pi_T - \overline{m}_T\|_2^2 + \frac{2}{T+1} < \overline{m}_T - \pi_T, \pi_T - \overline{m}_T > + \frac{2}{T+1} < \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \pi_T > + \frac{\|(m(p_{T+1},q_{T+1}) - \overline{m}_T)\|_2^2}{(T+1)^2}$$

$$D_{T+1}^2 \leqslant \|\pi_T - \overline{m}_T\|_2^2 - \frac{2}{T+1}\|\overline{m}_T - \pi_T\|_2^2 + \frac{2}{T+1} < \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \pi_T > + \frac{\|(m(p_{T+1},q_{T+1}) - \overline{m}_T)\|_2^2}{(T+1)^2}$$

$$D_{T+1}^2 \leqslant (1 - \frac{2}{T+1})D_T^2 + \frac{2}{T+1} < \overline{m}_T - \pi_T, m(p_{T+1},q_{T+1}) - \pi_T > + \frac{\|(m(p_{T+1},q_{T+1}) - \overline{m}_T)\|_2^2}{(T+1)^2}.$$

Or le produit scalaire est négatif d'après (1). De plus, par inégalité triangulaire,

$$\|m(p_{T+1},q_{T+1}) - \overline{m}_T\|_2 \leqslant 2M.$$

Finalement on a :

$$D_{T+1}^2 \leqslant (1 - \frac{2}{T+1})D_T^2 + \frac{4M^2}{(T+1)^2}.$$

On démontre alors par récurrence que $D_T^2 \leqslant \frac{4M^2}{T}$ pour tout T strictement positif.

Initialisation : On sait que π_1 réalise $\min_{x\in C}\|x - m(p_1,q_1)\|_2^2$, et donc si l'on considère p_1' un élément quelconque tel que $m(p_1',q_1) \in C$ (qui existe par hypothèse)

$$D_1^2 := \|\pi_1 - m(p_1,q_1)\|_2^2 \leqslant \|m(p_1',q_1) - m(p_1,q_1)\|_2^2 \leqslant 4M^2 .$$

Hérédité :

$$\begin{aligned}\left(1-\frac{2}{T+1}\right)\frac{1}{T}+\frac{1}{(T+1)^2} &= \frac{(T+1)^2-2(T+1)+T}{T(T+1)^2}\\ &= \frac{T^2-1+T}{T(T+1)^2}\\ &\leqslant \frac{T^2+T}{T(T+1)^2}\\ &\leqslant \frac{1}{T+1}\ ,\end{aligned}$$

et le résultat en découle alors immédiatement, ce qui achève la preuve. □

2 Stratégies aléatoires

On se place à présent dans le cadre d'un jeu où un facteur aléatoire va apparaître : dans le cas des stratégies aléatoires, les joueurs 1 et 2 choisissent à chaque instant t une loi de probabilité p_t et q_t dans les ensembles $\Delta(\mathcal{A})$ et $\Delta(\mathcal{B})$, mais cette fois-ci ils tirent leurs actions $I_t \in \mathcal{A}$ (respectivement $J_t \in \mathcal{B}$) selon les lois p_t (respectivement q_t). A la fin du tour, l'action J_t du joueur 2, et donc le gain du joueur 1 $m(I_t, J_t)$ sont révélés au joueur 1.

Dans cette situation, I_t (resp. J_t) est une variable aléatoire à valeurs dans $\mathcal{A}$ (resp. $\mathcal{B}$) de loi conditionnelle à $\mathcal{F}_{t-1}$ p_t (resp. q_t).

Dans toute la suite on notera $\mathcal{F}_T := \sigma\left(\{I_t, J_t \ / \ 1 \leqslant t \leqslant T\} \bigcup \{J_{T+1}\}\right)$, la tribu par rapport à laquelle p_t, I_t, J_t pour $t \leqslant T$ et J_{T+1} sont mesurables. En effet p_T est alors $\mathcal{F}_T$-mesurable car chacune de ses coordonnées $(p_T)_k$ s'exprime comme la probabilité (l'intégrale de la fonction caractériqtique) de $\{I_T = k\}$, qui appartient à $\mathcal{F}_T$.

Définition 3. *On appelle stratégie (pour le joueur 1) une suite de fonctions $f_T : (\mathcal{A} \times \mathcal{B})^{(T-1)} \mapsto \Delta(\mathcal{A})$, qui à partir des tirages passés (I_t, J_t pour $t \leqslant T-1$) donne le coup suivant, p_T pour le joueur 1 (ou $q_T \in \Delta(\mathcal{B})$ pour le joueur 2).*

On note alors D_T la distance du gain moyen à l'instant T à l'ensemble C, c'est-à-dire

$$D_T := \inf_{c \in C} \left\| c - \frac{1}{T} \sum_{t=1}^{T} m(I_t, J_t) \right\|_2 .$$

D_T est donc une variable aléatoire à valeurs réelles, $\mathcal{F}_T$-mesurable.

Définition 4. *On dira qu'un ensemble $C \subseteq \mathbb{R}^d$ est m-approchable par des stratégies aléatoires s'il existe une stratégie pour le joueur 1 telle que pour toute stratégie du joueur 2,*

$$\lim_{T \to \infty} D_T = 0 \quad ps.$$

Cette définition est la même que pour les stratégies déterministes, mais avec l'introduction du caractère probabiliste.

La démonstration du théorème de Blackwell utilise la démonstration du cas déterministe, qui va permettre de contrôler le comportement de la moyenne. Cependant, un argument de déviation (qui décrit l'écartement à la moyenne) est nécessaire. On utilisera donc l'inégalité d'Hoeffding-Azuma, démontrée en annexe.

Théorème 2 (inégalité d'Hoeffding-Azuma)**.**
Soit $(\mathcal{F}_n)_{n \geqslant 0}$ une filtration, (c_n) et (ℓ_n) deux suites réelles et $(Y_n)_{n \geqslant 1}$ une suite de variables aléatoires telles que $\mathbb{E}[Y_n | \mathcal{F}_{n-1}] = 0$ et $c_n \leqslant Y_n \leqslant c_n + \ell_n$ ps. Alors pour tout $\lambda \geqslant 0$,

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} Y_i \right| \geqslant \lambda \right) \leqslant 2 \exp\left(- \frac{2\lambda^2}{\sum_{i=1}^{n} \ell_i^2} \right).$$

En particulier, si $|Y_n| < c$ ps, alors si $0 < \delta < 1$, on obtient qu'avec probabilité au moins $1 - \delta$,

$$\left| \sum_{i=1}^{n} Y_i \right| \leqslant c \sqrt{2n \ln\left(\frac{2}{\delta}\right)}.$$

Théorème 3 (Blackwell)**.**
Soit C un convexe fermé. Alors C est approchable si et seulement si

$$\forall q \in \Delta(\mathcal{B}), \exists p \in \Delta(\mathcal{A}), \qquad m(p,q) \in C$$

et en jouant une stratégie semblable à celle du cas déterministe, si l'on note $M = \sup_{(i,j)\in\mathcal{A}\times\mathcal{B}} \|m(i,j)\|_2$*, alors avec probabilité au moins* $1-\delta$ *:*

$$D_T \leqslant \frac{2M}{\sqrt{T}}\left(1 + 2\sqrt{d\ln\left(\frac{2d}{\delta}\right)}\right).$$

Cas particulier : si le joueur 2 joue selon une stratégie fixe connue (par exemple iid) (cas qui ne rentre pas dans le cadre du théorème), ou si la loi du gain ne dépend pas du coup de 2, alors le résultat d'approchabilité découle de la loi des grands nombres.
En effet le joueur 1 peut choisir un $p_0 \in \Delta(\mathcal{A})$ tel que pour un $q_0 \in \Delta(\mathcal{B})$ on aît $m(p_0,q_0) \in C$ et donc en jouant la stratégie constante p_0 on aura $m(I_T,J_T) = m(I_0,J_0)$ en loi. Et $\mathbb{E}\left(m(I_T,J_T)\right) = m(p_0,q_0) \in C$. Les gains sont donc une suite de variables indépendantes et de même loi, bornés donc L^1. Par la loi forte des grands nombres, on aura :

$$\frac{1}{T}\sum_{t=1}^{T} m(I_t,J_t) \to m(p_0,q_0) \text{ ps.}$$

C'est-à-dire $D_T \to 0$ ps.

Ce cas particulier recouvre par exemple le cas de l'exemple de "Pierre-Feuille-Ciseaux", dans lequel quelles que soient les stratégies choisies le gain suit une loi répartie uniformément sur $\{-1,0,1\}$. Dans ce cas, on vérifie facilement qu'un convexe est approchable si et seulement s'il contient 0.

Si la norme de la matrice vaut 1 en dimension 2, et que la condition nécessaire et suffisante est vérifiée, le théorème donne par exemple qu'avec au moins 99% de chance, on aura que au bout de 1000 coups la distance sera inférieure à $\frac{1}{2}$ et pour savoir qu'avec au moins $99,99\%$ de chance on a $D_T \leqslant \frac{1}{2}$ il faut attendre 1600 coups environ. Une très bonne certitude n'est pas couteuse.

Démonstration. La démonstration du sens indirect se déduit du sens direct exactement de la même manière que dans le cas déterministe.

Pour le sens direct, on reprend à nouveau une démonstration de l'article [4]. Dans le cas des stratégies déterministes, la distance correspond à une distance réelle de la moyenne au convexe, ou encore à l'espérance conditionnelle de la distance dans le cas des stratégies aléatoires : en effet pour tout temps t, en utilisant :

1. le caractère $\mathcal{F}_T$-mesurable de J_{T+1},
2. le fait que la loi conditonelle de I_{T+1} sachant $\mathcal{F}_T$ est p_{T+1},

on a :

$$\mathbb{E}\left[m(I_{T+1},J_{T+1})|\mathcal{F}_T\right] = m(p_{T+1},J_{T+1}). \tag{2}$$

Cette égalité prouve que si on pose

$$Y_T := m(I_T,J_T) - m(p_T,J_T)\ ,$$

alors pour tout k tel que $1 \leqslant k \leqslant d$, en notant $(Y_T)_k$ la k-ième coordonnée de Y_T, $((Y_T)_k)_{T\in\mathbb{N}}$ est un accroissement de martingale par rapport à $(\mathcal{F}_T)_{T\geq 1}$ (c'est-à-dire $M_T := \sum_{i=1}^{T} Y_i$ est une martingale par rapport à $(\mathcal{F}_T)$). En effet, on a : $(Y_T)_k$ est $\mathcal{F}_T$-mesurable, et

$$\begin{aligned}\mathbb{E}[(Y_{T+1})_k|\mathcal{F}_T] &= \mathbb{E}\left[(m(I_{T+1},J_{T+1}))_k - (m(p_{T+1},J_{T+1}))_k|\mathcal{F}_T\right] \\ &= 0\end{aligned}$$

d'après (2) (en utilisant l'égalité pour chaque coordonnée).

Le théorème d'Hoeffding-Azuma donne des informations sur la vitesse de convergence : la suite de variables aléatoires $((Y_T)_k)_{T\in\mathbb{N}}$ vérifie les hypothèses nécessaires à l'application du théorème :

$$\begin{aligned}\forall T \in \mathbb{N}, \qquad |(m(I_T,J_T))_k - (m(p_T,q_T))_k| &\leqslant \|m(I_T,J_T) - m(p_T,J_T)\|_\infty \\ &\leqslant \|m(I_T,J_T) - m(p_T,J_T)\|_2 \\ &\leqslant 2M\ ,\end{aligned}$$

d'où

$$\forall T \in \mathbb{N}, \qquad |(Y_T)_k| \leqslant 2M \qquad \text{ps.}$$

On obtient donc directement :

$$\mathbb{P}\left(\left|\sum_{t=1}^{T}(Y_t)_k\right| \leqslant \sqrt{8TM^2\ln\left(\frac{2}{\delta}\right)}\right) \geqslant 1-\delta.$$

D'où, comme ce qui précède est vrai pour toutes les coordonnées, et en divisant par T :

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{\tau=1}^{T} m(I_\tau,J_\tau) - \frac{1}{T}\sum_{\tau=1}^{T} m(p_\tau,J_\tau)\right\|_\infty \leqslant 2M\sqrt{\frac{2\ln(\frac{2}{\delta})}{T}}\right) \geqslant 1-d\delta\ ,$$

c'est-à-dire, en remplaçant δ par $\frac{\delta}{d}$:

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{\tau=1}^{T} m(I_\tau,J_\tau) - \frac{1}{T}\sum_{\tau=1}^{T} m(p_\tau,J_\tau)\right\|_\infty \leqslant 2M\sqrt{\frac{2\ln(\frac{2d}{\delta})}{T}}\right) \geqslant 1-\delta.$$

Or $\|\cdot\|_2 \leqslant \sqrt{d}\|\cdot\|_\infty$, donc avec probabilité au moins $1-\delta$:

$$\left\|\frac{1}{T}\sum_{\tau=1}^{T} m(I_\tau,J_\tau) - \frac{1}{T}\sum_{\tau=1}^{T} m(p_\tau,J_\tau)\right\|_2 \leqslant 2M\sqrt{\frac{2d\ln(\frac{2d}{\delta})}{T}}.$$

On avait dans le cas déterministe :

$$\inf_{c\in C}\left\|c - \frac{1}{T}\sum_{t=1}^{T} m(p_t,q_t)\right\|_2 \leqslant \frac{2M}{\sqrt{T}}\ ,$$

en particulier si le joueur 2 joue la stratégie $q_T = \delta_{J_T}$, la stratégie p_t sera la même que si les J_T avaient été tirés successivement, car l'algorithme proposé n'utilise que le gain $\overline{m}_T$ pour déterminer le coup suivant, et l'inégalité obtenue est donc :

$$\inf_{c\in C}\left\|c - \frac{1}{T}\sum_{t=1}^{T} m(p_t,J_t)\right\|_2 \leqslant \frac{2M}{\sqrt{T}}.$$

En combinant les deux inégalités, on a qu'avec probabilité au moins $1-\delta$:

$$D_T = \inf_{c\in C} \left\| c - \frac{1}{T}\sum_{t=1}^{T} m(I_t, J_t) \right\|_2 \leqslant \frac{2M}{\sqrt{T}} \left(1 + \sqrt{2d \ln\left(\frac{2d}{\delta}\right)} \right).$$

On a donc prouvé que D_T tendait en probabilité vers 0, et donné une majoration de la vitesse de convergence. Il reste à prouver que D_T tend vers 0 presque sûrement. On va même prouver que $D_T \leqslant K\sqrt{\frac{\ln T}{T}}$ avec K une constante.

On utilise pour cela le lemme de Borel-Cantelli (Lemme 9.3.1 [3]) : on vient de voir qu'avec probabilité au moins $1-\frac{1}{T^2}$, pour tout $T \geqslant 2$:

$$\begin{aligned}
D_T &\leqslant \frac{2M}{\sqrt{T}} \left(1 + \sqrt{2d\ln(2dT^2)} \right) \\
&\leqslant \frac{2M}{\sqrt{T}} \left(1 + \sqrt{2d(2\ln(T) + \ln(2d))} \right) \\
&\leqslant \frac{2M}{\sqrt{T}} \left(1 + \sqrt{4d\ln(T)} + \sqrt{2d\ln(2d)} \right) car\ \sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b} \\
&\leqslant \frac{2M}{\sqrt{T}} \left(\left(1 + \sqrt{2d\ln(2d)} \right) \sqrt{\frac{\ln T}{\ln(2)}} + \sqrt{4d\ln(T)} \right) \\
&\leqslant K\sqrt{\frac{\ln T}{T}}\ ,
\end{aligned}$$

avec $K = 2M\left(\left(1 + \sqrt{2d\ln(2d)} \right) \frac{1}{\sqrt{\ln(2)}} + \sqrt{4d\ln(2dT)} \right)$.

Donc si on considère $A_T = \left\{ D_T \leqslant K\sqrt{\frac{\ln T}{T}} \right\}$ on a : $\mathbb{P}(A_T) \geqslant 1 - \frac{1}{T^2}$,

c'est-à-dire

$$\mathbb{P}({}^cA_T) \leqslant \frac{1}{T^2}\ ,$$

donc par Borel-Cantelli,

$$\mathbb{P}(\limsup\ {}^cA_T) = 0.$$

Ainsi ps il existe un nombre fini de T tels que $D_T > K\sqrt{\frac{\ln T}{T}}$.

Finalement ps, $\limsup_{T\to\infty} D_T \sqrt{\frac{T}{\ln T}} \leqslant K$.

□

3 Le cas bandits

Jusqu'à présent, dans les cas que nous avons vus, l'action q_T ou J_T était révélée au joueur 1 à la fin de chaque tour. Dans le cas déterministe, il suffisait de connaître $m(p_T, q_T)$ pour construire l'algorithme d'approchabilité. On aimerait avoir un résultat similaire dans le cas où l'action J_T n'est pas révélée. C'est ce qui se passe dans le cas "bandits" : on se place de nouveau avec des stratégies aléatoires, mais cette fois-ci, seul $m(I_T, J_T)$ est révélé au joueur 1 à la fin de chaque tour.

Un théorème de Blackwell existe toujours, mais la démonstration est légèrement différente.

Définition 5. *On appelle stratégie (pour le joueur 1), dans le cas bandits, une suite de fonctions $f_T : (\mathbb{R}^d)^{T-1} \mapsto \Delta(\mathcal{A})$, qui à partir des gains passés $m(I_t, J_t)$ $(t < T)$ donne le coup suivant, p_T pour le joueur 1 (ou $q_T \in \Delta(\mathcal{B})$ pour le joueur 2).*

La définition d'un ensemble approchable est la même que dans le cas des stratégies aléatoires. Cette fois, la tribu $\mathcal{F}_T$ est définie par avec $\mathcal{F}_T := \sigma(I_t, m(I_t, J_t), 1 \leqslant t \leqslant T)$.

Théorème 4 (Blackwell, cas bandit)**.**
Soit C un convexe fermé. Alors C est approchable si et seulement si

$$\forall q \in \Delta(\mathcal{B}), \exists p \in \Delta(\mathcal{A}), \qquad m(p, q) \in C.$$

Si on note $\overline{m}_T := \frac{1}{T}\sum_{\tau=1}^{T} m(I_\tau, J_\tau)$ et π_T sa projection sur C, la stratégie utilisée pour approcher C est la suivante : p_{T+1} résout

$$\min_{p\in\Delta(\mathcal{A})} \max_{q\in\Delta(\mathcal{B})} < \overline{m}_T - \pi_T, m(p, q) > ,$$

De plus, si l'on note $M = \sup_{(i,j)\in\mathcal{A}\times\mathcal{B}} \|m(i, j)\|_2$, alors on a :

$$\mathbb{E}[D_T^2] \leqslant \frac{4M}{T} \text{ et avec probabilité au moins } 1-\delta, \sup_{t\geqslant T} D_t \leqslant \sqrt{\frac{8M^2}{dT}}.$$

Remarque : dans le cas précédent, on obtenait la majoration "avec probabilité au moins $1 - \delta$, $D_T \leqslant \frac{2M}{\sqrt{T}}\left(1 + 2\sqrt{d \ln\left(\frac{2d}{\delta}\right)}\right)$".
La vitesse de convergence vis-à-vis de T est donc la même mais on a bien un résultat plus faible pour ce qui est de la précision en δ : si la norme de la matrice vaut 1 en dimension 2, et que la condition nécessaire et suffisante est vérifiée, le théorème donne par exemple qu'avec au moins 99% de chance, on aura pour $T = 3200$ coups, que $\sup_{t\geqslant T} D_t \leqslant \frac{1}{2}$ et pour savoir qu'avec au moins $99,99\%$ de chance on a $\sup_{t\geqslant T} D_t \leqslant \frac{1}{2}$, il faut attendre 320 000 coups environ. Une très bonne certitude est cette fois très couteuse. Néanmoins, la majoration porte cette fois sur le $\sup_{t\geqslant T}$.

Démonstration. On reprend ici le schéma de preuve donné dans [5]. Cette preuve est très semblable à celle donnée dans la partie consacrée aux stratégies déterministes, mais cette fois-ci, on travaille avec des variables aléatoires.

On constate que, d'après le choix de la stratégie,

$$\mathbb{E}\left[< \overline{m}_T - \pi_T, m(I_{T+1}, J_{T+1}) > |\mathcal{F}_T\right] \;\leqslant\; < \overline{m}_T - \pi_T, \pi_T >, \tag{3}$$

$$\begin{aligned} D_{T+1}^2 &\leqslant \|\overline{m}_{T+1} - \pi_T\|^2 \\ &= \|\overline{m}_T - \pi_T\|^2 + 2 < \overline{m}_T - \pi_T, \overline{m}_{T+1} - \overline{m}_T > + \|\overline{m}_{T+1} - \overline{m}_T\|^2. \end{aligned}$$

Or , tout comme en page 5 :

$$\begin{aligned}\overline{m}_{T+1} - \overline{m}_T &= \frac{1}{T+1}\left[m(I_{T+1}, J_{T+1}) - \overline{m}_T\right] \\ &= \frac{1}{T+1}\left[(m(I_{T+1}, J_{T+1}) - \pi_T) - (\overline{m}_T - \pi_T)\right].\end{aligned}$$

D'où :

$$D_{T+1}^2 \leqslant \|\overline{m}_T - \pi_T\|^2 + \|\overline{m}_{T+1} - \overline{m}_T\|^2 + \\ + \frac{2}{T+1}\left(< m(I_{T+1}, J_{T+1}) - \pi_T, \overline{m}_T - \pi_T > - \|\overline{m}_T - \pi_T\|^2\right)$$

$$D_{T+1}^2 \leqslant \left(1 - \frac{2}{T+1}\right)\|\overline{m}_T - \pi_T\|^2 + \|\overline{m}_{T+1} - \overline{m}_T\|^2 + \\ + \frac{2}{T+1} < m(I_{T+1}, J_{T+1}) - \pi_T, \overline{m}_T - \pi_T > . \quad (4)$$

On a donc :

1. D_T est $\mathcal{F}_T$-mesurable,
2. $\mathbb{E}\left[< m(I_{T+1}, J_{T+1}) - \pi_T, \overline{m}_T - \pi_T > |\mathcal{F}_T\right] \leqslant 0$ par l'inégalité (3),
3. $\overline{m}_{T+1} - \overline{m}_T = \frac{1}{T+1}\left[m(I_{T+1}, J_{T+1}) - \overline{m}_T\right]$.

Donc en prenant l'espérance conditionnelle selon $\mathcal{F}_T$ (qui passe sans problème aux inégalités ps) dans (4) on obtient :

$$\mathbb{E}\left[D_{T+1}^2|\mathcal{F}_T\right] \leqslant \left(1 - \frac{2}{T+1}\right)D_T^2 + \frac{1}{(T+1)^2}\mathbb{E}\left[\|m(I_{T+1}, J_{T+1}) - \overline{m}_T\|^2|\mathcal{F}_T\right].$$

On remarque que par inégalité triangulaire,

$$\mathbb{E}\left(\|m(I_{T+1}, J_{T+1}) - \overline{m}_T\|^2\right) \leqslant 4M^2.$$

On va prouver que pour toute suite de variables aléatoires (D_T) $\mathcal{F}_T$-mesurables vérifiant :

$$\mathbb{E}\left[D_{T+1}^2|\mathcal{F}_T\right] \leqslant \left(1 - \frac{2}{T+1}\right)D_T^2 + \frac{1}{(T+1)^2}\mathbb{E}\left[u_T|\mathcal{F}_T\right] , \quad (5)$$

avec (u_T) une suite de variables aléatoires telle que $\forall t,\ \mathbb{E}[u_t] \leqslant 4M^2$, on a la conclusion du théorème :

$$\mathbb{E}[D_T^2] \leqslant \frac{4M^2}{T} \text{ et avec probabilité au moins } 1-\delta, \sup_{t \geqslant T} D_t \leqslant \sqrt{\frac{8M^2}{\delta T}}.$$

La première conclusion s'obtient, tout comme à la partie I, par récurrence sur $\mathbb{E}[D_T^2]$, en passant aux espérances dans (5) et en utilisant le fait que pour toute variable aléatoire X et pour toute tribu $\mathcal{F}$, $\mathbb{E}\left[\mathbb{E}\left[X|\mathcal{F}\right]\right] = \mathbb{E}[X]$. En effet on obtient :

$$\mathbb{E}[D_{T+1}^2] \leqslant \left(1 - \frac{2}{T+1}\right)\mathbb{E}[D_T^2] + \frac{1}{(T+1)^2}4M^2 ,$$

la même relation de récurrence que dans la partie I. Donc $\mathbb{E}[D_T^2] \leqslant \frac{4M}{T}$, d'où $\mathbb{E}(D_T^2) \to 0$ et par l'inégalité de Markov, comme $D_T^2 \geqslant 0$, $D_T \to 0$ en probabilité.

Pour montrer la convergence presque sûre et déterminer un majorant de la vitesse de convergence, on va utiliser un argument de surmartingales. En effet, si on considère :

$$Z_T := D_T^2 + \mathbb{E}\left[\sum_{i=T}^{\infty} \frac{u_i}{(i+1)^2}\middle|\mathcal{F}_T\right],$$

on vérifie bien que :

$$\begin{aligned}
\mathbb{E}[Z_{T+1}|\mathcal{F}_T] &= \mathbb{E}\left[D_{T+1}^2 + \mathbb{E}\left[\sum_{i=T+1}^{\infty} \frac{u_i}{(i+1)^2}\middle|\mathcal{F}_{T+1}\right]\middle|\mathcal{F}_T\right] \\
&= \mathbb{E}\left[D_{T+1}^2|\mathcal{F}_T\right] + \mathbb{E}\left[\mathbb{E}\left[\sum_{i=T+1}^{\infty} \frac{u_i}{(i+1)^2}\middle|\mathcal{F}_{T+1}\right]\middle|\mathcal{F}_T\right] \\
&= \mathbb{E}\left[D_{T+1}^2|\mathcal{F}_T\right] + \mathbb{E}\left[\sum_{i=T+1}^{\infty} \frac{u_i}{(i+1)^2}\middle|\mathcal{F}_T\right] \text{ car } \mathcal{F}_T \subseteq \mathcal{F}_{T+1}, \\
&\leqslant \left(1 - \frac{2}{T+1}\right) D_T^2 + \mathbb{E}\left[\frac{u_T}{(T+1)^2}\middle|\mathcal{F}_T\right] + \mathbb{E}\left[\sum_{i=T+1}^{\infty} \frac{u_i}{(i+1)^2}\middle|\mathcal{F}_T\right] \text{ par (5)}, \\
&\leqslant Z_T,
\end{aligned}$$

et Z_T est bien $\mathcal{F}_T$-mesurable, car par définition d'une espérance conditionnelle, $\mathbb{E}[X|\mathcal{F}]$ est $\mathcal{F}$-mesurable. Donc $(Z_T)_{T\in\mathbb{N}}$ est bien une surmartingale positive. Donc $(Z_T)_{T\in\mathbb{N}}$ converge presque surement vers une variable aléatoire Z. En outre,

$$\begin{aligned}
\mathbb{E}[Z_T] &= \mathbb{E}[D_T^2] + \mathbb{E}\left[\sum_{i=T}^{\infty} \frac{u_i}{(i+1)^2}\right] \text{ donc par Fubini-Tonnelli} \\
\mathbb{E}[Z_T] &\leqslant \mathbb{E}[D_T^2] + 4M^2 \sum_{i=T}^{\infty} \frac{1}{(i+1)^2} \text{ car } \mathbb{E}[u_i] \leqslant 4M^2 \text{ et finalement} \\
\mathbb{E}[Z_T] &\leqslant \frac{4M^2}{T} + \frac{4M^2}{T} = \frac{8M^2}{T}.
\end{aligned}$$

Donc la limite est nulle ps (car par exemple, par le lemme de Fatou, Z vérifie $\mathbb{E}(Z) \leqslant \liminf \mathbb{E}(Z_t) = 0$ avec $Z \geqslant 0$), et comme $0 \leqslant D_T^2 \leqslant Z_T$, on a bien $D_T^2 \to 0$ ps.
Pour trouver la vitesse de convergence, on va utiliser l'inégalité maximale de Doob suivante, dont on donnera en annexe la démonstration et les liens avec l'inégalité maximale de Doob du cours d'intégration :

Théorème 5 (inégalité maximale de Doob)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une surmartingale positive, alors :*

$$\forall \lambda > 0, \qquad \lambda\, \mathbb{P}\left(\sup_{n\leqslant k} X_k \geqslant \lambda\right) \leqslant \mathbb{E}(X_n).$$

Appliqué à la surmartingale Z_T, cela donne :

$$\mathbb{P}(\sup_{T\leqslant t} Z_t \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon}\mathbb{E}(Z_T) \leqslant \frac{8M^2}{\varepsilon T},$$

c'est-à-dire avec probabilité au moins $1-\delta$:

$$\begin{aligned}
\sup_{T\leqslant t} Z_t &\leqslant \frac{8M^2}{\delta T}, \text{ d'où} \\
\sup_{T\leqslant t} D_t &\leqslant \sqrt{\frac{8M^2}{\delta T}} \text{ ce qui est le résultat désiré.}
\end{aligned}$$

□

4 Annexe

4.1 Le théorème du minimax

En théorie des jeux, le théorème du minimax de von Neumann joue un rôle fondamental. Nous nous proposons ici d'en donner une démonstration, issue de [6] (p.15). Pour cela, nous avons besoin du lemme suivant :

Lemme 1. *Soit $A = (a_{i,j})$ une matrice $m \times n$, alors l'une des deux propriétés suivantes est vérifiée :*

(i) le point $\underline{0}$ (de $\mathbb{R}^m$) est dans l'enveloppe convexe des n colonnes de A et des m vecteurs de la base canonique, c'est à dire :

$$a_1 = \begin{bmatrix} a_{1,1} \\ \vdots \\ a_{m,1} \end{bmatrix}, \quad \dots, \quad a_n = \begin{bmatrix} a_{1,n} \\ \vdots \\ a_{m,n} \end{bmatrix}$$

et

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots, \quad e_m = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

(ii) il existe $x_1, \dots, x_m \in \mathbb{R}$ tels que :
$\forall j, \quad x_j > 0$,
$\sum_{i=1}^m x_i = 1$,
$\sum_{i=1}^m a_{i,j} x_i > 0 \quad$ *pour tout j de 1 à n.*

Démonstration. On notera $x = (x_1, ..., x_m)$ les points de $\mathbb{R}^m$. Supposons que (i) soit faux, par le théorème de Hahn Banach ($2^{ème}$ forme géométrique), il existe un hyperplan H séparant strictement $\underline{0}$ de l'enveloppe convexe S des $m+n$ points précédents, c'est à dire :

$$\exists p \in \mathbb{R}^m, \exists a, \exists \varepsilon > 0, \quad \sum_{j=1}^m p_j \times 0 \leqslant a \quad \text{et} \quad \forall y \in S, \quad \sum_{j=1}^m p_j \times y_j \geqslant a + \varepsilon$$

(note : si les inégalités sont dans l'autre sens, c'est à dire

$$\exists p \in \mathbb{R}^m, \exists a, \exists \varepsilon > 0, \quad \sum_{j=1}^m p_j \times 0 \geqslant a + \varepsilon \quad \text{et} \quad \forall y \in S, \quad \sum_{j=1}^m p_j \times y_j \leqslant a,$$

on peut se ramener à ce cas en changeant $a + \varepsilon$ en $-a$ et p en $-p$).

On a donc $a \geqslant 0$ et

$$\forall y \in S, \qquad \sum_{j=1}^m p_j y_j > 0.$$

En particulier, quand y est l'un des $m+n$ vecteurs, on obtient :
$\forall j, \quad \sum_{i=1}^n a_{i,j} p_i > 0$,
$\forall i, \quad p_i > 0$.

On peut alors poser

$$x_i := p_i \Big/ \sum_{j=1}^n p_j.$$

Et on obtient donc
$x_i > 0$,
$\sum a_{i,j} x_i > 0$,
$\sum x_i = 1$.
Ce qui est le résultat attendu. □

Nous allons maintenant montrer le théorème suivant (dû à von Neumann).

Théorème 6 (Théorème du minimax)**.**
Soit $A = (a_{i,j})$ *une matrice* $m \times n$, $X = \{x \in \mathbb{R}^m \ / \ \sum x_i = 1 \text{ et } x_i \geqslant 0\}$ *et* $Y = \{y \in \mathbb{R}^n \ / \ \sum y_i = 1 \text{ et } y_i \geqslant 0\}$. *Alors si*

$$m_1 := \max_{x \in X} \min_{y \in Y} {}^t x A y,$$

$$m_2 := \min_{y \in Y} \max_{x \in X} {}^t x A y,$$

on a :

$$m_1 = m_2.$$

Démonstration. Tout d'abord, pour m_1, remarquons que ${}^t x A y = \sum_j y_j \sum_i x_i m_{i,j}$ et donc pour le minimiser, il suffit de choisir y_j valant 0 pour tout j, sauf là où $\sum_i x_i m_{i,j}$ est minimale. D'où $\min_{y \in Y} {}^t x A y = \min_j \sum_i x_i m_{i,j}$ et le *min* d'un nombre fini de fonctions continues est continu. Donc $x \mapsto \min_{y \in Y} {}^t x A y$ est continue et atteint son *max* sur le compact X, ce qui justifie le fait que les *inf* et les *sup* soient des *min* et des *max* (même raisonnement pour m_2).

On va raisonner par double inégalité :

1. L'inégalité $m_1 \leqslant m_2$ est évidente : il suffit de remarquer que pour $x \in X$ et $y \in Y$ on a

   $$\min_{y' \in Y} {}^t x A y' \leqslant {}^t x A y \leqslant \max_{x' \in X} {}^t x' A y$$

   et donc en passant au *max* sur x à gauche, on obtient :

   $$\max_{x \in X} \min_{y' \in Y} {}^t x A y' \leqslant \max_{x' \in X} {}^t x' A y,$$

   puis en passant au *min* sur y à droite, on a bien :

   $$\max_{x \in X} \min_{y' \in Y} {}^t x A y' \leqslant \min_{y \in Y} \max_{x' \in X} {}^t x' A y.$$

2. Pour l'autre inégalité, nous allons utiliser le lemme précédent pour montrer que selon que l'on a (i) ou (ii), on a soit $m_2 \leqslant 0$, soit $m_1 > 0$, et on en déduira que $m_1 \leqslant 0 < m_2$ est impossible puis, par translation, que $m_1 < k < m_2$ n'est vrai pour aucun k.
   Si (i) est vrai, $\underline{0}$ est combinaison convexe des $m + n$ vecteurs précédents, donc il existe $s_1, \ldots, s_{m+n}$ tels que

   $$\sum_{j=1}^{n} s_j a_j + \sum_{i=1}^{m} s_{n+i} e_i = \underline{0}\ ,$$

   $$\text{soit} \quad \sum_{j=1}^{n} s_j a_{i,j} + s_{n+i} = 0 \qquad \forall i = 1, \ldots, m$$

   $$s_j \geqslant 0 \qquad \forall j = 1, \ldots, m+n$$

   $$\sum_{j=1}^{m+n} s_j = 1.$$

$\underline{0}$ n'étant pas combinaison linéaire non triviale des $e_1, \ldots, e_m$, l'un des s_j, $1 \leqslant j \leqslant n$ est non nul, et donc $\sum_{j=1}^{n} s_j > 0$. Si l'on pose alors $y_j := s_j / \sum_{j=1}^{n} s_j$, on a :

$$y_j \geqslant 0$$
$$\sum_{j=1}^{n} y_j = 1$$
$$\sum_{j=1}^{n} a_{i,j} y_j = -s_{n+i} / \sum_{j=1}^{n} s_j \leqslant 0.$$

Donc chaque composante de Ay est négative, d'où :

$$m_2 := \min_{y' \in Y} \max_{x \in X} {}^t x A y' \leqslant \max_{x \in X} {}^t x A y \leqslant 0.$$

Si au contraire (ii) est vrai, en considérant $x \in \mathbb{R}$ donné par le lemme, on a alors :

$$m_1 := \max_{x' \in X} \min_{y \in Y} {}^t x' A y \geqslant \min_{y \in Y} {}^t x A y > 0,$$

car toutes les composantes de ${}^t x A$ sont strictement positives.

Donc il n'est pas possible d'avoir $m_1 \leqslant 0 < m_2$. Or en remplaçant la matrice A par la matrice $B := (a_{i,j} - k)$, $k \in \mathbb{R}$, on obtient pour $(x, y) \in X \times Y$

$${}^t x B y = {}^t x A y - k.$$

Donc

$$m_1(B) = m_1(A) - k,$$
$$m_2(B) = m_2(A) - k.$$

Puisque $m_1(B) < 0 < m_2(B)$ est impossible, on en déduit qu'on ne peut pas avoir

$$m_1(A) < k < m_2(A)$$

et ceci pour aucun $k \in \mathbb{R}$. Donc on a $m_1 \geqslant m_2$, ce qui conclut la preuve.

□

4.2 Inégalité de Hoeffding-Azuma

Les démonstrations du lemme et du théorème qui suivent peuvent être trouvées dans le livre [2].

Lemme 2 (Inégalité d'Hoeffding)**.**
Soit Y une variable aléatoire réelle bornée. Soient $\mathcal{F}$ une tribu, c une variable aléatoire $\mathcal{F}-$mesurable et $\ell > 0$ une constante. On suppose $\mathbb{E}[Y|\mathcal{F}] = 0$ et $\mathbb{P}(c \leqslant Y \leqslant c + \ell) = 1$. Alors pour tout réel s on a :

$$\mathbb{E}[e^{sY}|\mathcal{F}] \leqslant \exp\left(\frac{s^2 \ell^2}{8}\right) \quad ps.$$

Démonstration. On utilise tout d'abord la convexité de e^{st} sur $[c; c+\ell]$: en écrivant

$$t = \frac{c + \ell - t}{\ell} c + \frac{t - c}{\ell}(c + \ell),$$

on obtient

$$\forall t \in [c; c+\ell], \quad e^{st} \leqslant \frac{c + \ell - t}{\ell} e^{sc} + \frac{t - c}{\ell} e^{s(c+\ell)}.$$

En appliquant cette inégalité à Y (qui est tel que $c \leqslant Y \leqslant c+\ell$ p.s) et en passant à l'espérance conditionnelle, on a :

$$\mathbb{E}[e^{sY}|\mathcal{F}] \leqslant \frac{c+\ell}{\ell}e^{sc} - \frac{c}{\ell}e^{s(c+\ell)} =: f(s),$$

car $\mathbb{E}[Y|\mathcal{F}] = 0$. On pose alors :

$$\begin{aligned} u &= \ell s, \\ \psi(u) &= \ln\left(f(s)\right), \\ p &= \frac{-c}{\ell}, \\ 1-p &= \frac{c+\ell}{\ell}. \end{aligned}$$

Ainsi

$$\begin{aligned} \psi(u) &= \ln\left(f\left(\frac{u}{\ell}\right)\right) \\ &= \ln\left(\frac{c+\ell}{\ell}e^{\frac{uc}{\ell}} + \frac{-c}{\ell}e^{\frac{u(c+\ell)}{\ell}}\right) \\ &= \ln\left(e^{\frac{uc}{\ell}}\left(\frac{c+\ell}{\ell} + \frac{-c}{\ell}e^{u}\right)\right) \\ &= \frac{uc}{\ell} + \ln\left(1 + \frac{c}{\ell} - \frac{c}{\ell}e^{u}\right). \end{aligned}$$

En remplaçant :

$$\psi(u) = -pu + \ln(1-p+pe^u).$$

On constate que : $\psi(0) = 0$ et

$$\forall u, \quad \psi'(u) = -p + \frac{pe^u}{1-p+pe^u}.$$

Donc en particulier, $\psi'(0) = -p + \frac{pe^0}{1-p+pe^0} = -p + p = 0$ et de plus,

$$\begin{aligned} \psi''(u) &= \frac{pe^u(1-p+\cancel{pe^u}) - \cancel{pe^u}\cancel{pe^u}}{(1-p+pe^u)^2} \\ &= \frac{(1-p)e^u}{(1-p+pe^u)^2} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2} \text{ (avec des notations évidentes)} \\ &\leqslant \frac{1}{4} \text{ car} (\alpha+\beta)^2 - 4\alpha\beta = (\alpha-\beta)^2 \geqslant 0. \end{aligned}$$

Finalement, par l'inégalité de Taylor Lagrange, pour tout u il existe $\theta \in [0;1]$ tel que :

$$\psi(u) = \psi(0) + u\psi'(0) + \psi''(\theta u)\frac{u^2}{2} \leqslant \frac{u^2}{8},$$

ce qui se réécrit :

$$\mathbb{E}[e^{sY}|\mathcal{F}] \leqslant f(s) \leqslant \exp\left(\frac{s^2\ell^2}{8}\right).$$

□

Théorème 7 (inégalité d'Hoeffding-Azuma)**.**
Soit $(\mathcal{F}_n)_{n\geqslant 0}$ une filtration, (c_n) et (ℓ_n) deux suites réelles et $(Y_n)_{n\geqslant 1}$ une suite de variables aléatoires telles que $\mathbb{E}[Y_n|\mathcal{F}_{n-1}]=0$ et $c_n \leqslant Y_n \leqslant c_n+\ell_n$ ps. Alors pour tout $\lambda \geqslant 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \geqslant \lambda\right) \leqslant 2\exp\left(-\frac{2\lambda^2}{\sum_{i=1}^{n}\ell_i^2}\right).$$

En particulier, si $|Y_n|<c$ ps, alors si $0<\delta<1$, on obtient qu'avec probabilité au moins $1-\delta$,

$$\left|\sum_{i=1}^{n} Y_i\right| \leqslant c\sqrt{2n\ln\left(\frac{2}{\delta}\right)}.$$

Remarque : si on suppose de plus Y_n $\mathcal{F}_n$-mesurable et si on pose $M_0:=0$ et $M_n:=Y_1+\ldots+Y_n$ pour $n\geqslant 1$, on remarque que $\mathbb{E}[Y_n|\mathcal{F}_{n-1}]=0$ est équivalent à dire que (Y_n) est une accroissement de martingale (ie : $Y_n=M_n-M_{n-1}$ avec (M_n) une martingale).

Démonstration. Pour tout $s>0$ l'inégalité de Markov donne :

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \geqslant \lambda\right) &= \mathbb{P}\left(e^{s\sum_{i=1}^{n} Y_i} \geqslant e^{s\lambda}\right)\\ &\leqslant \mathbb{E}\left[e^{s\sum_{i=1}^{n} Y_i}\right]e^{-\lambda s}\\ &= \mathbb{E}\left[e^{s\sum_{i=1}^{n} Y_i}e^{-\lambda s}\right]\end{aligned}$$

On va chercher à majorer $\mathbb{E}[e^{sY_m}|F_{m-1}]$. On a

$$\mathbb{E}[Y_n|\mathcal{F}_{n-1}]=0,$$

$$\mathbb{P}(c_n\leqslant Y_n\leqslant c_n+\ell_n|\mathcal{F}_{n-1})=1.$$

Or la variable c_n est $\mathcal{F}_{n-1}$-mesurable, donc on peut appliquer le Lemme 2 et obtenir la majoration désirée :

$$\mathbb{E}\left[e^{sY_m}\middle|F_{m-1}\right] \leqslant \exp\left(\frac{s^2\ell_m^2}{8}\right).$$

Par suite :

$$\begin{aligned}\mathbb{E}\left[e^{s(Y_1+Y_2+\ldots+Y_n)}\right] &= \mathbb{E}\left[\mathbb{E}[e^{s(Y_1+Y_2+\ldots+Y_n)}\middle|\mathcal{F}_{n-1}]\right]\\ &= \mathbb{E}\left[e^{s(Y_1+Y_2+\ldots+Y_{n-1})}\mathbb{E}[e^{s(Y_n)}\middle|\mathcal{F}_{n-1}]\right]\\ &\leqslant \mathbb{E}\left[e^{s(Y_1+Y_2+\ldots+Y_{n-1})}\right]e^{\frac{s^2\ell_m^2}{8}}.\end{aligned}$$

Puis on obtient immédiatement par récurrence :

$$\mathbb{E}\left[e^{s(Y_1+Y_2+\ldots+Y_n)}\right] \leqslant \exp\left(\frac{1}{8}\sum_{i=1}^{n}\ell_i^2 s^2\right).$$

On a alors prouvé :

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \geqslant \lambda\right) \leqslant \exp\left(\frac{1}{8}\sum_{i=1}^{n}\ell_i^2 s^2\right)e^{-\lambda s},$$

inégalité vraie pour tout $s > 0$. On va donc chercher un s_0 pour lequel la majoration est optimale. La fonction exponentielle étant croissante il suffit de trouver le point s_0 où le polynôme en s : $\frac{1}{8}\sum_{i=1}^{n} \ell_i^2 s^2 - \lambda s$ atteint son minimum. On obtient immédiatement par dérivation :

$$s_0 = \frac{4\lambda}{\sum_{i=1}^{n} \ell_i^2},$$

d'où on tire la majoration :

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geqslant \lambda\right) \leqslant \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^{n} \ell_i^2}\right),$$

qui par symétrie donne en passant à des valeurs absolues :

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \geqslant \lambda\right) \leqslant 2\exp\left(-\frac{2\lambda^2}{\sum_{i=1}^{n} \ell_i^2}\right),$$

ce qui est l'inégalité d'Hoeffding-Azuma.

Pour la deuxième partie du théorème, il s'agit du cas particulier où $\ell_i = 2c$. On a alors

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \geqslant \lambda\right) \leqslant 2\exp\left(-\frac{\lambda^2}{2nc^2}\right),$$

il suffit alors de poser

$$\lambda := c\sqrt{2n\ln\left(\frac{2}{\delta}\right)},$$

et on a bien avec probabilité au moins $1-\delta$,

$$\left|\sum_{i=1}^{n} Y_i\right| \leqslant c\sqrt{2n\ln\left(\frac{2}{\delta}\right)}.$$

□

4.3 Inégalités maximales de Doob

Dans le cas Bandits, une inégalité maximale de Doob est nécessaire. Ces inégalités permettent de décrire le comportement du $\sup_{m\leqslant t\leqslant M}$ d'une martingale par rapport au comportement de la martingale en m ou M.

L'inégalité la plus classique, dont on trouvera une preuve dans [3] (Théorème 12.4.2), concerne les sous-martingales. Elle décrit la déviation (probabilité d'un écart à 0 plus grand qu'un λ) du *sup* entre 0 et n par rapport à l'état en n.

Théorème 8 (inégalité maximale de Doob, sous-martingales)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une sous-martingale, alors :*

$$\forall \lambda > 0,\ \forall n \in \mathbb{N}, \qquad \lambda\ \mathbb{P}\left(\sup_{0\leqslant k\leqslant n} X_k \geqslant \lambda\right) \leqslant \mathbb{E}\left[X_n \mathbb{1}_{\left\{\sup_{0\leqslant k\leqslant n} X_k \geqslant \lambda\right\}}\right] \leqslant \mathbb{E}(X_n^+).$$

L'inégalité dont nous avons besoin concerne les surmartingales et est beaucoup moins connue. La démonstration donnée est donc inspirée de celle du théorème ci-dessus. Cette inégalité décrit la déviation du *sup* après n par rapport à l'état en n.

Théorème 9 (inégalité maximale de Doob, surmartingales)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une surmartingale positive, alors :*

$$\forall \lambda > 0,\ \forall n \in \mathbb{N}, \qquad \lambda\ \mathbb{P}\left(\sup_{n\leqslant k} X_k \geqslant \lambda\right) \leqslant \mathbb{E}(X_n).$$

Démonstration.
Soit (X_n) une surmartingale (ie : $\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] \leqslant 0$), et $\lambda > 0$.

On pose $T := inf\{n \geqslant 0/X_n \geqslant \lambda\}$ et $A := \{\max_{0\leqslant k\leqslant n} X_k \geqslant \lambda\} = \{T \leqslant n\}$. Alors :

$$X_{T\wedge n} = X_0 + \sum_{k=1}^{n}(X_k - X_{k-1})\mathbb{1}_{T\geqslant k}.$$

Or $\mathbb{1}_{T\geqslant k} = 1 - \mathbb{1}_{T\leqslant k-1}$ est $\mathcal{F}_{k-1}$-mesurable et donc :

$$\begin{aligned}
\mathbb{E}[X_{T\wedge n}] &= \mathbb{E}[X_0] + \sum_{k=1}^{n}\mathbb{E}[(X_k - X_{k-1})\mathbb{1}_{T\geqslant k}] \\
&= \mathbb{E}[X_0] + \sum_{k=1}^{n}\mathbb{E}[\mathbb{E}[(X_k - X_{k-1})|\mathcal{F}_{k-1}]\mathbb{1}_{T\geqslant k}] \\
&\leqslant \mathbb{E}[X_0].
\end{aligned}$$

Or

$$X_{T\wedge n} \geqslant \lambda\mathbb{1}_A + X_n\mathbb{1}_{A^c}.$$

D'où :

$$\mathbb{E}[X_0] \geqslant \mathbb{E}[X_{T\wedge n}] \geqslant \mathbb{E}[\lambda\mathbb{1}_A + X_n\mathbb{1}_{A^c}] \geqslant \lambda\mathbb{P}(A).$$

car $X_n \geqslant 0$. On a donc :

$$\lambda\ \mathbb{P}\left(\max_{0\leqslant k\leqslant n} X_k \geqslant \lambda\right) \leqslant \mathbb{E}[X_0],$$

puis par suite (en appliquant le résultat à $(X_{p+n})_{p\geqslant 0}$), pour $n \leqslant N$:

$$\lambda\ \mathbb{P}\left(\max_{n\leqslant k\leqslant N} X_k \geqslant \lambda\right) \leqslant \mathbb{E}[X_n].$$

Cette inégalité étant vraie pour tout N, on a :

$$\begin{aligned}
\mathbb{P}\left(\max_{n\leqslant k} X_k \geqslant \lambda\right) &= \mathbb{P}\left(\bigcup_{N\in\mathbb{N}, N\geqslant n}\left\{\max_{n\leqslant k\leqslant N} X_k \geq \lambda\right\}\right) \\
&= lim\uparrow \mathbb{P}\left(\max_{n\leqslant k\leqslant N} X_k \geqslant \lambda\right) \text{ car il s'agit d'une union croissante,} \\
&\leqslant \frac{1}{\lambda}\,\mathbb{E}[X_n], \text{ l'inégalité est vraie pour chaque terme de la suite.}
\end{aligned}$$

Finalement on a bien :

$$\mathbb{P}\left(\sup_{n\leqslant k} X_k \geqslant \lambda\right) \leqslant \frac{1}{\lambda}\,\mathbb{E}[X_n].$$

□

Références

[1] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1) :1–8, 1956.

[2] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge Univ Pr, 2006.

[3] J.F. Le Gall. Intégration, probabilités et processus aléatoires. *Ecole Normale Supérieure de Paris*, 2006.

[4] S. Mannor, V. Perchet, and G. Stoltz. Robust approachability and regret minimization in games with partial monitoring. *Arxiv preprint arXiv :1105.4995*, 2011.

[5] J.F. Mertens, S. Sorin, and S. Zamir. Repeated games. part a : Background material. *CORE Discussion Papers*, 1994.

[6] Guillermo Owen. *Game theorie second edition*. Academic press, 1982.

[7] F. Paulin. Topologie, analyse et calcul différentiel. *Notes de cours, École Normale Supérieure http ://www. fimfa. ens. fr/fimfa/IMG/File/cours/polyanalyse12009. pdf*.

Examen du Cours de Concentration et Sélection

Aymeric Dieuleveut

30/01/2013

Table des matières

Introduction

Le but de cette présentation est de présenter un exposé relativement concis des résultats exposés dans l'article *Model Selection for Simplicial Application* de Claire Caillerie et Bertrand Michel et de s'intéresser dans un cadre différent de celui de l'article à une des notions clés qu'est "l'heuristique de pente". Outre cet article, la référence principale est *Concentration inequalities and Model Selection*, notes du Cours donné à Saint Flour en 2003 par Pascal Massart, ainsi que pour la dernière partie les notes du cours donné par Sylvain Arlot au cours Peccot en 2011, *Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique*.

Ces notes reprennent la structure de l'exposé ainsi que les références des principaux théorèmes exposés.

Les auteurs de cet article s'intéressent au problème de la reconstitution d'un objet géométrique à partir d'un nuage de point "mesurés" sur celui-ci, c'est à dire d'un ensemble de points séparés de l'objet par une erreur de mesure morale, qui prend la forme d'une gaussienne. Le problème sera vu du point de vue de la sélection d'estimateur, et utilisera un grand nombre de résultats issus issus de la théorie générale, résultats que l'on rappellera et resituera dans la première partie. On présentera alors le cadre statitique proposé dans l'article, ainsi que le théorème principal, que l'on démontrera. Ce sera l'objet de la partie 2. Enfin, la dernière partie sera consacrée à la présentation des résultats obtenus par les auteurs de l'article, puis à une tentative de mettre en oeuvre la technique de l'heuristique de pente dans le cas de la régression homoscédastique. On cherchera à illustrer ce cas par des simulations informatiques.

Cette approche, qui reprend un certain nombre de résultats extérieurs à l'article, et se termine en cherchant à extrapoler une notion découverte dans l'article correspond à une "organisation rationnelle" de celle que j'ai suivie : il m'a fallu aller et venir entre l'article et les notes de Saint Flour pour comprendre l'article - ces notes comprenant des résultats que je n'envisagerai pas vraiment d'ommettre car ils ne me sont pas encore assez familiers-, et celui ci m'a incité à aller voir si certains résulats que j'y découvrais pouvaient être utilisés dans d'autres cadres.

Ces notes consignent principalement le développement lié à la seconde partie, qui concerne le coeur théorique de l'article. Seules des remarques et la structure des résultats des autres parties sont proposées.

1 Rappels sur la théorie de la sélection de modèles

Pour bien comprendre le problème de la sélection dans le cadre géométrique, la connaissance des résultats dans un cadre plus classique est indispensable. C'est pourquoi on présentera un certain nombre de ces résultats et évoquera un cas particulier simple pour illustrer les théorèmes énoncés.

- **Le cadre des modèles linéaires gaussiens,**
- La procédure des moindres carrés,
- Le problème de la sélection linéaire et l'heuristique de Mallows qui permet de comprendre pourquoi on doit travailler sur la pénalité.
- Le théorème dans ce cas particulier et l'inégalité oracle qui en découle.
- Le théorème le plus général, ci dessous.

Theorem 4.18 *Let $\{S_m\}_{m\in\mathcal{M}}$ be some finite or countable collection of subsets of $\mathbb{H}$. We assume that for any $m\in\mathcal{M}$, there exists some a.s. continuous version W of the isonormal process on S_m. Assume furthermore the existence of some positive and nondecreasing continuous function ϕ_m defined on $(0,+\infty)$ such that $\phi_m(x)/x$ is nonincreasing and*

$$2\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(u)}{\|t-u\|^2+x^2}\right)\right]\leq x^{-2}\phi_m(x) \tag{4.71}$$

for any positive x and any point u in S_m. Let us define $\tau_m=1$ if S_m is closed and convex and $\tau_m=2$ otherwise. Let us define $D_m>0$ such that

$$\phi_m\left(\tau_m\varepsilon\sqrt{D_m}\right)=\varepsilon D_m \tag{4.72}$$

and consider some family of weights $\{x_m\}_{m\in\mathcal{M}}$ such that

$$\sum_{m\in\mathcal{M}}e^{-x_m}=\Sigma<\infty.$$

Let K be some constant with $K>1$ and take

$$\mathrm{pen}(m)\geq K\varepsilon^2\left(\sqrt{D_m}+\sqrt{2x_m}\right)^2. \tag{4.73}$$

We set for all $t\in\mathbb{H}$, $\gamma_\varepsilon(t)=\|t\|^2-2Y_\varepsilon(t)$ and consider some collection of ρ-LSEs $(\widehat{s}_m)_{m\in\mathcal{M}}$ i.e., for any $m\in\mathcal{M}$,

$$\gamma_\varepsilon(\widehat{s}_m)\leq\gamma_\varepsilon(t)+\rho,\ \textit{for all } t\in S_m.$$

Then, almost surely, there exists some minimizer $\widehat{m}$ of $\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)$ over $\mathcal{M}$. Defining a penalized ρ-LSE as $\widetilde{s}=\widehat{s}_{\widehat{m}}$, the following risk bound holds for all $s\in\mathbb{H}$

$$\mathbb{E}_s\left[\|\widetilde{s}-s\|^2\right]\leq C(K)\left[\inf_{m\in\mathcal{M}}\left(d^2(s,S_m)+\mathrm{pen}(m)\right)+\varepsilon^2(\Sigma+1)+\rho\right] \tag{4.74}$$

Ce théorème présente l'avantage fondamental de pouvoir s'appliquer à des modèles non linéaires, et nous procure deux éléments fondamentaux :

- la forme de la pénalité à utiliser,
- une inégalité sur le rique de l'estimateur $\hat{s}_{\hat{m}}$.

On verra par la suite que c'est le premier résultat qui est le plus utile, le second est avant tout une garantie théorique.

De plus, ce théorème ne fait pas l'hypothèse que s appartient aux modèles proposés, ce qui est un avantage et n'est pas systématique.

La difficulté majeure pour pouvoir appliquer ce théorème est de déterminer la fonction Φ_m. C'est facile dans le cas linéaire, cela demande du travail dans le cas des ellipsoïdes (voir cours).

2 Modèle statistique et résultat théorique principal

On définit le ps : $< x, y >= \frac{1}{Q} \sum_{i=1}^{Q} x_i y_i$ dans $\mathbb{R}^Q$

2.1 Cadre statistique

On définit un ensemble de points $X_1, ..., X_n \in (\mathbb{R}^D)^n$ que l'on suppose proche d'un objet géométrique $\mathcal{G}$. Les point X_i sont percçus comme des observations, c'est à dire

$$\forall i \qquad X_i = \bar{x}_i + \sigma \xi_i \qquad \xi_i \sim \mathcal{N}(0, Id_D) \quad iid$$

avec bien entendu ξ_i inconnu. On note

$$X = (X_1^t, ..., X_n^t)^t$$

et on a alors

$$X = \bar{x} + \sigma \xi, \qquad \xi \sim \mathcal{N}(0, Id_{nD})$$

.

On pourra donc considérer dans la suite que l'on observe :

$$Y_\varepsilon(t) =< \bar{x}, t > +\varepsilon W(t)$$

avec

$$W(t) = \sqrt{Q} < \xi, t >, \quad Q = nD$$

processus isonormal et

$$\varepsilon = \frac{\sigma}{\sqrt{Q}}.$$

Pour approcher $\mathcal{G}$, les bons objets à considérer sont les **complexes simpliciaux**, dont on donne la définition.

Pour un complexe simplicial $\mathcal{C}_\alpha$, on définit **le meilleur point d'approximation** de X dans $\mathcal{C}_\alpha$ par

$$\hat{x}_{\mathcal{C}_\alpha} = \arg\min_{t \in \mathcal{C}_\alpha} ||X - t||^2$$

Cette définition mérite commentaire.

Bien évidemment, comme on l'a anoncé précedemment, on ne va pas procéder avec un seul complexe, mais avec une collection, indexée par un paramètre caratéristique de leur taille : α. Plus α sera grand et plus le complexe contiendra des simplexes de grandes tailles. On perçoit déjà le compromis qu'il faudra effectuer entre :

- Un complexe avec trop peu de simplexes qui approchera mal (sous apprentissage, biais grand).
- Un complexe avec trop de simplexes trop grands qui approchera mal (sur apprentissage, variance grande).

On donne également :

- La définition du **risque de l'estimateur** $\hat{x}_\alpha$.
- le rôle de la pénalité et la forme de $\hat{\alpha}$

On a vu ci dessus qu'on était dans le bon cadre pour utiliser le théorème de la partie 1. Il va s'agir pour ce faire de déterminer la fonction Φ_m, qui est liée à la taille du modèle. A cette fin, on redonne la définition de **l'entropie métrique** d'un compact S :

$$H(S, ||.||, r) = \ln N(S, ||.||, r)$$

avec $N(S, ||.||, r)$ le cardinal d'un r-set, ainsi que les inégalité liant N et $N'(S, ||.||, r)$ nombre minimal de boules de rayons r pour recouvrir S.

2.2 Résultats théoriques

2.2.1 1er théorème

Le travail effectué jusqu'ici va enfin payer ces fruits, et on est a même d'énoncer le premier résultat conséquent de l'article (pour être précis, il faudrait avoir déjà la proposition 1 pour connaitre l'intégrabilité de l'entropie métrique en 0. l'article fait le choix de présenter d'abord ce théorème dans le cadre ci dessus sans lien avec les complexes simpliciaux, puis de prendre $Q = nD$, et $C_\alpha := \mathcal{C}_\alpha^n$, ce qui est sans doute un peu plus pertinent que la présentation qui suit.)

For all $\alpha \in \mathcal{A}$, the auxiliary entropic function Φ_α is defined by

$$\Phi_\alpha(u) = \kappa \int_0^u \sqrt{\mathrm{H}(C_\alpha, \|\cdot\|, r)}\, dr.$$

In the sequel, the constant κ can be taken greater or equal to 96 although this value is not optimal for the application of Theorem 1. For all $\alpha \in \mathcal{A}$ let d_α defined by the equation (if it exits)

$$\Phi_\alpha\left(\frac{2\sigma\sqrt{d_\alpha}}{\sqrt{Q}}\right) = \frac{\sigma d_\alpha}{\sqrt{Q}}. \tag{7}$$

Also suppose that some weights w_α fulfills

$$\sum_{\alpha\in\mathcal{A}} e^{-w_\alpha} = \Sigma < \infty. \tag{8}$$

Under the previous hypotheses, Theorem 4.18 in [28] can be rewritten as follows:

Theorem 1. *Let $\eta > 1$ and take*

$$\mathrm{pen}(\alpha) \geq \eta \frac{\sigma^2}{Q}\left(\sqrt{d_\alpha} + \sqrt{2w_\alpha}\right)^2. \tag{9}$$

Then, almost surely, there exists a minimizer $\hat{\alpha}$ of the penalized criterion

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|^2 + \mathrm{pen}(\alpha).$$

Defining the penalized estimator by $\hat{\mathbf{x}}_{\hat{\alpha}}$, the following risk bound holds for all $\bar{\mathbf{x}} \in \mathbb{R}^Q$

$$\mathbb{E}_{\bar{\mathbf{x}}}\|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \leq c_\eta\left[\inf_{\alpha\in\mathcal{A}}\left\{d(\bar{\mathbf{x}}, C_\alpha)^2 + \mathrm{pen}(\alpha)\right\} + \frac{\sigma^2}{Q}(\Sigma + 1)\right] \tag{10}$$

where c_η depends only on η and $d(\bar{\mathbf{x}}, C_\alpha) := \inf_{\mathbf{y}\in C_\alpha}\|\bar{\mathbf{x}} - \mathbf{y}\|$.

Les auteurs indiquent que ce théorème est une ré-écriture du théorème de la partie 1. Cependant, un peu de travail est nécessaire pour prouver que la fonction Φ_α vérifie les hypothèses voulues : **on veut montrer l'hypothèse (4.71).**

Le premier argument a été vu en cours au début du mois de décembre : c'est le critère de Dudley, donc on rappelle l'énoncé dit "des bonnes épiceries" ci dessous.

Dudley's criterion

In his landmark paper [56], Dudley has established the following metric entropy criterion for the sample continuity of some version of a Gaussian process.

Theorem 3.18 *Let $(X(t))_{t\in T}$ be some centered Gaussian process and d be the covariance pseudo-metric of $(X(t))_{t\in T}$. Assume that (T, d) is totally bounded and denote by $H(\delta, T)$ the δ-entropy number of (T, d), for all positive δ. If $\sqrt{H(., T)}$ is integrable at 0, then $(X(t))_{t\in T}$ admits a version which is almost surely uniformly continuous on (T, d). Moreover, if $(X(t))_{t\in T}$ is almost surely continuous on (T, d), then*

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \leq 12\int_0^\sigma \sqrt{H(x, T)}dx,$$

where $\sigma = \left(\sup_{t\in T}\mathbb{E}\left[X^2(t)\right]\right)^{1/2}$.

En particulier, ce critère peut s'appliquer pour les proceesus isonormaux. Le premier point consiste à **déterminer la distance d associée au processus gaussien** $X(t) := W(t)$. Fort heureusement, on à $d(s,t) = ||t-s||$, donc la d-entropie métrique utilisée dans la théorème est bien celle qu'on a utilisée pour définir la fonction Φ.

De plus, en considérant, pour tout u l'ensemble $T_\sigma(u) := \{t, ||t-u|| \leqslant \sigma\}$, et

$$V(t) := (X(t) - X(u))_{||t-u|| \leqslant \sigma}$$

est un processus gaussien. On déduit que :

$$\forall u, \sigma, \quad \sup_{t \in T_\sigma} \mathbb{E}\left[V(t)\right] \leqslant \sup_{t \in T_\sigma} ||t-u|| \leqslant \sigma \quad !$$

Et par le lemme de Dudley :

$$\forall u, \sigma, \quad \mathbb{E}\left[\sup_{||t-u|| \leqslant \sigma, t \in \mathcal{C}_\alpha} (W(t) - W(u))\right] \leqslant 12\Phi_\alpha(\sigma)$$

C'est la première partie du résultat. Pour obtenir la majoration voulue, on utilise le **Lemme de Pealing et sa conséquence**, énoncés ci après :

Lemma 4.23 *(Pealing lemma) Let S be some countable set, $u \in S$ and $a : S \to \mathbb{R}_+$ such that $a(u) = \inf_{t \in S} a(t)$. Let Z be some process indexed by S and assume that the nonnegative random variable $\sup_{t \in \mathcal{B}(\sigma)} [Z(t) - Z(u)]$ has finite expectation for any positive number σ, where*

$$\mathcal{B}(\sigma) = \{t \in S,\, a(t) \leq \sigma\}.$$

Then, for any function ψ on $\mathbb{R}_+$ such that $\psi(x)/x$ is nonincreasing on $\mathbb{R}_+$ and satisfies to

$$\mathbb{E}\left[\sup_{t \in \mathcal{B}(\sigma)} Z(t) - Z(u)\right] \leq \psi(\sigma), \quad \text{for any } \sigma \geq \sigma_* \geq 0$$

one has for any positive number $x \geq \sigma_*$

$$\mathbb{E}\left[\sup_{t \in S}\left[\frac{Z(t) - Z(u)}{a^2(t) + x^2}\right]\right] \leq 4x^{-2}\psi(x).$$

This Lemma warrants that in order to check assumption (4.71) it is enough to consider some nondecreasing function ϕ_m such that $\phi_m(x)/x$ is nonincreasing and satisfies

$$\mathbb{E}\left[\sup_{t \in S_m, \|u-t\| \leq \sigma} [W(t) - W(u)]\right] \leq \phi_m(\sigma)/8 \qquad (4.101)$$

for every $u \in S_m$ and any positive σ.

Ce qui tout donne donc exactement la condition (4.71) pour notre fonction Φ_α, à condition que $\kappa \geq 8 \times 12 = 96$! On explique la constante du texte.

En résumé, pour prouver ce premier théorème, on a utilisé :

- des conséquences du caractère isonormal de W.
- Le critère de Dudley.
- Le lemme de Pealing.

2.2.2 2nd théorème

Le second théorème à principalement pour but d'exprimer les résultats du premier dans le cadre de notre modèle. En particulier, pour Φ_α donnée, qui est d_α ? On commence par une proposition qui nous permet de controler l'entropie métrique d'un modèle $\mathcal{C}_\alpha$

For a k-simplex s in $\mathbb{R}^D$, let Δ_s be the diameter of the smallest including ball of s for the normalized norm (1) in $\mathbb{R}^D$. Then, for a k-homogeneous simplicial complex $\mathcal{C}$ in $\mathbb{R}^D$, let $|\mathcal{C}|_k := (\sum_{s\in\mathcal{C}^+} \Delta_s^k)^{1/k}$ and $\delta_\mathcal{C} := \inf_{s\in\mathcal{C}^+} \Delta_s$ where $\mathcal{C}^+$ is the subset of simplices of $\mathcal{C}$ of maximal dimension k. We start with the following entropic result on simplicial complexes.

Proposition 1. *For all k-homogeneous simplicial complex $\mathcal{C}$ of $\mathbb{R}^D$ and all $r \le \delta_\mathcal{C}$*

$$N(\mathcal{C}^n, \|\cdot\|, r) \le \left(\frac{4|\mathcal{C}|_k}{r}\right)^{nk}.$$

On prouve cette proposition. De cette majoration va découler assez naturellement le théorème suivant : dans l'impossibilité de déterminer précisémment une solution à l'équation (4.72), on ma majorer Φ_α par une fonction ϕ_α trouver une solution de (4.72) pour ϕ_α.

On arrive ainsi au théorème suivant :

Theorem 2. *Under the previous hypotheses, also suppose that for all $\alpha \in \mathcal{A}$,*

$$\sigma \le \delta_{\mathcal{C}_\alpha} \sqrt{\frac{D}{k}} \left[4\kappa \left(\sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi}\right)\right]^{-1}. \tag{11}$$

There exists some absolute constants c_1 and c_2 such that for all $\eta > 1$, if

$$\mathrm{pen}(\alpha) \ge \eta \frac{\sigma^2}{Q} \left(c_1 nk \left[\ln \frac{|\mathcal{C}_\alpha|_k \sqrt{D}}{\sigma\sqrt{k}} + c_2\right] + 4w_\alpha\right), \tag{12}$$

then, almost surely, there exists a minimizer $\hat{\alpha}$ of the penalized criterion

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|^2 + \mathrm{pen}(\alpha)$$

and the penalized estimator $\hat{\mathbf{x}}_{\hat{\alpha}}$ satisfies the following risk bound

$$\mathbb{E}_{\bar{\mathbf{x}}} \|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \le c_\eta \left[\inf_{\alpha\in\mathcal{A}} \left\{d(\bar{\mathbf{x}}, \mathcal{C}_\alpha^n)^2 + \mathrm{pen}(\alpha)\right\} + \frac{\sigma^2}{Q}(\Sigma + 1)\right]. \tag{13}$$

Remarques :

- La preuve donne les valeurs des constantes, qui sont des "vrais nombres".
- La condition de majoration de la variance vise à ne pas avoir de simplexe qui soit petit face à la variance, cas dans lequel les approximations n'auraient pas de sens.
- De même les marqueurs qui servent à définir les modèles doivent être suffisament éloignés les uns des autres.
- **L'intérêt majeur de ce théorème, outre la garantie théorique de performance est qu'il fournit la forme de la pénalité. Il suffit d'une petite manipulation inspirée du cas linéaire pour voir qu'elle est proportionnelle à** $\ln |\mathcal{C}_\alpha|_k$.

On passe à présent à une brêve analyse des résultats du papier et à un travail complémantaire sur l'heuristique de pente.

3 Interprétation des résultats et heuristique de pente.

3.1 Interprétation des résultats

La partie précédente de l'article ne nous donne pas des résultats directemnt utilisables. En effet la pénalité n'est connue qu'à une constante inconnue près.

On va supposer les complexes k-homogènes, et n, D fixés. Les complexes déterminés à partir d'un ensemble de points "landmarks" du paramètre de taille par lesquels on les indexe.

La méthode de l'heuristique de pente se résume ainsi :

1. **Pour chaque complexe (de taille α), on regarde le rique empirique** $Re(\alpha) = ||\hat{x}_\alpha - X||^2$.
2. **On plot** $Re(\alpha)$ **en fonction de** $\ln |\mathcal{C}_\alpha|_k$ **et on espère observer une tendance linéaire.**
3. **On infère sa pente** β**.**
4. **On utilise le critère pénalisé suivant :** $crit(\alpha) = ||\hat{x}_\alpha - X||^2 - 2\beta \ln |\mathcal{C}_\alpha|_k$**.**

Cet algorithme est testé sur un cas particulier, et les résultats sont présentés dans l'article. On peut faire un certain nombre de remarques :

- Les résulats semblent *a priori* très satisafisants.
- La rendance linéaire apparait sans difficulté, sur différents exemples, ainsi que sur des données réelles.
- L'expérience fonctionne mieux dans le cas idéal où les "landmarks" sont choisis sur l'objet géométrique, ce qui n'est pas réalisable en pratique. Ce point justifie de travailler plus avant sur les méthodes de choix de ces "landmarks", ce qui est fait dans la dernière partie de l'article.

3.2 Conclusion sur l'article

La reconstruction d'un objet géométrique sur lequel on mesure des points bruités peut se faire à partir d'un point de vue statistique. Le problème est alors vu comme un problème de sélection d'un modèle, les modèles étant choisis parmis des complexes simpliciaux qui s'avèrent être de bons outils pour approcher un objet géométrique à partir de points. La théorie générale peut s'appliquer à cet exemple, et nous fournit un théorème qui donne un résultat théorique, et la forme de la pénalité à utiliser. Cependant l'algorithme proposé n'est pas programmable directement, car certaines constantes sont inconnues, mais l'utilisation d'une heuristique de pente nous donne un algorithme simple, qui semble preformant sur les exemples testés.

3.3 L'heuristique de pente dans un cadre totalement différent.

L'heuristique de pente est elle un phénomène excepionnel.. ? je suis sûr d'avoir déjà entendu ça quelque part. Y aurait-il moyen d'illustrer cette technique relativement simplement ? Quelques recherches effectuées, le second cours de Sylvain Arlot au cours Peccot travaille sur ce point... Voyons ce qu'il y a moyen de faire...