Méthodes à noyaux en apprentissage statistique : déterminer des paramètres qui controlent la complexité d'un problème et construire des algorithmes rapides associés

Ulysse Marteau-Ferey, sous la direction de Francis Bach et Alessandro Rudi

le 5 octobre 2018

Table des matières

1 Introduction

L'apprentissage supervisé est un domaine dont l'objectif est de pouvoir prédire une sortie (par exemple la race d'un chien) à partir d'une entrée (par exemple une photo de chien). Pour ce faire, un algorithme apprend une fonction de prédiction à partir d'un nombre fini d'exemples de couples entrée-sortie (c'est pour cela que l'on parle d'apprentissage *supervisé*).

Le cadre théorique de l'apprentissage supervisé a été développé à partir des années 90 dans le cadre statistique habituel d'un dilemme biais-variance : il s'agit en effet d'apprendre avec un modèle suffisamment complexe pour que le biais ne soit pas trop grand, i.e. que le modèle ne soit pas trop simple, et en même temps que le modèle ne soit pas trop gros par rapport au nombre de données, sans quoi les observations ne seraient plus significatives.

Cependant, à ce cadre statistique vient s'ajouter un problème d'optimisation : en effet, il s'agit de pouvoir calculer de manière effective une fonction de prédiction ; le temps et la complexité en mémoire de ce calcul est donc également à prendre en compte.

De manière générique, il s'agit donc de traiter simultanément trois problèmes.

1. Tout d'abord un problème de modèle : quel est l'erreur liée au modèle ? À la paramétrisation ? Quelle est la complexité du problème d'apprentissage et comment la caractériser ?
2. Ensuite, un probleme statistique lié à la variance au sein de ce modèle. Ce problème, tout comme le précédent, est abordé par des moyens de statistiques classiques, inegalités de concentration, analyse non-asymptotique...

3. Enfin, un problème d'optimisation : comment trouver un bon prédicteur de manière computationellement efficiente ?

Cette introduction au domaine de recherche vise a introduire la cadre théorique de l'apprentissage supervisé et à essayer, dans le cadre des méthodes à noyaux, d'identifier des quantités théoriques qui caractérisent la difficulté du problème. Il s'agira ensuite de montrer comment ces quantités impactent la difficulté statistique et algorithmique du problème.

Cette introduction au domaine de recherche sera organisée en trois parties. Dans la partie, nous présenterons une partie de la théorie de l'apprentissage supervisé ainsi que les grands domaines qui y sont liés. Dans la deuxième partie, nous verrons le cadre des méthodes à noyau, ainsi que les objects qui permettent de caractériser la difficulté d'un problème dans un cas simple. Dans la troisième partie, nous situerons notre travail de recherche et en décrirons, dans les grandes lignes, les objectifs.

2 Théorie de l'apprentissage

Dans cette partie, nous parcourons de très loin une partie la théorie de l'apprentissage ainsi que ses principaux enjeux : proposer une cadre à la rencontre des statistiques, de l'analyse fonctionnelle et de l'optimisation pour modéliser et contrôler la complexité à la fois statistique et computationnelle d'un algorithme d'apprentissage.

2.1 Apprentissage statistique

L'apprentissage supervisé a été théorisé à partir des années 1990 (voir les travaux de Vapnik [18] ou encore le livre de référence par Hastie et al. [9]) ; il s'agit de prédire une sortie $Y \in \mathcal{Y}$ à partir de données $X \in \mathcal{X}$ que l'on suppose liées. On différencie habituellement :

- les problèmes de *classification*, où $\mathcal{Y}$ est un espace discret et où l'objectif est d'associer à chaque donnée $x \in \mathcal{X}$ une catégorie $y \in \mathcal{Y}$ (on se place très souvent dans le cadre de la classification binaire où $\mathcal{Y} = \{-1, 1\}$) ;
- les problèmes de *régression*, où $\mathcal{Y} \subset \mathbb{R}$.

Nous nous plaçons dans une cadre probabiliste où nous supposons que $Z = (X, Y)$ est une variable aléatoire de loi ρ. Il s'agit ici de trouver une bonne *fonction de prédiction* $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ où $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ désigne l'ensemble des applications mesurables de $\mathcal{X}$ dans $\mathcal{Y}$. Pour mesurer la qualité de cette fonction de prédiction, on introduit le plus souvent une *fonction de perte*

$$\ell : \mathcal{Z} \times \mathcal{M}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_+$$

où $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\ell(z, f) = \ell_z(f) = \ell_{(x,y)}(f)$ est une mesure de l'erreur entre $f(x)$ et y. Le *risque* ou encore *risque de généralisation* ou *risque en espérance* est l'espérance de la perte lorsque (x, y) suit la loi de (X, Y) que l'on note $\mathcal{R}(f) := \mathbb{E}_{z\sim Z}[\ell_z(f)]$. L'objectif général de l'apprentissage supervisé est de trouver une fonction de prédiction f de risque petit, c'est à dire tel que

$$\mathcal{R}(f) \approx \inf_{f \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(f)$$

Si l'infimum est atteint en une fonction f_ρ, l'objectif est donc de calculer f_ρ. Les fonctions de pertes les plus utilisées sont le plus souvent de la forme $\ell_z(f) = l(y, f(x))$ où $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, et l'on a les exemples suivants :

- la perte quadratique $\ell((x, y), f) = \frac{1}{2}|y - f(x)|^2$, qui sert surtout en régression ;
- la perte binaire $\ell((x, y), f) = \mathbf{1}_{f(x)\neq y}$ en classification. Cette perte est souvent relaxée dans le cadre de la classification binaire et on préfère des pertes plus lisses comme la perte logistique $\ell((x, y), f) := \log(1 + \exp(-yf(x)))$ ou plus généralement des pertes de la forme $\varphi(yf(x))$ où φ est convexe décroissante.

En apprentissage comme en statistiques, nous avons accès à la loi du couple (X, Y) uniquement à travers un nombre fini $n \in \mathbb{N}$ d'observations $(x_k, y_k)_{1\leq k\leq n} \in (\mathcal{X} \times \mathcal{Y})^n$. On fait l'hypothèse que ces observations sont indépendantes et identiquement distribuées selon la loi du couple (X, Y). Formellement, un algorithme d'apprentissage est une statistique sur $\mathcal{X} \times \mathcal{Y}$, soit la donnée, pour tout $n \in \mathbb{N}$, d'une application mesurable

$$\hat{f}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}(\mathcal{X}, \mathcal{Y}) \quad \text{(ou la classe des espaces probabilisés sur } \mathcal{M}(\mathcal{X}, \mathcal{Y}))$$

qui à tout échantillon de taille n $(x_k, y_k)_{1\leq k\leq n}$ associe une fonction (X, Y) mesurable $\hat{f}_n((x_k, y_k)_{1\leq k\leq n}) \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, (potentiellement elle-même stochastique ; formellement, il faut associer un espace probabilisé et une variable aléatoire sur cet espace à chaque échantillon). Ainsi, l'algorithme peut être vu comme une suite de variables alétoires $\hat{f}_n$ sur $((\mathcal{X} \times \mathcal{Y})^n, \rho^{\otimes n})$ à valeurs dans $\mathcal{M}(\mathcal{X}, \mathcal{Y})$.

2.2 Minimisation du risque empirique et dilemme biais-variance

Étant donné n échantillons $(x_k, y_k)_{1\leq k\leq n}$, le premier problème pour définir une bonne statistique $\hat{f}_n$ est que nous n'avons pas accès à la fonction de risque $\mathcal{R}$ à minimiser, mais uniquement à la fonction de perte ℓ. On minimise donc le plus souvent le *risque empirique* comme substitut du vrai risque :

$$\hat{f}_n \in \arg\min_{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})} \hat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell((x_i, y_i), f)$$

Pour la perte quadratique, c'est la régression des moindres carrés, introduite formellement par Legendre en 1805 [11] et Gauss en 1809 [8] dans leur travaux sur les corps célestes.

Minimiser sur l'espace $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ est beaucoup trop fort ; toute fonction telle que $f(x_i) = y_i$ est un minimiseur du risque empirique, et peut avoir un très mauvais comportement en risque en espérance. Il s'agit donc de régulariser le problème. Cette régularisation peut prendre les deux formes suivantes (qui sont en fait équivalentes à la deuxième).

- Elle peut passer par la restriction de la minimisation du risque empirique à un sous-ensemble $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ de fonctions de $\mathcal{X}$ dans $\mathcal{Y}$:

  $$\hat{f}_n \in \arg\min_{f\in\mathcal{F}} \hat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell((x_i, y_i), f)$$

  Ce sous-ensemble de fonctions $\mathcal{F}$ est souvent défini comme un modèle linéaire. Pour cela, on se donne une *représentation* ou *changement de représentation* $\phi : \mathcal{X} \to \mathcal{H}$ qui est une application qui envoie les données $\mathcal{X}$ dans un espace de Hilbert $\mathcal{H}$ dans lequel, la géométrie euclidienne de $\mathcal{H}$ restreinte à $\phi(\mathcal{X})$ est discriminante pour l'estimation de la sortie Y. On identifie ensuite l'espace $\mathcal{F}$ à l'espace de Hilbert $\mathcal{H}$ en posant $f(x) = \langle\theta, \phi(\cdot)\rangle_{\mathcal{H}}$ où $\theta \in \mathcal{H}$. Ainsi, f est une projection de la représentation des données dans une direction particulière θ. En minimisant le risque, on cherche alors la direction la plus discriminante pour estimer la sortie Y. Lorsque $\mathcal{H} = \mathbb{R}^d$, et que la représentation ϕ est donnée, on parle de modèle paramétrique car on représente alors les données $\mathcal{X}$ par un nombre fini d de paramètres. Comme nous le verrons plus tard, il est également possible de définir implicitement la représentation ϕ ; on est alors dans un cadre non-paramétrique.

  Ce qui peut paraître être un simple modèle linéaire pour les fonctions f peut être très complexe ; cette complexité dépend de celle de la représentation ϕ des données. Depuis les années 2010, c'est l'apprentissage d'une bonne représentation qui est au coeur de la partie expérimentale de l'apprentissage par réseaux de neurones profond, où l'objectif est d'apprendre une représentation ϕ parmi une classe paramétrée de représentations (ϕ_α) ; l'algorithme minimise alors le risque empirique $\frac{1}{n}\sum_{i=1}^{n} l((x_i, y_i), \langle\theta, \phi_\alpha(\cdot)\rangle_{\mathcal{H}})$ non seulement en fonction de θ mais aussi en fonction de α, afin de trouver à la fois la meilleure représentation ϕ_α et la direction θ la plus discriminante.

- La régularisation peut également passer par la *pénalisation* des fonctions au mauvais comportement avec un régulariseur $\Omega : \mathcal{F} \to \mathbb{R}_+$

  $$\hat{f}_n^\lambda \in \arg\min_{f\in\mathcal{F}} \hat{R}_n^\lambda(f) := \hat{R}_n(f) + \lambda\Omega(f)$$

  Ω peut être une norme ou une norme au carré (norme L^1 si l'on souhaite des fonctions parcimonieuses, norme L^2 si on souhaite des fonctions de faible variance...). Ce régulariseur ajoute un biais à notre estimateur qu'il faut controler grâce au paramètre λ.

Dans tous les cas, il s'agit de se restreindre à des fonctions dotées d'une certaine régularité, et donc en un sens à des espaces de fonctions plus petit. Dans ce qui suit, nous allons motiver un peu plus le besoin de définir des notions de régularité et de taille des ensembles de fonctions sur lesquels nous voulons travailler.

Dilemme biais-variance

Supposons que nous ayions calculé notre estimateur $\hat{f}_n^\lambda$ comme minimiseur du risque empirique (régularisé) sur une classe de fonctions $\mathcal{F}$:

$$\hat{f}_n^\lambda \in \arg\min_{f\in\mathcal{F}} \hat{R}_n^\lambda(f) := \frac{1}{n}\sum_{i=1}^{n} \ell((x_i, y_i), f) + \lambda\Omega(f)$$

et que nous voulons maintenant borner son risque. Une conséquence des définitions est l'inégalité suivante :

Proposition 1 (décomposition de l'erreur)**.**

$$\mathcal{R}(\hat{f}_n^\lambda) - \inf_{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})} \mathcal{R}(f) \leq \underbrace{\inf_{f\in\mathcal{F}} \mathcal{R}^\lambda(f) - \inf_{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})} \mathcal{R}(f)}_{\textit{erreur d'approximation}} + \underbrace{\mathcal{R}^\lambda(\hat{f}_n^\lambda) - \inf_{f\in\mathcal{F}} \mathcal{R}^\lambda(f)}_{\textit{erreur d'estimation}}$$

où $\mathcal{R}^\lambda = \mathcal{R} + \lambda\Omega$ *et on a également la borne suivante :*

$$\underbrace{\mathcal{R}^\lambda(\hat{f}_n^\lambda) - \inf_{f\in\mathcal{F}} \mathcal{R}^\lambda(f)}_{\textit{erreur d'estimation}} \leq \underbrace{2\sup_{h\in\mathcal{F}} \left|\mathcal{R}(h) - \hat{R}_n(h)\right|}_{\textit{erreur de fluctuation}}$$

- l'erreur d'approximation est l'erreur que l'on commet en se restreignant à $\mathcal{F}$ et en régularisant par λ ; elle décroit au fur et à mesure que la "taille" de $\mathcal{F}$ augmente et que λ diminue ;
- l'erreur d'estimation est l'erreur que l'on commet en minimisant le risque emprique (régularisé) et non le vrai risque. Cette erreur a tendance à croitre au fur et à mesure que la taille de $\mathcal{F}$ augmente, comme le suggère la borne ;
- l'erreur de fluctuation est une borne uniforme sur la différence entre le vrai risque et le risque empirique $\left|\mathcal{R}^\lambda - \hat{R}_n^\lambda\right|$; il croit donc au fur et à mesure que la taille de $\mathcal{F}$ augmente.

Nous avons donc ici un dilemme biais-variance ou biais-fluctuation ; il faut faire un arbitrage entre les deux pour que ces deux termes soient idéalement du même ordre. Lorsque l'erreur de fluctuation est trop importante, on parle de *sur-apprentissage* ; le minimiseur "colle" trop aux données sans capturer la régularité du phénomène sous-jacent qui va permettre de bien généraliser. Lorsdque l'erreur d'approximation ou erreur de modèle est trop importante, c'est que le modèle est trop simple.

2.3 Bornes sur l'erreur de fluctuation

De manière générale, l'étude de l'erreur de fluctuation comprend deux éléments :

- une notion de "taille"/"dimension"/"complexité" du modèle, liée à la classe de fonction $\mathcal{F}$ et au régulariseur $\lambda\Omega$;
- au nombre d'échantillons n.

Le premier exemple très simple est de considérer que l'espace $\mathcal{H}$ est fini et que la perte l est à valeurs dans $[0,1]$. Une application directe du lemme de Hoeffding (voir par exemple [5]) montre que pour tout $\delta > 0$

$$\mathbb{P}\left(\max_{h\in\mathcal{F}} \left|\mathcal{R}(h) - \hat{R}_n(h)\right| > \epsilon\right) \leq 1-\delta, \quad \text{avec } \epsilon^2 = \frac{\log(|\mathcal{F}|) + \log\frac{2}{\delta}}{2n}$$

Cette borne très naïve contient pourtant deux éléments essentiels : en effet, l'erreur de fluctuation, et donc a fortiori l'erreur d'estimation, serait de l'ordre $\sqrt{\frac{\log(|\mathcal{F}|)}{n}}$ avec forte probabilité. De manière, générale, la majoration de l'erreur d'estimation par une borne uniforme mène à ce que l'on appelle des *taux lents*, typiquement de l'ordre de $\sqrt{1/n}$.

De façon plus générale, pour obtenir des taux lents, on a le résultat fondamental suivant (voir par exemple [4])

Théorème 1. *Supposons ici que l'on se place dans le cadre de prédicteurs linéaires* $f(x) = f\cdot\phi(x)$ *où* $\phi : \mathcal{X} \to \mathcal{H}$ *avec* $\mathcal{H}$ *un Hilbert, que la fonction de perte* l *est de la forme* $l(y, f(x))$ *et est* $G-$*lipschitz en la deuxième variable, que les données sont bornées* ($||\phi(X)|| \leq R$) *et que l'on minimize le risque empirique en ajoutant la contrainte* $||f||_{\mathcal{H}} \leq D$*, c'est à dire sur l'espace des fonctions* $\mathcal{F} := \{f \mid ||f||_{\mathcal{H}} \leq D\}$ *, alors avec forte probabilité, on a*

$$\sup_{f\in\mathcal{F}} \left|\mathcal{R}(f) - \hat{R}_n(f)\right| \leq \frac{GRD}{\sqrt{n}}$$

Démonstration. Le premier element essentiel de cette preuve est d'introduire la *complexité de Rademacher*, outil essentiel pour mesurer la taille de l'espace de fonctions et pour borner l'erreur de fluctuation. On définit cette complexité de la façon suivante

$$\mathcal{R}_n := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \epsilon_k \ell_{Z_k}(f)\right)\right] \text{ où les } \epsilon_k \sim 2\mathcal{B}(1/2) - 1 \text{ sont i.i.d. et indépendantes des } Z_k := (X_k, Y_k)$$

Pour controler l'espérance du terme de fluctuation (voir par exemple [4]), on utilise une méthode de symétrisation pour obtenir

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathcal{R}(f)-\hat{R}_n(f)\right|\right]\leq 2\mathcal{R}_n$$

Ensuite, on utilise le lemme de Ledoux-Talagrand ([10]) pour borner la complexité de Rademacher $\mathcal{R}_n \leq \frac{RGD}{\sqrt{n}}$ en sortant la constante de lipschitz de l'espérance.

Enfin, on contrôle l'écart de $\sup_{f\in\mathcal{F}}\left|\mathcal{R}(f)-\hat{R}_n(f)\right|$ à sa moyenne en utilisant une inégalité de concentration du type MacDiarmid (voir [5]). □

Ici, la complexité de Rademacher remplace $\sqrt{\log|\mathcal{F}|/n}$ pour définir la taille de la classe de fonctions $\mathcal{F}$ par rapport aux données ; il existe d'autres façcons de mesurer cette taille comme l'entropie de Vapnik (voir [18]), ainsi que des façons d'incorporer la régularization par pénalité comme nous le verrons plus loin.

Les taux de décroissance de l'erreur de fluctuation l'ordre de $\frac{1}{\sqrt{n}}$ sont qualifiés de "lents" ; en effet, ils peuvent être obtenus sans hypothèse particulière sur les fonctions de perte $\mathcal{R}$. En particulier, dès lors que l'on suppose que la fonction de risque est μ-fortement convexe (on dit que F est μ-fortement convexe si $F-\frac{\mu}{2}||\cdot||^2$ est convexe), alors on obtient des taux rapides de l'ordre de $O(\frac{1}{n\mu})$ (voir par exemple [6]). Ainsi de façon générale, l'objectif général est de trouver des conditions de régularité sur la classe de fonctions $\mathcal{F}$ et sur la fonction de perte l telle ques que l'on obtienne de meilleurs taux de décroissance.

2.4 Optimisation

En apprentissage statistique, la dimension de l'espace de fonction $\mathcal{F}$ peut être très grande, du même ordre que le nombre d'échantillons, voire infinie dans le cas d'un espace de Hilbert, et le nombre d'échantillons n peut être lui aussi très grand, de l'ordre de 10^6 10^8. Cela rend le calcul exact du minimiseur du risque empirique trop couteux comme le montre l'exemple suivant.

Plaçons-nous dans le cas simple d'un modèle linéaire en dimension d ; on se donne $\phi : \mathcal{X} \to \mathbb{R}^d$ une représentation des données. Supposons que l'on se place dans le cadre de la perte quadratique. Si $\Phi \in \mathbb{R}^{n\times d}$ est la matrice dont la i-ème ligne est la représentation $\phi(x_i)^T$ du vecteur x_i, le risque empirique est de la forme $\hat{R}_n(\theta)=\frac{1}{n}\sum_{i=1}^n|[y-\Phi\theta]_i|^2$ et donc son minimiseur est $\hat{\theta}_n=\left(\Phi^T\Phi\right)^{-1}\Phi^T\hat{y}$.

Nous voyons bien que si d est grand, l'inversion de matrice a une complexité de $O(d^3)$ ce qui est beaucoup trop dès lors que d est de l'ordre de n. Ainsi, il faut trouver une alternative à la simple minimisation du risque empirique et trouver des estimateurs aux bonnes propriétés statistiques qui ont également de bonnes propriétés computationnelles. Dans [**?**], Bottou et Bousquets rappellent deux grands principe : que l'objectif n'est pas de minimiser le risque empirique mais bien l'erreur de généralisation et qu'il est inutile d'optimiser au delà du taux statistique.

Dès lors que la fonction de perte est convexe en θ, les problèmes de minimisation du risque et du risque empirique sont des problèmes convexe : le problème de minimisation du risque empirique est donc un problème d'optimisation convexe.

Les problèmes d'optimisation venant de l'apprentissage supervisé ont deux caractéristiques particulières :

1. Ils sont en très grande dimension.
2. Il ne faut les résoudre que jusqu'au moment où l'on atteint l'optimum statistique.

Ces deux propriétés ont induit un changement de paradigme majeur : tandis qu'avant, les algorithmes étaient des méthodes du second ordre faisant intervenir la Hessienne, les algorithmes pour l'apprentissage ne peuvent pas se permettre de faire rentrer la Hessienne en mémoire et donc sont majoritairement des algorithmes du premier ordre. Nous ne nous étendrons pas sur toutes ces méthodes mais mentionneront simplement deux approches majeures.

1. La première est apparu dans les années 1980 dans l'école russe d'optimisation emmenée par Nesterov et Nemirovski. Elle a visé a donner un cadre théorique à la convergence des méthodes d'optimisation convexe, comme les méthodes de Newton ou les méthodes de gradients. Par exemple, un résultat typique est le suivant

Proposition 2 (Voir [14]). *Soit $f \in \mathcal{F}_L^1(\mathbb{R}^d)$ l'ensemble des fonctions convexe L-lisse, c'est à dire deux-fois dérivables et dont les Hessienne sont majorées par LI_d. Alors si étant donné θ_0, on définit*

$$\theta_n = \theta_{n-1} - \frac{1}{L} f'(\theta_{n-1})$$

Alors cette méthode de gradient satisfait

$$f(\theta_n) - f(\theta_*) \leq \frac{2L||\theta_0 - \theta_*||}{n+4}$$

De plus, si on rajoute de la stricte convexité de paramètre μ, alors les mêmes itérées convergent linéairement vers l'optimum :

$$f(\theta_n) - f(\theta_*) \leq \left(1 - \frac{\mu}{L}\right)^n (f(\theta_0) - f(\theta_*))$$

Le but est donc de trouver des critères de régularité des fonctions convexes pour pouvoir contrôler leur convergence. Ces méthodes peuvent être accélérées (voire également [14]).

2. La deuxième approche majeure est l'approximation stochastique. Elle a d'abord été introduite par Robbins et Monroe ([**?**]) et consiste à construire la suite aléatoire suivante :

$$\theta_n \leftarrow \theta_{n-1} - \gamma_n \left(h(\theta_{n-1}) + \epsilon_n\right)$$

où l'on a une filtration $\mathcal{F}_n$ associée à θ_n ; on suppose que $\mathbb{E}[\epsilon_n | \mathcal{F}_{n-1}] = 0$. Dans el cas de la minimisation d'une fonction convexe, on prendra typiquement $h = f'$ même si le cadre de ces méthodes ne se limite pas du tout aux fonctions convexe. Dans cas, on parle de descente de gradient stochastique (SGD).

Ces deux approches sont évidemment reliées et ont formé un duo très puissant pour s'attaquer à la minimisation de fonctions de perte de la forme $f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$, omni-présentes en apprentissage (comme on le voit, le risque empirique prend naturellement cette forme). Cela a mené à de nombreux algoritmes de descente de gradient stochastique comme SGD, SAG, SAGA (voir par exemple [**?**]).

3 Méthodes à noyaux

Dans cette partie, nous présentons les méthodes à noyaux qui forment une classe importante de méthodes d'apprentissage non-paramétriques. Nous commencerons par définir ce qu'est un espace à noyau, puis nous regarderons comment le problème de minimisation du risque empirique se traduit dans ce cadre. Nous verrons ensuite, dans le cadre spécifique de la régression linéaire, une manière de mesurer la taille de ces espaces ainsi qu'une façon de définir une notion riche de régularité des fonctions dans ces espace. Enfin, nous mentionnerons et fairont l'ébauche des bornes que l'on peut obtenir sur l'erreur d'apprentissage en fonction des différents paramètres de régularité que nous aurons introduit avant.

3.1 Espaces à noyau reproduisant (E.N.R.)

Afin d'avoir une erreur d'approximation faible, il est important de pouvoir minimiser le risque empirique sur une classe de fonction suffisamment grande mais qui possède une certaine régularité. Les modèles paramétriques, où la classe de fonction est de la forme $(f_\theta)_{\theta\in\Theta}$ où $\Theta \subset \mathbb{R}^d$ sont parfois trop restrictifs ; en un sens, la régularité est trop forte ou du moins trop dépendante de cette paramétrisation. Une autre approche pour avoir des espaces de fonctions $\mathcal{H}$ réguliers est de considérer des espaces de Hilbert de fonctions : ils sont plus riche et la régularité n'est pas conditionnée par la paramétrisation. Plus specifiquement, nous allons ici considérer des espaces de Hilbert particuliers, les espaces à noyau reproduisant (ENR).

Ces espaces et leur définition peut être abordée de plusieurs manières. Le premier point de vue est le suivant :

Définition 1 (ENR par la continuité de l 'évaluation). *Soit $\mathcal{X}$ un espace, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ un espace de Hilbert de fonctions de $\mathcal{X}$ à valeurs dans $\mathbb{R}$. On dit de $\mathcal{H}$ qu'il est à noyau reproduisant si*

$$\forall x \in \mathcal{X},\ L_x : f \in \mathcal{H} \mapsto f(x) \text{ est continue}$$

De manière équivalente,

$$\exists K_x \in \mathcal{H},\ \forall f \in \mathcal{H},\ \langle f, K_x \rangle = f(x)$$

et l'on a $||L_x||_{\mathcal{H}^*} = ||K_x||_{\mathcal{H}}$.

- *La fonction* $K : (x, y) \in \mathcal{X} \times \mathcal{X} \longmapsto K_x(y) \in \mathbb{R}$ *est appelé noyau ou noyau reproduisant associé à l'espace* $\mathcal{H}$ *et est symétrique définie positive, c'est à dire que pour tout* $(x, x') \in \mathcal{X}^2$, $K(x, x') = K(x', x)$ *et pour tout* $(x_1, ..., x_n) \in \mathcal{X}^n$ *et* $(\alpha_1, ..., \alpha_n) \in \mathbb{R}^n$, $\sum_{i,j=1}^n \alpha_i K(x_i, x_j)\alpha_j \geq 0$.
- *La fonction* $\phi : x \in \mathcal{X} \longmapsto K_x \in \mathcal{H}$ *est appelée représentation de* $\mathcal{X}$ *dans* $\mathcal{H}$.

Par exemple, les espaces de Sobolev $H^s(\mathbb{R}^d)$ pour $s > \frac{d}{2}$ sont des espaces à noyau reproduisant. Ce point de vue est centré sur les espaces de Hilbert de fonctions eux-même, et se sert de ces propriétés d'espaces à noyau comme outil pour étudier ces espaces de fonctions. On les voit par exemple apparaitre dans les travaux de Bergmann ([3]).

L'autre point de vue est de partir d'un noyau $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, c'est à dire une fonction à deux variables symétrique semi-définie positive comme dans la définition précédente.

Définition 2 (ENR par le noyau semi-défini positif)**.** *Supposons que l'on ait un noyau semi-défini positif* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *et définissons*

$$\forall x,\ K_x : y \in \mathcal{X} \mapsto K(x, y) \in \mathbb{R}$$

Soit V_K *l'espace vectoriel* $\{\sum_{i=1}^p \alpha_i K_{x_i} \mid p \in \mathbb{N},\ (\alpha_i) \in \mathbb{R}^p,\ (x_i) \in \mathcal{X}^p\}$ *le sous-espace vectoriel de* $\mathbb{R}^{\mathcal{X}}$ *par engendré par* $(K_x)_{x\in\mathcal{X}}$ *que l'on muni du produit scalaire suivant*

$$\sum_{i=1}^p \alpha_i K_{x_i} \cdot \sum_{j=1}^q \beta_j K_{y_j} = \sum_{i,j} \alpha_i K(x_i, y_j)\beta_j$$

Alors le théorème d'Aronszajn [1] motre qu'il existe un unique espace à noyau reproduisant de noyau K *(à isomorphisme près) et que c'est le complété de* V_K.

Ici, le point de vue est centré sur les noyaux eux-même : l'objectif est de construire, à partir de K que l'on peut voir comme une mesure de similarité sur $\mathcal{X}$, un espace de Hilbert qui reflète cette mesure de similarité sur les données. Géométriquement, cela signifie que l'on plonge $\mathcal{X}$ dans un espace où la géométrie Hilbertienne reflète cette mesure de similarité. Les noyaux semi-définis positifs ont d'abord été introduits et ťudiés par Mercer [12] et l'existence d'un unique Hilbert à tout noyau a été prouvée dans le cas général par Aronszajn en 1950 (voir [1]).

Les exemples les plus classiques de noyau sur $\mathbb{R}^d$ sont les suivants :

- les noyaux linéaires, $K(x, y) = (x \cdot y)^m, m \in \mathbb{N}$;
- les noyaux exponentiels $K(x, y) = \exp\left(-\frac{||x-y||}{\sigma}\right)$ avec $\sigma > 0$;
- les noyaux gaussiens $K(x, y) = \exp\left(-\frac{||x-y||^2}{\sigma^2}\right)$ avec $\sigma > 0$.

3.2 Minimization du risque empirique dans les espaces à noyau

Supposons maintenenant que la fonction ℓ de perte soit de la forme $\ell_z(f) := l(y, f(x))$ où $l : (y, y') \in \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_+$ est une application convexe en la deuxième variable et plaçons nous dans le cas où l'espace de restreint de fonctions $\mathcal{F}$ est un ENR $\mathcal{H}$ de noyau K. Le problème de minimisation du risque de généralisation s'écrit toujours

$$\inf_{f\in\mathcal{M}(\mathcal{X},\mathcal{Y})} \mathbb{E}\left[l(Y, f(X))\right]$$

La façon la plus classique de trouver des estimateurs est de consdiérer la minimisation du risque empirique régularisé (par régularisation de Tikhonov aussi appelée régression Ridge) :

$$\inf_{f\in\mathcal{H}} \hat{R}_n^\lambda(f) := \frac{1}{n}\sum_{i=1} l(y_i, \langle f, K_{x_i}\rangle) + \frac{\lambda}{2}||f||_{\mathcal{H}}^2$$

On a alors la propoissition suivante :

Proposition 3. *Il existe un unique minimiseur du risque régularisé* $\hat{f}_n^\lambda$. *De plus, il appartient à* $\mathcal{H}_n := span(K_{x_i})_{1\leq i\leq n}$.

Démonstration. L'existence d'un unique minimiseur vient du fait que le problème régularisé est fortement convexe. Le fait que la solution soit dans $\mathcal{H}_n$ vient du théorème de représentants, c'est à dire que si $P_{\mathcal{H}_n}$ désigne le projecteur orthogonal sur $\mathcal{H}_n$, pour tout $f \in \mathcal{H}$ et $1 \leq i \leq n$, $f(x_i) = P_{\mathcal{H}_n}(f)(x_i)$ et $P_{\mathcal{H}_n^\perp}(f)(x_i) = 0$, de qui montre que $\hat{R}_n(f) = \hat{R}_n(P_{\mathcal{H}_n}(f))$. □

Cette proposition montre que le problème de minimisation du risque empirique régularisé est en fait un problème de dimension finie : si on cherche la solution sous la forme $\hat{f}_n^\lambda = \sum_{i=1}^n \alpha_i K_{x_i}$, alors α est la solution au problème

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1} l(y_i, [K_{nn}\alpha]_i) + \frac{\lambda}{2} \alpha^T K_{nn} \alpha \tag{1}$$

où K_{nn} est la matrice $K_{nn} := (K(x_i, x_j))_{1 \leq i,j \leq n}$.

Perte quadratique

Dans le cas de la perte quadratique $l(y, y') = \frac{1}{2}|y - y'|^2$, on peut faire les remarques suivantes.

- Si Y admet un moment d'ordre deux, i.e. $\mathbb{E}[|Y|^2] < \infty$, alors le problème de généralisation

$$\inf_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathbb{E}\left[|Y - f(X)|^2\right]$$

  admet une solution $f_\rho = \mathbb{E}[Y|X]$ qui est l'espérance conditionnelle de Y sachant X. De plus, cette solution f_ρ est dans $L^2(\mathcal{X}, \rho_X)$ où ρ_X est la loi marginale de ρ sur X.
- Le minimiseur $\hat{\alpha}$ du risque empirique (solution de (1)) est de la forme $\hat{\alpha} = (K_{nn} + \lambda n I)^{-1} \hat{Y}$, où $\hat{Y} \in \mathbb{R}^n$ est le vecteur des y_i.

Deux questions se posent alors naturellement dans ce cas simple :

- Quelle sont les erreurs d'approximation et d'estimation des estimateurs $\hat{f}_n^\lambda$? Plus précisément, comment choisir le bon λ ?
- Est-il possible de faire un algorithme efficace pour calculer une approximation de $\hat{f}_n^\lambda$ aux bonnes propriétés statistiques ? Plus spécifiquement, peut-on faire mieux que $O(n^3)$ que l'on aurait en résolvant naivement le système précédent ?

3.3 Theorème de Mercer et opérateur de co-variance

Continuons de nous placer dans le cadre de la perte des moindres carrés. Comme nous l'avons vu, le problème de généralisation a, sous certaines conditions simples sur Y, une solution dans $L^2(\mathcal{X}, \rho_X)$. La première question naturelle est de savoir si $\mathcal{H}$ est dense dans $L^2(\mathcal{X}, \rho_X)$; c'est le point de vue des noyaux universels, voir [13]. Plus généralement, on souhaite comparer les espaces $\mathcal{H}$ et $L^2(\mathcal{X}, \rho_X)$, et ceux pour deux raisons : 1) pour caractériser la régularité de la solution f_ρ de façon plus subtile que juste savoir qu'elle est dans L^2 et 2) pour comprendre précisément comment la norme de $\mathcal{H}$ se comporte vis-à-vis de la norme de L^2.

Dans cette partie, on suppose que $\mathcal{X}$ est séparable, et que K est continu et borné. Cela implique que l'espace $\mathcal{H}$ s'injecte continuement dans $L^2(\mathcal{X}, \rho_X)$ en posant $S : f \in \mathcal{H} \mapsto f \cdot K_{(\cdot)} \in L^2$ et que l'application conjuguée

$$S^* : \psi \in L^2(\mathcal{X}, \rho_X) \longmapsto \int_{\mathcal{X}} \psi(x) K_x d\rho_X \in \mathcal{H} \subset L^2(\mathcal{X}, \rho_X)$$

est une application surjective continue de $L^2(\mathcal{X}, \rho_X)$ dans $\mathcal{H}$. Pour comprendre le lien entre $L^2(\mathcal{X}, \rho_X)$ et $\mathcal{H}$, il est naturel d'introduire des *espaces d'interpolation* entre ces deux espaces, c'est à dire une suite $(\mathcal{H}^r)_{0 \leq r \leq 1}$ d'espaces de Hilbert tel que $\mathcal{H}^s \hookrightarrow \mathcal{H}^r,\ s < r$ et tel que $\mathcal{H}^0 = L^2(\mathcal{X}, \rho_X)$ et $\mathcal{H}^{1/2} = \mathcal{H}$. Ainsi, plus r est grand, plus les éléments de $\mathcal{H}^r$ sont réguliers et donc $\mathcal{H}^r$ est en un sens petit.

On utilise l'opérateur de noyau K T pour définir ces espaces :

$$T = SS^* \text{ tel que } T(\psi) = z \mapsto \int_{\mathcal{X}} \psi(x) K(x, z) d\rho_X$$

Définition 3. *En supposant toujours que K est borné et que $\mathcal{X}$ est séparable, on a les propriétés suivantes :*

- *T est auto-adjoint et compact ;*
- *Le suppléntaire orthogonal $Ker(T)^\perp$ du noyau de T dans $L^2(\mathcal{X}, \rho_X)$ admet donc une base orthonormée de vecteurs propres $(\phi_i)_{i \in I} \in L^2(\mathcal{X}, \rho_X)$ associés aux valeurs propres $(\mu_i)_{i \in I} > 0$ que l'on peut supposer rangées en ordre décroissant (en effet on peut prendre $I = [1, n]$ si T est de rang fini et $I = \mathbb{N}$ sinon) ;*
- *$T = \sum_{i \in I} \mu_i \phi_i \otimes_{L^2} \phi_i$*
- *Les $\frac{1}{\mu_i}\phi_i$ forment une base orthonormée de $\mathcal{H}$*

Pour $r \in]0,1]$, on définit T^r :

$$T^r : \psi = \sum_{i \in I} a_i \phi_i \in L^2 \longmapsto \sum_{i \in I} \mu_i^r a_i \phi_i$$

et $\mathcal{H}^r := Im(T^r) = \left\{ \sum_{i \in I} a_i \phi_i \mid \sum_{i \in I} \frac{a_i^2}{\mu_i^{2r}} < \infty \right\}$ muni du produit scalaire induit par T^r. De plus, $\mathcal{H} = \mathcal{H}^{1/2}$.

Grâce à ces espaces d'interpolation, on peut caractériser la régularité de la fonction objectif f_ρ pour ensuite montrer l'influence de cette régularité sur les taux statistiques des estimateurs ainsi que sur la complexité des algorithmes.

Hypotèse 1 (Condition de source)**.** *On fait souvent l'hypothèse suivante :*

$$f \in \mathcal{H}^r \text{ pour un certain } r \in [0,1]$$

Cela revient donc à faire une hypothèse de régularité sur la solution au problème de généralisation.

Si $f_\rho \in L^2$, cela correspond à $r = 0$. Si $f_\rho \in \mathcal{H}$, alors la condition de source sera vérifiée pour $r = \frac{1}{2}$.

Remarquons que l'opérateur $C = S^*S = \mathbb{E}[K_X \otimes K_X] \in \mathcal{L}(\mathcal{H})$ est caractérisé par la relation suivante :

$$\forall f, g \in \mathcal{H},\ \langle f, Cg \rangle_{\mathcal{H}} = \langle f, g \rangle_{L^2(\mathcal{X}, \rho_X)}$$

et que l'on a $C = \sum_{i \in I} \mu_i \left(\frac{1}{\mu_i} \phi_i \right) \otimes_{\mathcal{H}} \left(\frac{1}{\mu_i} \phi_i \right)$. On voit que les μ_i caractérisent la taille de l'espace $\mathcal{H}$ vis à vis de L^2 ; un grand μ_i correspond à une direction ϕ_i / μ_i dans $\mathcal{H}$ ayant une grande influence sur la norme L^2 alors que de petits μ_i correspondent à des directions dans $\mathcal{H}$ ayant peu d'impact sur la norme L^2.

Il est donc naturel de caractériser la taillle de l'espace $\mathcal{H}$ par rapport à L^2 par la condition suivante :

Hypotèse 2. *Le taux de décroissance des valeurs propres μ_i est de l'ordre $\mu_i \leq \frac{s^2}{i^\alpha}$ pour une paramètre $\alpha \in [1, +\infty]$. Alternativement, on peut voir cette condition de la manière suivante :*

$$Tr((C + \lambda I)^{-1} C) \leq Q^2 \lambda^{-1/\alpha}$$

L'introduction de l'opérateur T et des epaces $\mathcal{H}^r$ est donc motivé pour deux raisons :

— pour trouver une façon simple de caractèriser la taille de l'espace $\mathcal{H}$ grâce au paramètre α ; ceci est lié à l'étude du terme de fluctuation ou d'estimation ;

— pour trouver une façcon simple de caractériser la régularité de la solution au problème de généralisation grace au paramètre r ; ceci est liée à l'étude du terme d'approximation.

3.4 Résultats et fluctuations

Ces critères de régularité ont permis d'obtenir des taux dans le cas de la régression des moindres carrés avec régularisation. Soit $\mathcal{R}^\lambda$ le risque régularisé et f^λ le minimiseur de ce risque.

Dans ce cas, l'estimateur

$$\hat{f}_n^\lambda \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|^2 + \lambda ||f||^2$$

a les caractéristiques suivantes :

— le biais ou erreur d'approximation $\mathcal{R}(f^\lambda) - \mathcal{R}(f_\rho)$ est de l'ordre de λ^{2r}

— l'erreur d'estimation $\mathcal{R}^\lambda(\hat{f}_n^\lambda) - \mathcal{R}^\lambda(f^\lambda)$ est de l'ordre de $n^{-1}\lambda^{-1/\alpha}$

Donnons une idée des "preuves" de ces résultats.

Démonstration. 1. Établissons une décomposition biais-variance un peu différente de celles énoncées précédemment. Remarquons d'abord que le risque $\mathcal{R}$ peut s'exprimer de la manière suivante (grâce à la caractérisation de l'espérance conditionnelle) :

$$\forall f \in L^2(\mathcal{X}, \rho_X),\ \mathcal{R}(f) - \mathcal{R}(f_\rho) = \frac{1}{2} ||f - f_\rho||^2_{L^2(\mathcal{X}, \rho_X)}$$

Et donc

$$\forall f \in \mathcal{H},\ \mathcal{R}(f) - \mathcal{R}(f_\rho) = \frac{1}{2} ||Sf - f_\rho||^2_{L^2} = \frac{1}{2} \langle f, Cf \rangle_{\mathcal{H}} - \langle f, S^* f_\rho \rangle_{\mathcal{H}}$$

On peut donc obtenir la décomposition biais-variance suivante :

$$\left(\mathcal{R}(\hat{f}_n^\lambda) - \mathcal{R}(f_\rho)\right)^{1/2} \leq ||S\left(\hat{f}_n^\lambda - f^\lambda\right)||_{L^2} + ||Sf^\lambda - f_\rho||_{L^2} \leq ||C^{1/2}\left(\hat{f}_n^\lambda - f^\lambda\right)||_{\mathcal{H}} + ||Sf^\lambda - f_\rho||_{L^2}$$
$$\leq ||(C+\lambda I)^{1/2}\left(\hat{f}_n^\lambda - f^\lambda\right)||_{\mathcal{H}} + ||Sf^\lambda - f_\rho||_{L^2} = \left(\mathcal{R}^\lambda(\hat{f}_n^\lambda) - \mathcal{R}^\lambda(f^\lambda)\right)^{1/2} + \left(\mathcal{R}(f^\lambda) - \mathcal{R}(f_\rho)\right)^{1/2}$$

où C est l'opérateur de co-variance.

2. Pour borner l'erreur d'approximation, on remarque simplement que $f^\lambda = (C+\lambda I)^{-1}S^* f_\rho$, ce qui montre que vu comme une fonction de $L^2(\mathcal{X}, \rho_X)$, on a $f^\lambda = T(T+\lambda)^{-1}f_\rho$.
Puis :

$$||f^\lambda - f_\rho||^2_{L^2(\mathcal{X},\rho_X)} = ||(T+\lambda)^{-1}Tf_\rho - f_\rho||^2_{L^2} = \lambda^2||(T+\lambda I)^{-1}f_\rho||^2 \leq \lambda^2 \left\|(T+\lambda I)^{-1}T^r\right\|^2 ||f_\rho||^2_{\mathcal{H}^r} \leq \lambda^{2r}||f_\rho||^2_{\mathcal{H}^r}$$

3. Pour borner l'erreur de fluctuation, on remarque que la dimension effective du problème est $\mathrm{Tr}((C+\lambda I)^{-1}C)$, et donc que les l'opérateur de co-variance régularisé $C+\lambda I$ et l'opérateur de covariance empirique régularisé $\hat{C}_n + \lambda I = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}$ sont proches dès que $n \geq \mathrm{Tr}((C+\lambda I)^{-1}C)$ (on utilise une inégalité de concentration sur les opérateurs au moyen de leur dimension effective). C'est dans l'estimation de l'opérateur $C+\lambda I$ par $\hat{C}_n + \lambda I$ que la quantité $n^{-1}\mathrm{Tr}((C+\lambda I)^{-1}C)$ apparaît.

□

En combinant ces deux erreur et en optimisant en λ pour qu'elle soient du même ordre, on voit que l'on peut espérer, pour $\lambda = n^{-\frac{\alpha}{2r\alpha+1}}$ et donc un exces de risque de l'ordre de $n^{-\frac{2r\alpha}{2r\alpha+1}}$

Dans [7], Caponnetto et De Vito montrent que ces taux sont optimaux (au sens du minimax), c'est à dire qu'avec les hypothèses considérées, avec n données, on ne peut pas faire mieux que $n^{-\frac{2r\alpha}{2r\alpha+1}}$. L'estimateur du minimum du risque empirique est donc optimal. Cependant, comme expliqué plus haut, son calcul est trop couteux a priori. Il est donc nécéssaire de concevoir des estimateurs qui ont les mêmes propriétés statistiques, mais dont le calcul est plus aisée.

4 Domaine de recherche, travaux et directions futures

Dans cette partie, nous montrons le point précis où se place notre domaine de recherche actuel. Dans la première partie, nous introduisons une méthode algorithmique qui permet de guarantir une bonne complexité dans le cadre de la partie précédente, sous réserve que l'on ait $r \geq 1/2$. Dans la deuxième et troisième partie, nous présenterons les travaux actuels (inachevés) et direction futur pour étendre certains résultats statistiques à une classe plus vaste de fonctions de perte ainsi que de trouver des algorithmes performants atteignant ces taux.

4.1 Algorithmes rapides dans le cas $r \geq 1/2$

Rappelons que dans le cas de la régression Ridge, l'estimateur $\hat{f}_n^\lambda$ minimisant le risque empirique régularisé est optimal pour un certain λ bien choisi. Rappelons que cet estimateur est de la forme

$$\hat{f}_n^\lambda = \sum_{i=1}^n \alpha_i K_{x_i} \text{ où } (K_{nn} + \lambda n I)\,\alpha = \hat{y}$$

Résoudre ce système de manière exacte est beaucoup trop coûteux. Le point de départ de notre stage de M2 a été l'article [16] qui propose un nouvel estimateur atteignant les mêmes taux que celui-ci mais calculable en temps raisonnable. Cet algorithme repose sur deux idées.

1. La première est de réduire la dimension de l'espace dans lequel on optimize par *projections aléatoires* (ou projections de Nÿstrom, voir par exemple [15] ou [17]). Le principe est de tirer un sous-échantillon de taille M des données aléatoirement $(\tilde{x}_j)_{1\leq j\leq M}$ et de minimiser le risque régularisé sur l'espace vectoriel $\mathcal{H}_M$ engendré par les $K_{\tilde{x}_j}$. On peut interpréter cette m éthode comme une méthode de projection des données (K_{x_i}) sur un espace de dimension plus petite $M << n$. Le problème devient alors le suivant :

$$\hat{f}_M^\lambda = \sum_{j=1}^M \alpha_j K_{\tilde{x}_j} \text{ où } \left(K_{nM}^T K_{nM} + \lambda n K_{MM}\right)\alpha = K_{nM}^T \hat{y}$$

où K_{MM} est la matrice des $K(\tilde{x}_i, \tilde{x}_j)$ et K_{nM} est la matrice des $K(x_i, \tilde{x}_j)$. Remarquablement, cet estimateur a les mêmes propriétés statistiques que $\hat{f}_n^\lambda$ sous réserve que M soit assez grand, typiquement de l'ordre de $\frac{1}{\lambda}$. Dans ce cas, on peut montrer qu'avec forte probabilité, on a

$$||(C + \lambda I)^{1/2}(I - P_M)||^2 \leq 2\lambda$$

Cette inégalité montre que l'erreur induite par cette projection est essentiellement du même ordre que celle induite par le régulariseur, ce qui ne change donc pas l'erreur finale.

2. La deuxième est de résoudre le système précédent de manière efficace ; en effet, si on résoud le système directement, le calcul de la matrice $K_{nM}^T K_{nM}$ a un cout de l'ordre de nM^2 qui est donc de l'ordre de n^2 lorsque l'on se place dans le cadre statistique a taux lents ($\lambda = n^{-1/2}, \mathcal{R}(\hat{f}_n^\lambda) \approx n^{-1/2}$).

   L'idée est de ne pas résoudre directement ce système mais de le résoudre par une méthode iterative, i.e. de voir le système $Ax = b$ (où $A \in \mathcal{S}_n^+$)comme la minimisation de la fonctionnelle quadratique $x^T Ax - 2b^T x$, que l'on peut chercher à minimiser par descente de gradient (le gradient en x est ici $Ax - b$). Ces méthodes ont l'avantage de ne pas devoir calculer la matrice A explicitement mais uniquement des produits de la forme Ax, ce qui nous permet d'éviter le problème précédent. En revanche, le nombre d'itérations de cette méthode de gradient dépend du *conditionnement* de la matrice A, c'est à dire le rapport entre sa plus grande et sa plus petite valeur propre $\text{Cond}(A) := \frac{\sigma_{\max}(A}{\sigma_{\min}(A)}$. Il s'agit donc de *pré-conditionner* la matrice A, c'est à dire de trouver une matrice B telle que $B^T AB$ ait un bon conditionnement. Ensuite, il suffit de résoudre le système $B^T ABx' = B^T b$ rapidement et de recouvrir la solution approximative x au problème initial en posant $x = Bx'$.

   Ceci est faisable dans notre problème si l'on utilise comme matrice B une approximation de la racine carrée de $\left(K_{nM}^T K_{nM} + \lambda n K_{MM}\right)$, que l'on obtient en prenant la racine carrée de $\left(\frac{n}{M} K_{MM}^T K_{MM} + \lambda n K_{MM}\right)$. On a alors, avec forte probabilité, une borne sur le conditionnement du nouveau système ce qui garantit la convergence de l'algorithme en temps fini. Le cout devient donc de l'ordre de nM, soit $n\sqrt{n}$ dans le cas des taux lents.

Cette accélération significative et optimal de la régression ridge nous pousse à nous intéresser à de nouveaux problèmes à noyaux. Notamment, nous nous posons deux questions principales 1) est-il possible d'obtenir des taux pour les estimateurs obtenus par mininisation du risque empirique pour des pertes convexes plus générales ? 2) Est-il possible de créer des algorithmes qui atteignent rapidement ces taux ?

4.2 Généralisation des taux à d'autres fonctions de perte

Il est possible de montrer que les estimateurs obtenus par minimisation du risque empirique régularisé pour certaines fonctions convexes ont des propriétés proches de celles de la mininisation de la perte des moindres carrés. Le problème principal est que le risque ne peux plus s'écrire comme une fonction quadratique, et donc qu'il n'y a plus forcément de fonction de co-variance naturelle. Cependant, en se plaçant au voisinage de l'optimum, on voit que certaines classes de fonctions convexes se comportent comme des fonctions quadratiques et que l'on peut donc définir des paramètre semblables à ceux définis plus haut.

Par exemple si l'on fait les hypothèse suivantes :

Hypotèse 3.
- *La fonction de perte est de la forme $l(y, f(x))$ où l est trois fois dérivable et $l^{(3)} \leq C_l l''$*
- *l est M_1 lipschitz et sa Hessienne est M_2 lisse*
- *La fonction qui minimize le risque de généralisation $\mathcal{R}(f) = \mathbb{E}[l(Y, f(X))]$ f_ρ est dans $\mathcal{H}$. H désigne la Hessienne de $\mathcal{R}$ vu comme fonction sur $\mathcal{H}$ à l'optimum f_ρ*
- *La fonction f_ρ satisfait une condition de source de la forme $||H^{-r} f_\rho||_{L^2} = R < \infty$ pour un $r \in [1/2, 1]$.*
- *Les données sont bien conditionnées par rapport à H, c'est à dire qu'il existe $\alpha \in [0, 1]$ tel que $||(H + \lambda)^{-1/2}\phi(X)||_{\mathcal{H}}^2 \leq Q^2 \lambda^{-\frac{1}{\alpha}}$ (analogue de la décroissance des valeurs probres de l'opérateur de covariance)*

alors on a le théorème suivant, qui montre les mêmes taux que dans le cas de la régression des moindres carrés :

Théorème 2. *Soit $\delta > 0$. Soit $\tilde{C} := R + 4QM_1 \log \frac{2}{\delta}$. Supposons que l'on ait de la régularité, c'est à dire que $2r\alpha > 1$. Alors il existe un n_0 tel que pour tout $n \geq n_0$, avec probabilité supérieure à $1 - 2\delta$,*

$$\mathcal{R}(\hat{f}_n^\lambda) - \mathcal{R}(f_\rho) \leq 71\tilde{C}^2 n^{-\frac{2r\alpha}{1+2r\alpha}}$$

La méthode de preuve, plus technique, se base sur des propriétés d'estimation des fonctions convexe proche de l'optimum, en controlant les variation de la Hessienne grâce au contrôle sur la dérivée troisième.

4.3 Trouver des algorithme rapides qui atteignent ces taux

La méthode présentée dans [16] et profondémment quadratique ; un premier pas de notre travail a été de généraliser cet algorithme à n'importe quelle perte quadratique. Ceci peut être exploitable si l'on minimize le risque empirique par une méthode d'ordre deux, c'est à dire en calculant des pas de Newton successifs. Le problème de ces méthodes est qu'elles ne convergent vite que dans une petite région de l'espace autour de l'optimum ; nous souhaiterions donc trouver une façon intelligente de se retrouver dans une bonne région rapidement, en appliquant un algorithme de descente de gradient stochastique avant se faire un pas de Newton, comme expliqué dans l'article de Bach et Moulines [2]. Nous espérons pouvoir arriver à des complexités en temps de l'ordre de $n\sqrt{n}$ ou du moins inférieures à n^2.

Références

[1] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society **68** (1950), 337–404.

[2] Francis R. Bach and Eric Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n)*, CoRR **abs/1306.2119** (2013).

[3] St. Bergmann, *Uber die entwicklung der harmonischen funktionen der ebene und des raumes nach orthogonalfunktionen*, Mathematische Annalen **86** (1922), 238–271.

[4] S. Boucheron, O. Bousquet, and G. Lugosi, *Theory of classification : A survey of some recent advances*, ESAIM : Probability and Statistics **9** (2005), 323.

[5] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities. A nonasymptotic theory of independence*, Oxford University Press, 2013.

[6] Stéphane Boucheron and Pascal Massart, *A high-dimensional wilks phenomenon*, Probability Theory and Related Fields **150** (2011), no. 3, 405–433.

[7] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Foundations of Computational Mathematics **7** (2007), no. 3, 331–368.

[8] C.F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, Carl Friedrich Gauss Werke, sumtibus F. Perthes et I. H. Besser, 1809.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[10] M. Ledoux and M. Talagrand, *Probability in banach spaces : Isoperimetry and processes*, A Series of Modern Surveys in Mathematics Series, Springer, 1991.

[11] A.M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*, Nineteenth Century Collections Online (NCCO) : Science, Technology, and Medicine : 1780-1925, F. Didot, 1805.

[12] J. Mercer, *Functions of positive and negative type, and their connection with the theory of integral equations*, Philosophical Transactions of the Royal Society, London **209** (1909), 415–446.

[13] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang, *Universal kernels*, J. Mach. Learn. Res. **7** (2006), 2651–2667.

[14] Yurii Nesterov, *Introductory lectures on convex optimization : A basic course*, 1 ed., Springer Publishing Company, Incorporated, 2014.

[15] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco, *Less is more : Nyström computational regularization*, Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (Cambridge, MA, USA), NIPS'15, MIT Press, 2015, pp. 1657–1665.

[16] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco, *FALKON : an optimal large scale kernel method*, Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 3891–3901.

[17] Alessandro Rudi and Lorenzo Rosasco, *Generalization properties of learning with random features*, Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., 2017, pp. 3215–3225.

[18] Vladimir N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, Berlin, Heidelberg, 1995.