Performances statistiques de méthodes à noyaux

Sébastien Loustau

UNIVERSITÉ DE PROVENCE
U.F.R. M.I.M.
ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE E.D. 184

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PROVENCE

Spécialité : Mathématiques

par

Sébastien LOUSTAU

sous la direction du Pr. Laurent CAVALIER

Titre :

Performances statistiques de méthodes à noyaux

soutenue publiquement le 28 novembre 2008

JURY

M. Peter L. BARTLETT	University of Berkeley	Rapporteur
M. Philippe BERTHET	Université Paul Sabatier	Examinateur
M. Gérard BIAU	Université Pierre et Marie Curie	Rapporteur
M. Gilles BLANCHARD	FIRST Institute	Examinateur
M. Laurent CAVALIER	Université de Provence	Directeur
M. Oleg LEPSKI	Université de Provence	Examinateur
M. Liva RALAIVOLA	Université de Provence	Examinateur

Remerciements

Je tiens avant tout à remercier mon directeur de thèse Laurent Cavalier pour de multiples raisons. Tout d'abord, merci de m'avoir lancer sur ce sujet particulièrement riche, à l'intersection de plusieurs disciplines, des mathématiques théoriques et appliquées. Merci pour ton intuition statistique sur les problèmes d'apprentissages qui m'a guidé sans faille durant ces trois années, cela sans jamais toucher à ma liberté et mon autonomie. Merci enfin pour toutes ces discussions dans ton bureau qui m'ont permis d'acquérir une certaine culture statistique.

Je suis très heureux que Peter Bartlett et Gérard Biau aient accepté de rapporter cette thèse. Je les remercie du temps passé à la relecture du manuscrit dans leur emploi du temps très chargé. Je tiens ensuite à remercier les membres du jury Gilles Blanchard, Philippe Berthet, Oleg Lepski et Liva Ralaivola d'avoir accepté de participer à la soutenance.

J'aimerai aussi remercier tous les membres de l'équipe proba-stats du CMI, pour leur sympathie et les discussions échangées. L'accueil a été chaleureux et les discussions très intéressantes. Merci également à l'équipe BDAA du LIF pour les groupes de travail et l'organisation de l'EPIT, qui m'ont permis de lier contact avec la communauté de l'apprentissage. Plus largement, je voudrais aussi remercier tout le personnel du CMI, indispensable au quotidien d'un thésard.

Le quotidien de ces trois années de thèse a été très agréable grâce à l'ambiance du bureau 114 du CMI. Je voudrais remercier particulièrement Clément Marteau et Lionel Paris pour ces années de travail dans la bonne humeur. Merci plus largement à tous les doctorants du CMI pour les pauses repas, cafés (chocolat ou truffine ?), séminaires bucoliques ou autres. Dans le petit monde des doctorants, je voulais aussi remercier les thésards croisés "on the road", avec qui j'ai pu passer de très bons moments.

Je tiens à exprimer toute ma reconnaissance à ma famille pour son accompagnement dans ce long parcours, de Pau à Marseille, en passant par Bordeaux et Séville. La distance qui nous sépare en valait peut-être la peine...

Je tiens enfin à remercier tout particulièrement ma femme Sarah pour la confiance dont elle m'a témoigné en quittant son emploi pour suivre un étudiant à 600 kms. Son soutien a été indispensable à certains moments. Merci pour ta contribution à cette thèse. Merci enfin pour toutes les grandes choses déjà vécues, et sans doute à venir...

Table des matières

Chapitre 1

Classification, mathématique et statistique

Résumé

Ce chapitre suit deux lignes directrices : introduire les outils mathématiques indispensables à la compréhension de ce manuscrit et mettre en relief les enjeux statistiques de cette thèse en apprentissage. Les notions abordées sont donc assez variées et empruntées à plusieurs domaines des mathématiques, de la statistique et de l'apprentissage.
Dans un premier temps, nous présenterons brièvement le modèle de la classification et les premiers résultats statistiques, de la théorie de Vapnik à des développements plus récents. Nous décrirons ensuite l'algorithme souvent considéré dans cette thèse : les Machines à Vecteurs de support (SVM). Cette procédure de régularisation suscite de nombreuses investigations dans tous les domaines de l'apprentissage. On présentera sa version noyau et un point de vue fonctionnel des méthodes à noyaux. Ces précisions nous permettront de mieux comprendre les résultats obtenus dans les Chapitres 2, 3 et 4.

1.1 Introduction générale à la classification

L'apprentissage est un axe de recherche très vaste qui intéresse plusieurs communautés de scientifiques. D'une part, les praticiens font face à des problèmes réels de décision, de discrimination ou de prédiction. Ils veulent utiliser des outils performants pour résoudre leurs problèmes spécifiques. D'autre part, les théoriciens disposent de modèles qui s'adaptent à un grand nombre de problèmes. Leur but est de proposer des algorithmes capables de prédire, de décider ou de classer, sous certaines hypothèses.
La théorie de l'apprentissage regroupe un très grand nombre de modèles, que l'on peut décrire de manière générale de la façon suivante. On dispose d'un ensemble d'observations sur un phénomène mal connu. Ces observations se présentent sous la forme de couples (x_i,y_i) où x_i est une variable d'entrée et y_i la sortie ou réponse correspondante. Cet ensemble d'observations est appelé ensemble d'apprentissage. A partir de cet échantillon, l'objectif d'un algorithme d'apprentissage est de prédire la réponse y d'une nouvelle entrée x. On dit qu'il généralise.

1.1.1 Illustrations et modélisation

Illustrations

Cette thèse se restreint à un cadre relativement simple : la classification binaire. Pour illustrer ce problème, considérons un exemple de la vie courante : le problème des spams (pourriel). Chaque utilisateur du réseau internet possède une adresse e-mail (ou courriel) et une boîte de réception. Cette dernière regroupe toutes les communications électroniques reçues et envoyées par l'internaute. Le spam désigne une communication électronique non souhaitée expédiée en masse à des fins malhonnêtes. Ainsi, il n'est pas rare d'ouvrir sa boîte mail et de supprimer, chaque jour, plusieurs spams au beau milieu de messages valables. Aujourd'hui ce tri n'est plus seulement réalisé par l'utilisateur mais par un filtre anti-spam (ou modérateur). Ce filtre est bien l'exemple d'un algorithme de classification. Ayant une certaine connaissance des communications électroniques de l'utilisateur, via sa boîte de réception, il a pour mission de classer chaque nouveau mail reçu selon deux catégories : spam ou non-spam. Cet automate associe bien à un ensemble d'apprentissage une règle de décision. Il permet ainsi à l'utilisateur de minimiser les dégâts causés par ce phénomène : il n'aura plus qu'à corriger les erreurs éventuelles de l'algorithme. Elles sont de deux types : un spam peut apparaître au milieu des messages valables alors qu'un message sain peut être classé parmi les spams. Dans ce manuscrit, on mesure de manière statistique l'erreur de généralisation des algorithmes de classification.

On peut illustrer le pouvoir d'un algorithme d'apprentissage par un jeu simple : le jeu de pierre-papier-ciseau. Face à face, deux joueurs proposent simultanément la pierre, le papier ou le ciseau. Le papier bat la pierre, la pierre bat le ciseau et le ciseau bat le papier. Ce jeu est ainsi strictement basé sur le hasard. Néanmoins on peut construire un algorithme d'apprentissage capable de bonnes performances. Si un être humain joue contre cet algorithme, son score s'éloignera de celui de la machine au cours du temps. La raison est la suivante : le cerveau humain est incapable de générer de façon indépendante et uniforme sa séquence de pierre, papier, ciseau. La machine va ainsi apprendre au fur et à mesure du jeu le comportement du joueur et prédire relativement bien ses coups à l'avance. La Figure 1.1 montre l'évolution du score d'un joueur face à une machine[1], jusqu'à 1000 coups. On observe un score qui décroît assez significativement pour atteindre environ -80 après 1000 coups. Cependant, le comportement du joueur pouvant varier pendant la partie, ce problème est vraiment délicat. C'est pourquoi la courbe a une tendance "en dent de scie".

Modélisation

On propose de modéliser le problème de classification binaire de la manière suivante. Les notations qui suivent seront respectées dans ce premier chapitre. On considère que l'ensemble d'apprentissage $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$ est une suite de couples de variables aléatoires telles que :

- (X_i,Y_i), pour $i = 1,\dots,n$ sont indépendantes et identiquement distribuées (i.i.d.) de loi P inconnue,
- $X_i \in \mathcal{X}$ pour $i = 1,\dots,n$ sont appelées *variables d'entrées* (inputs),

1. L'expérience a été réalisée en utilisant l'algorithme d'intelligence artificielle WWW Roshambot de Perry Friedman, de l'Université de Stanford.

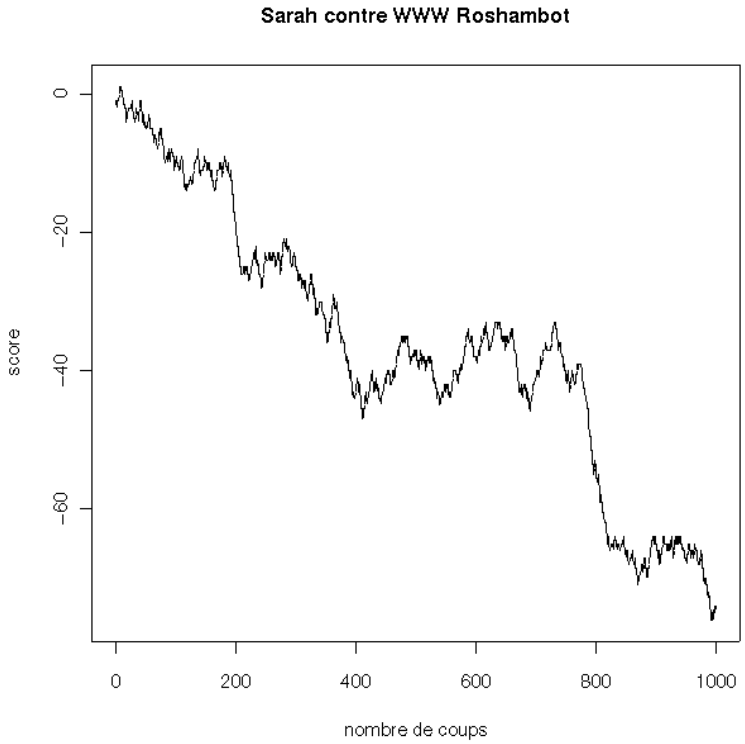

Figure 1.1 : Evolution du score du joueur face à la machine.

• $Y_i \in \{+1, -1\}$ pour $i = 1, \ldots, n$ sont les *classes* correspondantes (outputs).

De manière générale on ne fait aucune restriction sur $\mathcal{X}$. Cependant par la suite on supposera que $\mathcal{X} \subset \mathbb{R}^d$. Chaque entrée $x \in \mathcal{X}$ représente un vecteur de $\mathbb{R}^d$ où chaque composante est une caractéristique de x. On notera la loi marginale des X_i par P_X et la loi de l'ensemble d'apprentissage par P^n, qui est le produit tensoriel de n lois P. Etant donné cet ensemble d'apprentissage D_n, le but de la classification binaire est de prédire la classe Y d'une nouvelle observation X où (X,Y) est de loi P indépendante de D_n. Autrement dit, on cherche à construire une règle de décision $f : \mathcal{X} \to \{+1, -1\}$.

Définition 1.1 *On appelle règle de Bayes, que l'on note f^*, l'application définie sur $\mathcal{X}$ par :*

$$f^*(x) = \text{sign}(2\eta(x) - 1),$$

où $\eta(x) = P(Y = 1|X = x)$ et $\text{sign}(x) = 1$ pour $x \geq 0$ et -1 sinon.

Il est clair que f^* est le meilleur prédicteur possible. Il consiste à répondre 1 à x lorsque $P(Y = 1|X = x) > 1/2$ et -1 sinon. Malheureusement, il dépend de la probabilité conditionelle η et donc de la loi inconnue P. Il n'est donc pas calculable. On s'intéresse alors à construire à partir des observations un classifieur ou estimateur qui imite f^*. Pour cela, il faut mesurer la qualité d'une règle de décision.

Définition 1.2 *On appelle erreur de généralisation d'un classifieur f la quantité :*

$$R(f) = P(f(X) \neq Y).$$

Cette erreur est la plus répandue en classification. Elle mesure la probabilité d'erreur de f sur le couple (X,Y) de loi P.

Lemme 1.1 *Soit f^* la règle de Bayes. On a alors :*

$$R(f^*) = \min_{f \text{ mesurable}} R(f).$$

La règle de Bayes est bien la meilleure règle de décision. Elle minimise l'erreur de généralisation. On mesure alors la qualité d'un classifieur par rapport à celle du Bayes par l'excès de risque.

Définition 1.3 *On appelle excès de risque d'un classifieur $\hat{f}_n$ la quantité aléatoire :*

$$R(\hat{f}_n, f^*) = P(\hat{f}_n(X) \neq Y | X_1, \ldots, X_n, Y_1, \ldots, Y_n) - P(f^*(X) \neq Y).$$

La théorie statistique de l'apprentissage s'intéresse à la loi de probabilité de $R(\hat{f}_n, f^*)$. On peut notamment obtenir des bornes sur $R(\hat{f}_n, f^*)$ qui ont lieu avec grande probabilité. Cela revient à étudier les grandes déviations de cette quantité. On peut aussi se pencher sur le comportement moyen de l'excès de risque en l'intégrant par rapport à la mesure P^n de l'échantillon. On parle alors d'excès de risque intégré. Il permet de définir deux notions essentielles en statistique : la consistance et la vitesse de convergence.

Définition 1.4 *Soit $\hat{f}_n$ un classifieur.*

1. *On dit que $\hat{f}_n$ est consistant par rapport à P si :*

$$\mathbb{E}_{P^n} R(\hat{f}_n, f^*) \xrightarrow[n \to \infty]{} 0.$$

2. *On dit que $\hat{f}_n$ atteint la vitesse de convergence $(\psi_n)_{n\in\mathbb{N}^*}$ s'il existe une constante $C>0$ telle que pour tout n :*

$$\mathbb{E}_{P^n} R(\hat{f}_n, f^*) \leq C\psi_n. \tag{1.1}$$

Dans ce manuscrit, on s'intéresse principalement à l'étude de l'excès de risque intégré. On veut obtenir des inégalités du type (1.1), c'est-à-dire des vitesses de convergence vers la règle de Bayes. Par abus de notation, on notera parfois $\mathbb{E}$ l'espérance par rapport à P^n.

Pertes et marge

Plus généralement dans cette thèse, on considère des fonctions à valeurs réelles dont le signe détermine la classe. On peut dans ce cas définir la perte d'une fonction $f : \mathcal{X} \to \mathbb{R}$ et son risque associé de la manière suivante.

Définition 1.5 *1. On appelle perte toute application l définie sur $\{-1, +1\} \times \mathbb{R}$ à valeurs dans $\mathbb{R}^+$ pour laquelle $l(y,f(x))$ mesure la perte de f au point (x,y).*

2. On appelle l-Risque la quantité :

$$R_l(f) = \mathbb{E}_P l(Y, f(X)).$$

Il est clair que pour f à valeurs dans $\mathbb{R}$, l'erreur de généralisation s'écrit :

$$R(f) = \mathbb{E}_P \, \mathbb{1}(\text{sign}(f(X)) \neq Y). \tag{1.2}$$

Ainsi $R(f)$ est un l-Risque où $l(y,f(x)) = \mathbb{1}(\text{sign}(f(x)) \neq y)$ est appelée perte 0-1 ou perte dure. Cette perte vaut 1 si $\text{sign}(f(x)) \neq y$ et 0 sinon. Comme fonction de $yf(x)$, elle n'est donc ni convexe ni continue. Ces irrégularités engendrent des problèmes algorithmiques. C'est pourquoi on introduit souvent d'autres pertes.

De nombreuses pertes ont été proposées dans la communauté de l'apprentissage. Généralement ces pertes sont basées sur la marge. En considérant des fonctions à valeurs réelles, on peut mesurer la marge de la fonction f au point (x,y) par $\alpha = yf(x)$. Cette quantité est au centre de plusieurs algorithmes très influents. On peut citer notamment l'algorithme SVM (Boser et al. [28]) ou Adaboost (Freund and Schapire [56]). Ces deux algorithmes utilisent des pertes basées sur la marge, vérifiant :

$$l(y,f(x)) = \phi(yf(x)), \tag{1.3}$$

où ϕ est définie par

- $\phi(\alpha) = (1-\alpha)_+$ pour la perte SVM (ou perte charnière, perte douce),
- $\phi(\alpha) = \exp(-\alpha)$ pour la perte Adaboost.

Il existe d'autres pertes comme la perte q-SVM définie par $l(y,f(x)) = (1-yf(x))_+^q$. D'un point de vue statistique, la perte utilisée va influencer la nature du problème. La fonction cible, c'est-à-dire la fonction minimisant le l-Risque, peut varier de forme. Dans ce manuscrit, on s'intéresse particulièrement à la perte SVM. Cette perte possède les propriétés statistiques suivantes.

Lemme 1.2 *On considère $l(y,f(x)) = (1-yf(x))_+$ la perte SVM.*

1. l est dite Bayes-consistante :

$$\min_{f:\mathcal{X}\to\mathbb{R}} R_l(f) = R_l(f^*),$$

où f^* est la règle de Bayes.

2. Pour toute fonction $f : \mathcal{X} \to \mathbb{R}$,

$$R(f,f^*) \leq R_l(f,f^*) := R_l(f) - R_l(f^*).$$

La perte douce est fidèle au problème initial de classification. En effet, le minimiseur du risque associé à cette perte correspond à la règle de Bayes. De plus, un contrôle de $R_l(f,f^*)$ implique un contrôle de l'excès de risque de f. Ces propriétés nous permettrons d'obtenir des vitesses de convergence de la forme (1.1) à partir d'algorithmes minimisant la perte douce.
Le résultat suivant montre que ces propriétés ne sont plus vérifiées par la perte q-SVM.

Lemme 1.3 *On considère $l_q(y,f(x)) = (1 - yf(x))_+^q$ la perte q-SVM, où $q > 1$.*

1. Le minimum $f_q^* = \arg\min_f R_{l_q}(f)$ vérifie :

$$f_q^*(x) = \frac{\eta(x)^{\frac{1}{q-1}} - (1-\eta(x))^{\frac{1}{q-1}}}{\eta(x)^{\frac{1}{q-1}} + (1-\eta(x))^{\frac{1}{q-1}}},$$

où $\eta(x) = P(Y = 1|X = x)$.

2. Pour toute fonction $f : \mathcal{X} \to \mathbb{R}$,

$$R(f,f^*) \leq C_q \sqrt{R_{l_q}(f,f_q^*)},$$

où $C(q) \leq \sqrt{2}$ pour $q > 1$.

D'un point de vue statistique, la perte q-SVM est très différente de la perte SVM. Le meilleur classifieur n'est plus la règle de Bayes mais une fonctionnelle de η. Par exemple, si $q = 2$, $f_2^*(x) = 2\eta(x) - 1$. Cette fonction admet les mêmes propriétés de classification que la règle de Bayes, c'est-à-dire $\text{sign}(f_q^*) = f^*$. Mais cette fonction est à valeurs dans l'intervalle $[-1,1]$. De plus, un contrôle du risque associé à la perte q-SVM n'assure pas un contrôle de l'erreur de généralisation. Cela entraîne une perte dans les vitesses de convergence.
Plus généralement, Bartlett et al. [14] étudie l'utilisation de pertes du type (1.3) et leurs propriétés statistiques. Il obtient des conditions suffisantes assurant que la cible f_l^* vérifie $\text{sign}(f_l^*) = f^*$. En particulier, il suffit que la fonction ϕ dans (1.3) soit différentiable en zéro et que $\phi'(0) < 0$.

Régression et classification

Le modèle de régression non-paramétrique est un modèle fondamental des statistiques mathématiques. On dispose de n couples de variables (X_i,Y_i), $i = 1,\dots,n$ i.i.d. telles que :

$$Y_i = f(X_i) + \epsilon_i, \tag{1.4}$$

où les variables ϵ_i vérifient $\mathbb{E}\epsilon_i = 0$. Le problème de régression non-paramétrique est celui de l'estimation de la fonction f lorsqu'on sait a priori que cette fonction appartient à un espace fonctionnel de dimension infinie. Les résultats statistiques dans le modèle (1.4) sont nombreux. On peut citer Tsybakov [135] pour une introduction à l'estimation fonctionnelle.
Dans sa forme la plus générale, la théorie statistique de l'apprentissage dispose de couples (X_i,Y_i), $i = 1,\dots,n$. On peut ainsi inclure des modèles du type (1.4), où dans ce cas, $Y_i \in \mathcal{Y} \subset \mathbb{R}$ (Vapnik [139],Cucker and Smale [44]).

Dans le modèle de régression, les variables ϵ_i sont souvent supposées i.i.d. de loi normale $\mathcal{N}(0,\sigma^2)$. Dans le Chapitre 4, on étend une méthode de statistique mathématique au cadre de la classification. Pour cela, on se ramène à un modèle du type (1.4) de la manière suivante. Etant donné un ensemble d'apprentissage (X_i,Y_i), $i=1,\dots,n$ où $Y_i \in \{-1,+1\}$, on peut écrire :

$$Y_i = f_\eta(X_i) + \epsilon_i, \tag{1.5}$$

où $f_\eta(x) = 2\eta(x) - 1$ est la fonction de régression du modèle de classification. Les variables aléatoires $\epsilon_i = Y_i - f_\eta(X_i)$ représentent le bruit dans les observations. Elles sont bien centrées. Cependant la distribution de ϵ_i dépend des points X_i. D'un point de vue statistique, le modèle de classification paraît assez éloigné des modèles usuels en régression.

1.1.2 Principe de minimisation du risque empirique

La théorie statistique de l'apprentissage débute avec les travaux de Vapnik et Chervonenkis. L'idée est de minimiser un critère empirique basé sur l'ensemble d'apprentissage. Il s'agit du principe de minimisation du risque empirique.

La théorie de Vapnik-Chervonenkis

Soit $\mathcal{F}$ une collection de fonctions f définies sur $\mathcal{X}$ et à valeurs dans $\{-1,+1\}$. On cherche à estimer le meilleur classifieur de la collection $\mathcal{F}$, appelé *oracle* défini par :

$$f^*_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} R(f). \tag{1.6}$$

Ce classifieur n'est pas mesurable puisqu'il dépend de la loi P inconnue. L'idée naturelle proposée par Vapnik et Chervonenkis dans les années 70 (Vapnik [138]) est de considérer le risque empirique, défini par :

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f(X_i) \neq Y_i). \tag{1.7}$$

Le classifieur ERM (Empirical Risk Minimizer), s'il existe, sera noté $\hat{f}_n$ et minimise (1.7) sur la classe $\mathcal{F}$. Intuitivement, pour un nombre d'observations suffisamment grand, $\hat{f}_n$ s'approche de $f^*_{\mathcal{F}}$. Pour une fonction f fixée, $R_n(f) \to R(f)$ d'après la loi des grands nombres. La consistance de cet estimateur revient donc à établir une loi des grands nombres fonctionnelle, c'est-à-dire uniforme sur $\mathcal{F}$. En effet :

$$R(\hat{f}_n,f^*_{\mathcal{F}}) \leq 2\sup_{f\in\mathcal{F}} |R(f) - R_n(f)|. \tag{1.8}$$

Pour obtenir un contrôle de l'excès de risque de $\hat{f}_n$ par rapport à $f^*_{\mathcal{F}}$, on se ramène à l'étude du supremum d'un processus empirique. L'inégalité de Vapnik-Chervonenkis est la pierre angulaire de cette théorie. Elle permet de contrôler le membre de droite de (1.8) avec forte probabilité. Cette borne dépend de la complexité de la classe $\mathcal{F}$ considérée.

Définition 1.6 *Soit $\mathcal{F}$ un ensemble de fonctions de $\mathcal{X}$ à valeurs dans $\{-1,+1\}$. On définit $\mathcal{A}$ la classe de sous-ensembles de $\mathcal{X}$ qui s'écrivent :*

$$\{\{x\in\mathcal{X} : f(x)=1\}\times\{-1\}\}\bigcup\{\{x\in\mathcal{X} : f(x)=-1\}\times\{1\}\}, \forall f\in\mathcal{F}.$$

1. *On appelle coefficient d'éclatement de $\mathcal{F}$ la quantité :*

$$\mathcal{S}(\mathcal{F},n) = \max_{(x_1,\dots x_n)\in\mathcal{X}^n} N_{\mathcal{A}}(x_1,\dots x_n),$$

où $N_{\mathcal{A}}(x_1,\dots x_n)$ est le nombre d'espaces différents dans $\{\{x_1,\dots x_n\}\cap A, A\in\mathcal{A}\}$.

2. *On appelle dimension de Vapnik-Chervonenkis de $\mathcal{F}$, notée $V_{\mathcal{F}}$ le plus grand entier $k\geq 1$ tel que $\mathcal{S}(\mathcal{F},k)=2^k$.*
Si $\mathcal{S}(\mathcal{F},n)=2^n$, $\forall n$, alors par définition $V_{\mathcal{F}}=\infty$.

Le coefficient d'éclatement mesure la richesse de la classe $\mathcal{F}$. On a clairement $\mathcal{S}(\mathcal{F},n)\leq 2^n$ et on dira que $\mathcal{F}$ éclate $(x_1,\dots x_n)$ si $\mathcal{S}(\mathcal{F},n)=2^n$. La dimension de Vapnik représente le nombre minimal de points que $\mathcal{F}$ ne peut éclater. Cette notion de complexité est centrale dans la théorie de Vapnik.

Théorème 1.1 (Inégalité de Vapnik (1971)) *Soit $\mathcal{F}$ une classe de classifieurs et P une mesure de probabilité sur $\mathcal{X}\times\{-1,+1\}$. Alors pour tout n et $\epsilon>0$, on a :*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|R(f)-R_n(f)|>\epsilon\right)\leq 8\mathcal{S}(\mathcal{F},n)\exp\left(-\frac{n\epsilon^2}{32}\right).$$

La preuve de cette inégalité est présentée dans Devroye et al. [51]. Elle contient des idées essentielles, comme l'argument de symétrisation.

A partir de cette inégalité, on obtient les premières vitesses de convergence de la forme (1.1) dans le cas particulier où $f^*\in\mathcal{F}$. En effet, d'après (1.8), si $f^*=f^*_{\mathcal{F}}$:

$$\mathbb{E}R(\hat{f}_n,f^*)\leq 16\sqrt{\frac{\log(8e\mathcal{S}(\mathcal{F},n))}{2n}}.$$

Ce résultat est non-asymptotique. Cependant il n'est pas toujours satisfaisant. Il dépend de la croissance du coefficient d'éclatement en fonction de n. Si cette croissance est polynomiale, on obtient une vitesse de convergence en $\sqrt{\log n/n}$. Si $\mathcal{S}(\mathcal{F},n)=2^n$ pour tout n, cette inégalité n'a plus aucune valeur. La dimension de Vapnik-Chervonenkis permet donc de distinguer ces deux cas. Si $V_{\mathcal{F}}$ est finie, on obtient :

$$\mathbb{E}R(\hat{f}_n,f^*)\leq 16\sqrt{\frac{V_{\mathcal{F}}\log n+4}{2n}}. \tag{1.9}$$

On peut ajouter une situation où la vitesse de convergence est améliorée. Il s'agit du cas sans bruit où $R(f^*)=0$. Dans ce cas, il existe une version plus fine du Théorème 1.1. On obtient alors une borne de l'excès de risque de la forme :

$$\mathbb{E}R(\hat{f}_n,f^*)\leq\frac{2V_{\mathcal{F}}\log 2n+4}{n}. \tag{1.10}$$

Ces deux résultats traitent deux situations très différentes. La vitesse $n^{-1/2}$ est atteinte sans aucune hypothèse sur la distribution P. C'est donc un résultat universel en P. Par contre, la vitesse n^{-1} est atteinte sous une hypothèse très forte. Il faut s'assurer que $R(f^*)=0$, ce qui correspond au cas sans bruit où $\{x\in\mathcal{X}:\eta(x)\in(0,1)\}$ est de mesure nulle.

Récemment des résultats ont permis d'obtenir une gamme complète de vitesses entre $n^{-1/2}$ et n^{-1}. Ils s'appuient notamment sur une nouvelle hypothèse sur le bruit de classification : l'hypothèse de marge. Cette hypothèse porte sur le comportement de η au niveau $1/2$. Elle sera utilisée dans les chapitres suivants pour obtenir des vitesses de convergence.

Hypothèse de marge

L'hypothèse de marge provient de l'analyse discriminante. L'analyse discriminante de deux échantillons de densités f et g par rapport à une mesure Q repose sur l'estimation de l'ensemble $G = \{x : f(x) \geq g(x)\}$. Mammen and Tsybakov [95] propose des vitesses de convergence rapides (plus rapides que $n^{-1/2}$) vers l'ensemble G, s'il existe un réel $\alpha > 0$ tel que :

$$Q\left(\{x : |f(x) - g(x)| \leq t\}\right) \leq Bt^{\alpha}, \tag{1.11}$$

pour tout t suffisamment petit. Cette hypothèse porte sur le comportement de la différence $f(x) - g(x)$ au voisinage de la frontière de G.
Un problème de classification s'apparente à un problème d'analyse discriminante où l'on cherche à estimer l'ensemble $G^* = \{\eta(x) \geq 1 - \eta(x)\}$. Ainsi (1.11) se formule de la manière suivante en classification.

Définition 1.7 *On dit que P a un paramètre de marge $\alpha > 0$ s'il existe $B > 0$ tel que pour $t \geq 0$ suffisamment petit,*

$$P\left(\{x : |2\eta(x) - 1| \leq t\}\right) \leq Bt^{\alpha}. \tag{1.12}$$

On voit clairement la relation entre (1.12) et le comportement de η au voisinage de $1/2$. En particulier, s'il existe $h > 0$ tel que :

$$|2\eta(x) - 1| \geq h, \tag{1.13}$$

η réalise un saut au niveau $1/2$. Alors $P\left(\{x : |2\eta(x) - 1| \leq t\}\right) = 0$ pour $t \leq h$ et (1.12) est vérifiée quelque soit $\alpha > 0$. (1.13) correspond au meilleur cas possible. On dira que P a un paramètre de marge $\alpha = +\infty$. On parlera parfois d'hypothèse de marge forte.

Vitesses rapides

Tsybakov [136] et Massart and Nédélec [102] ont étudié les performances des ERM dans le cadre minimax, sur une classe de distributions satisfaisant une hypothèse de marge. On obtient des vitesses rapides (entre $n^{-1/2}$ et n^{-1}) qui sont optimales au sens minimax.
Tsybakov [136] considère une classe de distributions $P \in \mathcal{P}(\alpha,\rho)$ telles que :

1. P a un paramètre de marge $\alpha > 0$,
2. $f^* \in \mathcal{F}$ ensemble de classifieurs,
3. $\mathcal{F}$ a une complexité [2] d'indice $\rho \in]0,1[$.

Si on considère l'ERM $\hat{f}_n = \arg\min_{f \in \mathcal{F}} R_n(f)$, on obtient :

$$\sup_{P \in \mathcal{P}(,\alpha,\rho)} \mathbb{E}_{P^n} R(\hat{f}_n, f^*) = O(n^{-\frac{\alpha+1}{\alpha(\rho+1)+2}}), \text{ lorsque } n \to +\infty. \tag{1.14}$$

2. La complexité considérée dans Tsybakov [136] est issue de l'analyse discriminante. Elle porte sur la régularité de la frontière des règles de décision dans $\mathcal{F}$ en terme d'entropie :

$$\mathcal{H}_B(\delta, \mathcal{F}, d_\Delta) \leq A\delta^{-\rho}, \forall 0 < \delta \leq 1,$$

où $\mathcal{H}_B$ est l'entropie avec crochet et $d_\Delta(f,g) = \mathbb{E}(\mathbb{1}(f(X) \neq g(X))$ est une pseudo-distance.

Ce résultat est asymptotique. Il est valable lorsque $n \to +\infty$. Quand $\alpha \to 0$, la vitesse en (1.14) s'approche de $n^{-1/2}$ quelle que soit la complexité ρ du modèle. Lorsque $\alpha \to +\infty$ et $\rho \to 0$, la vitesse s'approche de n^{-1}. On a bien obtenu une gamme de vitesses entre $n^{-1/2}$ et n^{-1}. Tsybakov [136] montre que cette vitesse est optimale au sens minimax. Cela signifie que sous les mêmes hypothèses, aucun classifieur ne peut atteindre une meilleure vitesse pour le risque maximal défini en (1.14).

Massart and Nédélec [102] propose des résultats non-asymptotiques sous les hypothèses suivantes. On considère une classe de distributions $\mathcal{P}(h,V)$ telles que $\forall P \in \mathcal{P}(h,V)$:

1. $|2\eta(x) - 1| \geq h$, $\forall x \in \mathcal{X}$,
2. $f^* \in \mathcal{F}$,
3. $\mathcal{F}$ a une dimension de Vapnik $V_{\mathcal{F}}$ finie.

Dans ce cas, il existe une constante $C > 0$ telle que l'ERM $\hat{f}_n$ sur la classe $\mathcal{F}$ vérifie :

$$\sup_{P \in \mathcal{P}(h,V)} \mathbb{E}_{P^n} R(\hat{f}_n, f^*) \leq C\sqrt{\frac{V}{n}}, \text{ lorsque } h \leq \sqrt{\frac{V}{n}},$$

et de plus,

$$\sup_{P \in \mathcal{P}(h,V)} \mathbb{E}_{P^n} R(\hat{f}_n, f^*) \leq C\frac{V}{nh}\left(1 + \log\frac{nh^2}{V}\right), \text{ lorsque } h > \sqrt{\frac{V}{n}}. \tag{1.15}$$

On distingue deux cas : lorsque la marge h est suffisamment grande, la vitesse est plus rapide que $n^{-1/2}$. Elle est de l'ordre de n^{-1} (à un facteur logarithmique près) lorsque $h > c_0 > 0$. Par contre, on retrouve la vitesse (1.9) lorsque la marge est petite. Massart and Nédélec [102] montre que ces résultats sont optimaux, au facteur logarithmique près.

Ces résultats assurent que le principe de minimisation du risque empirique est optimal au sens minimax, sous certaines hypothèses sur la distribution P. Malgré tout, ils ne sont pas toujours satisfaisants. Tout comme les résultats de Vapnik et Chervonenkis du paragraphe 1.2, les vitesses de convergence (1.14) et (1.15) ont lieu à condition que $f^* \in \mathcal{F}$. Cette hypothèse est très forte et permet d'annuler la quantité $R(f^*_{\mathcal{F}}, f^*)$ appelée erreur d'approximation. Afin de l'éviter, il faut s'assurer que $f^*_{\mathcal{F}}$ est un bon classifieur. De plus (1.14) est un résultat asymptotique. Il a lieu lorsque $n \to +\infty$. Or, en pratique, nous ne disposons que d'un nombre fini d'observations. Enfin, d'un point de vue algorithmique, il n'est pas toujours possible de minimiser le risque empirique (1.7) sur la collection $\mathcal{F}$. C'est pourquoi on introduit des méthodes de régularisation. La majeure partie de cette thèse est dédiée à l'étude de ces méthodes de régularisation.

1.1.3 Régularisation et sélection de modèles

Soit $\mathcal{F}$ un ensemble de classifieurs, n fixé et D_n un ensemble d'apprentissage. Si $\mathcal{F}$ est suffisamment riche, le minimiseur du risque empirique $\hat{f}_n$ vérifie :

$$R_n(\hat{f}_n) = 0. \tag{1.16}$$

Le comportement d'une telle solution est schématisé dans la Figure 1.2. Il tient compte de toutes les observations de l'ensemble d'apprentissage. Par conséquent, il se peut que $\hat{f}_n$ ait un faible pouvoir de généralisation. On dit qu'il *sur-apprend* (phénomène d'*overfitting*).

Il semble plus naturel de considérer le séparateur de la Figure 1.3, où l'on a régularisé la solution initiale. Cette régularisation est liée à la faible connaissance du phénomène observé. Nous ne disposons que d'un nombre fini d'observations et devons donc faire preuve de prudence. Par la suite nous considèrerons plusieurs méthodes de régularisation. Se pose alors le problème suivant : établir un compromis entre fidélité aux données et régularisation.

Minimisation du risque empirique pénalisé

Les méthodes de régularisation sont nombreuses. Pour tenir compte de la complexité de la solution, considérons un espace $\mathcal{F}$ suffisamment riche. On munit $\mathcal{F}$ d'une mesure de complexité $\mathcal{C}(f)$, pour tout $f \in \mathcal{F}$. Les procédures d'ERM pénalisé s'écrivent :

$$\min_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(X_i)) + \alpha_n \mathcal{C}(f) \right]. \tag{1.17}$$

On minimise un critère empirique en tenant compte de la complexité de la solution. Ainsi, on empêche la sélection de $\hat{f}_n$ vérifiant (1.16) puisqu'en général $\mathcal{C}(\hat{f}_n)$ est très grand. Le paramètre α_n est appelé paramètre de régularisation.

Exemple 1.1 (SVM) *Le Chapitre 2 étudie les performances statistiques des SVM. Présenté dans la section suivante, cet algorithme peut s'exprimer comme un ERM pénalisé de la façon suivante :*

$$\min_{f\in\mathcal{H}_K} \left(\frac{1}{n}\sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \alpha_n \|f\|_K^2 \right), \tag{1.18}$$

où $\mathcal{H}_K$ est un espace de Hilbert à noyau reproduisant (voir Définition 1.12). Le critère empirique utilise la perte douce. La complexité est mesurée par la norme dans l'espace $\mathcal{H}_K$. On peut montrer que la solution de (1.18) *est unique et s'écrit :*

$$\hat{f}_n = \sum_{i=1}^{n} v_i^* Y_i K(X_i, \cdot). \tag{1.19}$$

La perte douce utilisée dans (1.18) *entraîne de la parcimonie par rapport aux observations. Cela signifie que dans* (1.19)*, peu de coefficients v_i^* sont non nuls. Les observations X_i associées aux $v_i^* \neq 0$ sont appelées vecteurs supports.*

Exemple 1.2 (Minimisation de type LASSO) *L'algorithme du LASSO a été introduit par Tibshirani [128] dans le modèle de régression. L'idée est de pénaliser la solution par une quantité proportionnelle à la norme l^1 de la solution. Dans le cadre de la classification, on considère un dictionnaire $\mathcal{F}_M = \{f_1, \ldots, f_M\}$. On cherche une solution de la forme :*

$$f_\lambda = \sum_{k=1}^{M} \lambda_k f_k, \lambda \in \mathbb{R}^M. \tag{1.20}$$

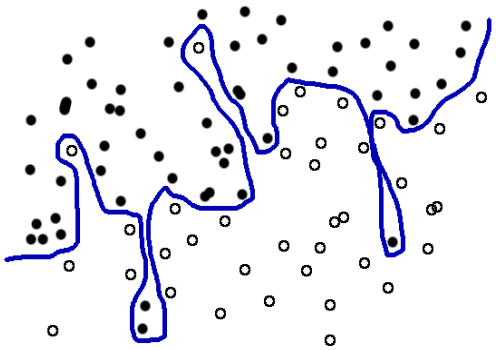

Figure 1.2 : Phénomène de sur-apprentissage (overfitting).

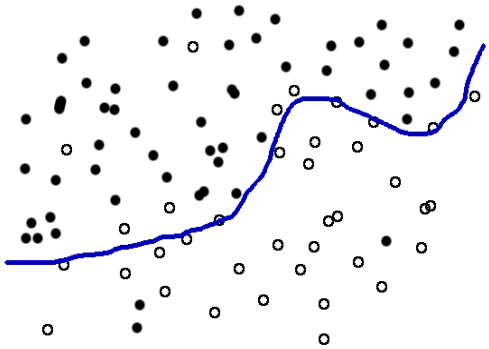

Figure 1.3 : Solution régularisée.

La procédure LASSO s'écrit alors :

$$\min_{\lambda\in\mathbb{R}^M}\left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i,f_\lambda(X_i)) + \alpha_n\|\lambda\|_{l^1}\right),$$

où l est une fonction de perte et $\|\lambda\|_{l^1} = \sum_{k=1}^{M}|\lambda_k|$. Cette pénalité génère une solution parcimonieuse en λ. Cela signifie que le nombre de coefficients non nuls dans (1.20) *est faible.*

Sélection de modèles pénalisée

On peut formuler le problème de régularisation différemment. L'équation (1.16) a lieu si la classe de fonctions $\mathcal{F}$ est suffisamment riche. Le problème revient à choisir la taille de $\mathcal{F}$ de manière optimale. Pour cela, on peut considérer $(\mathcal{F}_m)_{m\geq 1}$ une famille de modèles tels que $\mathcal{F}_m \subset \mathcal{F}_{m+1}$ pour tout m et $\bigcup \mathcal{F}_m = \mathcal{F}$ est une classe très riche. On note $\hat{f}_m$ l'ERM sur $\mathcal{F}_m$ et $f_m^* = \arg\min_{f\in\mathcal{F}_m}\mathbb{E}l(Y,f(X))$ l'oracle sur $\mathcal{F}_m$. Ainsi, l'excès de risque d'un classifieur $\hat{f}_m$ s'écrit :

$$R(\hat{f}_m,f^*) = R(\hat{f}_m,f_m^*) + R(f_m^*,f^*). \tag{1.21}$$

Le premier terme de (1.21) est appelé erreur d'estimation. Il mesure la qualité de $\hat{f}_m$ par rapport à celle de l'oracle. Cette erreur est due à l'utilisation d'un nombre fini d'observations. Elle est croissante avec la complexité du modèle $\mathcal{F}_m$. Le deuxième terme est l'erreur d'approximation. Elle mesure l'excès de risque de l'oracle f_m^*. Ces deux erreurs sont antagonistes. Un modèle $\mathcal{F}_m$ trop riche entraîne une erreur d'estimation trop grande alors qu'un modèle trop petit n'approcherait pas f^*. Le choix de m doit donc établir l'équilibre entre ces deux termes dans (1.21).
Une manière assez classique de sélectionner m est la sélection de modèles pénalisée. Le principe est le suivant : on introduit une pénalité qui va mesurer la complexité du modèle $\mathcal{F}_m$. Le choix optimal de m est alors :

$$\hat{m} = \arg\min_{m\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i,\hat{f}_m(X_i)) + \text{pen}(m)\right), \tag{1.22}$$

où $pen(m)$ est une fonction croissante de m.

Exemple 1.3 (Kernel Projection Machines) *On considère un espace de Hilbert $\mathcal{H}$ et $(\phi_k)_{k\geq 1}$ une base orthonormale de $\mathcal{H}$. On pose $\mathcal{F}_N = vect < \phi_1,\ldots,\phi_N >$ le sous-espace vectoriel de $\mathcal{H}$ de dimension N. On définit, pour tout entier $N \geq 1$:*

$$\hat{f}_N = \arg\min_{f\in\mathcal{F}_N}\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+. \tag{1.23}$$

Blanchard and Zwald [27] montre que le choix optimal de N s'écrit :

$$\widehat{N} = \arg\min_{N\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\hat{f}_N(X_i))_+ + \lambda N\right),$$

où λ est une constante à calibrer. Pour implémenter l'algorithme, Blanchard and Zwald [27] propose de considérer $\mathcal{H} = \mathcal{H}_K$ un espace de Hilbert à noyau reproduisant et $(\phi_k)_{k\geq 1}$ la base de fonctions propres de l'opérateur de covariance empirique $C_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)K(X_i,\cdot)$. Le Chapitre 4 propose d'étudier une méthode de sélection de modèles pénalisée pour choisir N de manière automatique. La famille de KPM utilisée sera de la forme (1.23) *où la perte douce est remplacée par la perte des moindres carrés $l(y,f(x)) = (y - f(x))^2$.*

Exemple 1.4 (SVM) *On peut interpréter la minimisation* (1.18) *comme une procédure de sélection de modèles pénalisée. Pour cela, on définit, pour $R > 0$, la collection $B_K(R) = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$. La solution $\hat{f}_n$ de* (1.18) *peut s'écrire $\hat{f}_n = \hat{f}_{\hat{R}}$ où on note :*

$$\hat{f}_R = \arg\min_{f\in B_K(R)} \frac{1}{n}\sum_{i=1}^n (1 - Y_i f(X_i))_+ \text{ et } \hat{R} = \arg\min_{R>0}\left(\frac{1}{n}\sum_{i=1}^n (1 - Y_i \hat{f}_R(X_i))_+ + \alpha_n R^2\right).$$

Cette représentation peut se généraliser à toute procédure d'ERM pénalisé, comme l'ERM pénalisé sur les espaces de Besov (Chapitre 3). Dans ce cas la collection de modèles sera un ensemble de boules dans l'espace de Besov.

Le problème de l'adaptation

L'introduction des méthodes de régularisation permet d'éviter le phénomène de surapprentissage des ERM. Cependant, il reste à traiter le problème suivant : comment régulariser la solution de manière automatique ?

Ce problème est appelé en statistique le problème de l'adaptation. A ce jour, les méthodes adaptatives pour choisir des paramètres de régularisation jouent un rôle très important en statistique théorique et appliquée. Les premiers résultats sont sans doute dus à H. Akaike (Akaike [2]) et C.L. Mallows (Mallows [93]). La littérature actuelle est très vaste à ce sujet. Dans la présentation des deux méthodes de régularisation ci-dessus, cet enjeu apparaît de la manière suivante :

- Comment choisir le paramètre de régularisation α_n dans (1.17) ?
- Comment définir la pénalité $pen(m)$ dans (1.22) ?

Un des axes principaux de cette thèse est l'utilisation de procédures adaptatives pour les méthodes de régularisation. D'un point de vue pratique, cela revient à rechercher des algorithmes sans hyper-paramètre à calibrer.

1.2 Machines à vecteurs de support (SVM)

La première formulation de l'algorithme SVM (Support Vector Machines) est due à Boser, Guyon et Vapnik (Boser et al. [28]) en 1992. L'idée principale est de maximiser la marge. Cet algorithme est à l'origine de fructueuses investigations, dans tous les domaines de l'apprentissage. En théorie statistique de l'apprentissage, on s'intéresse depuis quelques années à l'étude théorique des SVM. Les chapitres 2 et 3 s'inscrivent dans cette problématique. Ce paragraphe propose de décrire l'algorithme et de présenter le formalisme mathématique sous-jacent.

1.2.1 Description dans le cas linéaire

On suppose que $\mathcal{X} \subset \mathbb{R}^d$ est muni du produit scalaire usuel : $< x,y >= \sum_{i=1}^d x_i y_i, \forall x,y \in \mathcal{X}$.

Définition 1.8 *On dit qu'un ensemble d'apprentissage $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$ est linéairement séparable s'il existe $f_{w,b}(x) =< w,x > +b$ tel que $\forall i = 1,\dots,n$:*

$$Y_i f_{w,b}(X_i) \geq 0.$$

D_n est linéairement séparable s'il existe un hyperplan [3] (w,b) qui sépare les deux classes. Le classifieur $f_{w,b}$ est appelé classifieur linéaire et vérifie $R_n(f_{w,b}) = 0$.
Historiquement, le premier algorithme itératif produisant un classifieur dans le cadre linéaire est le perceptron (Rosenblatt [115]). Sa construction est relativement simple. On initialise un vecteur directeur w_0 (souvent $w_0 = 0$) et à chaque point (X_i,Y_i) mal classé, on met à jour le vecteur directeur de façon à ce que $Y_i f(X_i) \geq 0$.

Dans ce cadre la première version de l'algorithme SVM peut être vue comme un raffinement du perceptron, ou encore comme le perceptron optimal. Pour cela, on définit la marge d'un classifieur linéaire de la façon suivante :

Définition 1.9 *La marge d'un classifieur linéaire $f_{w,b}$ sur un ensemble d'apprentissage D_n est définie par la quantité :*

$$\gamma_{w,b} = \min_{i=1,\dots,n} Y_i f_{w,b}(X_i). \tag{1.24}$$

On s'intéresse au cas où $\gamma_{w,b} > 0$. Dans ce cas, $R_n(f_{w,b}) = 0$ et (1.24) mesure la distance entre l'hyperplan et l'ensemble d'apprentissage D_n.
La marge permet de définir la notion de classifieur optimal : il s'agit de l'hyperplan de marge maximale. L'idée pionnière des SVM est de maximiser la marge. Par conséquent, dans la littérature francophone, on appelle parfois le SVM le Séparateur à Vaste Marge. La Figure 1.4 nous montre un tel hyperplan dans le cas bi-dimensionnel. Intuitivement, le principe du classifieur de marge optimale consiste à se placer à mi-chemin entre les deux classes.
Si on considère un ensemble d'apprentissage linéairement séparable, le classifieur SVM f_{w^*,b^*} est défini par :

$$(w^*,b^*) = \arg \max_{w,b:\|w\|=1} \gamma_{w,b}.$$

En faisant le changement de variable $w' = \frac{w}{\gamma_{w,b}}$, on obtient :

$$\Leftrightarrow \begin{cases} \min_{w,b} \|w\| \\ \forall\, i = 1,..,n \; Y_i(< w,X_i > +b_0) \geq 1. \end{cases} \tag{1.25}$$

Cette écriture sera utile par la suite.

En pratique, il est assez rare que les données soient linéairement séparables dans le sens de la Définition 1.8. Un raffinement consiste alors à considérer des variables ressorts ξ_i définies de la façon suivante :

$$\xi_i = (\gamma_{w,b} - Y_i f_{w,b}(X_i))_+,$$

3. Dans toute la suite, on notera (w,b) l'hyperplan de vecteur normal w (normalisé à 1) défini par $\{x \in \mathcal{X} :< w,x > +b = 0\}$. Le classifieur $f_{w,b}$ consiste à séparer le domaine $\mathcal{X}$ en deux classes.

où $(x)_+$ représente la partie positive de $x \in \mathbb{R}$. Cette quantité est liée à la notion de marge. Elle mesure la localisation d'un couple (X_i,Y_i) par rapport à la marge du classifieur $f_{w,b}$. On distingue trois cas :

- si $\xi_i > \gamma_{w,b}$ alors le couple (X_i,Y_i) est mal classé par $f_{w,b}$;
- si $0 < \xi_i < \gamma_{w,b}$ alors le couple est bien classé mais proche de l'hyperplan (la distance à l'hyperplan est inférieure à la marge);
- si $\xi_i = 0$ alors le couple se trouve dans sa classe majoritaire (la distance à l'hyperplan est supérieure à la marge).

Dans la Figure 1.5, on a représenté ces variables ressorts.
L'introduction de ces variables va permettre de relâcher les contraintes sur la fidélité aux données. On s'autorise quelques points "intrus", c'est-à-dire des points que l'on étiquette $+1$ dans une région à forte concentration de -1 (et inversement). C'est un moyen de régulariser la solution. Dans l'algorithme, on ne doit plus seulement maximiser la marge, mais aussi minimiser ces variables ressorts. On obtient, à partir de (1.25), la minimisation suivante :

$$\Leftrightarrow \begin{cases} \min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i \right) \\ \\ \forall\, i = 1,..,n \;\; Y_i f_{w,b}(X_i) \geq 1 - \xi_i \,,\; \xi_i \geq 0, \end{cases} \tag{1.26}$$

où on a remplacé la norme $\|w\|$ dans (1.25) par son carré pour des raisons algorithmiques. Ce problème d'optimisation convexe admet une solution unique communément appelée SVM à marge molle (Soft-margin SVM). (1.26) est la formulation primale du SVM. Le paramètre C est un paramètre à calibrer. Il permet d'ajuster l'influence des variables ressorts dans la construction de l'hyperplan. Il correspond en statistique au paramètre de régularisation. L'introduction des variables ressorts est due à Cortes and Vapnik [42]. Cette avancée permet de définir les SVM comme une méthode de régularisation. A partir de (1.26), on peut en particulier interpréter les SVM comme une procédure de sélection de modèles pénalisée (voir Exemple 1.4 ou Blanchard et al. [24] pour les détails).

On peut exprimer ce problème d'optimisation de manière équivalente, dite duale, à l'aide des multiplicateurs de Lagrange :

$$\max_{\alpha:0\leq\alpha_i\leq C} L_D = \max_{\alpha:0\leq\alpha_i\leq C} \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} Y_i Y_{i'} < X_i, X_{i'} > \right), \tag{1.27}$$

sous la contrainte linéaire suivante :

$$\sum_{i=1}^n \alpha_i Y_i = 0.$$

On obtient une solution unique (w^*,b^*) où w^* s'écrit :

$$w^* = \sum_{i=1}^n \alpha_i^* Y_i X_i,$$

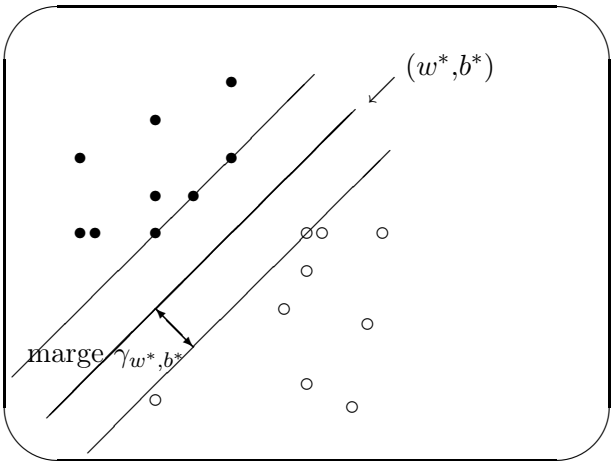

Figure 1.4 : Cas linéairement séparable.

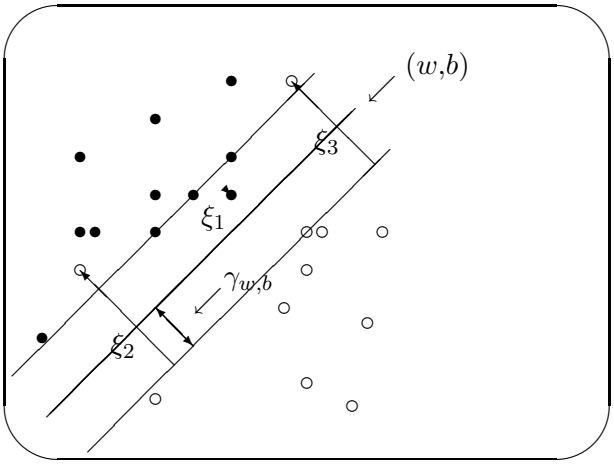

Figure 1.5 : Cas non séparable.

avec α^* solution de (1.27). Ainsi f_{w^*,b^*} peut s'écrire sous la forme suivante :

$$f_{w^*,b^*}(x) =< w^*,x > +b^* = \sum_{i=1}^{n} \alpha_i^* Y_i < X_i,x > +b^*. \tag{1.28}$$

Le vecteur $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ propose une autre représentation de la solution. La représentation (1.28) est dite représentation duale du classifieur. La principale caractéristique de ce vecteur est la suivante : un grand nombre de ces coefficients sont nuls. Le vecteur α^* est dit *parcimonieux* (*sparse*). Cette parcimonie assure qu'une quantité relativement faible des données sera utilisée par l'algorithme. D'un point de vue géométrique, les coefficients non nuls α_i^* correspondent aux points $X_i \in \mathcal{X}$ qui sont situés exactement sur la marge. Dans la Figure 1.4, on en compte 4. Ces observations sont appelées *vecteurs supports*. Ils caractérisent le vecteur normal w^*. L'influence des autres points dans la construction de w^* est nulle. Cette propriété de parcimonie est caractéristique des SVM. Elle entraîne des avantages algorithmiques importants, notamment pour traiter des masses de données. C'est sans doute aujourd'hui la raison principale de leur popularité. Cette thèse n'aborde pas le caractère parcimonieux des SVM. Bartlett and Tewari [19] relie la parcimonie des SVM à un problème purement statistique.

La représentation duale est souvent utilisée pour implémenter les SVM. Le principal avantage est que la dimension de l'espace $\mathcal{X}$ n'intervient pas dans l'expression de α^*. Les points de l'ensemble d'apprentissage n'apparaissent dans (1.27) que par l'intermédiaire du produit scalaire dans $\mathcal{X}$. Cette représentation est souvent utilisée pour résoudre des problèmes non-linéaires, par l'intermédiaire d'un noyau. Cependant il existe des programmes qui résolvent directement le problème primal (Chapelle [40]).

1.2.2 Cas non-linéaire : la version noyau des SVM

En pratique, un problème de classification n'est pas nécessairement linéairement séparable. En effet, il existe de nombreux cas où la solution de (1.26) n'est pas satisfaisante. Cela correspond à des problèmes non-linéaires, c'est-à-dire où un hyperplan ne suffit pas à séparer les données (Figure 1.6). Malgré tout, les méthodes à noyaux vont permettre d'utiliser l'algorithme issu du cadre linéaire dans un cadre non-linéaire.

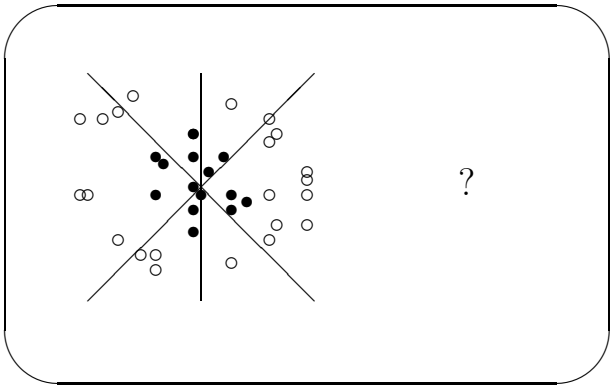

Figure 1.6 : Cas non-linéaire.

Le principe des méthodes à noyaux est le suivant : on va plonger les données X_i, $i = 1, \dots, n$ à l'aide d'une application Φ dans un ensemble $\Phi(\mathcal{X})$ tel que dim $\Phi(\mathcal{X}) >$ dim $\mathcal{X}$. L'intérêt d'une telle application est de disperser les observations de manière à obtenir des données $(\Phi(X_i),Y_i), i = 1, \dots, n$ linéairement séparables dans $\Phi(\mathcal{X})$. On peut ainsi appliquer une méthode linéaire dans l'espace $\Phi(\mathcal{X})$. D'un point de vue pratique, il ne sera pas nécessaire de définir explicitement l'application Φ, ni même l'espace d'arrivée $\Phi(\mathcal{X})$ sur lequel on plonge nos observations. Pour un panorama complet des méthodes à noyaux, on peut citer Schölkopf and Smola [117]. Dans cette section on se restreint à la version noyau de l'algorithme SVM.

Pour cela, on peut se placer dans le dual (1.27). Dans cette représentation, les points $X_i \in \mathcal{X}$ n'interviennent que sous la forme du produit scalaire $< X_i,X_j >$ dans $\mathbb{R}^d$. On peut ainsi résoudre ce programme d'optimisation dans l'espace $\Phi(\mathcal{X})$ en remplaçant le produit scalaire de $\mathbb{R}^d$ par le produit scalaire dans $\Phi(\mathcal{X})$, en munissant au préalable $\Phi(\mathcal{X})$ d'un produit scalaire idoine. Il reste à trouver un moyen de calculer le produit scalaire $< \Phi(x),\Phi(y) >_{\Phi(\mathcal{X})}$ pour $x,y \in \mathcal{X}$. Ce calcul est assuré par une application K appelée noyau. De manière générale, celui-ci est défini de la façon suivante :

Définition 1.10 *On appelle noyau une application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que $\forall x,y \in \mathcal{X}$:*

$$K(x,y) =< \Phi(x),\Phi(y) >_{\Phi(\mathcal{X})} . \tag{1.29}$$

La relation (1.29) est souvent appelée *kernel trick* en apprentissage. C'est la pierre angulaire des méthodes à noyaux. Elle permet de définir implicitement, à partir d'un noyau K, un plongement des données X_i de $\mathcal{X}$ dans un espace $\Phi(\mathcal{X})$ et ainsi de résoudre des problèmes non-linéaires dans $\mathcal{X}$. L'utilisation du kernel trick est très attractive. Au lieu de définir une application Φ et un espace d'arrivée $\Phi(\mathcal{X})$, on définira en pratique directement un noyau sur l'espace $\mathcal{X}$. La principale difficulté de l'utilisateur de l'algorithme est le choix du noyau. Le Chapitre 2 propose une comparaison expérimentale des SVM utilisant deux noyaux différents.
Si on remplace dans (1.27) le produit scalaire dans $\mathcal{X}$ par l'évaluation du noyau K, on obtient une solution qui s'écrit :

$$f_{svm}(x) = \sum_{i=1}^{n} \alpha_i^* Y_i K(X_i,x) + b^*, \tag{1.30}$$

où α^* est solution de la version noyau de (1.27) et b^* est obtenu à partir des contraintes linéaires. Par la suite, on introduira une version simplifiée du SVM (1.30) où l'on fixe la constante $b^* = 0$. Cela correspond à n'utiliser que les hyperplans contenant l'origine. Cette simplification nous permettra d'éliminer certaines difficultés techniques.

Le paragraphe suivant propose un formalisme mathématique où l'espace $\Phi(\mathcal{X})$ et le noyau K vérifient l'égalité (1.29). Ce formalisme sera souvent considéré dans cette thèse. Il est très largement utilisé dans la communauté des méthodes à noyaux.

1.2.3 Espace de Hilbert à noyau reproduisant (EHNR)

Pour obtenir une application K vérifiant (1.29), la propriété essentielle est la positivité du noyau. Elle est nécessaire à la construction d'un produit scalaire.

Définition 1.11 *Une application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est dite définie positive (respectivement*

strictement définie positive) si $\forall N \geq 1$, $\forall a_1, \dots a_N \in \mathcal{X}$, $\forall x_1, \dots x_N \in \mathcal{X}$,

$$\sum_{i,j=1}^{N} a_i a_j K(x_i,x_j) \geq 0 \text{ (respectivement } >0 \text{).}$$

De manière équivalente, $K_N = (K(x_i,x_j))_{i,j=1,\dots,N}$ est définie positive, $\forall x_1, \dots, x_N \in \mathcal{X}$. La matrice K_N est appelée *matrice à noyau.*

Définition 1.12 *Soit $\mathcal{H}$ un espace de Hilbert de fonctions $f : \mathcal{X} \to \mathbb{R}$ définies point par point (on notera par la suite $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$). On muni cet espace d'un produit scalaire noté $< \cdot,\cdot >_{\mathcal{H}}$. Alors $\mathcal{H}$ est appelé espace de Hilbert à noyau reproduisant (EHNR) s'il existe une application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ telle que :*

1. $\forall x \in \mathcal{X}$, $K(x,\cdot) : y \mapsto K(x,y) \in \mathcal{H}$,
2. $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$,

$$f(x) =< f,K(x,\cdot) >_{\mathcal{H}} . \tag{1.31}$$

La propriété (1.31) *est la propriété reproduisante. La fonction K est appelée noyau reproduisant et on dira que $\mathcal{H}$ est l'EHNR de noyau K. Il sera noté $\mathcal{H}_K$.*

L'EHNR $\mathcal{H}_K$ est alors la complétion de l'espace vectoriel engendré par les fonctions $K(x,\cdot)$, $x \in \mathcal{X}$, par rapport à la norme :

$$\left\| \sum_{i=1}^{N} a_i K(x_i,\cdot) \right\|_K^2 = \sum_{i,j=1}^{N} a_i a_j K(x_i,x_j).$$

Cette norme est engendrée par le produit scalaire suivant :

$$< f,g >_K =< \sum_{i=1}^{N} a_i K(x_i,\cdot), \sum_{i=1}^{N} b_i K(y_i,\cdot) >_K = \sum_{i,j=1}^{N} a_i b_j K(x_i,y_j). \tag{1.32}$$

La représentation (1.32) est la représentation primale de l'EHNR $\mathcal{H}_K$.
La propriété reproduisante est la propriété principale des EHNR. Le théorème suivant permet d'obtenir une équivalence en terme de continuité d'une application.

Théorème 1.2 (Théorème de représentation de Riesz) *Soit $\mathcal{H}$ un espace de Hilbert muni de son produit scalaire $< \cdot,\cdot >$. Soit δ une forme linéaire continue sur $\mathcal{H}$. Il existe alors un unique $g \in \mathcal{H}$ tel que :*

$$< g,f >= \delta(f), \forall f \in \mathcal{H}.$$

On peut alors définir un EHNR à l'aide de la continuité de la fonction d'évaluation.

Proposition 1.1 *Soit $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ un espace de Hilbert. Les deux assertions suivantes sont équivalentes :*

1. *$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau reproduisant générant $\mathcal{H}$ comme EHNR.*
2. *$\forall y \in \mathcal{X}$, la fonction d'évaluation $\delta_y : f \mapsto f(y)$ définie de $\mathcal{H}$ dans $\mathbb{R}$ est continue.*

Preuve
$\Rightarrow \delta_y$ est linéaire, donc la continuité est équivalente à la continuité en 0. Soit $f \in \mathcal{H}$. Alors $\forall y \in \mathcal{X}$, par Cauchy-Schwarz :

$$|f(y)| = | < f,K(y,\cdot) >_{\mathcal{H}} | \leq K(y,y)^{1/2}\|f\|_{\mathcal{H}}.$$

$\Leftarrow$ On applique le théorème de représentation de Riesz. Puisque δ_y est linéaire continue sur l'espace de Hilbert $\mathcal{H}$, il existe $\rho_y \in \mathcal{H}$ telle que :

$$\delta_y(f) = f(y) =< f,\rho_y >_{\mathcal{H}} .$$

En posant $\rho_y = K(y,\cdot)$, on obtient le résultat où $\mathcal{H}$ est l'EHNR de noyau reproduisant K. □

Théorème 1.3 *Une application $K : \mathcal{X}^2 \to \mathbb{R}$ est symétrique définie positive si et seulement si il existe un unique EHNR $\mathcal{H}_K$ de noyau reproduisant K.*

Ce théorème est dû à Aronszajn [5]. Le cas $\mathcal{X}$ compact et K continue remonte à Mercer (1909).

Cette notion d'espace à noyau permet d'obtenir explicitement une méthode à noyau de la manière suivante. Etant donnée une application $K : \mathcal{X}^2 \to \mathbb{R}$ symétrique définie positive, on pose $\Phi : \mathcal{X} \to \mathcal{H}_K$ telle que :

$$\Phi(x) = K(x,\cdot). \tag{1.33}$$

Alors Φ est bien définie et d'après la propriété reproduisante :

$$< \Phi(x),\Phi(y) >_{\Phi(\mathcal{X})} =< K(x,\cdot),K(y,\cdot) >_{\mathcal{H}_k} = K(x,y), \forall x,y \in \mathcal{X}.$$

K est bien un noyau au sens de la Définition 1.10. En vertu du Théorème 1.3, le terme de noyau décrit souvent une application symétrique définie positive. Cette terminologie suppose qu'on se restreint au cadre des EHNR. On discutera par la suite l'extension de ces résultats à un cadre non-hilbertien.

1.3 Théorie fonctionnelle des noyaux

Une étude statistique des méthodes à noyaux nécessite un regard précis sur les espaces à noyaux. Ce paragraphe propose une vision fonctionnelle de la théorie des noyaux. On cherche à expliquer la régularité des fonctions composant ces espaces. Ceci permet notamment de mieux comprendre l'influence du noyau dans la régularisation des méthodes à noyaux.

1.3.1 Noyaux de Mercer et ellipsoïdes

Théorie de Mercer

On considère un noyau K générant un EHNR $\mathcal{H}_K$. Dans ce paragraphe, le terme noyau décrit une application symétrique définie positive. La théorie de Mercer s'intéresse à une classe particulière de noyaux.

Définition 1.13 *1. Un noyau $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est appelé noyau de Mercer si les hypothèses suivantes sont vérifiées :*

- *$\mathcal{X}$ est compact,*

- *K est continue comme application définie sur $\mathcal{X} \times \mathcal{X}$.*

2. *On appelle opérateur intégral de noyau K l'opérateur $L_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ défini par :*

$$L_K(f)(x) = \int_{\mathcal{X}} K(x,y) f(y) d\mu(y),$$

où μ mesure de Borel quelconque sur $\mathcal{X}$.

L'opérateur intégral est l'opérateur associé au noyau K. Si K est un noyau de Mercer, il existe une constante C_K telle que :

$$\sup_{x,y \in \mathcal{X}} |K(x,y)| \leq C_K.$$

Alors l'opérateur L_K est bien défini. De plus, on peut montrer que L_K est linéaire et borné (par continuité de K sur $\mathcal{X}$ compact), auto-adjoint et positif (par symétrie et positivité de K), et compact. La compacité de l'opérateur associé à un noyau de Mercer est essentielle dans la théorie de Mercer. D'après le théorème spectral (Théorème 1.5), il existe :

1. $(\phi_k)_{k \geq 1}$ base orthonormée de $L^2_\mu(\mathcal{X})$ composée des fonctions propres de L_K,
2. $(\lambda_k)_{k \geq 1}$ valeurs propres de L_K positives ou nulles vérifiant $\text{card}(\{k : \lambda_k > 0\})$ est finie ou bien $\lambda_k \to 0$, lorsque $k \to +\infty$.

Le théorème suivant est le principal résultat concernant la théorie de Mercer.

Théorème 1.4 (Mercer, 1909) *Soit $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ noyau de Mercer. Soit $(\lambda_k)_{k \geq 1}$ les valeurs propres de L_K et $(\phi_k)_{k \geq 1}$ les fonctions propres de L_K. Alors, $\forall x,y \in \mathcal{X}$:*

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_k \phi_k(x) \phi_k(y),$$

où la convergence est absolue (pour tout couple $(x,y) \in \mathcal{X} \times \mathcal{X}$) et uniforme sur $\mathcal{X} \times \mathcal{X}$.

Ce théorème permet de définir l'EHNR $\mathcal{H}_K$ à partir des valeurs propres de l'opérateur intégral de la manière suivante :

Corollaire 1.1 *Soit K un noyau de Mercer et $\mathcal{H}_K$ l'EHNR associé. Alors :*

$$\mathcal{H}_K = \left\{ f \in L^2_\mu(\mathcal{X}) : f = \sum_{k=1}^{\infty} a_k \phi_k \text{ vérifie } \left(\frac{a_k}{\sqrt{\lambda_k}} \right)_{k \geq 1} \in l^2(\mathbb{N}) \right\}, \tag{1.34}$$

et le produit scalaire $< \cdot, \cdot >_K$ s'écrit :

$$< \sum_{k=1}^{\infty} a_k \phi_k, \sum_{k=1}^{\infty} b_k \phi_k >_K = \sum_{k=1}^{\infty} \frac{a_k b_k}{\lambda_k}.$$

La représentation (1.34) est appelée représentation spectrale de $\mathcal{H}_K$. Elle est indépendante de la mesure μ. L'EHNR s'écrit comme une ellipsoïde dont la norme dépend des valeurs propres de L_K. Ce corollaire permet d'exprimer la régularité d'une fonction $f \in \mathcal{H}_K$ en terme du comportement asymptotique de ses coefficients dans la base $(\phi_k)_{k \geq 1}$. On donne ci-dessous un exemple de noyau K générant la base de Fourier dans le domaine spectral. Cela entraîne une représentation des EHNR dans le domaine de Fourier.

Noyaux de convolution et analyse de Fourier

On introduit un type de noyau très souvent utilisé en théorie statistique de l'apprentissage : les noyaux de convolution [4]. Dans la théorie de Mercer, ils correspondent à un choix particulier de base orthonormale $(\phi_k)_{k\geq 1}$: la base de Fourier.

Définition 1.14 *Un noyau K est appelé noyau de convolution s'il existe une fonction $\phi : \mathcal{X} \to \mathbb{R}$ telle que :*

$$\forall x,y \in \mathcal{X}, K(x,y) = \phi(x-y).$$

La fonction ϕ est appelée fonction de base du noyau K.
Si on considère un noyau de convolution de fonction de base ϕ, il est clair que l'opérateur intégral est l'opérateur de convolution circulaire suivant :

$$L_K(f) = \int_{\mathcal{X}} \phi(x-y)f(y)d\mu(y) = \phi * f.$$

Alors d'après le théorème spectral, si $\mathcal{X}$ est compact :

1. ses fonctions propres $(e_k)_{k\geq 1}$ correspondent à la base de Fourier des fonctions périodiques dans $L^2(\mathcal{X})$.
2. ses valeurs propres $(\hat{\phi}_k)_{k\geq 1}$ correspondent aux coefficients de Fourier de la fonction ϕ.

D'après le Corollaire 1.1, l'EHNR de noyau reproduisant K noyau de Mercer de convolution s'écrit :

$$\mathcal{H}_K = \{f \in L^2(\mathcal{X}) : \frac{\hat{f}_k}{\sqrt{\hat{\phi}_k}} \in l^2(\mathbb{N})\},$$

où $\hat{f}_k = \int_{\mathcal{X}} f(x)e_k(x)dx$ sont les coefficients de Fourier de f. Dans ce cas, la régularité s'exprime en terme de décroissance des coefficients de Fourier. Ce critère de régularité est classique en analyse. L'exemple suivant illustre le cas particulier des espaces de Sobolev périodiques.

Exemple 1.5 *On considère, pour $m \in \mathbb{N}$, l'espace de Sobolev $\mathcal{W}_m([0,1])$ défini par :*

$$\mathcal{W}_m([0,1]) = \{f : f^{(k)} \text{ abs. cont. } \forall k = 0,\dots,m-1 \text{ et } f^{(m)} \in L^2([0,1])\},$$

où $f^{(k)}$ est la $k^{ième}$ dérivée au sens des distributions. On définit $\mathcal{W}_m(per)$ le sous-ensemble de $\mathcal{W}_m([0,1])$ des fonctions vérifiant :

$$\forall k = 0,\dots m-1, f^{(k)}(1) = f^{(k)}(0).$$

Ainsi on peut étendre toute fonction de $\mathcal{W}_m(per)$ à une fonction continue, périodique sur $\mathbb{R}$. On munit cet espace de la norme :

$$\|f\|^2 = \left(\int_0^1 f(u)du\right)^2 + \int_0^1 f^{(m)}(u)^2du.$$

4. ou noyaux invariant par translation, ou encore noyaux RBF (Radial Basis Function).

Alors, on peut montrer que $\mathcal{W}_m(per)$ est un EHNR de noyau reproduisant K défini par :

$$K(x,y) = 1 + \sum_{k=1}^{\infty} \frac{2}{(2k\pi)^{2m}} \cos(2k\pi(x-y)).$$

Les valeurs propres de l'opérateur intégral sont $\lambda_0 = 1$ (avec multiplicité 1) et $\lambda_k = \frac{1}{(2\pi k)^{2m}}$, pour $k \geq 1$ (avec multiplicité 2). Les fonctions propres sont données par la suite :

$$(\phi_k(x), k \geq 0) = (1, \sqrt{2}\sin 2\pi x, \sqrt{2}\cos 2\pi x, \sqrt{2}\sin 4\pi x, \ldots, \sqrt{2}\sin 2k\pi x, \sqrt{2}\cos 2k\pi x, \ldots),$$

correspondant à la base trigonométrique de $L^2([0,1])$. D'après le Corollaire 1.1, on peut écrire :

$$\mathcal{W}_m(per) = \left\{ f(x) = a_0 + \sum_{k=1}^{\infty} a_k \cos(2k\pi x) + \sum_{k=1}^{\infty} b_k \sin(2k\pi x) : \sum_{k=1}^{\infty} (a_k^2 + b_k^2)((2k\pi)^{2m}) < \infty \right\}.$$

On obtient la représentation d'un espace de Sobolev sous la forme d'une ellipsoïde.

1.3.2 Généralisation au cas non compact

La théorie de Mercer se restreint au cas où le noyau K est défini sur un ensemble $\mathcal{X}$ compact. Cela implique la compacité de l'opérateur associé au noyau, assurant l'existence d'une base orthonormale $(\phi_k)_{k\geq 1}$ de fonctions propres de L_K et un spectre discret $(\lambda_k)_{k\geq 1}$. Le théorème spectral permet d'obtenir un résultat plus général, pour des opérateurs linéaires bornés auto-adjoints quelconques.

Théorème 1.5 (Théorème spectral (Halmos [67])) *Soit $A : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ un opérateur linéaire borné et auto-adjoint. Alors il existe un espace mesuré (S,ν) et $U : L^2_\mu(\mathcal{X}) \to L^2_\nu(S)$ isométrique telle que :*

$$UAU^{-1} = M_\rho,$$

où $M_\rho : L^2_\nu(S) \to L^2_\nu(S)$ est la multiplication suivante dans le domaine spectral :

$$M_\rho f = \rho \times f.$$

ρ est une fonction bornée à valeurs réelles appelée spectre de A.

Le théorème spectral peut se résumer ainsi : tout opérateur linéaire borné et auto-adjoint est unitairement équivalent à une multiplication. Le cas où A est compact permet simplement d'obtenir un spectre discret. Dans ce cas, $L^2_\nu(S) = l^2(\mathbb{N})$ et le théorème spectral représente la diagonalisation de la matrice (de taille infinie) associée à l'opérateur.

On peut appliquer ce théorème à notre cadre pour généraliser la théorie de Mercer. On s'intéresse ici au cas particulier des noyaux de convolution. Cela entraîne l'utilisation d'une isométrie particulière : la transformée de Fourier.

On considère l'espace $L^2(\mathbb{R}^d)$ (respectivement $L^1(\mathbb{R}^d)$) des fonctions de carrés intégrables (respectivement intégrables) par rapport à la mesure de Lebesgue. On définit la transformée de Fourier de $f \in L^1(\mathbb{R}^d)$ par la relation suivante :

$$\mathcal{F}[f](\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-i\omega.t} dt, \forall \omega \in \mathbb{R}^d,$$

où $x.y$ est le produit scalaire de $\mathbb{R}^d$. On étend $\mathcal{F}$ de $L^1(\mathbb{R}^d)$ à $L^2(\mathbb{R}^d)$ avec le Théorème de Plancherel pour obtenir une isométrie de $L^2(\mathbb{R}^d)$.
On considère un noyau $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ de convolution (Définition 1.14). Alors il existe $\phi : \mathbb{R}^d \to \mathbb{R}$ vérifiant :

$$K(x,y) = \phi(x-y), \forall x,y \in \mathbb{R}^d.$$

L'opérateur intégral associé à ce noyau est l'opérateur de convolution par ϕ sur $\mathbb{R}^d$. D'après le théorème spectral, L_K vérifie :

$$\mathcal{F} L_K \mathcal{F}^{-1} = M_{\mathcal{F}[\phi]},$$

où $M_{\mathcal{F}[\phi]}$ est la multiplication par la transformée de Fourier de ϕ. On peut montrer que dans ce cas, sous certaines hypothèses de régularité sur ϕ, l'EHNR de noyau K s'écrit [5] :

$$\mathcal{H}_K = \{f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[\phi](\omega)} d\omega < \infty\}. \tag{1.35}$$

Exemple 1.6 (Le cas Sobolev) ***Définition 1.15*** *Soit $s \in \mathbb{R}^+$. L'espace de Sobolev $\mathcal{W}_s^2(\mathbb{R}^d)$ s'écrit :*

$$\mathcal{W}_s^2(\mathbb{R}^d) := \{f \in L^2(\mathbb{R}^d) : \|f\|_s^2 = \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 (1+\|\omega\|^2)^s d\omega < \infty\}. \tag{1.36}$$

Muni du produit scalaire suivant,

$$< f,g >_s = \int_{\mathbb{R}^d} \mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}(1+\|\omega\|^2)^s d\omega,$$

$\mathcal{W}_s^2(\mathbb{R}^d)$ est un espace de Hilbert.

On peut citer Triebel [132] ou Adams [1] pour une présentation de ces espaces fonctionnels. Afin de les relier à la théorie des noyaux, il suffit de considérer un noyau de convolution dont la fonction de base vérifie :

$$\mathcal{F}[\phi](\omega) = \frac{C}{(c+\|\omega\|^2)^s}, \forall \omega \in \mathbb{R}^d,$$

où C, $c > 0$ sont deux constantes. D'après (1.35), l'espace à noyau associé est un espace de Sobolev. La classe des noyaux Laplace, définie par :

$$K(x,y) = \exp(-\sigma\|x-y\|), \forall x,y \in \mathbb{R}^d,$$

pour $\sigma > 0$ vérifie cette propriété.
Dans le Chapitre 2, on étudie les performances statistiques des SVM utilisant les espaces de Sobolev. La classe des noyaux Laplace sera utilisée dans la partie expérimentale pour illustrer les résultats théoriques.

5. Ici, on suppose notamment que le support de $\mathcal{F}[\phi]$ est $\mathbb{R}^d$ tout entier pour alléger l'écriture du Théorème 2.1 du Chapitre 2.

Exemple 1.7 (Les noyaux gaussiens) *La classe des noyaux gaussiens s'écrit :*

$$K(x,y) = \exp(-\sigma\|x-y\|^2), \forall x,y \in \mathbb{R}^d,$$

où $\sigma > 0$ est un paramètre appelé fenêtre. Les noyaux gaussiens sont des noyaux de convolution. On peut interpréter σ comme un paramètre de régularisation. En effet, la transformée de Fourier de $\psi(x) = \exp(-\sigma\|x\|^2)$ s'écrit (Williamson et al. [145] pour les détails) :

$$\mathcal{F}[\psi](\omega) = \frac{1}{(\sqrt{2}\sigma)^d}\exp(-\frac{\|\omega\|^2}{4\sigma^2}).$$

En utilisant la représentation (1.35)*, on obtient :*

$$\mathcal{H}_\sigma = \{f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \sigma^d \exp\left(\frac{\|\omega\|^2}{4\sigma^2}\right) d\omega < \infty\}.$$

Cet espace est constitué de fonctions dont la transformée de Fourier décroît de façon exponentielle. Ces fonctions sont dites super-régulières. Le paramètre σ intervient dans cette représentation dans la définition de la norme. On peut remarquer que $\sigma < \sigma' \Rightarrow \mathcal{H}_\sigma \subset \mathcal{H}_{\sigma'}$. La valeur de σ influence la régularité des fonctions $f \in \mathcal{H}_\sigma$.
Steinwart et al. [123] propose une description détaillée des EHNR gaussiens.

Les noyaux de convolution génèrent des EHNR exprimés dans le domaine de Fourier. La régularité des fonctions dans ces espaces fonctionnels est exprimée en fonction de la décroissance asymptotique du module de leur transformée de Fourier. Or la décroissance de $\mathcal{F}[f]$ à l'infini dépend des variations de f sur tout $\mathbb{R}$. Ce critère est global.
Etudions plus particulièrement la dérivabilité d'une fonction f définie sur $\mathbb{R}$. On sait que la transformée de Fourier de la k$^{\text{ième}}$ dérivée de f s'écrit $\mathcal{F}[f^{(k)}](\omega) = (i\omega)^k\mathcal{F}[f](\omega)$. En appliquant la transformée de Fourier inverse, on obtient :

$$|f^{(k)}(t)| \leq \int |e^{i\omega t}(i\omega)^k \mathcal{F}[f](\omega)|dt \leq \int |\omega|^k |\mathcal{F}[f](\omega)|dt.$$

Ce résultat montre que si la fonction $\omega \mapsto \omega^k\mathcal{F}[f](\omega) \in L^1(\mathbb{R})$, la fonction $f \in \mathcal{C}^p$. Par exemple pour la fonction $f(x) = \mathbb{1}_{[0,1]}(x)$, la discontinuité en 0 et 1 entraîne une décroissance lente du module de la transformée de Fourier (en $|\omega|^{-1}$).
Dans ce cas, il peut être important de savoir que f est régulière en dehors des points 0 et 1. Pour caractériser la régularité locale d'un signal f, il est nécessaire de la décomposer sur une famille de fonctions qui sont localisées dans le temps. Ce n'est pas le cas des sinusoïdes $e^{i\omega t}$. C'est pourquoi on propose de se pencher sur une analyse différente, appelée analyse multirésolution. Cela entraîne l'utilisation d'outils plus récents, les bases d'ondelettes. Ces bases vont permettrent de faire un zoom sur les irrégularités d'une fonction, à l'aide d'un niveau de résolution croissant. Elles seront utilisées dans le Chapitre 3.

1.3.3 Analyse multirésolution et noyaux d'ondelettes

Analyse multirésolution

Les fonctions de base de l'analyse ont longtemps été les fonctions sinus et cosinus. La théorie des ondelettes[6] propose une alternative à l'analyse de Fourier. Initiée dans

6. Pour une introduction vulgarisée de ce nouveau langage, on peut citer "Parlez-vous wavelets?" d'Ingrid Daubechies.

les années 80, ses applications sont aujourd'hui très nombreuses, en traitement du signal ou en statistique. Pour une introduction mathématique, on peut citer Meyer [105]. Mallat [91] propose une exploration des ondelettes en traitement du signal. Dans le domaine statistique, Härdle et al. [68] est un survol complet de l'aspect approximation et l'aspect purement statistique des ondelettes. Il s'inspire notamment d'articles récents en statistiques mathématiques concernant l'estimation de densité par des méthodes de seuillage d'ondelettes (Kerkyacharian and Picard [75], Donoho et al. [53]).

Définition 1.16 *On appelle analyse multirésolution de $L^2(\mathbb{R})$ toute suite croissante $(V_j)_{j\in\mathbb{Z}}$ de sous-espaces fermés de $L^2(\mathbb{R})$ vérifiant les propriétés suivantes :*

1. *il existe une fonction $\phi \in V_0$ telle que $\{\phi(\cdot - k), k \in \mathbb{Z}\}$ soit une base orthonormée de V_0,*
2. $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$,
3. $\bigcup_{j\in\mathbb{Z}} V_j$ *est dense dans* $L^2(\mathbb{R})$,
4. $\forall f \in L^2(\mathbb{R})$, $\forall j \in \mathbb{Z}$, $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$,
5. $\forall f \in L^2(\mathbb{R})$, $\forall k \in \mathbb{Z}$, $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$.

La fonction ϕ est alors appelée fonction d'échelle et on dira que ϕ génère l'analyse multirésolution $(V_j)_{j\in\mathbb{Z}}$.

Pour l'existence d'une telle fonction ϕ, on peut citer Meyer [105]. Il existe plusieurs exemples de fonctions d'échelles. Ces fonctions sont plus ou moins lisses et génèrent des analyses multirésolutions plus ou moins régulières. Par la suite, on dira que l'analyse est r-régulière si ϕ est de classe $\mathcal{C}^r$ et chaque dérivée jusqu'à l'ordre r est rapidement décroissante (Donoho et al. [53] pour les détails).

D'après la Définition 1.16, on peut écrire $V_j = vect\{\phi_{jk}, k \in \mathbb{Z}\}$ où $\{\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k), k \in \mathbb{Z}\}$ est une base orthonormée de V_j obtenue par translation et dilatation de la fonction d'échelle ϕ. De plus, on introduit les espaces W_j, $j \in \mathbb{Z}$ vérifiant :

$$V_{j+1} = V_j \oplus W_j \,, \forall j \in \mathbb{Z}.$$

Alors, il existe une fonction ψ appelée ondelette mère telle que $W_j = vect\{\psi_{jk}, k \in \mathbb{Z}\}$ où $\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), k \in \mathbb{Z}\}$ est une base orthonormée de W_j issue de l'ondelette mère ψ. Ainsi, pour tout entier j_0, on obtient la décomposition suivante de $L^2(\mathbb{R})$:

$$L^2(\mathbb{R}) = V_{j_0} \oplus \bigoplus_{j=j_0}^{\infty} W_j.$$

En particulier, toute fonction $f \in L^2(\mathbb{R})$ peut s'écrire :

$$f = \sum_{k\in\mathbb{Z}} \alpha_{j_0 k}\phi_{j_0 k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}, \tag{1.37}$$

où pour tout $j \geq j_0$, $k \in \mathbb{Z}$, les coefficients d'ondelettes sont définis par :

$$\alpha_{j_0 k} = \int_{-\infty}^{+\infty} f(x)\phi_{j_0 k}(x)dx \text{ et } \beta_{jk} = \int_{-\infty}^{+\infty} f(x)\psi_{jk}(x)dx.$$

Exemple 1.8 (Base de Haar) *Dans sa thèse publiée en 1909 sous la direction de D. Hilbert, A. Haar construit la base orthonormée de* $L^2([0,1])$ *suivante : cette base est composée de la fonction constante égale à* 1 *et des fonctions* ψ_{jk}, $j \geq 0$, $k = 0, \dots 2^{j-1}$ *définies précédemment où :*

$$\psi(x) = \mathbb{1}_{[0,1/2]}(x) - \mathbb{1}_{[1/2,1]}(x).$$

En prenant toutes les valeurs entières j, k, *on obtient une base orthonormée de* $L^2(\mathbb{R})$.

Cette base est constituée de fonctions non continues. Il existe de nombreuses bases d'ondelettes plus régulières (Daubechies [45] par exemple).

Noyaux d'ondelettes

Cette analyse offre un choix de base très large pour construire des EHNR. On peut montrer que sous certaines hypothèses, un espace de Hilbert généré par une base d'ondelettes est un EHNR. Amato et al. [3] considère une base orthonormale d'ondelettes $\{\psi_g, g \in G\}$ de $L^2([0,1])$ à support compact. Sous certaines hypothèses sur la suite $\{\Gamma(g), g \in G\}$, l'espace :

$$\mathcal{H}_\Gamma = \{f \in L^2([0,1[) : f = \sum_{g \in G} f_g \psi_g \text{ et } \sum_{g \in G} \frac{|f_g|^2}{\Gamma(g)} < \infty\}$$

est un EHNR de noyau reproduisant :

$$K(x,y) = \sum_{g \in G} \Gamma(g) \psi_g(x) \psi_g(y). \tag{1.38}$$

Le noyau (1.38) est appelé noyau d'ondelettes ou noyau multi-échelle. On peut généraliser cette approche à la notion de repère (Gao et al. [58], Rakotomamonjy and Canu [110]), offrant un choix de noyaux encore plus varié. Des applications numériques utilisant des noyaux multi-échelles existent en estimation fonctionnelle (Zhang et al. [151], Zeng and Zhao [150], Rakotomamonjy and Canu [110]).

Dans le Chapitre 3, on s'intéresse aux bases d'ondelettes. Elles permettent de caractériser de nombreux espaces fonctionnels classiques, comme les espaces de Besov.

1.3.4 Espaces de Besov

Dans les années 60, O.V. Besov combine l'approche Sobolev et les idées de A.Zygmund qui généralise les espaces de Hölder (on peut citer Triebel [132] pour une introduction historique). Les espaces de Besov constituent une large classe d'espaces fonctionnels, contenant notamment les espaces de Sobolev et les espaces de Hölder.

Définition

On note, pour x, $h \in \mathbb{R}$, le module de continuité de f au point x par :

$$\Delta_h f(x) = f(x-h) - f(x).$$

Suivant l'idée de A. Zygmund, les espaces de Besov se définissent en terme de module de continuité du deuxième ordre $\Delta_h^2 f = \Delta_h(\Delta_h f)$.

Définition 1.17 *1. Soit $0 < s \leq 1$, $p \geq 1$, $q \leq \infty$. Une fonction f appartient à l'espace de Besov $\mathcal{B}_{spq}(\mathbb{R})$ si $f \in L^p(\mathbb{R})$ est telle que :*

$$\gamma_{spq}(f) = \left(\int_{\mathbb{R}} \left(\frac{\|\Delta_h^2 f\|_p}{|h|^s} \right)^q \frac{dh}{|h|} \right)^{1/q} < +\infty$$

pour $q < \infty$ et :

$$\gamma_{sp\infty}(f) = \sup_{h \in \mathbb{R}^*} \frac{\|\Delta_h^2 f\|_p}{|h|^s} < +\infty$$

pour $q = \infty$.
On munit cet espace de la norme :

$$\|f\|_{spq} = \|f\|_p + \gamma_{spq}(f).$$

2. Soit $s = [s] + \alpha$, avec $[s] \in \mathbb{N}$ et $0 < \alpha \leq 1$. On dit que f appartient à l'espace de Besov $\mathcal{B}_{spq}(\mathbb{R})$ si $f^{(m)} \in \mathcal{B}_{\alpha pq}(\mathbb{R})$ pour tout $m \leq [s]$.
On munit cet espace de la norme :

$$\|f\|_{spq} = \|f\|_p + \sum_{m \leq [s]} \gamma_{\alpha pq}(f^{(m)}).$$

Les espaces de Besov constituent une très grande famille d'espaces fonctionnels. En particulier, dans le cas $p = q = 2$, l'espace $\mathcal{B}_{s22}(\mathbb{R})$ correspond précisemment à l'espace de Sobolev $\mathcal{W}_s^2(\mathbb{R})$ vu dans le paragraphe précédent. Les espaces de Hölder sont aussi inclus dans cette famille. On s'intéresse dans le Chapitre 3 au cas $p < 2$. L'espace $\mathcal{B}_{spq}(\mathbb{R})$ pour $p < 2$ contient des fonctions dont la régularité est inhomogène. Cela permet d'étendre la notion de régularité homogène des espaces de Sobolev.

Propriétés d'approximation dans les espaces de Besov

Il existe plusieurs caractérisations des espaces de Besov, empruntées à la théorie de l'approximation. Les performances statistiques établies dans les Chapitres 2 et 3 utilisent les résultats suivants.

Considérons une analyse multirésolution $(V_j)_{j \in \mathbb{Z}}$. Notons P_j les projections orthogonales sur V_j, $j \geq 0$. Sous certaines conditions de régularité sur la fonction d'échelle ϕ générant l'analyse multirésolution, $f \in \mathcal{B}_{spq}(\mathbb{R})$ si et seulement si $f \in L^p(\mathbb{R})$ et s'il existe une suite positive $(\epsilon_j) \in l^q(\mathbb{N})$ telle que :

$$\|f - P_j(f)\|_p \leq 2^{-js} \epsilon_j.$$

Considérons ϕ fonction d'échelle r-régulière, avec $r > s$. Notons $D_j = P_{j+1} - P_j$. Alors une fonction $f \in \mathcal{B}_{spq}(\mathbb{R})$ si et seulement si $f \in L^p(\mathbb{R})$ et s'il existe une suite positive $(\epsilon_j) \in l^q(\mathbb{N})$ telle que :

$$\|D_j(f)\|_p \leq 2^{-js} \epsilon_j. \tag{1.39}$$

Il est alors possible de caractériser les espaces de Besov en terme de coefficients d'ondelettes de la manière suivante.

Proposition 1.2 *Soit (ϕ,ψ) un système d'ondelettes r-régulier, $r > s$ et soit $f \in L^p(\mathbb{R})$. Alors $f \in \mathcal{B}_{spq}(\mathbb{R})$ si et seulement si :*

$$\|f\|_{spq} = \left(\sum_{k\in\mathbb{Z}} |\alpha_k|^p\right)^{\frac{1}{p}} + \left(\sum_{j\in\mathbb{N}} \left(2^{j\left(s+\frac{1}{2}-\frac{1}{p}\right)} \left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\right)^{\frac{1}{p}}\right)^q\right)^{\frac{1}{q}} < +\infty,$$

où les coefficients d'ondelettes s'écrivent :

$$\alpha_k = \int_{\mathbb{R}} f(x)\phi_{0k}(x)dx \text{ et } \beta_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)dx.$$

De plus lorsque $r > s$, la définition ne dépend pas de la base choisie.

Le premier terme de cette norme correspond à la norme L^p de la projection de f sur V_0. Le deuxième terme représente la norme l^q de $2^{js}\|D_j(f)\|_p$. En effet, on peut montrer (Meyer [105]) que :

$$\|D_j(f)\|_p \sim 2^{j\left(\frac{1}{2}-\frac{1}{p}\right)} \left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\right)^{\frac{1}{p}}.$$

La caractérisation (1.39) des espaces de Besov a lieu en dimension $d > 1$. Cela permet de généraliser la Proposition 1.2 aux espaces de Besov $\mathcal{B}_{spq}(\mathbb{R}^d)$. Cette généralisation est présentée dans le Chapitre 3 et permettra de contrôler une mesure de la richesse des espaces de Besov.

Enfin, les espaces de Besov possèdent des résultats remarquables en théorie de l'interpolation (Proposition 1.3). Cela permet de contrôler une vitesse d'approximation définie ci-dessous.

Notons $(\mathcal{F},\mathcal{G})_{\theta,q}$ l'espace d'interpolation entre deux espaces de Banach $\mathcal{F}$ et $\mathcal{G}$ (voir le Chapitre 2 pour une définition précise). On peut démontrer (Smale and Zhou [118]) :

$$f \in (\mathcal{F},\mathcal{G})_{\theta,\infty} \Rightarrow \inf_{g\in\mathcal{G}:\|g\|_{\mathcal{G}}\leq R} \|f-g\|_{\mathcal{F}} \leq \|f\|_{\theta,\infty}^{\frac{1}{1-\theta}} \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}},$$

où $\|\cdot\|_{\theta,\infty}$ est la norme dans l'espace $(\mathcal{F},\mathcal{G})_{\theta,\infty}$. Ainsi une fonction $f \in (\mathcal{F},\mathcal{G})_{\theta,\infty}$ peut être approchée par une boule de rayon R dans $\mathcal{G}$ à une vitesse polynomiale de R. Le résultat suivant concerne la stabilité des espaces de Besov par rapport à la notion d'interpolation.

Proposition 1.3 *Soit $-\infty < s_0,s_1 < +\infty$, $1 < p < \infty$, $1 \leq q_0,q_1 \leq +\infty$ et $0 < \theta < 1$. Alors :*

$$(\mathcal{B}_{s_0pq_0}(\mathbb{R}^d),\mathcal{B}_{s_1pq_1}(\mathbb{R}^d))_{\theta,q} = \mathcal{B}_{spq}(\mathbb{R}^d), \tag{1.40}$$

où $s = (1-\theta)s_0 + \theta s_1$.

En particulier,

$$(L^2(\mathbb{R}^d),\mathcal{W}_s^2(\mathbb{R}^d))_{\theta,\infty} = \mathcal{B}_{r2\infty}(\mathbb{R}^d), \tag{1.41}$$

avec $r = \theta s$.

La relation (1.41) sera utilisée pour contrôler la qualité d'approximation des SVM utilisant les espaces de Sobolev (Chapitre 2), alors que (1.40) permettra de généraliser ces résultats aux espaces de Besov (Chapitre 3).

1.3.5 Retour aux noyaux

La théorie des noyaux présentée ci-dessus se restreint au cadre des espaces de Hilbert à noyaux reproduisant. D'après la Définition 1.10, un noyau est une application qui engendre un produit scalaire dans l'espace $\Phi(\mathcal{X})$. Le formalisme des EHNR permet d'identifier dans ce cas un noyau à une application symétrique définie positive. Cependant, certains noyaux utilisés en pratique n'ont pas ces propriétés. Par exemple, le noyau tangente hyperbolique $K(x,y) = \tanh(x.y + b)$ n'est pas positif. Certains algorithmes utilisent des régularisations de normes L^1 (comme le LASSO), qui ne sont pas hilbertiennes. On peut alors se pencher sur une généralisation de la théorie des noyaux à un cadre non-hilbertien.
Canu et al. [32] énonce les principes fondamentaux de l'espace $\Phi(\mathcal{X}) = \mathcal{H}$ appelé espace d'hypothèses. Cet espace doit être muni d'un critère de convergence. De plus, $\mathcal{H}$ doit contenir des fonctions définies point par point ($\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$). Cette propriété assure que tout élément de $\mathcal{H}$ peut être mesuré en chaque point de l'ensemble d'apprentissage. De plus, la fonction d'évaluation $\delta_x : \mathcal{H} \to \mathbb{R}$ définie par $\delta_x(f) = f(x)$ doit être continue, pour tout $x \in \mathcal{X}$. Ainsi la convergence dans $\mathcal{H}$ implique la convergence ponctuelle. Ces propriétés sont vérifiées par les EHNR. En effet, un EHNR est inclus dans $\mathbb{R}^{\mathcal{X}}$ par définition. De plus, d'après la Proposition 1.1, la continuité de la fonction d'évaluation est équivalente à la propriété reproduisante. Par contre, la structure hilbertienne n'est pas indispensable.

Définition 1.18 *Soit $\mathcal{H}$ un espace vectoriel normé tel que $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. $\mathcal{H}$ est un espace d'évaluation si et seulement si pour tout $x \in \mathcal{X}$, δ_x est continue.*

Dans le cas hilbertien, d'après le théorème de Riesz, pour tout $x \in \mathcal{X}$, il existe une unique fonction $g_x \in \mathcal{H}$ vérifiant $\delta_x(f) =< f,g_x >_{\mathcal{H}}$. g_x est appelée noyau reproduisant et toute fonction de $\mathcal{H}$ s'exprime comme combinaison linéaire de ce noyau.
Généraliser ces propriétés aux espaces d'évaluations non-hilbertiens nécessite l'existence d'une forme bilinéaire qui se substitue au produit scalaire. Ce rôle peut être joué par le crochet de dualité. Pour cela, on considère des espaces vectoriels en dualité contenus dans $\mathbb{R}^{\mathcal{X}}$, appelés sous-dualités d'évaluations (Mary et al. [99]). De manière analogue aux EHNR, on peut associer à chaque sous-dualité d'évaluation un noyau reproduisant de manière unique. Dans ce cas, le noyau obtenu n'est plus symétrique ni défini positif. Cette généralisation au cas non-hilbertien est à relier aux résultats du Chapitre 3, où l'espace d'hypothèses considéré est un espace de Besov.

1.4 Présentation des résultats

1.4.1 Vitesses de convergence

L'étude des performances statistiques d'algorithmes de type SVM est le premier axe de cette thèse. Dans la littérature, plusieurs travaux établissent des vitesses de convergence pour les SVM. Wu and Zhou [147] obtient des vitesses très lentes (logarithmiques en fonction de n) pour les SVM utilisant un noyau gaussien à fenêtre fixée. Chen et al. [41] étudie une version modifiée des SVM utilisant la perte q-SVM $l(y,f(x)) = (1 - yf(x))_+^q$, $q > 1$. Des vitesses sont obtenues sous des hypothèses de régularité sur f_q^* (voir Lemme 1.3). Wu et al. [146] suppose que la vitesse d'approximation est connue et propose de considérer des SVM utilisant plusieurs noyaux. Steinwart and Scovel [126] contient une étude complète des propriétés statistiques de l'algorithme. L'erreur d'approximation des espaces à noyaux

gaussiens est contrôlée en utilisant une hypothèse sur la mesure P. A l'aide d'une inégalité oracle pour l'estimation, des vitesses de convergence rapides (plus rapides que $n^{-1/2}$) sont établies. Steinwart et al. [125] se concentre sur l'erreur d'estimation et propose une inégalité oracle plus fine. Blanchard et al. [24] utilise la représentation de l'Exemple 1.4 des SVM afin d'établir une inégalité oracle pour l'estimation.

Les chapitres 2 et 3 établissent des vitesses de convergence de méthodes de régularisation de type SVM.

Cas Sobolev

Approximation Les travaux de I. Steinwart et C. Scovel portent principalement sur les performances des SVM utilisant les noyaux gaussiens. Le Chapitre 2 s'intéresse à une autre classe de noyaux que les noyaux gaussiens, souvent utilisés en apprentissage statistique. On a vu précédemment que le noyau gaussien génère un EHNR constitué de fonctions super-régulières. Cela vient de la régularité du noyau de convolution gaussien $K_\sigma(x,y) = \exp(-\sigma\|x-y\|^2)$, $\sigma > 0$. Une alternative consiste à considérer une classe de noyau K_r, $r > 0$ générant un espace de Sobolev $\mathcal{W}^2_{r/2}(\mathbb{R}^d)$ comme EHNR. Cette classe est appelée classe de noyaux Sobolev d'exposant r. Le noyau Laplace $K(x,y) = \exp(-\sigma\|x-y\|)$ est un noyau Sobolev d'exposant $r = d+1$, puisque le module de sa transformée de Fourier décroît en $\dfrac{C}{\|\omega\|^{d+1}}$ lorsque $\omega \to +\infty$. Cela est dû à l'irrégularité de sa fonction de base en l'origine. Ce noyau engendre un espace à noyau $\mathcal{H}_{K_r}$ vérifiant :

$$\mathcal{H}_{K_r} \supset \mathcal{H}_{K_\sigma}, \forall \sigma > 0.$$

L'intérêt de considérer un noyau moins régularisant porte principalement sur le contrôle de l'erreur d'approximation. On appelle fonction d'approximation la quantité :

$$a(\alpha_n) = \inf_{f \in \mathcal{H}_K} \left(R_l(f,f^*) + \alpha_n \|f\|_K^2 \right). \tag{1.42}$$

Dans le cas d'un noyau gaussien, Steinwart and Scovel [126] obtient (sous l'hypothèse de bruit géométrique) :

$$a(\alpha_n) \leq C \alpha_n^{-\frac{1}{d(\delta+1)}}, \tag{1.43}$$

où δ est le bruit géométrique. Pour obtenir ce résultat, la fenêtre du noyau gaussien σ est calibrée en fonction du paramètre de régularisation α_n. Plus précisemment, (1.43) a lieu lorsque $\sigma = \sigma_n \to +\infty$ pour $n \to \infty$. De plus, le choix de la fenêtre dépend de δ.

Dans le Chapitre 2, on considère un noyau Sobolev. Sous une hypothèse de régularité sur f^*, on obtient :

$$a(\alpha_n) \leq C \alpha_n^{\frac{s}{r-s}}, \tag{1.44}$$

où s est la régularité de f^* et r l'exposant du noyau. Le contrôle de la fonction d'approximation a lieu pour un noyau fixé K_r, où l'exposant $r > 0$ ne dépend pas de n. Ce résultat distingue les noyaux Sobolev par rapport à la classe des noyaux gaussiens.

Vitesse de convergence Le Chapitre 2 propose des vitesses de convergence de la forme :

$$n^{-\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}, \tag{1.45}$$

où

- r est l'exposant du noyau Sobolev,
- q est le paramètre de marge,
- d est la dimension de $\mathcal{X} = \mathbb{R}^d$,
- s est la régularité de f^*.

Afin d'obtenir (1.45), on suppose que $f^* \in \mathcal{B}_{s2\infty}(\mathbb{R}^d)$. Cette hypothèse assure qu'en utilisant un noyau Sobolev, la vitesse d'approximation des SVM vers f^* est polynomiale. Cette hypothèse n'est pas vérifiée pour de grandes valeurs de s. En effet, il est clair que f^* n'est pas continue. Alors, d'après les propriétés de l'espace de Besov $\mathcal{B}_{s2\infty}(\mathbb{R}^d)$, $f^* \in \mathcal{B}_{s2\infty}(\mathbb{R}^d)$ seulement si $s < \frac{d}{2}$. Cela entraîne dans (1.45) une gamme de vitesses vérifiant :

$$n^{-\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}} > n^{-\frac{1}{2}}.$$

Par conséquent, le Chapitre 2 ne permet pas réellement d'obtenir des vitesses de convergence rapides pour les SVM. Cela est dû principalement à la partie approximation. Ce problème est abordé dans le Chapitre 3 où l'on utilise des espaces fonctionnels plus vastes.

Généralisation aux espaces de Besov

Le Chapitre 3 propose de considérer les espaces de Besov comme alternative aux EHNR. On a vu dans le paragraphe 1.3.3 que les espaces de Besov $\mathcal{B}_{spq}(\mathbb{R}^d)$ étaient une large classe d'espaces fonctionnels. Dans le cas $p = q = 2$, $\mathcal{B}_{spq}(\mathbb{R}^d)$ correspond à l'espace de Sobolev $\mathcal{W}_s^2(\mathbb{R}^d)$. La régularité d'une fonction $f \in \mathcal{B}_{s22}(\mathbb{R}^d)$ s'exprime en fonction du comportement asymptotique de sa transformée de Fourier. Ce critère est global et une irrégularité locale entraîne une décroissance lente de sa transformée de Fourier. Lorsque $p < 2$, la régularité des fonctions $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ s'exprime différemment. Ces espaces sont adaptés à des fonctions globalement régulières ayant des irrégularités locales.
Le Chapitre 3 étudie notamment les performances statistiques de la procédure suivante :

$$\min_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \alpha_n \|f\|_{spq}^2 \right), \tag{1.46}$$

pour $p \geq 1$. Cette minimisation s'apparente aux SVM. Cependant, la norme $\|\cdot\|_{spq}$ n'est pas induite par un noyau. De plus, l'espace d'hypothèses n'est pas un EHNR ($\mathcal{B}_{spq}(\mathbb{R}^d)$ n'est pas un espace de Hilbert). C'est pourquoi on parle de minimisation de risque empirique pénalisé. On peut dire que cette méthode généralise le Chapitre 2 puisque $\mathcal{B}_{spq}(\mathbb{R}^d)$ correspond à $\mathcal{W}_s^2(\mathbb{R}^d)$ pour $p = q = 2$. En considérant le cas $p < 2$, on s'éloigne des SVM (disparition du produit scalaire, et donc du noyau). Par contre, on dispose d'espaces fonctionnels où la régularité est inhomogène, c'est-à-dire plus adaptée à notre fonction cible. On obtient dans le Chapitre 3 des vitesses de convergence de la forme :

$$n^{-\frac{r}{2s-r}\frac{2u}{2u+d}}, \tag{1.47}$$

où $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$. La principale avancée réside dans l'hypothèse sur la régularité de f^*. Pour contrôler l'approximation (1.42), on suppose que $f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d)$, où on peut autoriser $p < 2$. Cela entraîne dans (1.47) une vitesse plus rapide que $n^{-1/2}$ dans certains cas.

1.4.2 Adaptation et sélection de modèles

Les vitesses de convergence (1.45) et (1.47) sont obtenues sous des hypothèses a priori sur la distribution P. Ces hypothèses portent principalement sur la régularité de la règle de Bayes et sur le bruit de classification (hypothèses de marge). Elles ne sont pas vérifiables à partir des observations mais peuvent intervenir dans la construction de la solution. Dans ce cas, on dira que notre procédure est non-adaptative. Le problème statistique de l'adaptation revient à construire un estimateur directement calculable à partir des observations, sans paramètre à calibrer. Le Chapitre 2 utilise une méthode appelée agrégation alors que le Chapitre 3 applique un théorème général de sélection de modèles. Le Chapitre 4 est entièrement consacré à l'étude d'une nouvelle méthode de sélection de modèles : la minimisation de l'enveloppe du risque.

Agrégation

Régularisation non-adaptative Dans le Chapitre 2, on s'intéresse à l'algorithme SVM, que l'on peut écrire :

$$\min_{f\in\mathcal{H}_K}\left[\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ + \alpha_n\|f\|_K^2\right]. \tag{1.48}$$

Le paramètre α_n est le paramètre de régularisation. Il doit équilibrer l'influence des données et la régularité de la solution. Cet équilibre est obtenu dans le Chapitre 2 en minimisant une borne supérieure de l'excès de risque.
Pour cela, on propose de contrôler l'excès de risque de la solution de (1.48) en deux étapes. La première étape consiste à obtenir une inégalité oracle de la forme :

$$\mathbb{E}R_l(\hat{f}_n,f^*) \leq C \inf_{f\in\mathcal{H}_K}\left[R_l(f,f^*) + \alpha_n\|f\|_K^2\right] + \delta(\alpha_n). \tag{1.49}$$

Cette inégalité assure que le classifieur $\hat{f}_n$ est comparable au meilleur classifieur de l'espace d'hypothèses $\mathcal{H}_K$, à une constante $C > 0$ près et à un terme résiduel près $\delta(\alpha_n)$. Cette inégalité concerne l'erreur d'estimation. Elle nécessite principalement un contrôle de la complexité de l'espace d'hypothèses $\mathcal{H}_K$.
La deuxième étape consiste à majorer le membre de droite de l'inégalité oracle. Le premier terme représente la fonction d'approximation $a(\alpha_n)$ définie précédemment. On peut écrire d'après (1.44) :

$$\mathbb{E}R(\hat{f}_n,f^*) \leq C\alpha_n^{\frac{s}{r-s}} + \delta(\alpha_n).$$

On peut équilibrer ces deux termes en fonction de α_n. On obtient le choix optimal suivant :

$$\alpha_n = n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}. \tag{1.50}$$

Ce choix dépend notamment des paramètres q et r qui représentent le paramètre de marge (1.12) et la régularité de f^*. Ces paramètres ne sont pas connus. La solution de (1.48) avec α_n vérifiant (1.50) est alors dite non-adaptative.

Agrégation La méthode d'agrégation a été introduite dans le cadre général par Nemirovski [107]. Cette méthode propose une approche oracle du problème. Il peut se formuler de manière générale de la façon suivante. On dispose d'une famille d'estimateurs $\mathcal{F} = \{f_1, \dots, f_M\}$ appelés estimateurs faibles. Le but de l'agrégation est de construire à partir d'observations i.i.d. $Z_1, \dots, Z_n$ un agrégat $\tilde{f}_n$ dont le risque est aussi proche que possible de celui de l'oracle $f_{or} = \arg \min_{i=1\dots M} R(f_i, f)$.
Cette méthode d'agrégation permet de résoudre des problèmes d'adaptation. On peut citer les travaux de Lecué [78]. Le principe est relativement simple : on va séparer l'échantillon en deux parties. Avec la première partie, on calcule une famille d'estimateurs faibles. La deuxième partie permet d'agréger ces estimateurs. Dans notre cas, on distingue 4 étapes :

- L'échantillon est divisé en deux parties D_n^1 et D_n^2.
- On calcule à l'aide de D_n^1 une famille de classifieurs faibles $\mathcal{F} = \{\hat{f}_1, \dots, \hat{f}_M\}$, où l'on fait varier le paramètre α_n à ajuster dans une grille à M éléments définie au préalable.
- On construit une famille de poids $\{w_1, \dots, w_M\}$ qui dépend des performances de chaque classifieur faible sur D_n^2.
- L'agrégat $\tilde{f}_n$, combinaison linéaire des classifieurs faibles, est défini par :

$$\tilde{f}_n = \sum_{i \in 1}^{M} w_i \hat{f}_i.$$

L'estimateur $\tilde{f}_n$ est construit seulement à partir des observations D_n. Il s'adapte à la fois au paramètre de marge et à la régularité de la règle de Bayes. Il atteint les mêmes vitesses de convergence que l'estimateur non-adaptatif.
Le principal avantage de cette méthode réside dans le choix de la grille. Cette dernière est construite à partir d'arguments théoriques. Cela permet de restreindre sa taille et de réduire les calculs (contrairement à des méthodes de validation croisée où le choix de la grille est arbitraire).
On peut illustrer ces performances statistiques à partir de données réelles de classification. Cela nous permet de comparer les résultats de généralisation des SVM utilisant un noyau Sobolev avec l'approche de Steinwart and Scovel [126] pour les noyaux gaussiens.

Sélection de modèles pénalisée de Birgé-Massart

Dans le Chapitre 3, on utilise une autre méthode d'adaptation. Afin de choisir le paramètre α_n dans (1.46), on interprète la solution $\hat{f}_n$ comme solution du problème de sélection de modèles suivant. On peut écrire $\hat{f}_n = \hat{f}_{\widehat{R}}$ où :

$$\hat{f}_R = \arg \min_{f \in \mathcal{B}(R)} \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i)) \text{ et } \widehat{R} = \arg\min_{R} \left(\frac{1}{n} \sum_{i=1}^{n} l(Y_i, \hat{f}_R(X_i)) + pen(R) \right),$$

avec $\mathcal{B}(R) = \{f \in \mathcal{B}_{spq}(\mathbb{R}^d) : \|f\|_{spq} \leq R\}$, l la perte douce et $pen(R) = \alpha_n R^2$. Ainsi, le choix de α_n est remplacé par le choix de la pénalité $pen(R)$. Le théorème suivant propose, dans un cadre très général, un choix optimal de pénalité afin d'obtenir une inégalité oracle pour l'estimation.

Théorème 1.6 *Soit l une fonction de perte telle que $g^* \in \arg\min_{f \in L^2(P_X)} \mathbb{E} l(Y, f(X))$. Soit $(\mathcal{G}_m)_{m \in \mathcal{M}}$ une collection de modèles vérifiant $\mathcal{G}_m \subset L^2(P_X)$, $\forall m \in \mathcal{M}$. On suppose*

qu'il existe une pseudo-distance d sur $L^2(P_X)$, une suite de fonctions racines $(\phi_m)_{m\in\mathcal{M}}$, et deux suites positives $(b_m)_{m\in\mathcal{M}}$ et $(C_m)_{m\in\mathcal{M}}$ telles que :

(H1) $\forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m, \|l(g)\|_\infty \leq b_m$.

(H2) $\forall g,g' \in L^2(P_X), Var\left(l(Y,g(X)) - l(Y,g'(X))\right) \leq d^2(g,g')$.

(H3) $\forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m, d^2(g,g^*) \leq C_m \mathbb{E}\left(l(Y,g(X)) - l(Y,g(X)^*)\right)$.

(H4) $\forall m \in \mathcal{M}, \forall g_0 \in \mathcal{G}_m, \forall r \geq r_m^*$:

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_m : d(g,g_0)^2 \leq r} (\mathbb{E} - \hat{E}_n)(l(Y,g(X)) - l(Y,g(X)_0))\right] \leq \phi_m(r),$$

où r_m^* vérifie $\phi_m(r_m^*) = r/C_m$ et $\hat{\mathbb{E}}_n X = \frac{1}{n}\sum X_i$.

Soit $(x_m)_{m\in\mathcal{M}}$ une suite de réels vérifiant $\sum_{m\in\mathcal{M}} e^{-x_m} \leq 1$ et telle que :

$$\forall m,m' \in \mathcal{M}, x_m \leq x_{m'} \Rightarrow b_m \leq b_{m'} \text{ et } C_m \leq C_{m'}.$$

On suppose qu'il existe $\tilde{g}$ vérifiant $\tilde{g} = \hat{g}_{\hat{m}}$ où :

$$\hat{g}_m = \arg\min_{f\in\mathcal{G}_m} \frac{1}{n}\sum_{i=1}^n l(Y_i,g(X_i)) \text{ et } \hat{m} = \arg\min_{m\in\mathcal{M}} \left(\frac{1}{n}\sum_{i=1}^n l(Y_i,\hat{g}_m(X_i)) + pen(m)\right).$$

Alors, si la pénalité satisfait, pour tout $m \in \mathcal{M}$:

$$pen(m) \geq 250K\frac{r_m^*}{C_m} + \frac{B_m(x_m + \log 2)}{3n} + \frac{B_m \log B_m}{n}, \tag{1.51}$$

où $B_m = 75KC_m + 28b_m$, on obtient :

$$\mathbb{E}R_l(\tilde{g},g^*) \leq \frac{K+\frac{1}{5}}{K-1}\inf_{m\in\mathcal{M}}\left(\inf_{g\in\mathcal{G}_m} R_l(g,g^*) + 2pen(m)\right) + \frac{2}{n}. \tag{1.52}$$

Le principal avantage du Théorème 1.6 réside dans sa généralité. Il peut s'adapter à un très grand nombre de modèles statistiques. Il permet de résoudre des problèmes d'adaptation en régression, en estimation de densité, en classification. Il est issu de la théorie de sélection de modèles pénalisée de Birgé et Massart (Barron et al. [10]).

Appliqué au problème (1.46), ce théorème propose une pénalité minimale de la forme $pen(R) = \alpha_n R$ seulement. La régularisation minimale est donc linéaire en $\|\cdot\|_{spq}$. Dans l'algorithme (1.46), la régularisation est de l'ordre de $\|\cdot\|_{spq}^2$. En utilisant le Théorème 1.6, ces deux pénalités aboutissent à des vitesses de convergence différentes.

La relation (1.51) entraîne la condition suivante sur α_n :

$$\alpha_n \geq c_1 n^{-\frac{2u}{2u+d}} + \eta_1^{-1}\left(c_2\frac{\log n}{n} + c_3\frac{\log\log n}{n} + \frac{c_4}{n}\right), \tag{1.53}$$

où $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$. Cette relation ne dépend pas de la régularité de f^*. Le choix de α_n s'adapte à la régularité de la règle de Bayes.

Cependant, (1.51) dépend d'une constante K inconnue. Ainsi le Théorème 1.6 propose un choix de pénalité explicite, à une constante près. En pratique, la validation croisée ou plus récemment l'heuristique de pente sont utilisées pour choisir automatiquement la constante dans la pénalité. La méthode de sélection de modèles étudiée dans le dernier chapitre propose un autre choix de pénalité.

Minimisation de l'enveloppe du risque

Modèle des suites gaussiennes Le Chapitre 4 se consacre à l'étude d'une nouvelle méthode de sélection de modèles : la minimisation de l'enveloppe du risque (RHM : Risk Hull Minimization). Cette méthode a été introduite par Cavalier and Golubev [39] dans le modèle des suites gaussiennes :

$$y_k = b_k\theta_k + \epsilon\xi_k, k \in \mathbb{N}^*, \tag{1.54}$$

où (θ_k) sont les coefficients d'une fonction f à estimer. ξ_k sont des variables aléatoires modélisant le bruit dans les observations. Elles sont supposées indépendantes et identiquement distribuées $\mathcal{N}(0,1)$. $\epsilon > 0$ est le niveau de bruit, supposé connu et (b_k) est une suite de coefficients connus qui tend vers 0 lorsque $k \to \infty$. Ainsi, pour de grandes valeurs de k, l'observation y_k contient essentiellement du bruit. Sous ces hypothèses, Cavalier and Golubev [39] considère la famille d'estimateurs par projection $\{\hat{\theta}(N), N \geq 1\}$ définie par :

$$\hat{\theta}_k(N) = \frac{y_k}{b_k}\ \mathbb{1}(k \leq N), k \in \mathbb{N}^*. \tag{1.55}$$

Cette méthode de régularisation consiste à ne conserver que les N premiers termes $y_1, \ldots, y_N$. Le paramètre N est appelé la fenêtre. Le problème de sélection de modèles est alors le choix de N. On veut, comme précédemment, une méthode automatique pour choisir N. Pour comprendre l'influence de N dans ce modèle, on peut écrire le risque quadratique de l'estimateur $\hat{\theta}(N)$:

$$R(\theta,N) = \mathbb{E}\|\theta - \hat{\theta}(N)\|^2 = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2}. \tag{1.56}$$

Le premier terme de (1.56) est appelé biais. Il provient de l'utilisation d'un nombre fini N d'observations pour approcher la suite de coefficients $(\theta_k)_{k\geq 1}$. Il s'agit d'une erreur d'approximation qui décroît lorsque N croît. Le deuxième terme est appelé variance ou terme stochastique. Cette erreur est due au bruit dans les observations. Lorsque N grandit, cette erreur grandit. Ainsi, pour minimiser le risque, il faut choisir N établissant le compromis biais-variance.

Dans ce modèle, on peut choisir la fenêtre par un critère de sélection de modèles pénalisée. La méthode d'estimation du risque empirique pénalisée s'écrit :

$$\widehat{N} = \arg\min_{N\geq 1} \left[-\sum_{k=1}^{N} y_k^2 b_k^{-2} + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} + pen(N) \right],$$

où $pen(N)$ est une pénalité croissante en fonction de N. Le terme $-\sum_{k=1}^{N} y_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2}$ est issu de l'estimation du biais $\sum_{k>N} \theta_k^2$ où on remplace chaque θ_k^2 par son estimateur sans biais $b_k^{-2}(y_k^2 - \epsilon^2)$. La pénalité permet de contrôler la variabilité de la solution. On distingue plusieurs types de pénalité dans la littérature :

- $pen(N) = \epsilon^2 \sum_{k=1}^{N} b_k^{-2}$ (critère d'Akaike). Ce choix correspond à minimiser un estimateur sans biais du risque (1.56). Il s'agit de la méthode d'estimation du risque sans biais. Elle génère un estimateur $\hat{\theta}(\widehat{N})$ assez instable.

- $pen(N) = (1+K)\epsilon^2 \sum_{k=1}^{N} b_k^{-2}$ (pénalité de Birgé-Massart). Cette pénalité est issue de l'approche de sélection de modèles présentée dans le paragraphe précédent. La constante K assure une pénalisation supérieure à Akaike, et ainsi une solution plus stable. En pratique K est choisie par validation croisée ou heuristique de pente.
- $pen(N) = \epsilon^2 \sum_{k=1}^{N} b_k^{-2} + (1+\alpha)\sqrt{2\sigma(N)^2 \log \frac{\sigma(N)^2}{\pi \epsilon^4 b_k^{-4}}}$ (enveloppe du risque). Le deuxième terme de cette pénalité est proportionnel à l'écart type du processus η_N. Ainsi on ne tient plus compte seulement de l'espérance du critère (méthode d'Akaike) mais aussi de sa variance. Les expérimentations montrent une amélioration significative par rapport à Akaike.

L'idée principale de l'enveloppe du risque est la suivante : au lieu de minimiser un estimateur du risque (1.56), on se concentre sur la perte aléatoire :

$$r(\theta,N) = \|\theta - \hat{\theta}(N)\|^2 = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} \xi_k^2.$$

Pour minimiser cette perte, on va contrôler les variations du terme stochastique $\epsilon^2 \sum_{k=1}^{N} b_k^{-2}\xi_k^2$ uniformément en N. On cherche une fonction déterministe $V(N)$ telle que :

$$\mathbb{E} \sup_{N \geq 1} \left[\epsilon^2 \sum_{k=1}^{N} b_k^{-2} \xi_k^2 - V(N) \right] \leq 0. \tag{1.57}$$

Ainsi on obtient, pour un choix aléatoire $\tilde{N}$ quelconque :

$$\mathbb{E}\|\theta - \hat{\theta}(\tilde{N})\|^2 \leq \mathbb{E}\left[\sum_{k>N} \theta_k^2 + V(\tilde{N}) \right] := \mathbb{E} l(\theta,\tilde{N}). \tag{1.58}$$

On appelle $l(\theta,N)$ enveloppe du risque. Cette quantité est une borne supérieure du risque. Elle sera plus facile à minimiser que le risque, qui lui est trop instable.
Pour obtenir une enveloppe du risque, il faut déterminer $V(N)$ dans (1.57). Cela revient à étudier le processus suivant :

$$\eta_N = \epsilon^2 \sum_{k=1}^{N} b_k^{-2}(\xi_k^2 - 1).$$

Dans le modèle de suites gaussiennes, ce processus est un processus de Wiener. On peut alors obtenir (1.57) en utilisant des inégalités exponentielles sur les déviations du processus η_N.
Enfin, pour minimiser $l(\theta,N)$ à partir des observations, on remplace chaque θ_k^2 par son estimateur sans biais $b_k^{-2}(y_k^2 - \epsilon^2)$. On obtient un choix automatique de N, noté N_{rhm} qui minimise l'enveloppe du risque $l(\theta,N)$ sans pertes significatives. Avec (1.58), on obtient :

$$\mathbb{E}\|\theta - \hat{\theta}(N_{rhm})\|^2 \leq C \min_{N \geq 1} \left[\sum_{k>N} \theta_k^2 + V(N) \right] + \text{ résidus}, \tag{1.59}$$

où $C > 0$ est une constante arbitrairement proche de 1. Cette inégalité oracle assure que l'estimateur $\hat{\theta}(N_{rhm})$ a un risque comparable à celui du meilleur estimateur de la famille

(1.55), à un terme résiduel près. Ce terme représente le prix à payer dû au choix adaptatif de N_{rhm} minimisant un estimateur de $l(\theta,N)$. Cavalier and Golubev [39] montre qu'il existe une fonction $V(N)$ telle que ce terme résiduel soit petit, caractérisant la stabilité de la méthode.

Classification L'application de cette méthode en classification est l'enjeu du Chapitre 4. On considère, pour $N \geq 1$, la famille de classifieurs $\hat{f}_N$ définie par :

$$\hat{f}_N = \arg\min_{f \in \mathcal{F}_N} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

où $\mathcal{F}_N$ est un sous-espace vectoriel de dimension N d'un EHNR $\mathcal{H}_K$. Cette famille de classifieurs, appelée KPM (Kernel Projection Machines) a été étudiée par Zwald [154] (voir Exemple 1.3). Le problème de sélection de modèles est le choix du paramètre N, la dimension du sous-espace de $\mathcal{H}_K$. Zwald [154] propose d'utiliser un théorème général de sélection de modèles (Théorème 1.6). Dans cette dernière partie, on réécrit le modèle de classification comme un modèle séquentiel de la forme (1.54). Cela nous permet, sous certaines hypothèses, d'obtenir une enveloppe du risque $R(\widehat{f}_N,f_\eta) = \mathbb{E}(\hat{f}_{\widehat{N}}(X) - f_\eta(X))^2$ où $f_\eta(x) = 2\eta(x) - 1$. On en déduit une inégalité oracle de la forme :

$$\mathbb{E}R(\hat{f}_{\widehat{N}},f_\eta) \leq C \min_{N \geq 1} \Big[R(\hat{f}_N,f_\eta) + pen(N) \Big] + \text{ résidus}, \tag{1.60}$$

où $C > 0$ proche de 1 et $pen(N)$ est une fonction croissante de N. Le choix de $\hat{N}$ adaptatif s'écrit :

$$\hat{N} = \arg\min_{N \geq 1} \left[-\sum_{k=1}^{N} y_k^2 + 2\frac{(1+h)(1-h)}{n} N + \alpha \frac{(1+h)(1-h)}{n} \Sigma(N)^p \right],$$

où les y_k, $k = 1 \dots n$ sont calculés à partir des observations (X_i,Y_i), $i = 1, \dots n$, h est un paramètre de marge et $\Sigma(N)$ est l'écart type d'un processus aléatoire.

D'un point de vue probabiliste, le Chapitre 4 est une généralisation de Cavalier and Golubev [39] à l'étude de processus aléatoires plus généraux. Pour obtenir une inégalité oracle de la forme (1.60), on doit contrôler uniformément en N les variations du processus :

$$\zeta(N) = \epsilon^2 \sum_{k=1}^{N} (\xi_k^2 - 1),$$

où ξ_k sont des variables aléatoires telles que $\mathbb{E}\xi_k = 0$ et $\mathbb{E}\xi_k\xi_l = \delta_{kl}$. Contrairement au modèle des suites gaussiennes, ces variables ne sont pas gaussiennes i.i.d. Pour obtenir une enveloppe du risque, on considère alors une classe de processus plus générale que les processus de Wiener : les processus ordonnés. Ces processus sont présentés dans la dernière partie du Chapitre 4.

Chapitre 2

Aggregation of SVM classifiers using Sobolev spaces

The material of this chapter has been published in the *Journal of Machine Learning Research*, 2008.

Abstract

This chapter investigates statistical performances of Support Vector Machines (SVM) and considers the problem of adaptation to the margin parameter and to complexity. In particular we provide a classifier with no tuning parameter. It is a combination of SVM classifiers. The contribution is two-fold: (1) we propose learning rates for SVM using Sobolev spaces and build a numerically realizable aggregate that converges with same rate; (2) we present practical experiments of this method of aggregation for SVM using both Sobolev spaces and Gaussian kernels.

2.1 Introduction

We consider the binary classification setting. Let $\mathcal{X} \times \{-1,1\}$ be a measurable space endowed with P an unknown probability distribution on $\mathcal{X}\times\{-1,1\}$. Let $D_n = \{(X_i,Y_i),i = 1,\dots n\}$ be n realizations of a random variable (X,Y) with law P (in the sequel we also write P_X for the marginal distribution of X). Given this training set D_n, the goal of Learning is to predict class Y of new observation X. In other words, a classification algorithm builds a decision rule from $\mathcal{X}$ to $\{-1,1\}$ or more generally a function f from $\mathcal{X}$ to $\mathbb{R}$ where the sign of $f(x)$ determines the class of an input x.
The efficiency of a classifier is measured by the *generalization error*

$$R(f) := \mathbb{P}(\text{sign}(f(X)) \neq Y),$$

where sign(y) denotes the sign of $y \in \mathbb{R}$ with the convention sign$(0) = 1$. A well-known minimizer over all measurable functions of the generalization error is called the *Bayes rule*, defined by

$$f^*(x) := \text{sign}(2\eta(x) - 1)$$

where $\eta(x) := \mathbb{P}(Y = 1|X = x)$ for all $x \in \mathcal{X}$. Unfortunately, the dependence of f^* on the unknown conditional probability function η makes it uncomputable in practice.
A natural way to overcome this difficulty is to provide an empirical decision rule or classifier based on the data D_n. It has to mimic the Bayes. The way one measures the efficiency of a classifier $\hat{f}_n := \hat{f}_n(D_n)$ is via its *excess risk*:

$$R(\hat{f}_n,f^*) := R(\hat{f}_n) - R(f^*), \tag{2.1}$$

where here $R(\hat{f}_n) := \mathbb{P}(\text{sign}(\hat{f}_n(X)) \neq Y|D_n)$. Given P, we hence say that a classifier $\hat{f}_n$ is consistent if the expectation of (3.2) with respect to $P^{\otimes n}$ (the distribution of the training set) goes to zero as n goes to infinity. Finally, we can look for a way of quantifying this convergence. A classifier $\hat{f}_n$ learns with rate $(\psi_n)_{n\in\mathbb{N}}$ if there exists an absolute constant $C > 0$ such that for all integer n,

$$\mathbb{E}R(\hat{f}_n,f^*) \leq C\psi_n, \tag{2.2}$$

where in the sequel $\mathbb{E}$ is the expectation with respect to $P^{\otimes n}$. Of course (3.3) ensures consistency of $\hat{f}_n$ whenever (ψ_n) goes to zero with n.

It has been shown in Devroye [50] that no classifier can learn with a given rate for all distributions P. However several authors propose different rates reached by restricting the class of joint distributions. Pioneering works of Vapnik [140, 141] investigate the statistical procedure called Empirical Risk Minimization (ERM). The ERM estimator consists in searching for a classifier that minimizes the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\text{sign}(f(X_i))\neq Y_i\}}, \tag{2.3}$$

over a class of prediction rules $\mathcal{F}$, where $\mathbb{1}_A$ denotes the indicator function of the set A. If we suppose that the class of decision rules $\mathcal{F}$ has finite VC dimension, ERM reaches the parametric rate $n^{-\frac{1}{2}}$ in (3.3) when f^* belongs to the class $\mathcal{F}$. Moreover, if P is noise-free (i.e. $R(f^*) = 0$), the rate becomes n^{-1}. This is a fast rate.
More recently, Tsybakov [136] describes intermediate situations using a margin assumption. This assumption adds a control on the behavior of the conditional probability function η at the level $\frac{1}{2}$ (see (2.10) below). Under this condition, Tsybakov [136] gets minimax fast rates of convergence for classification with ERM estimators over a class $\mathcal{F}$ with controlled complexity (in terms of entropy). These rates depend on two parameters: the margin parameter and the complexity of the class of candidates f^* (see also Massart and Nédélec [102]). Another study of the behavior of ERM is presented in Bartlett and Mendelson [17].
It is well known, however, that minimizing (3.4) is computationally intractable for many non trivial classes of functions [6]. It comes from the non convexity of the functional (3.4). It suggests that we must use a convex surrogate Φ for the loss. The main idea is to minimize an empirical Φ-risk

$$A_n^{\Phi}(f) = \frac{1}{n}\sum_{i=1}^{n} \Phi(Y_i f(X_i)),$$

over a class $\mathcal{F}$ of real-valued functions. Then $\hat{f}_n = \text{sign}(\hat{F}_n)$ where $\hat{F}_n \in \text{Arg}\min_{f\in\mathcal{F}} A_n^{\Phi}(f)$ has a small excess risk. Recently a number of methods have been proposed, such as boosting [55] or Support Vector Machines. The statistical consequences of choosing a convex

surrogate is well treated by Zhang [152] and Bartlett et al. [14]. In this chapter it is proposed to use the hinge loss $\Phi(v) = (1-v)_+$ (where $(\cdot)_+$ denotes the positive part) as surrogate, i.e. to focus on the SVM algorithm.

SVM was first proposed by Boser et al. [28] for pattern recognition. It consists in minimizing a regularized empirical Φ-risk over a Reproducing Kernel Hilbert Space (RKHS for short in the sequel). Given a training set D_n, the SVM optimization problem without offset can be written:

$$\min_{f\in\mathcal{H}_K} \left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i,f(X_i)) + \alpha_n \|f\|_K^2 \right), \tag{2.4}$$

where in the sequel:

1. The functional l is called the hinge loss and is now written $l(y,f(x)) = (1-yf(x))_+$. The first term of the minimization (2.4) is then the empirical Φ-risk A_n^Φ for $\Phi(v) = (1-v)_+$.
2. The space $\mathcal{H}_K$ is a RKHS with reproducing kernel K. Under some mild conditions over K, it consists of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ or $\mathbb{C}$ with the reproducing property:

$$\forall f \in \mathcal{H}_K, \forall x \in \mathcal{X}, f(x) =< K(x,\cdot), f >_{\mathcal{H}_K} .$$

   Recall that every positive definite kernel has an essentially unique RKHS [5].
3. The sequence α_n is a decreasing sequence that depends on n. This smoothing parameter has to be determined explicitly. Such a problem will be studied in this work.
4. The norm $\|.\|_K$ is the norm associated to the inner product in the Hilbert space $\mathcal{H}_K$.

For a survey on this kernel method we refer to Cristianini and Shawe-Taylor [43].

This algorithm is at the heart of many theoretical considerations. However, its good practical performances are not yet completely understood. The study of statistical consistency of the algorithm and approximation properties of kernels can be found in Steinwart [120] or more recently in Steinwart [122]. Blanchard et al. [24] propose a model selection point of view for SVM. Finally, several authors provide learning rates to the Bayes for SVM [147, 146, 126]. In these papers, both approximation power of kernels and estimation results are presented. Wu and Zhou [147] state slow rates (logarithmic with the sample size) for SVM using a Gaussian kernel with fixed width. It holds under no margin assumption for Bayes rule with a given regularity. Steinwart and Scovel [126] give, under a margin assumption, fast rates for SVM using a decreasing width (which depends on the sample size). An additional geometric hypothesis over the joint distribution is necessary to get a control of the approximation using Gaussian kernels.

These results focus on SVM using Gaussian kernels. The goal of this work is to clarify both practical and theoretical performances of the algorithm using two different classes of kernels. In a first theoretical part, we consider a family of kernels generating Sobolev spaces as RKHS. It gives an alternative to the extensively studied Gaussian kernels. We quantify the approximation power of these kernels. It depends on the regularity of the Bayes prediction rule in terms of Besov space. Then under the margin assumption, we give learning rates of convergence for SVM using Sobolev spaces. It holds by choosing optimally the tuning parameter α_n in (2.4). This choice strongly depends on the regularity assumption

over the Bayes and the margin assumption. As a result, it is non-adaptive. Then we turn out into more practical considerations. Following Lecué [81], we give a procedure to construct directly from the data a classifier with similar statistical performances. It uses a method called aggregation with exponential weights. Finally, we show practical performances of this aggregate and compare it with a similar classifier using Gaussian kernels and results of Steinwart and Scovel [126].

The chapter is organized as follows. In Section 2, we give statistical performances of SVM using Sobolev spaces. Section 3 presents the adaptive procedure of aggregation and show the performances of the data-dependent aggregate. This procedure does not damage the learning rates stated in Section 2. We show practical experiments in Section 4 and conclude in Section 5 with a discussion. Section 6 is devoted to the proofs.

2.2 Statistical Performances

As a regularization procedure, minimization (2.4) generates two types of errors: the estimation error and the approximation error. The use of a finite sample size produces the estimation error. The approximation error can be seen as the distance between the hypothesis space and the Bayes decision rule. It comes from the use of a RKHS of continuous functions in the minimization whereas the Bayes is not continous. The first one is random and depends on the fluctuation of the training set. The second one is deterministic and depends on the size of the RKHS. We can see coarselly that these errors are antagonist. Theorem 2.3 gives a choice of the regularization parameter α_n that makes the trade-off between these two errors.
For the estimation error, we will state an oracle-type inequality of the form:

$$\mathbb{E}R_l(\hat{f}_n,f^*) \leq C \inf_{f\in\mathcal{H}_K} \left(R_l(f,f^*) + \alpha_n\|f\|_K^2\right) + \epsilon_n, \tag{2.5}$$

where $R_l(f,f^*) := \mathbb{E}_P l(Y,f(X)) - \mathbb{E}_P l(Y,f^*(X))$ is the excess l-risk of f. The term ϵ_n must be a residual term and satisfies:

$$\epsilon_n \leq C' \inf_{f\in\mathcal{H}_K} \left(R_l(f,f^*) + \alpha_n\|f\|_K^2\right),$$

where $C' > 0$. Inequality (2.5) deals with the estimation error. It depends on the complexity of the class of functions $\mathcal{H}_K$ and the difficulty of the problem.
Hence it remains to control the infimum in the right hand side (RHS for short) of (2.5). Steinwart and Scovel [126] define the approximation error function as:

$$a(\alpha_n) := \inf_{f\in\mathcal{H}_K} \left(R_l(f,f^*) + \alpha_n\|f\|_K^2\right). \tag{2.6}$$

This function represents the theoretical version of the empirical minimization (2.4). It depends on the chosen $\mathcal{H}_K$ and the behaviour of α_n as a function of n.

Using this approach, Steinwart and Scovel [126] study the statistical performances of SVM minimization (2.4) with the parametric family of Gaussian kernels. For $\sigma \in \mathbb{R}$, we define the Gaussian kernel $K_\sigma(x,y) = \exp\left(-\sigma^2\|x-y\|^2\right)$ on the closed unit ball of $\mathbb{R}^d$ (denoted $\mathcal{X}$). The parameter σ^{-1} is called the width of the Gaussian kernel. In this paper, under a margin assumption and a geometric assumption over the distribution, they

state fast learning rates for SVM. These rates hold under some specific choices of tuning parameters recalled in Sect. 4. Following Lecué [81], we will use this result and more precisely these choices of tuning parameters to implement the aggregate using Gaussian kernels.

2.2.1 Sobolev Smooth Kernels

In this chapter we propose to deal with other class of kernels than the Gaussian kernels. First we need to introduce some notations. Let us consider the set of complex-valued and integrable (resp. square-integrable) functions on $\mathbb{R}^d$ denoted as $L^1(\mathbb{R}^d)$ (resp. $L^2(\mathbb{R}^d)$). On this set, we define the Fourier transform of f to be:

$$\hat{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-i\omega.t} dt, \forall \omega \in \mathbb{R}^d,$$

where $x.y$ denotes the usual scalar product of $\mathbb{R}^d$ between two points $x,y \in \mathbb{R}^d$.
After the usual extension from $L^1(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$ with Plancherel, this operator is an isometry on $L^2(\mathbb{R}^d)$. It allows us to define, for any $s \in \mathbb{R}^+$, the Sobolev space $\mathcal{W}_s^2$ (often called fractional Sobolev space) as the following subspace of $L^2(\mathbb{R}^d)$ [92]:

$$\mathcal{W}_s^2 := \{f \in L^2(\mathbb{R}^d) : \|f\|_s^2 = \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 (1 + \|\omega\|^2)^s d\omega < \infty\}. \tag{2.7}$$

We refer to Triebel [132] or Adams [1] for a large study of this well-known functional space. With such a norm, $\mathcal{W}_s^2$ is a Hilbert space endowed with the inner product defined as:

$$< f,g >_s = \int_{\mathbb{R}^d} \hat{f}(\omega) \overline{\hat{g}(\omega)} (1 + \|\omega\|^2)^s d\omega,$$

where $\bar{z}$ is the complex conjugate of z in $\mathbb{C}$. Moreover it is a Hilbert space of continuous functions for any $s > \frac{d}{2}$ (due to the embedding between $\mathcal{W}_s^2$ and $C(\mathbb{R}^d)$ for any $s > \frac{d}{2}$). It can be seen as a RKHS.

In this framework, a kernel is a symmetric and positive definite function $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{C}$. For $r \in \mathbb{R}^+$, a kernel K_r will be called *Sobolev smooth kernel* with exponent $r > d$ if the associated RKHS $\mathcal{H}_{K_r}$ is such that

$$\mathcal{H}_{K_r} = \mathcal{W}_{\frac{r}{2}}^2,$$

where $\mathcal{W}_{\frac{r}{2}}^2$ is defined in (2.7). The restriction $r > d$ ensures that the RKHS consists of continuous functions from $\mathbb{R}^d$ to $\mathbb{C}$. Corollary 2.1 provides a way of constructing such a kernel.

We say that a kernel K is a *translation invariant kernel* (or RBF kernel), if for all $x,y \in \mathbb{R}^d$,

$$K(x,y) = \Phi(x - y) \tag{2.8}$$

for a given $\Phi : \mathbb{R}^d \to \mathbb{C}$. Function Φ is often called RB function for Radial Basis function. The most popular example of translation invariant kernel is the Gaussian kernel $K_\sigma(x,y) = \exp(-\sigma^2 \|x - y\|^2)$. This kernel is not a Sobolev smooth kernel (see below).

Under suitable assumptions on Φ, the following theorem gives a Fourier representation of a RKHS associated to a translation invariant kernel. The proof is given in Section 6.

Theorem 2.1 *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be a translation invariant kernel where in (2.8) Φ belongs to $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and such that $\widehat{\Phi}$ is integrable. Then the RKHS associated to K can be written*

$$\mathcal{H}_K = \{f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega < \infty \text{ and } \hat{f} = 0 \text{ on } \mathbb{R}^d \backslash S\}$$

with the inner product

$$< f,g >_K = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega,$$

where $S := \{\omega \in \mathbb{R}^d : \widehat{\Phi}(\omega) \neq 0\}$ is the support of $\widehat{\Phi}$.

Sufficient conditions to have a Sobolev smooth kernel are:

Corollary 2.1 *Let K satisfying assumptions of Theorem 2.1. Suppose moreover that there exist constants C,$c > 0$ and a real number $s > \frac{d}{2}$ such that*

$$\widehat{\Phi}(\omega) = \frac{C}{(c + \|\omega\|^2)^s}, \forall \omega \in \mathbb{R}^d. \tag{2.9}$$

Then K is a Sobolev smooth kernel with exponent $r = 2s > d$.

In Section 5 we propose an example of Sobolev smooth kernel and use it into the SVM procedure.

Remark 2.1 (Gaussian kernels are not Sobolev smooth) *Theorem 2.1 can be used to define Gaussian kernels in terms of Fourier transform. Indeed, the Gaussian kernel defined above is a translation invariant kernel with RB function $\Phi(x) = \exp(-\sigma^2\|x\|^2)$. Its Fourier transform is given by*

$$\widehat{\Phi}(\omega) = \frac{1}{(\sqrt{2}\sigma)^d} \exp(-\frac{\|\omega\|^2}{4\sigma^2}).$$

Then Φ satisfies assumptions of Theorem 2.1. The Fourier representation of $\mathcal{H}_\sigma$ is given by:

$$\mathcal{H}_\sigma = \{f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 \sigma^d \exp(\frac{\|\omega\|^2}{4\sigma^2}) d\omega < \infty\}.$$

From definition (2.7), it is clear that $\mathcal{H}_\sigma$ is not a Sobolev space. This integral representation of a Gaussian RKHS illustrates the smoothness of functions $f \in \mathcal{H}_\sigma$. Indeed we can see trivially that $\mathcal{H}_\sigma \subset \mathcal{H}_{K_r}$ for any fixed σ,$r > 0$ (because the Fourier transform of Φ is rapidly decreasing in this case). Moreover the parameter σ can be seen as a regularization parameter: the fewer is σ, the smoother are the functions in $\mathcal{H}_\sigma$. More precisely, $\sigma < \sigma'$ entails $\mathcal{H}_\sigma \subset \mathcal{H}_{\sigma'}$.

2.2.2 Approximation Efficiency of Sobolev Smooth Kernels

Here we are interested in approximation properties of $\mathcal{H}_{K_r}$. We aim at bounding the approximation function $a(\alpha_n)$ defined in (2.6) for the procedure (2.4). The best case appears when $f^* \in \mathcal{H}_K$. Then we get coarselly $a(\alpha_n) \leq C\alpha_n$ where C is an absolute constant. This case is not realizable considering a continuous RKHS since the Bayes classifier is not. In this chapter, we get a control of the approximation function when f^* does not belong to the RKHS. Theorem 2.2 provides such a result using a Sobolev smooth kernel.

Theorem 2.2 *Consider the approximation function $a(\alpha_n)$ defined in (2.6), with Sobolev smooth kernel K_r such that $r > 2s > 0$. Suppose P_X satisfies $\frac{dP_X}{dx} \leq C_0$.*
Then if $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$, we have:

$$a(\alpha_n) \leq C_0^{\frac{r-2s}{r-s}} \|f^*\|_{s2\infty}^{\frac{r}{r-s}} \alpha_n^{\frac{s}{r-s}},$$

where $\|.\|_{s2\infty}$ defines the norm in the Besov space $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.
The proof is detailled in Section 6 where we define explicitly Besov spaces $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.

Remark 2.2 *(*Bayes regularity*) Here we get a control of the approximation function under an assumption on the smoothness of the Bayes classifier. Of course large values of s are not possible because $f^*(x) = \text{sign}(2\eta(x) - 1)$ is not even continuous (except for the trivial case $\eta(x) < \frac{1}{2}$ a.s. or $\eta(x) > \frac{1}{2}$). More precisely, the Besov space $\mathcal{B}^p_{s,q}(\mathbb{R}^d)$ is included in the space of continuous functions for $s > \frac{d}{p}$ and $q > 1$. Here $p = 2$ then parameter s must satisfy $s < \frac{d}{2}$ to have $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$. In Remark 2.7 we give an example of Bayes rule verifying this smoothness assumption.*

Remark 2.3 *(*comparison with Steinwart and Scovel, 2007*) Steinwart and Scovel [126] propose a same type of result using Gaussian kernels. Under a geometric assumption over the distribution, they get*

$$a(\alpha_n) \leq C\alpha_n^{\frac{\gamma}{\gamma+1}},$$

where γ is the geometric noise exponent. Here we propose a same type of result under a regularity assumption over the possible f^*. Theorem 2.5 in Section 6 shows that this result can be generalized to any other kernel, using interpolation spaces.

2.2.3 Learning Rates

In this work, we restrict the class of considered distributions P. We add a control on the local slope of the conditional probability function η at the level $\frac{1}{2}$. This margin assumption (we often call $|\eta - \frac{1}{2}|$ the margin) is originally due to Mammen and Tsybakov [95] for discriminant analysis. We will use throughout this chapter the following formulation: we say that P has *margin parameter* $q > 0$ if there exists a constant $c_0 > 0$ such that

$$\mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^q, \tag{2.10}$$

for all sufficiently small t.
According to Boucheron et al. [29], this hypothesis is equivalent to the low noise or margin assumption in Tsybakov [136]. Best situation for learning appears when the conditional

probability makes a jump at the level $\frac{1}{2}$. Hence (2.10) holds true for any positive q. It corresponds to a margin parameter $q = +\infty$, i.e. $\kappa = 1$ in the sense of Tsybakov [136].
Finally, last step of modelling consists in clipping the solution of minimization (2.4). For any classifier $\hat{f}$, we hence define the *clipped version* $\hat{f}^C$ with values in $[-1,1]$ by

$$\hat{f}^C(x) = \begin{cases} -1 \text{ for } x : \hat{f}(x) < -1, \\ f(x) \text{ for } x : \hat{f}(x) \in [-1,1], \\ 1 \text{ for } x : \hat{f}(x) > 1. \end{cases}$$

This operation does not modify the classification property of $\hat{f}$ since $\text{sign}(\hat{f}) = \text{sign}(\hat{f}^C)$. It produces classifiers with bounded norm $\|.\|_\infty$. It appears in several works [11, 125]. We stress that the clip does not modify the algorithm. It is done after the training as a part of the theoretical study of the algorithm. We are now on time to state the main result of this section.

Theorem 2.3 *Let P be a distribution over $\mathbb{R}^d \times \{-1,1\}$ such that P_X satisfies $\frac{dP_X}{dx} \leq C_0$ and (2.10) holds for $q \in [0, +\infty]$. Let $s > 0$ and suppose $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.*
Consider the SVM minimization (2.4) with Sobolev smooth kernel K_r, with $r > 2s \vee d$, built on the i.i.d. sequence $(X_i,Y_i), i = 1 \dots n$ according to P.
If we choose α_n such that

$$\alpha_n \sim n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}, \tag{2.11}$$

then there exists a constant C which depends on r,s,d,c_0,q and C_0 such that

$$\mathbb{E}R(\hat{f}^C_n, f^*) \leq Cn^{-\gamma(q,s)},$$

where

$$\gamma(q,s) = \frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}. \tag{2.12}$$

The proof of this theorem is given in Section 6.

Remark 2.4 *(*FAST RATES*) Rate (2.12) is a fast rate (i.e. faster than $n^{-\frac{1}{2}}$) if $\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)} > \frac{1}{2}$. In particular, for $q = +\infty$, it corresponds to $s > \frac{rd}{r+d}$. The presence of fast rates depends on the regularity of the Bayes classifier. Unfortunately the behaviour of f^* (see Remark 2.2) entails $s < \frac{d}{2}$. As a result, $\frac{sr}{sr+d(r-s)} < \frac{1}{2}$ and fast rates can not be reached.*

Remark 2.5 *(*COMPARISON WITH STEINWART AND SCOVEL, 2007*) This theorem gives performances of SVM using a fixed kernel. On the contrary, according to Steinwart and Scovel [126], the bandwidth of the kernel has to be chosen as a function of n. Nevertheless, rates of convergence are fast for sufficiently large geometric noise parameter. Here we cannot get fast rates for reasonable assumption over f^*.*

Remark 2.6 *(*OPTIMAL SMOOTHING PARAMETER*) Theorem 2.3 provides a particular choice of α_n to reach rates* (2.12)*. Other definitions for the sequence α_n give other rates of convergence. We only mention the best possible rates. It holds for a regularization parameter optimizing the statistical performances. Indeed, α_n in* (2.11) *makes the balance between the estimation error and the approximation error.*

Remark 2.7 (EXAMPLE) *Consider the one-dimensional case where $\mathcal{X} = \mathbb{R}$. Suppose f^* is such that:*

$$card\{x \in \mathbb{R} : f^* \text{ jumps at } x\} = N < \infty. \tag{2.13}$$

It means that the Bayes rule changes only a finite number of times over the real line. Using standard analysis, we get

$$\|f^*\|_{TV} = \int_{\mathbb{R}} |Df^*(x)|dx = 2N,$$

where Df^* is the generalized derivative of f^*. Moreover, for any f, $|\hat{f}(\omega)| \leq \|f\|_{VT}/|\omega|$. Then f^* belongs to $\mathcal{W}_{s,2}$ only for $s < 1/2$. Finally, with basic properties of Besov spaces [132], we have $\mathcal{W}_{s,2} = \mathcal{B}^2_{s,2} \subset \mathcal{B}^2_{s,\infty}$.
Consequently, f^* verifying (3.23) *belongs to $\mathcal{B}^2_{s,\infty}$ for any $s < \frac{1}{2}$. If we consider a margin parameter $q = +\infty$, we hence cannot reach the rate of convergence*

$$n^{-\frac{r}{3r-1}},$$

which corresponds to a regularity $s = \frac{1}{2}$ in the Besov space. Then the SVM using Sobolev smooth kernel H_{K_r} with $r > 1$ cannot learn with fast rate in this simple case.

2.3 Aggregation

Theorem 2.3 provides the optimal value of α_n to reach rates of convergence (2.12) in the context of Sobolev spaces. It holds under two ad-hoc assumptions: a margin assumption over the distribution and a regularity assumption over the Bayes rule. Hence the choice of the smoothing parameter depends on two unknown parameters: the margin parameter q and the exponent s in the Besov space. Consequently the classifier $\hat{f}_n$ of Theorem 2.3 cannot be constructed from the data. It is called non-adaptive.

The goal of this section is to overcome this difficulty. We propose a classifier that adapts automatically both to the margin and to regularity. In other words, we will build a decision rule from D_n which does not depend on the unknown parameters s and q. Moreover, Theorem 2.4 shows that this procedure of adaptation will not damage the learning rates of Theorem 2.3.
We use a technique called aggregation [107, 149]. We apply the method presented in Lecué [81] to our framework of Sobolev smooth kernel. It consists of splitting the data into two parts: the first part in used to construct a family of classifiers. The second part is used to make a convex combination of these classifiers. We obtain an adaptive decision rule which mimics the best one over the family. Let us first describe the method.
Denote $D^1_{n_1}$ (resp. $D^2_{n_2}$) the first subsample of size n_1 (resp. second subsample of size n_2)

with $n_1 + n_2 = n$. The choice of n_1 and n_2 will be discussed later. We construct a set of classifiers $(\hat{f}^{\alpha}_{n_1})_{\alpha \in \mathcal{G}(n_2)}$ defined by $\hat{f}^{\alpha}_{n_1} = sign\left(\hat{F}^{\alpha}_{n_1}\right)$ where

$$\hat{F}^{\alpha}_{n_1} := \arg\min_{f \in \mathcal{H}_{Kr}} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} l(Y_i, f(X_i)) + \alpha \|f\|^2_K \right).$$

The grid $\mathcal{G}(n_2)$ is defined by

$$\mathcal{G}(n_2) := \left\{ \alpha_k = n_2^{-\phi_k} : \phi_k = \frac{1}{2} + k\Delta^{-1}, k = 0, \dots, \lfloor \frac{(2r-d)\Delta}{2d} \rfloor \right\}, \tag{2.14}$$

with $\Delta = n_2^b$ for some $b > 0$. We hence have $\left\lfloor \frac{(2r-d)\Delta}{2d} \right\rfloor + 1$ classifiers to aggregate.
The procedure of aggregation uses the second subsample $D^2_{n_2}$ to construct a convex combination with exponential weights. Namely, the aggregate $\tilde{f}_n$ is defined by

$$\tilde{f}_n = \sum_{\alpha \in \mathcal{G}(n_2)} \omega^{(n)}_{\alpha} \hat{f}^{\alpha}_{n_1}, \tag{2.15}$$

where

$$\omega^{(n)}_{\alpha} = \frac{\exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}^{\alpha}_{n_1}(X_i) \right)}{\sum_{\alpha' \in \mathcal{G}(n_2)} \exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}^{\alpha'}_{n_1}(X_i) \right)}.$$

We hence have the following result.

Theorem 2.4 *Consider the classifier $\tilde{f}_n$ defined in (2.15) where $n_2 = \lceil a \frac{n}{\log n} \rceil$ for $a > 0$. Let K a compact of $(0,\infty)^2$. Then there exists a constant C which depends on r,d,c_0,K,a,b,L and C_0 such that for all $(q,s) \in K$*

$$\sup_{P \in \mathcal{Q}_{q,s}} \mathbb{E} R(\tilde{f}_n, f^*) \leq C n^{-\gamma(q,s)},$$

where

$$\gamma(q,s) = \frac{rs(q+1)}{s(r(q+2)-d) + d(r-s)(q+1)}$$

and $\mathcal{Q}_{q,s}$ is the set of distributions P satisfying $\frac{dP_X}{dx} < C_0$, (2.10) with parameter q and such that $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d, L) = \{f \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d) : \|f\| \leq L\}$.

Remark 2.8 *Same rates as in Theorem 2.3 are attained. Here we deal with an implementable classifier. In Section 5 we sum up practical performances of this aggregate.*

Remark 2.9 *Instead of aggregating a power of n classifiers, only $\log n$ classifiers are enough to obtain this result. Lecué [80] states an oracle inequality such as (2.23) without any restriction on the number of estimators to aggregate.*

Remark 2.10 *(*AVERAGE OF AGGREGATES*) This method supposes, for a given n_1 and n_2, an arbitrary choice for the subsample $D_{n_1}^1$ and $D_{n_2}^2$. However we can use different splits of the training set. We get an average of aggregates, namely*

$$\bar{f}_n = \frac{1}{M}\sum_{k=1}^{M} \tilde{f}_n^k.$$

It does not depend on a particular split. Each $\tilde{f}_n^k$ is defined in (2.15) *for the split number k. With [81, Theorem 2.4], this average satisfies the oracle inequality* (2.23)*. Then Theorem 2.4 holds for $\bar{f}_n$ for any family of M splits, for $M \leq \mathcal{C}_n^{n_1}$.*

2.4 Practical Experiments

We now propose experiments illustrating performances of the aggregate of Section 3. We study SVM classifiers using both Sobolev spaces and Gaussian kernels. The aggregates were implemented in **R** using the free library *kernlab*. It contains implementations of support vector machines. For a description of this package for kernel-based learning methods in **R**, we refer to Karatzoglou et al. [74]. We use real world datasets from benchmark repository [1] used by Rätsch et al. [111]. We consider 9 datasets called "Banana", "Titanic", "Thyroid", "Diabetis", "Breast-Cancer", "Flare-solar", "Heart", "Image" and "Waveform". These datasets are explained in Table 1. For each dataset, we have several realizations of training and test set. The dimension of the input space is denoted by d whereas the number of observations for the training set is n. It follows the notations used in the previous sections. On each realization, we train and test our classifiers. The results presented in Table 2,3,4 show the average test errors over these realizations and the standard deviations.

Dataset	d	n	test sample	realizations
Banana	2	400	4900	100
Titanic	3	150	2051	100
Thyroid	5	140	75	100
Diabetis	8	468	300	100
Breast-cancer	9	200	77	100
Flare-solar	9	666	400	100
Heart	13	170	100	100
Image	18	1300	1010	20
Waveform	21	400	4600	100

Table 1: Description of the datasets.

2.4.1 SVM Using Sobolev Smooth Kernel

The first step is to pick up a Sobolev smooth kernel. Consider the following class of RBF kernels, with Radial Basis function Φ:

$$K(x,y) = \Phi(x-y) = \exp\left(-\sigma\|x-y\|\right), \forall \sigma \in \mathbb{R}. \tag{2.16}$$

1. available online at this address http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm.

For a given σ, this kernel is called a Laplacian kernel. It is clear that $\Phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Recall the Fourier transform of $\Phi : \mathbb{R}^d \mapsto \mathbb{R}$ (see Williamson et al. [145]):

$$\widehat{\Phi}(\omega) = 2^{\frac{d}{2}} \pi^{-\frac{1}{2}} \Gamma(\frac{d}{2}+1) \frac{\sigma}{(\sigma^2 + \|\omega\|^2)^{\frac{d+1}{2}}}, \forall \omega \in \mathbb{R}^d,$$

where $\Gamma(x) = \int_{\mathbb{R}^+} e^{-t} t^{x-1} dt$ is the Gamma function.
With Corollary 2.1, for any fixed σ, the Laplacian kernel defined in (2.16) is a Sobolev smooth kernel with exponent $r = d+1$. It satisfies assumptions of Theorem 2.3 and can be used in the implementation of the algorithm.
It is worth noticing that the parameter σ is constant. If we take a significantly small value for σ, as $\sigma = n^{-u}$, $u > 0$, (2.9) holds for C and c depending on n. Thus Corollary 2.1 does not hold. To avoid this problem, we choose in our aggregation step using this class of kernels a constant $\sigma = 5$. In the sequel the Laplacian kernel used is precisely $K(x,y) = \exp(-5\|x-y\|)$.
Table 2 shows the first experiments. For each realization of training set, we use previous section to build

- the set of classifiers $(\hat{f}^{\alpha}_{n_1})$ for α belonging to $\mathcal{G}(n_2)$;
- exponential weights $\omega^{(n)}_{\alpha}$ to deduce aggregate $\tilde{f}_n$.

Recall the definition of $\mathcal{G}(n_2)$ in this case:

$$\mathcal{G}(n_2) := \left\{ \alpha_k = n_2^{-\phi_k} : \phi_k = \frac{1}{2} + k\Delta^{-1}, k = 0, \dots, \lfloor \frac{(2r-d)\Delta}{2d} \rfloor \right\},$$

where $\Delta = n_2^b$. We take $b = 1$ in the construction. Instead of a step $\Delta = n_2^b$, it is possible to take only $\Delta = \log n_2$ (see Remark 2.9) . The value of b governs the size of the grid. The cardinal is given in Table 2 for each dataset. Note that growing b does not improve significantly the performances whereas it adds computing time. Indeed, whatever b, $\mathcal{G}(n_2)$ is contained in this case into $[n_2^{-\frac{d+1}{d}}, n_2^{-\frac{1}{2}}]$. This location is motivated by Theorem 2.3, namely equation (2.11). The value of b only deals with the distance between each point of $\mathcal{G}(n_2)$. It does not change the location of the grid.
Table 2 relates the average test errors and the standard deviations. We first collect the performances of the family of weak estimators $(\hat{f}^{\alpha}_{n_1})$,$\alpha \in \mathcal{G}(n_2)$. We mention in order the performances of the worst estimator, the mean over the family and the best over the family. It gives an idea of the estimators to aggregate. Then the performances of the aggregate using exponential weights are given in the last column.

Dataset	card$\mathcal{G}(n_2)$	max	mean	min	Laplace Aggregate
Banana	102	11.41±0.58	11.33± 0.57	11.12±0.59	11.31± 0.57
Titanic	38	22.80±1.16	22.80±1.14	22.77±1.13	22.77±1.13
Thyroid	31	5.97±2.61	5.45±2.56	4.77±2.63	5.45±2.68
Diabetis	72	29.56±2.03	28.40±2.00	27.33±1.96	28.34±2.27
Breast-cancer	35	35.10±5.34	33.26±5.06	31.49±5.05	32.74±5.16
Flare-solar	95	35.97±1.94	35.68±1.90	35.52±1.90	35.69±1.93
Heart	29	22.38±3.97	22.11±3.98	21.76±3.99	22.12±3.98
Image	152	4.35±0.87	4.06±0.74	3.79±0.74	3.95±0.74
Waveform	56	14.51±0.70	14.16±0.67	13.78±0.65	14.12±0.72

Table 2: Performances using Laplacian kernel.

Note that the amplitude in the family is not very important. It may be explain by its construction. Indeed, $\mathcal{G}(n_2)$ is motivated by Theorem 2.3, which gives the location of the grid (see above). This family has a mathematical justification. The test errors of the aggregate are located between the average over the family and the oracle of the family.
A temperature parameter usually appears in aggregation methods. It governs the variations of values $\omega_\alpha^{(n)}$, for $\alpha \in \mathcal{G}(n_2)$. In Table 2 the weak classifiers have almost the same performances. This could explain why no temperature parameter is needed here.

2.4.2 SVM Using Gaussian Kernels

Here we focus on the parametric class of Gaussian kernels $K_\sigma(x,y) = \exp\left(-\sigma^2\|x-y\|^2\right)$, for $\sigma \in \mathbb{R}$. We build an aggregate made of a convex combination of Gaussian SVM classifiers. In this case, the construction is not exactly the same. It comes from Steinwart and Scovel [126]. In this paper, they introduce a geometric noise assumption. This hypothesis deals with the concentration of the measure $|2\eta - 1|P_X$ near the decision boundary. It allows to control the approximation function (2.6). According to Steinwart and Scovel [126], suppose that the probability distribution P has a geometric noise $\gamma > 0$ and assumption (2.10) holds with margin parameter $q > 0$. Then if we choose

$$\alpha_n = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} \text{ if } \gamma \leq \frac{q+2}{2q}, \\ n^{-\frac{2(\gamma+1)(q+1)}{2\gamma(q+2)+3q+4}}, \text{ otherwise} \end{cases}$$

the solution of (2.4) using a Gaussian kernel K_σ with $\sigma = \alpha_n^{-\frac{1}{(\gamma+1)d}}$ learns with rates

$$\begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\epsilon} \text{ if } \gamma \leq \frac{q+2}{2q}, \\ n^{-\frac{2\gamma(q+1)}{2\gamma(q+2)+3q+4}+\epsilon} \text{ otherwise,} \end{cases}$$

for all $\epsilon > 0$.
We can see that the variance of the Gaussian kernels is not fixed. It has to be chosen as a function of the geometric noise exponent. As a result, parameter σ must be considered in the aggregation procedure, as the smoothing parameter α. It gives a two-dimensional grid of Gaussian SVM of the following form [81]:

$$\mathcal{N}(n_2) = \left\{(\sigma_{n_2,\phi},\alpha_{n_2,\psi}) = (n_2^{\phi/d},n_2^{-\psi}) : (\phi,\psi) \in \mathcal{M}(n_2)\right\}$$

where

$$\mathcal{M}(n_2) = \left\{(\phi_{n_2,p_1},\psi_{n_2,p_2}) = \left(\frac{p_1}{2\Delta},\frac{p_2}{\Delta}+\frac{1}{2}\right) : p_1 = 1,\dots,2\lfloor\Delta\rfloor; p_2 = 1,\dots,\lfloor\Delta/2\rfloor\right\},$$

for $\Delta = n_2^b$. Thus we have more classifiers to aggregate and needs more time to run. As a consequence, we choose constant $b = 0.5$ in our experiments. Such as the Sobolev case, the number of classifiers to aggregate is mentioned in Table 3 for each dataset.
Table 3 relates the generalization performances of the classifiers over the test samples. We

first give the performances of the family of Gaussian SVM (namely the worst, the mean and the oracle over the family). The performances of the aggregate using exponential weights are given in the last column.

Dataset	card$\mathcal{N}(n_2)$	max	mean	min	Gaussian aggregate
Banana	100	17.29± 3.08	12.27±0.89	10.85±0.63	11.43±0.84
Titanic	36	23.15±1.30	22.81±1.00	22.49±0.78	22.57±0.79
Thyroid	36	8.19±2.63	6.76±2.72	5.59±2.94	6.31±2.97
Diabetis	100	29.82±1.98	28.19±1.84	26.39±1.85	27.80±2.06
Breast-cancer	42	34.83±5.12	32.76±4.82	30.48±4.61	32.13±4.77
Flare-solar	144	39.06±1.92	36.01±1.54	34.09±1.69	34.87±1.82
Heart	42	23.1±3.80	22.60±3.71	21.99±3.59	22.62±3.77
Image	256	7.79±1.00	6.33±0.83	5.30±0.73	5.66±0.74
Waveform	100	15.41±0.80	15.08±0.78	14.72±0.77	15.04±0.79

Table 3: Performances using Gaussian kernels.

In this case the generalization errors in the family are more disparate. It comes from a two-dimensional grid of parameters. The performances of the Gaussian aggregate, as above, are located between the average of weak estimators and the best among the family.

2.4.3 Comparison With Rätsch et al. [111]

Table 4 combines the performances of the aggregates using Laplacian kernel and Gaussian kernels. The errors are comparable. Gaussian kernels and Laplacian kernel lead to similar performances. Then we mention the generalization errors of Rätsch et al. [111]. Rätsch et al. [111] proposes generalizations of the original Adaboost algorithm. However, extensive simulations are presented like experimental results for SVM using Gaussian kernels. The choice of the parameters (α_n,σ) are done by 5-fold-cross validation thanks to several training datasets. This approach has not any mathematical justification. Moreover their mathematical programming problems are distributed over 30 computers. We only use last column to have an idea of reasonable average test errors for these datasets.

Dataset	Laplace Aggregate	Gaussian Aggregate	Rätsch et al. [111]
Banana	11.31± 0.57	11.43±0.84	11.53±0.66
Titanic	22.77±1.13	22.57±0.79	22.42±1.02
Thyroid	5.45±2.68	6.31±2.97	4.80±2.19
Diabetis	28.34±2.27	27.80±2.06	23.53±1.76
Breast-cancer	32.74±5.16	32.13±4.77	26.04±4.74
Flare-solar	35.69±1.93	34.87±1.82	32.43±1.82
Heart	22.12±3.98	22.62±3.77	15.95±3.26
Image	3.95±0.74	5.66±0.74	2.96±0.6
Waveform	14.12±0.72	15.04±0.79	9.88±0.83

Table 4: Comparison with Rätsch et al. [111].

Table 4 illustrates good resistence of our aggregates when the dimension is not too large. Nevertheless, in the last columns, our estimators fail. This may have a theoretical explanation. In Theorem 2.3 and 2.4, a constant C appears in the upper bounds. This constant in

front of the rates of convergence depends on the dimension of the input space. Increasing d grows this constant C and may affect the performances. Moreover, the choice of the parameters in Rätsch et al. [111] are done with several training sets. In our approach, for each realization of a training set, we construct an adaptive classifier using n observations. The amount of information used is not the same. It may also explain this difference.

2.5 Conclusion

This chapter gives some insights into SVM algorithm, from both theoretical and practical point of view. We have tackled several important questions such as its statistical performances, the role of the kernel and the choice of the tuning parameters.

The first part of the chapter focuses on the statistical performances of the method. In this study, we consider Sobolev smooth kernels as an alternative to the Gaussian kernels. It allows us to bring out a functional class of Bayes rule (namely Besov spaces $\mathcal{B}^2_{s,\infty}$) ensuring good approximation properties for our hypothesis space. Explicit rates of convergence have been given depending on the margin and the regularity (Theorem 2.3). Nevertheless, this result was non-adaptive.

Then it has been necessary to consider the problem of adaptation. The aggregation method appeared suitable in this context to construct directly from the data a competitive decision rule: it has the same statistical performances as the non-adaptive classifier (Theorem 2.4). In this procedure, we use explicitly the theoretical part to choose the scale of tuning parameters. For completeness, we have finally implemented the method and gave practical performances over real benchmark datasets. These practical experiments are to be considered as preliminary. However it shows similar performances for SVM using Gaussian or non-Gaussian kernel. Moreover it illustrates rather well the importance of constructing a classifier with some mathematical background.

2.6 Proofs

This section contains proofs of the results presented in this chapter.

2.6.1 Proof of Theorem 2.1 and Corollary 2.1

We consider a translation invariant kernel $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{C}$ with RB function Φ satisfying assumptions of Theorem 2.1. The following lemma will be useful.

Lemma 2.1 *For any $y \in \mathbb{R}^d$, consider the function $k_y : x \mapsto K(x,y)$ defined in $\mathbb{R}^d$. Then we have the following statements:*

1. $k_y(x) = \overline{\hat{g}_y(x)}$ *where* $g_y(\omega) = e^{i\omega.y}\widehat{\Phi}(\omega)$.
2. $\hat{k}_y(\omega) = e^{-i\omega.y}\widehat{\Phi}(\omega)$.

Proof.

1. $\Phi \in L^2(\mathbb{R}^d)$ hence the inverse Fourier formula allows us to write:

$$\begin{aligned} k_y(x) = \Phi(x-y) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\omega.(x-y)} \widehat{\Phi}(\omega) d\omega \\ &= \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{-i\omega.x} e^{i\omega.y} \widehat{\Phi}(\omega) d\omega} \\ &= \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{-i\omega.x} g_y(\omega) d\omega}. \end{aligned}$$

2. Now using 1. one gets

$$\hat{k}_y(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\omega.x} k_y(x) dx = \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{i\omega.x} \hat{g}_y(x) dx}.$$

Gathering with the inverse Fourier transform of $g_y \in L^2(\mathbb{R}^d)$, we have

$$\hat{k}_y(\omega) = \overline{g_y(\omega)} = e^{-i\omega.y} \widehat{\Phi}(\omega). \square$$

Proof. (of Theorem 2.1)
We write

$$\mathcal{H}_0 = \{f \in L^2(\mathbb{R}^d) : \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega < \infty \text{ and } \hat{f} = 0 \text{ on } S\},$$

with the corresponding norm

$$\|f\|_{\mathcal{H}_0} := \sqrt{\frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega}.$$

We will show that $\mathcal{H}_0$ coincides with $\mathcal{H}_K$.
For a given $y \in \mathbb{R}^d$, from Lemma 2.1 it is clear that $\hat{k}_y(\omega) = 0$ for $\omega \in \mathbb{R}^d \backslash S$. Moreover using again Lemma 2.1:

$$\int_S \frac{|\hat{k}_y(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega = \int_S \widehat{\Phi}(\omega) d\omega < \infty$$

since $\widehat{\Phi}$ is integrable. Then $k_y \in \mathcal{H}_0$ for any $y \in \mathbb{R}^d$.
Now we have to establish that $\mathcal{H}_0$ is a Hilbert space. Following Matache and Matache [103], we can show that, for any $f \in \mathcal{H}_0$:

$$\|\hat{f}\|_1 \leq \sqrt{(2\pi)^{d/2} \|\widehat{\Phi}\|_1} \|f\|_{\mathcal{H}_0} \text{ and } \|\hat{f}\|_2 \leq \sqrt{(2\pi)^{d/2} \|\Phi\|_1} \|f\|_{\mathcal{H}_0},$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the norms in $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$.
Indeed, by Cauchy-Schwarz,

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)| d\omega \leq \sqrt{\frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega} \sqrt{(2\pi)^{d/2} \int_S \widehat{\Phi}(\omega) d\omega}.$$

Moreover, since $\|\widehat{\Phi}\|_\infty \leq \|\Phi\|_1$,

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 d\omega \leq \|\Phi\|_1 \int_S \frac{|\hat{f}(\omega|^2}{\widehat{\Phi}(\omega)} d\omega.$$

Then considering a Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{H}_0$ endowed with $\|\cdot\|_{\mathcal{H}_0}$, $(\hat{f}_n)_{n\in\mathbb{N}}$ will be a Cauchy sequence in both $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$. We conclude with Matache and Matache [103] that $(f_n)_n$ is convergent in $\mathcal{H}_0$. Then $\mathcal{H}_0$ is complete and becomes a Hilbert space endowed with the following inner product:

$$< f,g >_{\mathcal{H}_0} = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega.$$

Finally reproducing property holds. Indeed let $f \in \mathcal{H}_0$. Using again Lemma 2.1:

$$f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\omega.x} \hat{f}(\omega) d\omega = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)}{\widehat{\Phi}(\omega)} \overline{\hat{k}_x(\omega)} d\omega = < f,k_x >_{\mathcal{H}_0} .$$

We have already shown that $\forall x \in \mathbb{R}^d$, $k_x \in \mathcal{H}_0$. As a result, the unicity of the RKHS for a given kernel concludes the proof.□ *Proof. (of Corollary 2.1)*
First we have trivially that $\widehat{\Phi}$ is integrable since $s > \frac{1}{2}$. We can hence apply Theorem 2.1 to have

$$\mathcal{H}_K = \{f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 (c + \|\omega\|^2)^s d\omega < \infty\},$$

since the support of $\widehat{\Phi}$ is $\mathbb{R}^d$. This expression of the RKHS associated to K corresponds, up to a constant, to the Sobolev space $\mathcal{W}_s^2$ defined in (2.7). Then K is a Sobolev smooth kernel with exponent $r = 2s$.□

2.6.2 Proof of Theorem 2.2

First introduce the notion of interpolation space [20]. We restrict ourselves to a description of the real interpolation method. Let $(B,\|.\|_B)$ be a Banach space and $\mathcal{H}$ a Hilbert space dense in B. The Peetre's functional for the couple $(B,\mathcal{H})$ is defined by, for $t > 0$,

$$P(f,t,B,\mathcal{H}) := \inf \left\{ \|f_0\|_B + t\|f_1\|_{\mathcal{H}}, f = f_0 + f_1 \text{ such that } f_0 \in B, f_1 \in \mathcal{H} \right\}.$$

For fixed $t > 0$, the functional P defines a norm in the Banach space B. It is therefore a simple way to define the interpolation space between B and $\mathcal{H}$ entirely in terms of this functional.
Given $\theta \in]0,1[$ and $q \in [0,\infty]$, the space $(B,\mathcal{H})_{\theta,q}$ called *interpolation space between B and $\mathcal{H}$* consists of all $f \in B$ such that

$$\|f\|_{\theta,q} := \begin{cases} \left(\int_0^{+\infty} t^{-\theta q} P(f,t,B,\mathcal{H})^q \frac{dt}{t} \right)^{\frac{1}{q}} \text{ if } q < \infty, \\ \sup_{t>0} \left\{ t^{-\theta} P(f,t,B,\mathcal{H}) \right\} \text{ if } q = \infty \end{cases}$$

is finite.
Here we are interested in the case $q = \infty$ and the following geometric explanation of interpolation space [118, Theorem 3.1]:

$$f \in (B,\mathcal{H})_{\theta,\infty} \Longrightarrow \inf_{g \in B_{\mathcal{H}}(R)} \|f-g\|_B \leq \|f\|_{\theta,\infty}^{\frac{1}{1-\theta}} \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}}, \tag{2.17}$$

where $B_{\mathcal{H}}(R) := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$. Hence the interpolation space between $\mathcal{B}$ and $\mathcal{H}$ satisfies $\mathcal{H} \subset (\mathcal{B},\mathcal{H})_{\theta,\infty} \subset \mathcal{B}$. To be more precise it consists of functions located at a polynomial decreasing distance in $\mathcal{B}$ from a ball in $\mathcal{H}$ of radius R as a function of R. It would be useful to control the approximation error function in our framework.

Theorem 2.5 *Consider $a(\alpha_n)$ defined in (2.6). Suppose the marginal of X is such that $\frac{dP_X}{dx} \leq C_0$. Then if $f^* \in (L^2(\mathbb{R}^d),\mathcal{H}_K)_{\theta,\infty}$ we have:*

$$a(\alpha_n) \leq \|f^*\|_{\theta,\infty}^{\frac{2}{2-\theta}} \alpha_n^{\frac{\theta}{2-\theta}}.$$

Proof. By the lipschitz property of the hinge loss, we have clearly since $\frac{dP_X}{dx} \leq C_0$:

$$\begin{aligned} a(\alpha_n) &\leq \inf_{f \in \mathcal{H}_K} \left(\|f-f^*\|_{L^1(P_X)} + \alpha_n \|f\|_K^2\right) \\ &\leq \inf_{R>0} \left(C_0 \inf_{f \in B_{\mathcal{H}_K}(R)} \|f-f^*\|_{L^2(\mathbb{R}^d)} + \alpha_n R^2 \right). \end{aligned}$$

Now from (2.17), it follows that if $f^* \in (L^2(\mathbb{R}^d),\mathcal{H}_K)_{\theta,\infty}$,

$$a(\alpha_n) \leq \inf_{R>0} \left(\|f^*\|_{\theta,\infty}^{\frac{1}{1-\theta}} \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}} + \alpha_n R^2 \right).$$

Optimizing with respect to R leads to the conclusion. □
Let introduce Besov spaces $\mathcal{B}^p_{s,q}(\mathbb{R}^d)$. A Besov space is a collection of functions with common smoothness, in terms of modulus of continuity. This is a large class of functional spaces, including in particular the Sobolev spaces defined in (2.7) ($\mathcal{W}^2_s = \mathcal{B}^2_{s,2}(\mathbb{R}^d)$ for any $s > 0$) and the Hölder spaces ($H^s = \mathcal{B}^s_{\infty,\infty}(\mathbb{R}^d)$ for any $s > 0$). For a large study, we refer to Triebel [132].
Here we restrict ourselves to the spaces $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$. For any $h \in \mathbb{R}^d$, we write I for the identity operator, T_h for the translation operator ($T_h(f,x) = f(x+h)$) and $\Delta_h^r := (T_h - I)^r$ for the difference operator. The modulus of continuity of order r of a function $f \in L^2(\mathbb{R}^d)$ is then

$$\omega_r(f,t)_2 = \sup_{|h| \leq t} \|\Delta_h^r(f)\|_{L^2(\mathbb{R}^d)}.$$

Then the Besov space $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$ consists of all functions f such that the semi-norm

$$\|f\|_{s,\infty} = \sup_{t>0} t^{-s} \omega_r(f,t)_2$$

is finite.
If we add $\|f\|_{L^2(\mathbb{R}^d)}$ to this semi-norm, we obtain the usual norm of $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.
Lemma 2.2 *Let $s > 0$ and $0 < \theta < 1$. Then,*

$$(L^2(\mathbb{R}^d),\mathcal{W}^2_s)_{\theta,\infty} = \mathcal{B}^2_{\theta s,\infty}(\mathbb{R}^d).$$

A proof is presented by Triebel [131] in a more general framework.
Proof. (of Theorem 2.2)
From the definition of Sobolev smooth kernels, we have $\mathcal{H}_{K_r} = \mathcal{W}_{\frac{r}{2}}^2$. Hence we obtain with Lemma 2.2:

$$(L^2(\mathbb{R}^d),\mathcal{H}_{K_r})_{\theta,\infty} = \mathcal{B}^2_{\frac{\theta r}{2},\infty}(\mathbb{R}^d).$$

Applying Theorem 2.5 with $\theta = \frac{2s}{r}$, this ends up the proof since P_X satisfies $\frac{dP_X}{dx} < C_0$.□

2.6.3 Proof of Theorem 2.3

In order to control the generalization error, we have to state an inequality such as (2.5). We propose to use a stochastic oracle inequality from Steinwart et al. [125]. This result takes place under a margin assumption of the type (2.10) and a complexity assumption over the used RKHS.
We define the covering numbers of a subset A of a Banach space (E,d) as:

$$\mathcal{N}(A,\epsilon,E) = \min\{n \geq 1 : \exists x_1,\dots x_n \in E \text{ such that } A \subset \cup_{i=1}^n B_d(x_i,\epsilon)\}.$$

Furthermore, given a realization $T = \{(x_1,y_1),\dots (x_n,y_n)\}$ of the training set, we denote by $L^2(T_X)$ the space of all equivalence classes of functions $f : \mathcal{X} \to \mathbb{R}$ such that the norm

$$\|f\|_{L^2(T_X)} := \left(\frac{1}{n}\sum_{i=1}^n f(x_i)^2\right)^{1/2} \tag{2.18}$$

is finite. Then we can consider the behaviour of $\log\mathcal{N}(B_{\mathcal{H}_K},\epsilon,L^2(T_X))$ as a complexity measure for the used RKHS.

Proposition 2.1 (Steinwart and Scovel, 2007) *Let P be a distribution on $\mathcal{X}\times\{-1,1\}$ and $\mathcal{H}_K$ a RKHS of continuous functions on $\mathcal{X}$. Suppose*

1. *There exists $q \in [0,+\infty]$ and $c_0 > 0$ such that*

$$\mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^q, ,\forall t > 0.$$

2. *There exist $a \geq 1, 0 < p < 1$ such that*

$$\sup_{T\in(\mathcal{X}\times\mathcal{Y})^n} \log\mathcal{N}\left(B_{\mathcal{H}_K},\varepsilon,L^2(T_X)\right) \leq a\varepsilon^{-2p},\forall\epsilon > 0. \tag{2.19}$$

Then there exist constants $c \geq 1$, $\kappa,\kappa',\kappa'' > 0$ such that for all $x \geq 1$, the clipped version $\hat{f}_n^C$ of SVM classifier $\hat{f}_n$ satisfies, with probability larger than $1 - e^{-x}$,

$$\begin{aligned} R_l(\hat{f}_n^C,f^*) &\leq c\inf_{f\in\mathcal{H}_K}\left(\mathbb{E}_P\left(l(f) - l(f^*)\right) + \alpha_n\|f\|_K^2\right) + \frac{\kappa}{n\alpha_n^p} \\ &+ \left(\frac{\kappa}{n\alpha_n^p}\right)^{\frac{q+1}{q+2-p}} + \frac{\kappa'}{n^{\frac{q+1}{q+2}}} + \frac{\kappa''x}{n}. \end{aligned}$$

Proof. (of Theorem 2.3)
The hinge loss $l(y,f(x)) = (1 - yf(x))_+$ satisfies, for all classifier $\hat{f}$ [152]:

$$R(\hat{f},f^*) \leq R_l(\hat{f},f^*). \tag{2.20}$$

Therefore, to control the excess risk of a classifier, it is sufficient to control the RHS of (2.20).
We apply Proposition 2.1 for the stochastic part and Theorem 2.2 for the approximation part of the analysis.
Recall a standard result for covering numbers of Sobolev spaces [41]:

$$\log \mathcal{N}(B_{\mathcal{W}_r^2},\varepsilon,C(\mathbb{R}^d)) \leq a\varepsilon^{-\frac{d}{r}}, \tag{2.21}$$

where constant $a := a(d)$. From (2.18) we have $\|f\|_{L^2(T_X)} \leq \|f\|_\infty$ for any $f \in C(\mathbb{R}^d)$, $T \in (\mathcal{X} \times \mathcal{Y})^n$. Then (2.21) holds true for $\log \mathcal{N}(B_{\mathcal{W}_r^2},\varepsilon,L^2(T_X))$ uniformly over $T \in (\mathcal{X} \times \mathcal{Y})^n$. Gathering with $\mathcal{H}_{K_r} = \mathcal{W}^2_{r/2}$, the RKHS $\mathcal{H}_{K_r}$ satisfies (2.19) of Proposition 2.1 with $p = \frac{d}{r}$. Applying Proposition 2.1, there exist $c \geq 1$, $\kappa,\kappa',\kappa'' > 0$ such that, for all $x \geq 1$, with probability larger than $1 - e^{-x}$,

$$\begin{aligned} R_l(\hat{f}_n^C,f^*) \quad &\leq \quad c \inf_{f \in \mathcal{H}_K} \left(R_l(f,f^*) + \alpha_n \|f\|_K^2\right) + \frac{\kappa}{n\alpha_n^{\frac{d}{r}}} \\ &+ \quad \left(\frac{\kappa}{n\alpha_n^{\frac{d}{r}}}\right)^{\frac{q+1}{q+2-d/r}} + \frac{\kappa'}{n^{\frac{q+1}{q+2}}} + \frac{\kappa'' x}{n}. \end{aligned}$$

Since $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$, we get from Theorem 2.2 that with probability larger than $1 - e^{-x}$,

$$R_l(\hat{f}_n^C,f^*) \leq cC_0^{\frac{r}{r-s}} \|f^*\|_{s,\infty}^{\frac{r}{r-s}} \alpha_n^{\frac{s}{r-s}} + \frac{\kappa}{n\alpha_n^{\frac{d}{r}}} + \left(\frac{\kappa}{n\alpha_n^{\frac{d}{r}}}\right)^{\frac{q+1}{q+2-d/r}} + \frac{\kappa'}{n\alpha_n^{\frac{d}{r}}} + \frac{\kappa'' x}{n}.$$

The choice of α_n in (2.11) optimizes the RHS. Integrating with respect to the training set, one leads to the conclusion.□

2.6.4 Proof of Theorem 2.4

To prove Theorem 2.4, we use a general oracle inequality for aggregation. Let us first recall the general context of aggregation.
Suppose we have $M \geq 2$ differents classifiers $f_1,\ldots,f_M$ with values in $\{-1,1\}$. The method of aggregation consists in building a new classifier $\tilde{f}_n$ from D_n called aggregate which mimics the best among $f_1,\ldots f_M$. Our procedure is using exponential weights of the following form:

$$\omega_j^{(n)} = \frac{\exp\left(\sum_{i=1}^n Y_i f_j(X_i)\right)}{\sum_{k\in\{1\ldots M\}} \exp\left(\sum_{i=1}^n Y_i f_k(X_i)\right)}.$$

Then we define the following aggregate:

$$\tilde{f}_n = \sum_{j=1}^M \omega_j^{(n)} f_j. \tag{2.22}$$

Under the margin assumption (2.10), we have this oracle inequality:

Theorem 2.6 (Lecué, 2005) *Suppose* (2.10) *holds for some* $q \in (0,+\infty)$. *Assume we have at least a polynomial number of classifiers to aggregate (i.e. there exist* $a \geq 1$, $b > 0$ *such that* $M \geq an^b$*). Then the aggregate defined in* (2.22) *satisfies, for all integer* $n \geq 1$,

$$\mathbb{E}R(\tilde{f}_n,f^*) \leq (1+2\log^{-1/4} M)\left(2\min_{k\in\{1,\dots M\}} R(f_k,f^*) + Cn^{-\frac{q+1}{q+2}}\log^{7/4} M\right), \tag{2.23}$$

where C *depends on* a,b *and the constant* c_0 *appearing in* (2.10).

Proof. (of Theorem 2.4)
Let $(q_0,s_0) \in K$ and consider $0 < q_{min} < q_{max} < +\infty$ and $0 < s_{min} < s_{max} < +\infty$ such that $K \subset [q_{min},q_{max}] \times [s_{min},s_{max}]$. We consider the function

$$\Phi(q,s) = \frac{r(r-s)(q+1)}{s(r(q+2)-d)+(r-s)(q+1)d}$$

defined on $[0,+\infty[\times[0,+\infty[$ with value on $[\frac{1}{2},\frac{r}{d}]$. We denote by $k_0 \in \left\{0,\dots,\left\lfloor\frac{(2r-d)\Delta}{2d}\right\rfloor - 1\right\}$ the integer such that

$$\frac{1}{2} + k_0\Delta^{-1} \leq \Phi(q_0,s_0) \leq \frac{1}{2} + (k_0+1)\Delta^{-1}.$$

Since $q \mapsto \Phi(q,s)$ continuously increases on $\mathbb{R}^+$, for n greater than a constant depending on b, r, d and K, there exists $\bar{q}_0 \in \left[\frac{q_{min}}{2},q_{max}\right]$ such that $\bar{q}_0 \leq q_0$ and

$$\Phi(\bar{q}_0,s_0) = \frac{1}{2} + k_0\Delta^{-1}. \tag{2.24}$$

Now we can apply Theorem 2.6 for $\bar{q}_0$. Since $\Delta = n_2^b$, putting $M = \left\lfloor\frac{(2r-d)\Delta}{2d}\right\rfloor$ we have the following oracle inequality:

$$\mathbb{E}_{P^{\otimes n_2}}\left(R(\tilde{f}_n,f^*)|D^1_{n_1}\right) \leq (1+2\log^{-\frac{1}{4}} M)\left(2\min_{\alpha\in\mathcal{G}(n_2)}\left(R(\hat{f}^{\alpha}_{n_1},f^*)\right) + C_1 n_2^{-\frac{\bar{q}_0+1}{\bar{q}_0+2}}\log^{7/4} M\right),$$

where C_1 depends on c_0, K and b. Hence we have, integrating with respect to $D^1_{n_1}$,

$$\mathbb{E}\left(R(\tilde{f}_n,f^*)\right) \leq C_2\left(\mathbb{E}R(\hat{f}^{\alpha_{k_0}}_{n_1},f^*) + n_2^{-\frac{\bar{q}_0+1}{\bar{q}_0+2}}\log^{7/4} n_2\right),$$

where $\alpha_{k_0} = n_2^{-\Phi(\bar{q}_0,s_0)}$ with (2.24) and C_2 depends on K,b,r,d and c_0. Therefore we can apply Theorem 2.3 to the classifier $\hat{f}^{\alpha_k}_{n_1}$:

$$\mathbb{E}_{P^{\otimes n_1}} R(\hat{f}^{\alpha_{k_0}}_{n_1},f^*) \leq Cn_1^{-\frac{s_0}{r-s_0}\Phi(\bar{q}_0,s_0)},$$

where C depends on r,d and K. Remark that C does not depend on $\bar{q}_0$ and s_0 since $(\bar{q}_0,s_0) \in [\frac{q_{min}}{2},q_{max}] \times [s_{min},s_{max}]$. Moreover C is uniformly bounded over (q,s) belonging

to a compact in Theorem 2.3.
Finally suppose P satisfies (2.10) for q_0. Hence we obtain:

$$\mathbb{E}\left(R(\tilde{f}_n,f^*)\right) \leq C_3\left(n_1^{-\frac{s_0}{r-s_0}\Phi(\bar{q}_0,s_0)} + n_2^{-\frac{\bar{q}_0+1}{\bar{q}_0+2}}\log^{\frac{7}{4}} n_2\right)$$

for $C_3 := C_3(K,b,c_0,r,C_0,d)$. We have $n \geq n_2 \geq \frac{an}{\log n}$ and $n_1 \geq n(\frac{2}{3} - \frac{a}{\log 3})$. Then for n greater than a constant depending on β_{min}, a, and b, there exists $C_3' := C_3'(K,b,c_0,r,C_0,d)$ such that

$$\begin{aligned}\mathbb{E}\left(R(\tilde{f}_n,f^*)\right) &\leq C_3\left(n^{-\frac{s_0}{r-s_0}\Phi(\bar{q}_0,s_0)} + n^{-\frac{\bar{q}_0+1}{\bar{q}_0+2}}\log^{\frac{11}{4}} n\right)\\ &\leq C_3' n^{-\frac{s_0}{r-s_0}\Phi(\bar{q}_0,s_0)}.\end{aligned}$$

The construction of $\bar{q}_0$ and restrictions on r entail $\frac{s_0}{r-s_0}|\Phi(\bar{q}_0,s_0) - \Phi(q_0,s_0)| \leq \Delta^{-1} = n_2^{-b}$. We lead to the conclusion since the sequence $(n^{n_2^{-b}})_{n\in\mathbb{N}}$ is convergent.□

Chapitre 3

Penalized empirical risk minimization over Besov spaces

Abstract

Kernel methods are closely related to the notion of reproducing kernel Hilbert space (RKHS). A kernel machine is based on the minimization of an empirical cost and a stabilizer (usually the norm in the RKHS). In this chapter we propose to use Besov spaces as alternative hypothesis spaces. We study statistical performances of a penalized empirical risk minimization for classification where the stabilizer is a Besov norm. More precisely, we state fast rates of convergence to the Bayes rule. These rates are adaptive with respect to the regularity of the Bayes.

3.1 Introduction

3.1.1 Classification framework

We consider the binary classification setting. Let (X,Y) be a random variable with unknown probability distribution P over $\mathcal{X}\times\{-1,+1\}$. $X\in\mathcal{X}$ is called the *input* variable. It is a feature vector, whereas $Y\in\{-1,1\}$ is the corresponding *class* or *label*. The goal of classification is to predict class Y when only X is observed. In other words, a classification algorithm builds a decision rule from $\mathcal{X}$ to $\{-1,1\}$. More generally a classifier is a function $f:\mathcal{X}\to\mathbb{R}$ where the sign of $f(x)$ determines the class of an input x. The performance of a classifier is measured by the *generalization error*, given by:

$$R(f) := \mathbb{P}(\text{sign}(f(X)) \neq Y).$$

If we assume that the joint distribution P is known, the best classifier is defined by:

$$f^*(x) := 2\ \mathbb{1}_{\{\eta(x)\geq 1/2\}} - 1, \tag{3.1}$$

where $\eta(x) := \mathbb{P}(Y=1|X=x)$. Classifier (3.1) is called the Bayes rule. It is easy to see that it minimizes the generalization error.
Unfortunately, in practice η is unknown and then f^* is not available. A natural way to overcome this difficulty is to provide an empirical classifier based on training data. Suppose

we have at our disposal a *training set* $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$ made of i.i.d. realizations of the random variable (X,Y) of law P. Now classification can be seen as a standard statistical model where we have to estimate f^* from i.i.d. observations. The efficiency of an empirical classifier $\hat{f}_n$ is measured via its *excess risk*:

$$R(\hat{f}_n,f^*) := R(\hat{f}_n) - R(f^*), \tag{3.2}$$

where $R(\hat{f}_n) := \mathbb{P}(\text{sign}(\hat{f}_n(X)) \neq Y|D_n)$. Here we are interested in consistent classifier $\hat{f}_n$, i.e. such that (3.2) tends to zero as $n \to \infty$. Finally, a classifier $\hat{f}_n$ learns with rate $(\psi_n)_{n\in\mathbb{N}^*}$ if there exists an absolute constant $C > 0$ such that for all integer n,

$$\mathbb{E}R(\hat{f}_n,f^*) \leq C\psi_n, \tag{3.3}$$

where $\mathbb{E}$ is the expectation with respect to the training set.
Without any assumption over the joint distribution P, Devroye et al. [51] gives arbitrary slow rates. However several authors propose different rates restricting the class of distributions P. Pioneering works of Vapnik [140, 141] investigate the statistical performances of the Empirical Risk Minimization (ERM). The idea is very simple: we are looking at the minimizer of the empirical risk:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\text{sign}(f(X_i))\neq Y_i\}}. \tag{3.4}$$

If we suppose that the class of possible Bayes rules has finite VC dimension, ERM reaches the parametric rate $n^{-\frac{1}{2}}$ in (3.3). Moreover, if P is noise-free (i.e. $R(f^*) = 0$), the rate becomes n^{-1}. This is a fast rate. More recently, Tsybakov [136] or Massart and Nédélec [102] describes intermediate situations using margin assumptions. These assumptions add a control on the behaviour of the conditional probability function η at the level $\frac{1}{2}$. Under this condition, they get minimax fast rates of convergence between $n^{-\frac{1}{2}}$ and n^{-1} for ERM estimators in classification. At the present time, there exists a vast literature about the fast rates phenomenon. Fast rates have been obtained for different procedure such as Boosting (Blanchard et al. [26]), Plug-in rules (Audibert and Tsybakov [8]), SVM (Steinwart and Scovel [126]), or dyadic decision trees (Lecué [82]). In this work we propose to state fast rates of convergence for a penalized empirical risk minimization using the hinge loss.

3.1.2 SVM regularization

Support Vector Machines was first proposed by Boser, Guyon and Vapnik (Boser et al. [28]) for pattern recognition. Given a training set D_n, the SVM classifier without offset $\hat{f}_n$ solves the following minimization:

$$\min_{f\in H_K} \left(\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ + \alpha_n \|f\|_K^2 \right), \tag{3.5}$$

where H_K denotes the reproducing kernel Hilbert space (RKHS) associated to the kernel K. The first term in (3.5) is an empirical cost using the hinge loss $l(y,f(x)) := (1-yf(x))_+$. This term allows the solution to fit the data. The second term regularizes the solution. The

parameter α_n is called the smoothing parameter. It has to be determined explicitly to make the trade-off between these two terms. The regularization is done with $\|\cdot\|_K$, the norm in the RKHS. To get statistical performances of the method, we need to take a closer look at this stabilizer.

Consider $L_K : L^2(P_X) \to L^2(P_X)$ the integral operator defined as:

$$L_K : f \mapsto \int_{\mathcal{X}} K(x,\cdot) f(x) P_X(dx).$$

This operator is closely related to the kernel K. If $\mathcal{X}$ is compact and K is continuous (K is called a Mercer kernel), L_K is compact. From spectral theorem, there exist $(\phi_k)_{k\geq 1}$, orthonormal basis of $L^2(P_X)$ of eigenfunctions of L_K with $(\lambda_k)_{k\geq 1}$ corresponding eigenvalues. It allows us to get a representation of H_K in a sequence space as follows:

$$H_K = \left\{ f \in L^2(P_X) : f = \sum a_k \phi_k \, , \sum_{k\geq 1} \frac{a_k^2}{\lambda_k} < +\infty \right\}. \tag{3.6}$$

In this case, the regularization in (3.5) can be written:

$$\|f\|_K^2 = \sum_{k\geq 1} \frac{a_k^2}{\lambda_k}, \tag{3.7}$$

where $(a_k)_{k\geq 1}$ gives a representation of f in the basis $(\phi_k)_{k\geq 1}$. For instance, consider a convolution kernel $K(x,y) = \Phi(x-y)$. Then in (3.7) $(a_k)_{k\geq 1}$ are the Fourier coefficients of f whereas $(\lambda_k)_{k\geq 1}$ are the Fourier coefficients of Φ.
Representation (3.6) holds for Mercer kernels. One can generalize this representation to $\mathcal{X} = \mathbb{R}^d$ in the following case. Suppose $K(x,y) = \Phi(x-y)$ is a convolution kernel. If Φ has some mild properties, the RKHS associated to K can be written:

$$H_K = \left\{ f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Phi}(\omega)} d\omega < \infty \right\}, \tag{3.8}$$

where $\hat{\Phi}$ is the Fourier transform of Φ. In this case the regularity is expressed by the asymptotic behaviour of the Fourier transform. For example, if we suppose that $\hat{\phi}$ decreases polynomially with ω, H_K is a Sobolev space. This is not the case of the Gaussian kernels $K_\sigma(x,y) = \exp(-\sigma\|x-y\|^2)$. The associated RKHS contains smoother functions.

Loustau [87] uses representation (3.8) to state learning rates of the method for $H_K = \mathcal{W}_s^2(\mathbb{R}^d)$ with $s > d/2$. It corresponds to a Sobolev space of continuous functions as RKHS. If $f^* \in \mathcal{B}_{r2\infty}(\mathbb{R}^d)$ one gets:

$$\mathbb{E}R(\hat{f}_n, f^*) \leq C n^{-\beta(q,r,s)}, \tag{3.9}$$

where β is a function of:

- q the margin parameter,
- s the exponent of the Sobolev spaces $\mathcal{W}_s^2(\mathbb{R}^d)$,
- r the smoothness of $f^* \in \mathcal{B}_{r2\infty}(\mathbb{R}^d)$.

Parameter r describes the regularity of f^* in the Besov space $\mathcal{B}_{r2\infty}(\mathbb{R}^d)$. This hypothesis is strongly related to the use of Sobolev spaces as hypothesis space. It ensures a control of the approximation term. Unfortunately f^* is not continuous. As a result this assumption holds for small r and fast rates are not reached with this approach.

An alternative is to take into account the regularity of f^* : it is piecewise constant with local discontinuities. This can be done using a multiresolution analysis and considering Besov spaces as hypothesis spaces.

3.1.3 Besov regularization

It seems interesting to consider minimization (3.5) with more general hypothesis spaces. This is the purpose of this work. We propose to use Besov spaces as hypothesis spaces and study the minimization procedure:

$$\min_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ + \alpha_n \|f\|_{spq}^2 \right), \tag{3.10}$$

where $\|\cdot\|_{spq}$ denotes the norm in $\mathcal{B}_{spq}(\mathbb{R}^d)$. We replace H_K by Besov spaces. An explicit description of $\mathcal{B}_{spq}(\mathbb{R}^d)$ and $\|\cdot\|_{spq}$ is given in Section 2. There are several motivations to introduce Besov spaces in (3.5):

- We have $\mathcal{B}_{s22}(\mathbb{R}^d) = \mathcal{W}_s^2(\mathbb{R}^d)$ is a Sobolev space of order s. For $p = q = 2$, (3.10) corresponds to the standard SVM (without offset) using Sobolev spaces as RKHS. Then (3.10) generalizes the Sobolev case.
- Fast rates cannot be reached in Loustau [87] with Sobolev spaces because f^* is not continuous. $\mathcal{B}_{spq}(\mathbb{R}^d)$ with $p < 2$ gives more flexibility. It contains for instance piecewise regular functions. In this case it will be easier to approximate the Bayes, and leads to better rates of convergence.
- There is a large theory around Besov spaces, such as a characterization using wavelet coefficients. It gives a representation of the norm in (3.10) in a sequence space as follows:

$$\|f\|_{spq} = \left(\sum_{k\in\mathbb{Z}^d} |\alpha_k|^p \right)^{\frac{1}{p}} + \left(\sum_{j\in\mathbb{N}} \left(2^{j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k\in\mathbb{Z}^d} |\beta_{jkl}|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}},$$

  where (α_k) and (β_{jkq}) are the wavelet coefficients.

This representation can be compared to the sequence space representation (3.7) of a RKHS norm. In the standard SVM case, the regularization can be expressed with respect to the spectrum of L_K. It allows to control the complexity of the RKHS. Mendelson [104] or Blanchard et al. [24] control the local Rademacher of balls in RKHS in this sequence space. It depends on the asymptotic behaviour of the sequence $(\lambda_k)_{k\geq 1}$ and affects the statistical performances of the method. In this chapter we point out that similar facts can be derived for Besov spaces using a wavelet analysis.

Minimization (3.10) is strongly related to the SVM minimization. However as a kernel method, SVM uses a RKHS norm as regularization. It allows to define SVM as a large margin hyperplane in some Hilbert feature space. Here the hypothesis space is a Besov

space. Besov spaces are not Hilbertian, and then cannot be represented as RKHS. This penalized empirical risk minimization is not a SVM minimization. However an interesting open problem is to express (3.10) as a kernel method. This problem is connected to recent developments on the theory of reproducing kernels. In this direction, a short discussion is proposed at the end of this chapter (see also "Conclusion et perspectives").

The remainder of this chapter is organized as follows. In Section 2, we introduce the wavelet theory. We characterize Besov spaces in terms of wavelet coefficients. It reduces the control of the Rademacher to a problem in a sequence space, leading to very natural proofs. An oracle inequality is deduced in Section 3. It is a direct application of a general model selection theorem due to Blanchard et al. [24]. We finally control the approximation power of Besov balls to state fast learning rates for the procedure (3.10). The solution is adaptive with respect to the regularity of the Bayes. We conclude in Section 4 with a discussion. Section 5 is dedicated to the proofs of the main results.

3.2 Wavelet framework

For the mathematical aspects of wavelets, we refer for example to Meyer [105], while Mallat [91] proposes comprehensive expositions for signal processing. Wavelet applications in statistical settings are given for instance in Härdle et al. [68]. For a complete study of minimax rates of convergence for density estimation by wavelet thresholding we refer to Donoho et al. [53].

Here recall some definitions and notations for wavelets and Besov spaces. Going back to statistical learning theory, one proposes a control of the local Rademacher average of Besov balls.

3.2.1 Besov spaces and wavelets

Wavelet bases of $L^2(\mathbb{R}^d)$

For the one-dimensional case, we refer to the first chapter. To introduce the d-dimensional case, we begin with an example in dimension 2 using the tensor product. Write $(V_j^1)_{j\in\mathbb{Z}}$ a multiresolution analysis (MRA for short in the sequel) of $L^2(\mathbb{R})$ generated by ϕ. Write $V_j = V_j^1 \times V_j^1$ for all $j \in \mathbb{Z}$. Then the system $(\phi(x-k)\phi(y-l))_{k,l\in\mathbb{Z}}$ is an orthonormal basis of V_0 in $L^2(\mathbb{R}^2)$. Let consider W_j, $j \in \mathbb{Z}$ such that $V_{j+1} = V_j \oplus W_j$. Then we have for $j = 0$:

$$W_0 = \overline{V_0^1 \otimes W_0^1} \oplus \overline{W_0^1 \otimes V_0^1} \oplus \overline{W_0^1 \otimes W_0^1}.$$

A basis of W_0 is obtained with the three collections $\phi(x-k)\psi(y-l)$, $\psi(x-k)\phi(y-l)$ and $\psi(x-k)\psi(y-l)$ for $(k,l) \in \mathbb{Z}^2$. More generally for all $j \in \mathbb{Z}$:

$$W_j = \overline{V_j^1 \otimes W_j^1} \oplus \overline{W_j^1 \otimes V_j^1} \oplus \overline{W_j^1 \otimes W_j^1}.$$

Then the two-dimensional mother wavelets are $2^j\phi(2^jx-k)\psi(2^jy-l)$, $2^j\psi(2^jx-k)\phi(2^jy-l)$ and $2^j\psi(2^jx-k)\psi(2^jy-l)$ for $(k,l) \in \mathbb{Z}^2$. This means that there are three wavelets in the two-dimensional case. This fact is illustrated in Meyer [105] with a geometrical point of view. We can generalize this result in higher dimensions with the following lemma.

Lemma 3.1 *Let V_j, $j \in \mathbb{Z}$ a MRA r-regular of $L^2(\mathbb{R}^d)$. Then there exist $L = 2^d - 1$ functions $\psi_1, \dots \psi_L \in V_1$ such that:*

1. for all $l \in \{1 \dots L\}$, for all $\alpha \in \mathbb{N}^d : |\alpha| \leq r$, for all $x \in \mathbb{R}^d$ and $N \geq 1$,

$$|\partial^\alpha \psi_l(x)| \leq C_N(1 + |x|)^{-N};$$

2. the system $\{\psi_l(x-k), 1 \leq l \leq L,\ k \in \mathbb{Z}^d\}$ is an ONB of W_0.

As a result, the system:

$$2^{\frac{dj}{2}} \psi_l(2^j x - k), 1 \leq l \leq L, k \in \mathbb{Z}^d, j \in \mathbb{Z} \tag{3.11}$$

is an orthonormal basis of $L^2(\mathbb{R}^d)$.

This lemma generalizes the one-dimensional case. From a scaling function r-regular and rapidly decreasing generating a MRA, we can construct $2^d - 1$ mother wavelets with the same regularity. The existence of such a wavelet basis is proved in Meyer [105].
As a consequence, any $f \in L^2(\mathbb{R}^d)$ can be decomposed as:

$$f = \sum_{k \in \mathbb{Z}^d} \alpha_{0k}\phi_{0k} + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \beta_{jkl}\psi_{jkl}, \tag{3.12}$$

where:

$$\alpha_{0k} = \int_{\mathbb{R}^d} f(x)\phi_{0k}(x)dx \text{ and } \beta_{jkl} = \int_{\mathbb{R}^d} f(x)\psi_{jkl}(x)dx.$$

In the case of tensor product, we have for all $k \in \mathbb{Z}^d$ and $x \in \mathbb{R}^d$:

$$\phi_{0k}(x) = \phi(x_1 - k_1) \dots \phi(x_d - k_d).$$

Moreover for all $j \geq 0$, $k \in \mathbb{Z}^d$, $l \in \{1, \dots, 2^d - 1\}$ and $x \in \mathbb{R}^d$:

$$\psi_{jkl}(x) = 2^{\frac{dj}{2}} \psi^{e_1}(2^j x_1 - k_1) \dots \psi^{e_d}(2^j x_d - k_d),$$

for $e \in \{0,1\}^d \backslash 0_{\mathbb{R}^d} = E$ and where we write for simplicity $\psi^0 = \phi$ and $\psi^1 = \psi$.
Here we are interested in compactly supported wavelet bases. Daubechies [45] has shown that in dimension $d = 1$, there exists an orthonormal basis of compactly supported wavelets satisfying conditions of Lemma 3.1, for any integer $r \geq 1$ (for $r = 0$, it corresponds to the Haar basis). Using the tensor product, this result gives a compactly supported d-dimensional wavelet basis of $L^2(\mathbb{R}^d)$ (see Meyer [105] for details).

Besov spaces

Besov spaces were introduced by O.V. Besov in the 60s. We refer to Chapter 1 for an introduction of these functional spaces. Here we propose to characterize Besov spaces $\mathcal{B}_{spq}(\mathbb{R}^d)$ in terms of wavelet coefficients.
Recall $P_j : L^2(\mathbb{R}^d) \to V_j$ is the projection operator into V_j and $D_j = P_{j+1} - P_j$. We know that for $f \in L^p(\mathbb{R}^d)$, $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ if and only if $P_0(f) \in L^p(\mathbb{R}^d)$ and if there exists a positive sequence $(\epsilon_j)_{j \in \mathbb{N}}$ such that:

$$\|D_j(f)\|_p \leq 2^{-js}\epsilon_j. \tag{3.13}$$

To express the L^p-norm of $D_j(f)$ in terms of the AMR of $L^2(\mathbb{R}^d)$, we need the following lemma.

Lemma 3.2 *Let $g_1, \dots g_L$ compactly supported on $\mathbb{R}^d$ satisfying assumptions 1. and 2. of Lemma 3.1 for $L = 2^d - 1$. Let $f(x) = \sum_{l=1}^{L} \sum_{k\in\mathbb{Z}^d} \lambda_{kl} 2^{\frac{dj}{2}} g_l(2^j x - k)$. Then there exist $0 < c_1 < c_2$ such that for all $1 \leq p$,*

$$c_1 2^{dj\left(\frac{1}{2}-\frac{1}{p}\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k\in\mathbb{Z}^d} |\lambda_{kl}|^p \right)^{\frac{1}{p}} \leq \|f\|_p \leq c_2 2^{dj\left(\frac{1}{2}-\frac{1}{p}\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k\in\mathbb{Z}^d} |\lambda_{kl}|^p \right)^{\frac{1}{p}}.$$

Lemma 3.2 is a direct consequence of [105, Lemma 8], using the d-dimensional change of variables formula.

Gathering with (3.13), we arrive at the following characterization of Besov spaces.

Lemma 3.3 *Let $p \geq 1$ and $f \in L^p(\mathbb{R}^d)$. Then $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ if and only if:*

$$\left(\sum_{k\in\mathbb{Z}^d} |\alpha_{0k}|^p \right)^{\frac{1}{p}} + \left(\sum_{j\in\mathbb{N}} \left(2^{j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k\in\mathbb{Z}^d} |\beta_{jkl}|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}} < +\infty, \tag{3.14}$$

where,

$$\alpha_k = \int_{\mathbb{R}^d} f(x)\phi_{0k}(x)dx \text{ and } \beta_{jkl} = \int_{\mathbb{R}^d} f(x)\psi_{jkl}(x)dx.$$

First term in (3.14) corresponds to the L^p-norm of $P_0(f)$ whereas the second term corresponds to the l^q-norm of $2^{js}\|D_j(f)\|_p$.

This characterization of Besov spaces will be useful to control the complexity of $\mathcal{B}_{spq}(\mathbb{R}^d)$ in this sequence space. For other characterizations, we refer to Peetre [109] or Triebel [132], including the most usual definition in terms of modulus of continuity (see also Chapter 1).

3.2.2 Local complexity of Besov balls

First error bounds for empirical risk minimization go back to Vapnik (see Vapnik and Chervonenkis [141]). Consider an ERM estimator $\hat{f}_{ERM}$ over a collection of classifiers $\mathcal{F}$, Vapnik and Chervonenkis [141] states that:

$$R(\hat{f}_{ERM}) - \inf_{f\in\mathcal{F}} R(f) \leq 2 \sup_{f\in\mathcal{F}} |R_n(f) - R(f)|. \tag{3.15}$$

This leads to the study of the supremum of an empirical process. With concentration inequalities, this random process can be controled by its expectation, up to some residual terms. Recently, sharp bounds have been established using different localized versions of (3.15). It is now common to use localized averages. Considering the penalized empirical minimization (3.10) using the hinge loss $l(y,f(x)) = (1 - yf(x))_+$, we are interesting in:

$$\mathbb{E} \sup_{f\in B(R):\mathbb{E}f(X)^2\leq r} \left| \frac{1}{n}\sum_{i=1}^{n} l(Y_i,f(X_i)) - \mathbb{E}l(Y,f(X)) \right|, \tag{3.16}$$

where in the sequel $B(R) = \{f \in \mathcal{B}_{spq}(\mathbb{R}^d) : \|f\|_{spq} \leq R\}$. Since l is 1-Lipschitz, (3.16) can be bounded by:

$$\mathbb{E} \sup_{f \in B(R): \mathbb{E}f(X)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right|. \tag{3.17}$$

The parameter r allows us to identify the scale of richness of the function class. Small r yields to an error bound for the ERM principle. Otherwise, a useful comparison between the empirical risk and the true risk is not available.
From a simple symmetrization device (originally in Vapnik and Chervonenkis [140]), we can bound (3.17) as follows:

$$\mathbb{E} \sup_{f \in B(R): \mathbb{E}f(X)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \leq 2\mathbb{E} \sup_{f \in B(R): \mathbb{E}f^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|,$$

where ϵ_i, $i = 1, \ldots, n$ are i.i.d. with $P(\epsilon_1 = 1) = P(\epsilon_1 = -1) = \frac{1}{2}$. The ϵ_i are called Rademacher variables. Right hand side is called the local Rademacher average of $B(R)$. The use of Rademacher averages in Classification goes back to Koltchinskii [76] (see also Bartlett and Mendelson [16], Bartlett et al. [13, 12]). Mendelson [104] has proved that the local Rademacher average of a kernel class is determined by the spectrum of its integral operator (see also Blanchard et al. [24]). Under assumptions on the law of X, we propose a same type of result for Besov classes:

Theorem 3.1 *Suppose P_X admits a density p such that*

- *$a \leq p(x) \leq A$ for any $x \in \mathcal{X}$;*
- *p has compact support $\mathcal{P} = \{x \in \mathcal{X} : p(x) \neq 0\}$.*

Then if $s > \frac{d}{p}$ and $1 \leq p \leq 2$, there exists a constant c depending on a and A such that:

$$\forall r > 0, \mathbb{E} \sup_{f \in B(R): \mathbb{E}f(X)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq \frac{c}{\sqrt{n}} R^{\frac{d}{2u}} r^{\frac{s-\frac{d}{p}}{2u}},$$

where $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$.

Remark 3.1 *This result holds for parameter range of Besov spaces such that $s > \frac{d}{p}$. In this case, there exists a continuous embedding from $\mathcal{B}_{spq}(\mathbb{R}^d)$ into $C(\mathbb{R}^d)$. It ensures that the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous on $\mathcal{B}_{spq}(\mathbb{R}^d)$, exactly as in the RKHS case. In the sequel we consider Besov spaces with such a restriction.*

Remark 3.2 *Consider the Sobolev case in dimension 1. If we put $p = q = 2$ and $d = 1$, the upper bound becomes $R^{\frac{1}{2s}} r^{\frac{2s-1}{4s}}$. It corresponds to the upper bound of [104, Theorem 2.1] for eigenvalues of the integral operator such that $\lambda_k \leq k^{-2s}$. It illustrates that this result generalizes the Sobolev case $p = q = 2$.*

Remark 3.3 *This result is the meaty part to deduce statistical performances of minimization* (3.10)*. It allows us to control the local average* (3.16) *and to deduce Proposition 3.1.*

Remark 3.4 *A detailed proof is presented in Section 5. As mentionned in the introduction, we use wavelet theory presented above and precisely Lemma 3.3. This characterization of Besov spaces allows us to control* (3.16) *in a sequence space.*

Remark 3.5 *Here the localization is determined by the second order moment. Bartlett and Mendelson [17] proposes a sharper localization using the indexing set* $\{f : \mathbb{E}f = r\}$. *The upper bound of Theorem 3.1 can be improved since* $\{f \in B(R) : \mathbb{E}f = r\} \subset \{f \in B(R) : \mathbb{E}f(X)^2 \leq r\}$.

3.3 Statistical performances

To state learning rates to the Bayes, we act in two steps. First step is to state an oracle inequality of the form:

$$R_l(\hat{f}_n,f^*) \leq C \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f,f^*) + \alpha_n\|f\|^2_{spq}\right) + \delta_n.$$

The statistical sense of this inequality is rather transparent. It ensures classifier $\hat{f}_n$ to have comparable performances with the best SVM (which minimizes the true risk), up to a residual term δ_n such that $\delta_n \to 0$ as $n \to \infty$. Constant C has to be close to 1.
It remains to control the right hand side of the oracle inequality. The main term of this bound is called the approximation function, defined in this case as:

$$a(\alpha_n) = \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f,f^*) + \alpha_n\|f\|^2_{spq}\right).$$

Following Loustau [87], we use the theory of interpolation spaces to control this function. Under an assumption over the Bayes f^*, we deduce rates of convergence for the procedure.

3.3.1 Oracle inequality

To obtain good statistical properties, we need to restrict the class of considered distributions P. A standard way is to impose a margin hypothesis over the conditional probability function η. In this work we will assume that there exist η_0, $\eta_1 > 0$ such that:

$$\forall x \in \mathcal{X}, \left|\eta(x) - \frac{1}{2}\right| \geq \eta_0 \textbf{ and } \min_{x\in\mathcal{X}}(\eta(x),1-\eta(x)) \geq \eta_1. \tag{3.18}$$

This assumption is closely related to the margin assumption originally due to Tsybakov [136]. The first part ensures a jump of the probability η at the level $\frac{1}{2}$. The second part is not natural. It avoids the no noise case where $\eta(x) \in \{0,1\}$. It appears for some technical reasons discussed in Section 5 (see also Blanchard et al. [24]).

Proposition 3.1 (Oracle inequality) *Let* P *the joint distribution such that the marginal of* X *satisfies assumptions of Theorem 3.1. Suppose* (3.18) *holds for some* η_0, $\eta_1 > 0$. *Consider a non-decreasing function* ϕ *on* $\mathbb{R}^+$ *such that* $\phi(0) = 0$ *and* $\phi(x) \geq x$ *for* $x \geq \frac{1}{2}$. *Given* $(X_i,Y_i), i = 1\dots,n$ *i.i.d. from* P, *we denote:*

$$\hat{g}_n = \arg\min_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n}\sum_{i=1}^n l(Y_i,f(X_i)) + \alpha_n\phi(\|f\|_{spq})\right), \tag{3.19}$$

where $s > \frac{d}{p}$ *and* $1 \leq p \leq 2$. *If we choose* α_n *such that:*

$$\alpha_n \geq c_1 n^{-\frac{2u}{2u+d}} + \eta_1^{-1}\left(c_2\frac{\log n}{n} + c_3\frac{\log\log n}{n} + \frac{c_4}{n}\right), \tag{3.20}$$

then the estimator $\hat{g}_n$ *is such that:*

$$\begin{aligned}\mathbb{E}R_l(\hat{g}_n,f^*) &\leq 2\inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)}\left(R_l(f,f^*)+\alpha_n\phi(\|f\|_{spq})\right)\\ &+ 4\alpha_n\left(2\phi(2)+c\frac{\eta_1}{\eta_0}\right)+\frac{2}{n},\end{aligned}$$

where c, c_1, c_2, c_3 *and* c_4 *are absolute constants and* $u = s + d\left(\frac{1}{2}-\frac{1}{p}\right)$.

Remark 3.6 *This oracle inequality is a direct application of Theorem 1.6 in Chapter 1. We write empirical minimization* (3.10) *as a model selection procedure where models are balls in* $\mathcal{B}_{spq}(\mathbb{R}^d)$.

Remark 3.7 *It holds whatever* $\phi : \phi(0) = 0$ *and* $\phi(x) \geq x$ *for* $x \geq \frac{1}{2}$. *From the model selection approach, the minimum required regularization is of order* $\|f\|_{spq}$ *only. In the standard SVM, a regularization of order* $\|f\|_{spq}^2$ *is used. Thus we consider in Corollary 3.1 the two cases* $\phi(x) = x$ *and* $\phi(x) = 2x^2$. *These two orders of regularization will lead to different statistical performances.*

Remark 3.8 *Multiplicative constant 2 in front of the infimum can be taken arbitrarily close to 1. However the smoothing parameter* α_n *grows consequently. We choose constant 2 to simplify the result.*

Remark 3.9 *This inequality is independent of the approximation term. The choice of* α_n *in* (3.20) *only depends on the hypothesis set we consider. A control of the approximation power of Besov spaces will give adaptive learning rates.*

3.3.2 Rates of convergence

Last step is to control the approximation term in the oracle inequality of Proposition 3.1. The theory of interpolation spaces allows us to measure how well the hypothesis space approximate the target function f^*. We finally get the following rates of convergence.

Corollary 3.1 (Rates of convergence) *Let* P *satisfying assumptions of Proposition 3.1. Then for any* $1 \leq p \leq 2$ *and* $s > \frac{d}{p}$, *define the estimators:*

$$\hat{f}_n := \arg\min_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)}\left(\frac{1}{n}\sum_{i=1}^n l(Y_i,f(X_i)) + \alpha_n\|f\|_{spq}\right)$$

and:

$$\hat{g}_n := \arg\min_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)}\left(\frac{1}{n}\sum_{i=1}^n l(Y_i,f(X_i)) + 2\alpha_n\|f\|_{spq}^2\right).$$

Suppose that $f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d)$ *with* $r < s$. *Then there exist absolute constants* C, $C' > 0$ *such that:*

$$\mathbb{E}R(\hat{f}_n, f^*) \leq Cn^{-\frac{r}{s}\frac{2u}{2u+d}}, \tag{3.21}$$

and:

$$\mathbb{E}R(\hat{g}_n, f^*) \leq C'n^{-\frac{r}{2s-r}\frac{2u}{2u+d}}, \tag{3.22}$$

where we choose α_n *such that an equality holds in* (3.20).

Remark 3.10 *We consider two special cases for the function* ϕ *of Proposition 3.1. Estimator* $\hat{f}_n$ *is the penalized empirical minimizer using the weakest regularization (linear with respect to the norm) whereas* $\hat{g}_n$ *uses the standard SVM penalization (of order* $\|f\|^2$*). We can see coarsely that the rate of* $\hat{f}_n$ *outperforms the one of* $\hat{g}_n$ *since* $\frac{r}{s} > \frac{r}{2s-r}$. *With this approach, a lighter regularization results in a better bound.*

Remark 3.11 *The construction of these estimators does not depend on the regularity of the Bayes. The smoothing parameter* α_n *is chosen independently of the parameter* r *appearing in the assumption* $f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d)$. *As a result, estimators* $\hat{f}_n$ *and* $\hat{g}_n$ *are called adaptive. They adapt to the regularity of the Bayes.*

Remark 3.12 *Chapter 2 gives learning rates for SVM using Sobolev spaces. In particular, under a strong margin assumption, we obtain* $n^{-\frac{2rs}{2rs+d(2s-r)}}$. *We can compare this bound with* (3.22) *for* $p = q = 2$. *In this case we have* $n^{-\frac{r}{2s-r}\frac{2s}{2s+d}}$. *This rate is clearly slower than* $n^{-\frac{2rs}{2rs+d(2s-r)}}$ *since* $s > r$. *However it gives similar results when* $s \to r$.

Remark 3.13 (Fast rates) *Consider the one-dimensional case where* $\mathcal{X} = \mathbb{R}$. *Suppose* f^* *is such that:*

$$card\{x \in \mathbb{R} : f^* \text{ jumps at } x\} = N < \infty. \tag{3.23}$$

It means that the Bayes rule changes only a finite number of times over the real line. Under this assumption, SVM algorithm using Sobolev spaces cannot reach fast rates (see Loustau [87] or Chapter 2). From Corollary 3.1, we can consider a value of $p < 2$. *In this case, if* $1 \leq p = q < 2$, *we have using Mallat [91]:*

$$f^* \in \mathcal{B}_{rpq}(\mathbb{R}) \text{ for } r = \frac{1}{2} + \frac{1}{p}.$$

Consequently, f^* *such that* (3.23) *holds belongs to* $\mathcal{B}_{r11}(\mathbb{R}) \subset \mathcal{B}_{r1\infty}(\mathbb{R})$ *for* $r = 3/2$. *Substituing into* (3.22), *the rate becomes* $n^{-\frac{6s-3}{2s(4s-3)}}$ *which is a fast rate for* $s > r$ *small enough. This example illustrates the importance to consider Besov spaces with* $p < 2$ *as hypothesis space. For* $p < 2$, *these spaces contain piecewise regular functions with local discontinuities. It gives fast rates of convergence.*

3.4 Conclusion

We have studied a new procedure of penalized empirical risk minimization using Besov spaces. This method generalizes Chapter 2 where we consider SVM over Sobolev spaces. The introduction of Besov spaces gives more flexibility to study the approximation power of the procedure. For the estimation part of the analysis, we adopt the model selection approach of Blanchard et al. [24]. We propose a control of the local Rademacher average of Besov balls. We hence obtain fast learning rates to the Bayes. Moreover, the construction of these estimators does not depend on the regularity of the Bayes. They are adaptive with respect to the regularity of f^*.
From technical point of view, this chapter generalizes the control of Rademacher to a non Hilbertian functional space. It is well-known that local Rademacher of RKHS balls can be controled using RKHS formalism. Here we propose to use a wavelet analysis to get a similar result for Besov spaces. A compactly supported wavelet basis allows us to work in a sequence space.

This chapter could be compared with another introduction of wavelet theory in classification. Lecué [82] studies the statistical performances of the LASSO estimator, solving the minimization:

$$\min_{f\in\mathcal{F}^d}\left(\sum_{i=1}^{n}\mathbb{1}(f(X_i)\neq Y_i)+\alpha\|f\|_{L^1}\right).$$

The hypothesis space $\mathcal{F}^d$ is made of piecewise constant classifiers on a dyadic regular grid of $[0,1]^d$. It allows to decompose each classifier into a fundamental system of indicators on dyadic sets of $[0,1]^d$. This system is closely related to the wavelet tensor product of the Haar basis. As a consequence, in all the proofs, similitudes with the technics used in the wavelet literature are granted. From this point of view, the present work can be compared to Lecué [82].

Unfortunately, from practical point of view, the presence of Besov norms in our procedure leads to some computational problems. In Chapter 2, we pick up a kernel generating Sobolev spaces as RKHS. Besov spaces are not Hilbert spaces. The feature space is not a RKHS in our procedure. As a result, our method cannot be embedded into a kernel method and computed as SVM algorithm.
Recently, several authors investigate learning algorithm with non Hilbertian hypothesis space. Canu et al. [32] underlines the main principles of an hypothesis space in a learning problem. The hypothesis set must be composed of pointwise defined functions. Moreover the evaluation functional $\delta_x : f \mapsto f(x)$ must be continuous. Due to the embedding theorem, Besov spaces $\mathcal{B}_{spq}(\mathbb{R}^d)$ with $s > \frac{d}{p}$ have this property. In the RKHS case, it corresponds to the reproducing property. It gives a reproducing kernel lying in the RKHS. However the Hilbertian structure is not necessary. To generalize the notion of RKHS to RKS (Reproducing Kernel Space), we need a bilinear form corresponding to the scalar product for RKHS. It could be done with the duality map, considering a duality couple $(\mathcal{H},\mathcal{H}^*)$. Mary et al. [99] establishes an equivalence between particular dualities called evaluation subdualities and a set of weakly continuous applications called reproducing kernels. Canu et al. [32] also provides an explicit construction of both subdualities and the associated reproducing kernel. It is a generalization to the construction of RKHS using Carleman operator. The construction is based on the duality map.

Finally we know from Meyer [105] that $(\mathcal{B}_{spq}(\mathbb{R}^d),\mathcal{B}_{-sp'q'}(\mathbb{R}^d))$ are in duality through the duality map:

$$< f,g >_{\mathcal{B}_{spq}(\mathbb{R}^d),\mathcal{B}_{-sp'q'}(\mathbb{R}^d)}=< P_0(f),P_0(g) >_{L^2(\mathbb{R}^d)} + \sum_{j\geq 0} < D_j(f),D_j(g) >_{L^2(\mathbb{R}^d)} ,$$

where $\mathcal{B}_{-sp'q'}(\mathbb{R}^d))$ is the space of distributions such that (3.13) holds for $-s < 0$. As a result, it will be interesting in this direction to find a kernel generating Besov spaces as RKS. Last step would be to implement our procedure with such a kernel.

3.5 Proofs

This section is dedicated to the proofs of the main results of this paper. Throughout this section, c denotes a constant that may vary from line to line. For $p,q > 0$, we write p',q' such that $1/p+1/p' = 1/q+1/q' = 1$. Finally, with some abuse of notations, we write $\mathbb{E}f$ for $\mathbb{E}_{P_X} f(X)$ and $\mathbb{E}l(f)$ for $\mathbb{E}_P l(Y,f(X))$.

3.5.1 Proof of Theorem 3.1

The marginal of X admits a bounded density p with compact support $\mathcal{P}$. Then we have:

$$\left\{f \in B(R) : \mathbb{E}f^2 \leq r\right\} \subseteq \left\{f \in B(R) : \int_{\mathcal{P}} f^2(x)dx \leq \frac{r}{a}\right\} := \mathcal{F}(R,r).$$

Moreover $X_1, \dots X_n$ are i.i.d. from p. Then:

$$\mathbb{E} \sup_{f\in\mathcal{F}(R,r)} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| = \mathbb{E} \sup_{f\in B(R):\|f\|^2_{L^2(\mathbb{R}^d)}\leq \frac{r}{a}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|. \tag{3.24}$$

We then have to bound the RHS of (3.24).

Let begin with the one-dimensional case, i.e. when the input domain $\mathcal{X} \subset \mathbb{R}$. From wavelet decomposition, we can write $f \in L^2(\mathbb{R})$ as:

$$f = \sum_{k\in\mathbb{Z}} \alpha_{0k}\phi_{0k} + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk} = f_{\alpha,\beta}. \tag{3.25}$$

The description of Besov spaces using wavelets leads to the following equivalent norm:

$$\|f\|_{spq} = \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^p\right)^{\frac{1}{p}} + \left(\sum_{j\geq 0} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}.$$

Moreover from Lemma 3.2,

$$\|f\|_{L^2} \approx \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^2\right)^{\frac{1}{2}} + \sum_{j\geq 0}\left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^2\right)^{\frac{1}{2}},$$

where $x \approx y$ means there exist c,$C > 0$ such that $cy \leq x \leq Cy$. We hence obtain:

$$\mathbb{E} \sup_{f \in B(R): \|f\|_{L^2}^2 \leq \frac{r}{a}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \leq \mathbb{E} \sup_{(\alpha,\beta) \in \Gamma(R,r)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i) \right|,$$

where $f_{\alpha,\beta}$ is defined in (3.25) and:

$$\Gamma_d(R,r) = \{(\alpha,\beta) : \gamma_{pq}(\alpha,\beta) \leq R \text{ and } \gamma_2(\alpha,\beta) \leq \frac{\sqrt{r}}{\sqrt{a}c}\}$$

for:

$$\gamma_{pq}(\alpha,\beta) = \left(\sum_{k \in \mathbb{Z}^d} |\alpha_{0k}|^p \right)^{\frac{1}{p}} + \left(\sum_{j \geq 0} \left(2^{j\left(s + d(\frac{1}{2} - \frac{1}{p})\right)} \sum_{l=1}^{2^d - 1} \|\beta_{j \cdot l}\|_p \right)^q \right)^{\frac{1}{q}},$$

and:

$$\gamma_2(\alpha,\beta) = \left(\sum_{k \in \mathbb{Z}^d} |\alpha_k|^2 \right)^{\frac{1}{2}} + \sum_{j \geq 0} \sum_{l=1}^{2^d - 1} \left(\sum_{k \in \mathbb{Z}} |\beta_{jkl}|^2 \right)^{\frac{1}{2}}.$$

Hence we get for any integer d':

$$\begin{aligned}
\left| \sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i) \right| &= \left| \sum_{k \in \mathbb{Z}} \alpha_{0k} \sum_{i=1}^n \epsilon_i \phi_{0k}(X_i) + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}} \beta_{jk} \sum_{i=1}^n \epsilon_i \psi_{jk}(X_i) \right| \\
&\leq \left| \sum_{k \in \mathbb{Z}} \alpha_{0k} \sum_{i=1}^n \epsilon_i \phi_{0k}(X_i) + \sum_{j=0}^{d'} \sum_{k \in \mathbb{Z}} \beta_{jk} \sum_{i=1}^n \epsilon_i \psi_{jk}(X_i) \right| \\
&+ \left| \sum_{j > d'} \sum_{k \in \mathbb{Z}} \beta_{jk} \sum_{i=1}^n \epsilon_i \psi_{jk}(X_i) \right| := T_1 + T_2.
\end{aligned}$$

To prove the inequality, we will bound this two terms separately.
We begin applying Hölder (twice) and Jensen inequalities to T_2:

$$\begin{aligned}
\mathbb{E}[T_2] &\leq \mathbb{E} \sum_{j > d'} \left(\sum_{k \in \mathbb{Z}} |\beta_{jk}|^p \right)^{\frac{1}{p}} \left(\sum_{k \in \mathbb{Z}} |\sum_{i=1}^n \epsilon_i \psi_{jk}(X_i)|^{p'} \right)^{\frac{1}{p'}} \\
&\leq \sum_{j > d'} \left(\sum_{k \in \mathbb{Z}} |\beta_{jk}|^p \right)^{\frac{1}{p}} \left(\sum_{k \in \mathbb{Z}} \mathbb{E}|\sum_{i=1}^n \epsilon_i \psi_{jk}(X_i)|^{p'} \right)^{\frac{1}{p'}} \\
&\leq \left(\sum_{j > d'} \left(2^{j(s + \frac{1}{2} - \frac{1}{p})} \left(\sum_{k \in \mathbb{Z}} |\beta_{jk}|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}} \times \\
&\quad \left(\sum_{j > d'} \left(2^{-j(s + \frac{1}{2} - \frac{1}{p})} \left(\sum_{k \in \mathbb{Z}} \mathbb{E} \left| \sum_{i=1}^n \epsilon_i \psi_{jk}(X_i) \right|^{p'} \right)^{\frac{1}{p'}} \right)^{q'} \right)^{\frac{1}{q'}}.
\end{aligned}$$

The definition of $\Gamma(R,r)$ leads to:

$$\mathbb{E}[T_2] \leq R\left(\sum_{j>d'}\left(2^{-j(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right|^{p'}\right)^{\frac{1}{p'}}\right)^{q'}\right)^{\frac{1}{q'}}, \tag{3.26}$$

where $\frac{1}{p}+\frac{1}{p'}=\frac{1}{q}+\frac{1}{q'}=1$.
Next step is to control, for all $j>d'$, the serie:

$$\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right|^{p'}.$$

Lemma 3.4 *Let $Y_1,\dots Y_n$ i.i.d. with zero mean and σ^2 variance. Then for all $p\geq 2$, there exists $c_p>0$ such that:*

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}Y_i\right|^{p}\leq c_p\left[\frac{\sigma^p}{n^{\frac{p}{2}}}+\frac{\mathbb{E}|Y_1|^p}{n^{p-1}}\right].$$

This concentration inequality is due to Rosenthal (Rosenthal [116]).
Putting $Y_i=\epsilon_i\psi_{jk}(X_i)$, we have with Lemma 3.2, gathering with conditions on the density p:

$$\mathbb{E}|Y_i|^p\leq A\|\psi_{jk}\|_p^p\leq c2^{j\left(\frac{p}{2}-1\right)},$$

for an absolute constant c depending on A. As a result, applying Lemma 3.4 for $p'\geq 2$, we obtain:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right|^{p'}\leq c_{p'}n^{-\frac{p'}{2}}\left[c^{\frac{p'}{2}}+c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right].$$

Now it is worth noticing that since p and the wavelets functions ψ_{jk} are compactly supported, the quantity:

$$\mathbb{E}\psi_{jk}^{p'}=\int_{\mathbb{R}}|\psi_{jk}(x)|^{p'}p(x)dx$$

is zero whatever $k\in\mathcal{S}^C(j):=\{k\in\mathbb{Z}:\text{supp}\psi_{jk}\cap\mathcal{P}=\varnothing\}$. We know from Meyer [105] that there exists a constant $c>0$ such that $\sharp\mathcal{S}(j)\leq c2^j$. Then:

$$\begin{aligned}\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right|^{p'} &= \sum_{k\in\mathcal{S}(j)}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right|^{p'}\\ &\leq c2^jn^{-\frac{p'}{2}}\left[c^{\frac{p'}{2}}+c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right].\end{aligned}$$

Gathering with (3.26), we obtain:

$$\begin{aligned}
\mathbb{E}[T_2] &\leq c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-j\left(s+\frac{1}{2}-\frac{1}{p}\right)q'}\left(\sum_{k\in\mathcal{S}(j)} c^{\frac{p'}{2}} + c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}} \\
&\leq c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-jq'\left(s-\frac{1}{2}\right)}\left(c^{\frac{p'}{2}} + c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-jq'\left(s-\frac{1}{2}\right)} + 2^{-jq'\left(s-\frac{1}{p}\right)} n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{1}{2}\right)} + 2^{-d'q'\left(s-\frac{1}{p}\right)} n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}},
\end{aligned}$$

where the convergence of the geometric serie comes from the condition $s > \frac{1}{p}$. Moreover we have $p \leq 2$ then $s - \frac{1}{2} \geq s - \frac{1}{p}$ and $1 - \frac{p'}{2} \leq 0$. Then:

$$\begin{aligned}
\mathbb{E}[T_2] &\leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{1}{p}\right)}\left(1 + n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}} 2^{-d'\left(s-\frac{1}{p}\right)}.
\end{aligned}$$

Last step is to control T_1. For brievity, we put $\beta_{-1,k} = \alpha_{0k}$ and $\psi_{-1k} = \phi_{0k}$. Then we have, applying successively Cauchy-Schwarz and Jensen inequalities,

$$\begin{aligned}
\mathbb{E}[T_1] &\leq \mathbb{E}\left|\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}}\left(\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right)^2\right)^{\frac{1}{2}}\right| \\
&\leq \sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}\left(\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Besides, $\mathbb{E}\epsilon_i\epsilon_j = 0$, $\forall i \neq j$. Then:

$$\mathbb{E}\left(\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right)^2 = \sum_{i=1}^n \mathbb{E}\psi_{jk}(X_i)^2.$$

We have to control, for $j \in \{-1, \dots d'\}$ the serie:

$$\sum_{k\in\mathbb{Z}}\sum_{i=1}^n \mathbb{E}\psi_{jk}(X_i)^2.$$

As above, since p and the wavelet mother ψ are compactly supported,

$$\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2 = \sum_{k\in\mathcal{S}(j)}\sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2$$

where $\sharp\mathcal{S}(j) \leq c2^j$. Finally we have:

$$\begin{aligned}
\mathbb{E}[T_1] &\leq c\sqrt{n}\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}2^{\frac{j}{2}} \\
&\leq c\sqrt{n}2^{\frac{d'}{2}}\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}} \\
&\leq 2^{\frac{d'}{2}}c\frac{\sqrt{rn}}{\sqrt{a}},
\end{aligned}$$

where last line comes from the definition of $\Gamma(R,r)$.
Then there exists a constant $c>0$ depending on a and A such that:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\alpha,\beta}(X_i)\right| \leq \frac{c}{\sqrt{n}}\inf_{d'\in\mathbb{N}}\left(R2^{-d'\left(s-\frac{1}{p}\right)} + 2^{\frac{d'}{2}}\sqrt{\frac{r}{a}}\right).$$

Optimizing with respect to d', we obtain the following upper bound in dimension 1:

$$\frac{c}{\sqrt{n}}R^{\frac{1}{2\left(\frac{1}{2}+s-\frac{1}{p}\right)}}r^{\frac{s-\frac{1}{p}}{2\left(\frac{1}{2}+s-\frac{1}{p}\right)}}.$$

Now we turn out into the d-dimensional case. The principle of the proof follows the one dimensional case. From (3.12) we have, for any $f\in\mathcal{B}_{spq}(\mathbb{R}^d)$:

$$f = \sum_{k\in\mathbb{Z}^d}\alpha_{0k}\phi_{0k} + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\beta_{jkl}\psi_{jkl} = f_{\alpha,\beta}(x).$$

Then we can write:

$$\mathbb{E}\sup_{f\in B(R):\|f\|_{L^2}^2\leq\frac{r}{a}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right| \leq \mathbb{E}\sup_{(\alpha,\beta)\in\Gamma_d(R,r)}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\alpha,\beta}(X_i)\right|,$$

where now:

$$\Gamma_d(R,r) = \{(\alpha,\beta) : \gamma_{pq}^d(\alpha,\beta)\leq R \text{ and } \gamma_2^d(\alpha,\beta)\leq\frac{\sqrt{r}}{\sqrt{a}c}\},$$

for:

$$\gamma_{pq}^d(\alpha,\beta) = \left(\sum_{k\in\mathbb{Z}^d}|\alpha_k|^p\right)^{\frac{1}{p}} + \left(\sum_{j\geq 0}\left(2^{j\left(s+d(\frac{1}{2}-\frac{1}{p})\right)}\sum_{l=1}^{2^d-1}\left(\sum_{k\in\mathbb{Z}^d}|\beta_{jkl}|^p\right)^{\frac{1}{p}}\right)^q\right)^{\frac{1}{q}},$$

and:

$$\gamma_2^d(\alpha,\beta) = \left(\sum_{k\in\mathbb{Z}^d} |\alpha_k|^2\right)^{\frac{1}{2}} + \sum_{j\geq 0}\sum_{l=1}^{2^d-1}\left(\sum_{k\in\mathbb{Z}^d}|\beta_{jkl}|^2\right)^{\frac{1}{2}}.$$

We proceed as in dimension 1. For any integer d':

$$\begin{aligned}\left|\sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i)\right| &\leq \left|\sum_{k\in\mathbb{Z}^d}\alpha_{0k}\sum_{i=1}^n \epsilon_i\phi_{0k}(X_i) + \sum_{j=0}^{d'}\sum_{l=1}^{2^d-1}\sum_{k\in\mathbb{Z}^d}\beta_{jkl}\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right| \\ &+ \left|\sum_{j>d'}\sum_{l=1}^{2^d-1}\sum_{k\in\mathbb{Z}^d}\beta_{jkl}\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right| := T_3 + T_4.\end{aligned}$$

We begin applying Hölder (twice) and Jensen inequalities to T_4:

$$\mathbb{E}[T_4] \leq R\left(\sum_{j>d'}\left(2^{-j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right|^{p'}\right)^{\frac{1}{p'}}\right)^{q'}\right)^{\frac{1}{q'}}.$$

Next step is to control, for all $j > d'$, the serie:

$$\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right|^{p'}.$$

We have with Lemma 3.2:

$$\mathbb{E}|Y_i|^p \leq A\|\psi_{jkl}\|_p^p \leq c2^{dj\left(\frac{p}{2}-1\right)},$$

since ψ is compactly supported. As a result, applying Rosenthal inequality, we get:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right|^{p'} \leq c_{p'}n^{-\frac{p'}{2}}\left[c^{\frac{p'}{2}} + c\left(\frac{2^{dj}}{n}\right)^{\frac{p'}{2}-1}\right].$$

Now it is worth noticing that since p and the wavelets functions ψ_{jkl} are compactly supported, the quantity:

$$\mathbb{E}|\psi_{jkl}|^{p'} = \int_{\mathbb{R}}|\psi_{jkl}(x)|^{p'}p(x)dx$$

is zero whatever $k \notin \mathcal{S}_d(j) := \{k\in\mathbb{Z}^d : supp(\psi_{jkl})\cap\mathcal{P}\neq\varnothing\}$. There exists an absolute constant $c>0$ which only depends on d such that $\sharp\mathcal{S}_d(j) \leq c2^{dj}$. As a result,

$$\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right|^{p'} = \sum_{k\in\mathcal{S}_d(j)}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jkl}(X_i)\right|^{p'}.$$

With previous inequality, we hence have:

$$\begin{aligned}\mathbb{E}[T_4] &\leq c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'}2^{-j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)q'}\left(\sum_{k\in\mathcal{S}_d(j)}c^{\frac{p'}{2}}+c\left(\frac{2^{dj}}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}}\\ &\leq c\frac{R}{\sqrt{n}}\left(\sum_{j>d'}2^{-jq'\left(s-\frac{d}{2}\right)}+2^{-jq'\left(s-\frac{d}{p}\right)}n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}}\\ &\leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{d}{2}\right)}+2^{-d'q'\left(s-\frac{d}{p}\right)}n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}}.\end{aligned}$$

where the convergence of the geometric serie comes from the condition $s>\frac{d}{p}$. Moreover we have $p\leq 2$ then, as above:

$$\mathbb{E}[T_4]\leq c\frac{R}{\sqrt{n}}2^{-d'\left(s-\frac{d}{p}\right)}.$$

It remains to control T_3. For brievity, we put $\beta_{-1k1}=\alpha_k$, $\psi_{-1k1}=\phi_k$ and for any $l>1$ $\beta_{-1kl}=0$ $\psi_{-1kl}=0$. Then we have, applying successively Cauchy-Schwarz and Jensen:

$$\mathbb{E}\,T_3\leq\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\beta_{jkl}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left(\sum_{i=1}^{n}\epsilon_i\psi_{jkl}(X_i)\right)^2\right)^{\frac{1}{2}}.$$

With the same argument as in dimension one, we obtain:

$$\mathbb{E}\left(\sum_{i=1}^{n}\epsilon_i\psi_{jkl}(X_i)\right)^2=\sum_{i=1}^{n}\mathbb{E}\psi_{jkl}(X_i)^2.$$

We have to control, for $j\in\{-1,\dots d'\}$ the serie:

$$\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\sum_{i=1}^{n}\mathbb{E}\psi_{jkl}(X_i)^2.$$

As above, since f and the wavelet mother ψ are compactly supported,

$$\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\sum_{i=1}^{n}\mathbb{E}\psi_{jkl}(X_i)^2=\sum_{k\in\mathcal{S}_d(j)}\sum_{l=1}^{2^d-1}\sum_{i=1}^{n}\mathbb{E}\psi_{jkl}(X_i)^2$$

where $\sharp\mathcal{S}_d(j)\leq c2^{dj}$. Finally we have:

$$\mathbb{E}[T_3]\leq c\sqrt{n}\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}|\beta_{jkl}|^2\right)^{\frac{1}{2}}2^{\frac{dj}{2}}\leq\sqrt{nr}2^{\frac{dd'}{2}},$$

from the definition of $\Gamma_d(R,r)$.
Then the control of the two terms entails:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\alpha,\beta}(X_i)\right| \leq \frac{c}{\sqrt{n}}\inf_{d'\in\mathbb{N}}\left(R2^{-d'\left(s-\frac{d}{p}\right)}+2^{\frac{dd'}{2}}\sqrt{r}\right).$$

Optimizing with respect to d', we lead to the conclusion.

3.5.2 Proof of Proposition 3.1

To prove the oracle inequality, we use the model selection approach of Blanchard et al. [24]. It is straightforward that minimization (3.10) can be rewritten as $\hat{f}_n = \hat{f}_{\widehat{R}}$ where:

$$\hat{f}_R = \arg\min_{f\in B(R)}\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ \text{ and } \widehat{R} = \arg\min_{R>0}\left(\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\hat{f}_R(X_i))_+ + \alpha_n R^2\right),$$

where $B(R) = \{f \in \mathcal{B}_{spq}(\mathbb{R}^d) : \|f\|_{spq} \leq R\}$. This gives a model selection interpretation of classifier $\hat{f}_n$, where models are balls in $\mathcal{B}_{spq}(\mathbb{R}^d)$. We can then apply a general model selection theorem (Theorem 5 in Blanchard et al. [24]). In our setup, we have to find constant b_R,C_R, a subroot function ϕ_R and a distance d on $L^2(P_X)$ such that (see also Chapter 1):

(H1) $\forall R \in \mathbb{R}$,$\forall g \in B(R)$,$\|g\|_\infty \leq b_R$;
(H2) $\forall g,g' \in L^2(P_X)$,$Var\left(l(g)-l(g')\right) \leq d^2(g,g')$;
(H3) $\forall R \in \mathbb{R}$,$\forall g \in B(R)$,$d^2(g,f^*) \leq C_R\mathbb{E}\left(l(g)-l(f^*)\right)$;
(H4) $\forall R \in \mathbb{R}$,$\forall r > 0$, we have:

$$\mathbb{E}\sup_{f\in B(R):d(f,0)^2\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right| \leq \phi_R(r).$$

Once assumptions **(H1)**-**(H4)** are granted, next step is to discretize the continuous family of models $(B(R))_{R\in\mathbb{R}}$ over a certain family of values of the radii. Following Blanchard et al. [24], we consider the set of discretized radii:

$$\mathcal{R} = \{M^{-1}2^k, k \in \mathbb{N}, 0 \leq k \leq \lceil \log_2 n \rceil\}.$$

To apply the second part of Theorem 5 in Blanchard et al. [24], the penalty function should satisfy:

$$pen(R) \geq c_1\left(\frac{r^*}{C_R} + \frac{(C_R+b_R)(x_R+\log 2)}{3n} + \frac{(C_R+b_R)\log(C_R+b_R))}{n}\right),$$

where c_1 is a suitable constant. It can be checked that condition (3.20) on α_n ensures such an inequality for:

$$pen(R) = \alpha_n\left(\phi\left(\frac{MR}{2}\right)+\frac{\eta_1}{\eta_0}\right).$$

Last step is to forth between the discretized framework and the continuous framework. The residual term $\frac{2}{n}$ appearing in the oracle inequality comes from the use of the second part of Theorem 5 in Blanchard et al. [24].
It only remains to prove **(H1)**-**(H4)**.

Proof of (H1)

To verify this hypothesis, we need the following lemma:

Lemma 3.5 *Let $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ such that $\|f\|_{spq} \leq R$. If $s > \frac{d}{p}$, then there exists $c > 0$ such that, for any $x \in \mathbb{R}^d$,*

$$|f(x)| \leq cR.$$

Proof. We only consider the one dimensional case. The d-dimensional case follows the same lines. Let $x \in \mathbb{R}$. With wavelet expansion (3.12), gathering with Hölder inequality:

$$\begin{aligned} |f(x)| &\leq \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^p\right)^{\frac{1}{p}} \left(\sum_{k\in\mathbb{Z}} |\phi_k(x)|^{p'}\right)^{\frac{1}{p'}} \\ &+ \left(\sum_{j\geq 0} \left(2^{j(s+1/2-1/p)} \left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\right)^{\frac{1}{p}}\right)^q\right)^{\frac{1}{q}} \times \\ &\left(\sum_{j\geq 0} 2^{-jq'(s+1/2-1/p)} \left(\sum_{k\in\mathbb{Z}} |\psi_{jk}(x)|^{p'}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}}. \end{aligned}$$

We have, for the scaling function ϕ and the mother wavelet ψ:

$$\forall x \in \mathbb{R}, \forall N \geq 1, |\phi(x)| \leq \frac{C_N}{(1+|x|)^N} \text{ and } |\psi(x)| \leq \frac{C_N}{(1+|x|)^N}.$$

Then for a given $x \in \mathbb{R}$, $j \in \mathbb{Z}$, there exist $\alpha_1 = \alpha_1(x) \in [0,1)$ and $\alpha_2 = \alpha_2(x,j) \in [0,1)$ such that:

$$\sum_{k\in\mathbb{Z}} |\phi_k(x)|^{p'} \leq \sum_{k\in\mathbb{N}} \left|\frac{C_N}{(1+k+\alpha_1)^N}\right|^{p'} + \sum_{k\geq 1} \left|\frac{C_N}{(1+k-\alpha_1)^N}\right|^{p'} \leq 2\sum_{k\geq 1} \frac{C_N^{p'}}{k^{Np'}} \leq c$$

and:

$$\sum_{k\in\mathbb{Z}} |\psi_{jk}(x)|^{p'} \leq 2^{\frac{jp'}{2}} \sum_{k\in\mathbb{Z}} \left|\frac{C_N}{(1+k+\alpha_2)^N}\right|^{p'} + \sum_{k\in\mathbb{Z}} \left|\frac{C_N}{(1+k-\alpha_1)^N}\right|^{p'} \leq c2^{\frac{jp'}{2}}.$$

Hence, we obtain:

$$\begin{aligned} |f(x)| &\leq c\left(\sum_{k\in\mathbb{N}} |\alpha_{0k}|^p\right)^{\frac{1}{p}} \\ &+ c\left(\sum_{j\leq 0} 2^{-jq'(s-1/p)}\right)^{\frac{1}{q'}} \left(\sum_{j\leq 0} 2^{jq(s+1/2-1/p)} \left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}. \end{aligned}$$

For $s > \frac{1}{p}$, we hence have:

$$|f(x)| \leq cR.$$

This leads to the conclusion. The proof is similar in dimension d.□

Remark 3.14 *This proof could be shortened as follows. From the continuous embedding of $\mathcal{B}_{spq}(\mathbb{R}^d)$ into $C(\mathbb{R}^d)$ for $s > \frac{d}{p}$, one gets for any $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$:*

$$\|f\|_\infty \leq c\|f\|_{spq}.$$

This leads to the conclusion of the lemma.
We hence obtain (H_1) with $b_R = 1 + cR$ since $|l(y,f(x))| \leq 1 + |f(x)|$.

Proof of (H2)-(H3)

To check these assumptions, we have to choose a distance d in $L^2(P_X)$. This choice has already been done implicitely in Theorem 3.1. This theorem will prove **(H4)** with the usual distance $d(g,g') = \mathbb{E}(g - g')^2$, for any $g,g' \in L^2(P_X)$. It comes from Section 3.2.1 which allows us to write the L^2-norm of a function in $\mathcal{B}_{spq}(\mathbb{R}^d)$, using wavelet decomposition. Then we consider the same distance to check **(H2)** and **(H3)**.
(H2) is trivially satisfied because the hinge loss l is a Lipschitz function. Moreover with Lemma 11 of Blanchard et al. [24], hypothesis (3.18) ensures **(H3)** with constant $C_R = 2\left(\frac{MR}{\eta_1} + \frac{1}{\eta_0}\right)$. The choice of the distance above corresponds to the setting (S1) in Blanchard et al. [24].

Proof of (H4)

The proof of **(H4)** has been done in Section 3.2.2.

3.5.3 Proof of Corollary 3.1

We only treat the particular case $\phi(x) = x$. Recall $a(\alpha_n)$ is defined by:

$$a(\alpha_n) = \inf_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\mathbb{E}(l(f) - l(f^*)) + \alpha_n \|f\|_{spq}\right).$$

By the Lipschitz property of the hinge loss, gathering with assumptions on the marginal of X, we have, for any $p \geq 1$:

$$\begin{aligned} a(\alpha_n) &\leq \inf_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(A\|f - f^*\|_{L^p(\mathbb{R}^d)} + \alpha_n \|f\|_{spq}\right) \\ &= c \inf_{R \in \mathbb{R}} \left(\inf_{f \in B(R)} \|f - f^*\|_{L^p(\mathbb{R}^d)} + \alpha_n R\right), \end{aligned}$$

where c depends on A.
The following lemma gives us an assumption on the regularity of f^* to control the first term of the above minimization:
Lemma 3.6 *For any $r < s$,*

$$f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d) \Rightarrow \inf_{f \in B(R)} \|f - f^*\|_{L^p(\mathbb{R}^d)} \leq \|f^*\|_{rp\infty}^{\frac{s}{s-r}} \left(\frac{1}{R}\right)^{\frac{r}{s-r}}.$$

Proof For the definition of interpolation space, we refer to Chapter 2 (see Triebel [131] for completeness). For instance Besov spaces have the following property:

$$\forall 0 < \theta < 1, (L^p(\mathbb{R}^d), \mathcal{B}_{spq}(\mathbb{R}^d))_{\theta,\infty} = \mathcal{B}_{\gamma p\infty}(\mathbb{R}^d),$$

where $\gamma = \theta s$ and $(L^p(\mathbb{R}^d), \mathcal{B}_{spq}(\mathbb{R}^d))_{\theta,\infty}$ is the interpolation space between $L^p(\mathbb{R}^d)$ and $\mathcal{B}_{spq}(\mathbb{R}^d)$. Using [118, Theorem 3.1] with $\theta = \frac{r}{s}$, we have the result. □
Using this lemma and optimizing with respect to R leads to:

$$a(\alpha_n) \leq c \inf_{R\in\mathbb{R}} \left(\left(\frac{1}{R}\right)^{\frac{r}{s-r}} + \alpha_n R \right) \leq c\alpha_n^{\frac{r}{s}}.$$

Using Proposition 3.1, we arrive at:

$$\mathbb{E}(l(\hat{g}_n) - l(f^*)) \leq 2\alpha_n^{\frac{r}{s}} + 4\alpha_n \left(4 + c\frac{\eta_1}{\eta_0}\right) + \frac{2}{n}.$$

Choosing α_n such that an equality holds in (3.20) concludes the proof.

Chapitre 4

Risk hull method and Kernel Projection Machines in classification

Abstract

This chapter is devoted to a new procedure of model selection called the Risk Hull Minimization (RHM). Initiated by Cavalier and Golubev [39] for statistical inverse problems, RHM proposes to select from the data the bandwidth of a spectral cut-off. It is based on the study of the stochastic variations of the square loss.
We propose to apply this method to the context of classification. We consider a toy model of classification where the noise level is constant over the input domain. We study a family of Kernel Projection Machines in a sequence space. It allows us to control the stochastic variations of the square loss and to provide an oracle inequality for the mean square risk. This chapter is a work in progress and gives several future directions.

4.1 Introduction

4.1.1 RHM: the original problem

Rish Hull Minimization has been introduced by Cavalier and Golubev [39] to solve the statistical inverse problem $Y = Af + \epsilon\xi$, where f has to be reconstructed from indirect and noisy observations Y. The operator A is assumed to be known and compact. Let ξ be a Gaussian white noise and $\epsilon > 0$ the noise level. If we look at the Singular Value Decomposition (SVD) of A, one gets:

$$y_k = b_k\theta_k + \epsilon\xi_k, k = 1, \dots \tag{4.1}$$

where ξ_k are independent identically distributed (i.i.d.) $\mathcal{N}(0,1)$ random variables, b_k are the singular values and θ_k are the coefficients of the signal f in the basis of eigenfunctions of A^*A. Considering (4.1), the aim is to estimate the sequence $(\theta_k)_{k\geq 1}$ from noisy observations $(y_k)_{k\geq 1}$. Ill-posed inverse problems are characterized by the property that $b_k \to 0$ as $k \to \infty$. As a result, for large k, observation y_k contains a large amount of noise. To estimate f, a standard way is to use the first N terms of (4.1) and define:

$$\hat{\theta}_k(N) = \frac{y_k}{b_k}\ \mathbb{1}(k \leq N). \tag{4.2}$$

The family $\{\hat{\theta}(N), N \geq 1\}$ is called the family of projection estimators (or spectral cut-off). The regularization parameter N is called the bandwidth. A major statistical problem is to choose this parameter. RHM method proposes a data-driven choice of N.

The idea at the core of the method is to focus on the square loss:

$$\|\hat{\theta}(N) - \theta\|^2 = \sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} \xi_k^2.$$

More precisely, we want to control the stochastic variations of the random process $\epsilon^2 \sum_{k=1}^{N} b_k^{-2} \xi_k^2$. We are looking at a quantity $V(N)$ such that:

$$\mathbb{E} \sup_{N \geq 1} \left[\epsilon^2 \sum_{k=1}^{N} b_k^{-2} \xi_k^2 - V(N) \right] \leq 0. \tag{4.3}$$

The quantity $V(N)$ is non-random. It ensures that for any data-driven choice $\tilde{N}$:

$$\mathbb{E}\|\hat{\theta}(\tilde{N}) - \theta\|^2 \leq \mathbb{E} \left[\sum_{k>\tilde{N}} \theta_k^2 + V(\tilde{N}) \right] := \mathbb{E} l(\theta, \tilde{N}). \tag{4.4}$$

If (4.4) holds, $l(\theta,N)$ is called a risk hull. It allows us to easily control the risk of any estimator in the family (4.2). At the first glance it seems from (4.4) that the smaller $l(\theta,N)$, the better we can control the risk. Unfortunately risk hull which looks very good at first sight, may be absolutely impossible to stochastically minimize. The statistical problem becomes to choose a risk hull in the (vast) set of all risk hulls.
Cavalier and Golubev [39] proposes a compromise to choose a stable risk hull. From the probabilistic viewpoint, it requires the control of the centered random process:

$$\eta_N = \epsilon^2 \sum_{k=1}^{N} b_k^{-2} (\xi_k^2 - 1). \tag{4.5}$$

In the Gaussian white noise model (4.1), this random process is a Wiener process. Using exponential inequalities for the large deviations of the process (4.5), we obtain (4.3).
Last step is to minimize the hull thanks to the data. To get a data-driven choice of N, Cavalier and Golubev [39] proposes to replace θ_k^2 by its unbiased estimates $b_k^{-2}(y_k^2 - \epsilon^2)$. We get the following adaptive procedure:

$$\widehat{N} = \arg\min_{N \geq 1} \left(- \sum_{k=1}^{N} b_k^{-2} y_k^2 + 2\epsilon^2 \sum_{k=1}^{N} b_k^{-2} + (1+\alpha) U_0(N) \right),$$

where

$$U_0(N) = \inf\{t > 0 : \mathbb{E}\eta_N \ \mathbb{1}(\eta_N \geq t) \leq \epsilon^2 b_1^{-2}\}.$$

A cornerstone idea of the approach is that $\widehat{N}$ minimizes $l(\theta,N)$ without significant loss. Therefore, with (4.4), one gets an oracle inequality of the form:

$$\mathbb{E}\|\hat{\theta}(\widehat{N}) - \theta\|^2 \leq C \min_{N \geq 1} \left[\sum_{k>N} \theta_k^2 + V(N) \right] + \textit{small term},$$

where $C > 0$ is a constant which can be arbitrarily close to 1. This inequality ensures that $\hat{\theta}(\widehat{N})$ performs as well as the best projection estimator over the family, up to a residual term. This term is the price to pay for the adaptive selection of $\widehat{N}$.

RHM can be embedded into a more general approach called penalized empirical risk minimization (penalized ERM). In this context, this method proposes the bandwidth choice:

$$\tilde{N} = \arg\min_{N \geq 1} \left(-\sum_{k=1}^{N} b_k^{-2} y_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2} + pen(N) \right), \tag{4.6}$$

where $pen(N)$ is a penalty function to determine. There exist several penalties proposed in the literature. One of the well-known approaches is the principle of unbiased risk estimation (URE). This method is extensively used in non-parametric statistics (Akaike [2], Mallows [93]). It is motivated by the existence of an unbiased estimator of the risk. According to the principle of URE, the data-driven choice of the bandwidth is defined to minimize this estimator based on the observations. In the statistical inverse problem setting (4.1), it corresponds to the penalty function $pen(N) = \epsilon^2 \sum_{k=1}^{N} b_k^{-2}$ in (4.6).
However several authors have shown the instability of this method in many statistical problems (for instance Golubev [60] for estimating linear functionals or Cavalier and Golubev [39] in statistical inverse problems). Larger penalties has been proposed to improve the performances of URE. RHM can be written as a penalized ERM where:

$$pen(N) = \epsilon^2 \sum_{k=1}^{N} b_k^{-2} + (1+\alpha) U_0(N).$$

This penalty may improve substancially the performances of URE method. Moreover the principal advantage is that it can be approximated using the Monte-Carlo method. As a result, the method can be directly computed from the data. It offers an explicit choice for the penalty.

However Cavalier and Golubev [39] focuses on regularization by projections. A generalization to several regularization methods, with better performances, is proposed in Marteau [96]. It requires in particular the control of more general random processes than the Wiener process (4.5). The aim of this chapter is to extend RHM to the context of binary classification and to explore the robustness of the method in another setting.

4.1.2 Binary classification

We have at our disposal a training set $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$ of i.i.d. random pairs (X_i,Y_i) of law P. For $i = 1,\dots,n$, $X_i \in \mathcal{X}$ is an input variable with associated class $Y_i \in \{-1,+1\}$. Given these observations, the goal is to predict class Y of a new observation X, where $(X,Y) \sim P$ is independent of D_n. In other words, we have to build a decision rule or classifier $f : \mathcal{X} \to \{-1,+1\}$. It is clear that the best possible decision rule is given by:

$$f^*(x) = \text{sign}(2\eta(x) - 1) = \text{sign}(f_\eta(x)),$$

where $\eta(x) = P(Y = 1|X = x)$. f^* is called the Bayes rule. Unfortunately it depends on the conditional probability function $\eta(x)$. As a result, f^* is not measurable with respect

to the observations.
Vapnik and Chervonenkis propose in the 70s to minimize an empirical cost based on the data. Given a loss function l, the ERM (Empirical Risk Minimizer) estimator consists in minimizing:

$$R_n^l(f) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i,f(X_i)), \tag{4.7}$$

where $l(Y_i,f(X_i))$ measures the loss of f at the point (X_i,Y_i).
In this chapter, we focus on the least square loss defined by:

$$l(y,f(x)) = (y - f(x))^2.$$

The minimizer of $\mathbb{E}l(Y,f(X))$ is often called the regression function. It is defined by $\mathbb{E}(Y|X = x)$. In binary classification, it corresponds to the function $f_\eta(x) = 2\eta(x) - 1$. Practical performances of algorithms using the least square loss, such as Least Square SVM or Regularized Least Square Classification have been studied (see Rifkin et al. [112] or Zeng and Zhao [150]).

However, minimizing the empirical l-risk (4.7) can lead to bad generalization performances. It gives a classifier that fits exactly the data. It does not take into account the presence of noise. To avoid this phenomena of overfitting, a possible way is to measure the complexity of the solution. This is what we call a regularization method. Tikhonov and Arsenin (see Tikhonov and Arsenin [130]) propose a regularization scheme called Tikhonov regularization for solving inverse problems. It has been applied to learning with rather great success. The main example is the well-known SVM-type minimization:

$$\min_{f\in\mathcal{H}_K}\left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i,f(X_i)) + \alpha\|f\|^2_{\mathcal{H}_K}\right),$$

where $\mathcal{H}_K$ is commonly a reproducing kernel Hilbert space.
In this paper, we propose to study another regularization scheme: regularization by finite dimensional projections. This regularization method is an alternative to Tikhonov. It has been extensively studied in regression (see Baraud [9] or Birgé and Massart [23]) or in inverse problems (see above). An investigation in classification has been proposed in Blanchard and Zwald [27]. The KPM (for Kernel Projection Machines) algorithm minimizes an empirical risk over a finite dimensional subspace of a RKHS. From statistical point of view, it avoids some technical details compared to the SVM method. Here we propose to study the KPM algorithm using the least square loss. Given an orthonormal basis $(\phi_k)_{k\geq 1}$ of $\mathcal{H}_K$, we associate to each dimension $N \geq 1$ the following classifier:

$$\hat{f}_N = \arg\min_{f\in<\phi_1,\dots,\phi_N>} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2. \tag{4.8}$$

The regularization consists in projecting f into a N-dimensional subspace of $\mathcal{H}_K$. In practice, we choose $(\phi_k)_{k\geq 1}$ as the eigenfunctions of the empirical covariance operator associated to the kernel K. It leads to an interpretation of the selection of N as an optimal dimension in a dimensionality reduction method called KPCA (Kernel Principal Component Analysis).

Blanchard and Zwald [27] considers the problem of selecting the dimension N in the family of KPM, thanks to a model selection approach. The choice of N minimizes a penalized empirical criterion as follows:

$$\widehat{N} = \arg\min_{N\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i,\hat{f}_N(X_i)) + \lambda N\right),$$

where λ has to be suitably chosen. The order of the penalty (linear with the dimension) is justified theoretically by a general model selection theorem.

4.2 Working in a sequence space

The aim of this section is to link the RHM method presented in Section 4.1.1. with the binary classification set-up of this manuscript. We want to use the new model selection approach of the risk hull to solve the statistical problem of choosing N in (4.8). We act in two steps:

- We write observations Y_i as a function of the inputs X_i blurred by some discrete random noise.
- We make a linear transformation of these observations thanks to a kernel.

It gives a representation of the family of Kernel Projection Machines in the spectral domain. However, to perform RHM in classification, we make an additional assumption over the behavior of the conditional probability function η.

4.2.1 Classification toy model

We propose to adopt the regression point of view. Given the training set $D_n = \{(X_i,Y_i), i = 1,\dots,n\}$, we can write:

$$Y_i = f_\eta(X_i) + \sigma_i, \tag{4.9}$$

where $f_\eta(x) = 2\eta(x) - 1$. It gives new independent random variables $\sigma_i = Y_i - f_\eta(X_i)$, $i = 1,\dots,n$. These variables represent the noise in the observations. If we consider the zero error case (i.e. $R(f^*) = 0$), it is easy to see that in (4.9) $\sigma_i = 0$, $i = 1,\dots,n$. In this case there is no noise. If we make no assumption, we have, conditioning with respect to the X_i:

$$\begin{aligned}\mathbb{E}(\sigma_i|X_i) &= (1 - f_\eta(X_i))\eta(X_i) + (-1 - f_\eta(X_i))(1 - \eta(X_i)) \\ &= 2(1 - \eta(X_i))\eta(X_i) - 2(\eta(X_i)(1 - \eta(X_i)) = 0,\end{aligned}$$

and for the same reasons:

$$\mathbb{E}(\sigma_i^2|X_i) = 4(1 - \eta(X_i))^2\eta(X_i) + 4\eta(X_i)^2(1 - \eta(X_i)) = 4\eta(X_i)(1 - \eta(X_i)).$$

It is important to note that the law of σ_i depends on the location of the input X_i. In a region where η is close to $1/2$, the variance of σ_i will be close to 1. On the contrary if η is close to 1 or 0, the variance will be smaller. This is called in regression the heteroscedastic case. The level of noise is not constant with respect to the design points. To avoid this difficulty, we propose to introduce the following assumption on the distribution P:

$$\eta(x) = (1 \pm h)/2, \text{ for any } x \in \mathcal{X}, \tag{4.10}$$

for a given $0 < h < 1$. Parameter h is the classification ability of the model. If h is close to 0, $R(f^*) = 1/2$ and the problem is hopeless. For increasing values of h, the noise in the couple (X,Y) of law P decreases. This hypothesis is a particular case of the strong margin assumption usually used in classification (Blanchard et al. [24], Massart and Nédélec [102], Blanchard and Zwald [27]). The distribution P has strong margin h_0 if:

$$\left|\eta(x) - \frac{1}{2}\right| \geq h_0 > 0, \text{ for any } x \in \mathcal{X},$$

where $h_0 > 0$ is called the margin. (4.10) implies in particular a strong margin $h_0 = \frac{h}{2}$. For our interests, assumption (4.10) ensures the noise to be constant in $\mathcal{X}$. Under this hypothesis, σ_i in (4.9) is such that, for any $i = 1, \dots n$:

$$\mathbb{E}(\sigma_i^2 | X_i) = (1+h)(1-h).$$

The level of noise in the regression model (4.9) does not depend on the design points. It allows us to simplify numerous technical details in the proofs. Replacing (4.10) by the standard strong margin assumption is an interesting future direction though it is out of the scope of the present work.

4.2.2 Sequence space model for classification

To study the RHM method, we consider a linear transformation of (4.9). Given a family $\{\varphi_k(X_i), k = 1, \dots, n\}$, we define the sequence:

$$y_k := \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_k(X_i), k = 1, \dots, n. \tag{4.11}$$

Such a transformation has been done in Wahba [144] for smoothing splines (see also Cao and Golubev [33]). It gives a new representation of the data. Next result proposes a special choice of basis $\{\varphi_k(X_i), k = 1, \dots, n\}$, related to the orthonormal basis used in (4.8). It involves a spectral representation of the estimators $\hat{f}_N$.

Lemma 4.1 *Let $K : \mathcal{X}^2 \to \mathbb{R}$ be a symmetric and positive definite application. Let $(\mathcal{H}_K, <\cdot,\cdot>_K)$ the unique RKHS with reproducing kernel K. Consider the empirical covariance operator $C_n : \mathcal{H}_K \to \mathcal{H}_K$ defined by $C_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)K(X_i,\cdot)$. If we write $(\phi_k)_{k\geq 1}$ its eigenfunctions and $(\lambda_k)_{k\geq 1}$ the corresponding eigenvalues, we have:*

1. $$\frac{1}{n}\sum_{i=1}^{n} \frac{\phi_k(X_i)}{\sqrt{\lambda_k}} \frac{\phi_l(X_i)}{\sqrt{\lambda_l}} = \delta_{kl} \text{ and } <\phi_k, \phi_l>_K = \delta_{kl},$$

   where δ_{kl} is the Kronecker's delta.

2. *If we choose $\left\{\varphi_k(X_i) = \frac{\phi_k(X_i)}{\sqrt{\lambda_k}}, i = 1, \dots, n\right\}$ in (4.11), estimator $\hat{f}_N$ defined in (4.8) can be written in the basis φ_k as follows:*

$$\hat{f}_N = \sum_{k=1}^{N} y_k \frac{\phi_k}{\sqrt{\lambda_k}} = \sum_{k=1}^{n} y_k \ \mathbb{1}(k \leq N)\varphi_k.$$

As a result, we can identify the family of KPM estimators $\{\hat{f}_N, N \geq 1\}$ *to the family* $\left\{\hat{\theta}(N) = y_k \ \mathbb{1}(k \leq N), N \geq 1\right\}$.

Proof. 1. From basic RKHS theory, we have that $\phi_k(X_i) = \sqrt{n\lambda_k}V_k^i$ where $(V_k)_{k=1...,n}$ is an orthonormal basis of $\mathbb{R}^n$ made of eigenvectors of the kernel matrix $K_n = (K(X_i,X_j))_{i,j=1,..n}$.
2. From (4.8), we have:

$$\arg\min_{f\in<\phi_1,...,\phi_N>} \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 = \arg\min_{\alpha\in\mathbb{R}^N} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{k=1}^N \alpha_k\phi_k(X_i)\right)^2$$

$$= \arg\min_{\alpha\in\mathbb{R}^N} \frac{1}{n}\sum_{k=1}^N \left[-2\alpha_k \sum_{i=1}^n Y_i\phi_k(X_i) + \alpha_k \sum_{k'\neq k} \alpha_{k'} \sum_{i=1}^n \phi_k(X_i)\phi_{k'}(X_i) + \alpha_k^2 \sum_{i=1}^n \phi_k(X_i)^2\right].$$

Using 1. and the linear transformation (4.11):

$$\arg\min_{\alpha\in\mathbb{R}^N} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{k=1}^N \alpha_k\phi_k(X_i)\right)^2 = \arg\min_{\alpha\in\mathbb{R}^N} \sum_{k=1}^N \left[-2\alpha_k\sqrt{\lambda_k}y_k + \alpha_k^2\lambda_k\right].$$

Minimizing over α_k concludes the proof. □
Lemma 4.1 provides a representation of (4.8) in the spectral domain. As a result, we make the linear transformation (4.11) with $\left\{\varphi_k(X_i) = \frac{\phi_k(X_i)}{\sqrt{\lambda_k}}, i = 1,\dots,n\right\}$. Under (4.10), we can write:

$$y_k = \theta_k + \epsilon\xi_k, \tag{4.12}$$

where

$$\theta_k = \frac{1}{n}\sum_{i=1}^n f_\eta(X_i)\frac{\phi_k(X_i)}{\sqrt{\lambda_k}},\ \epsilon = \frac{\sqrt{(1+h)(1-h)}}{\sqrt{n}},$$

and

$$\xi_k = \frac{1}{\sqrt{n(1+h)(1-h)}}\sum_{i=1}^n \sigma_i\frac{\phi_k(X_i)}{\sqrt{\lambda_k}}.$$

It is important to note that in (4.12), ξ_k are zero mean with variance $\mathbb{E}\xi_k^2 = 1$. Indeed, from the proof of Lemma 4.1, we can write:

$$\xi_k = \sum_{k=1}^n V_k^i\sigma_i',$$

where $\sigma_i' = \frac{\sigma_i}{\sqrt{(1+h)(1-h)}}$ is zero mean with variance 1 and $\{V_k, k = 1,\dots,n\}$ is an ONB of $\mathbb{R}^n$. As a result, the sequence ξ_k, $k = 1,\dots,n$ is such that $\mathbb{E}\xi_k\xi_l = \delta_{kl}$. However in comparison with (4.1), $(\xi_k)_k$ are neither Gaussian nor independent random variables. This fact has strong statistical consequences. It illustrates rather well the difference between classification and more standard statistical models. In the appendix we give another interpretation of the sequence space model of classification (4.12) in terms of the diagonalization of the empirical covariance operator associated to K.

4.2.3 RHM for classification

Given the family of KPM $\{\hat{f}_N, N \geq 1\}$, we are looking at a data-driven choice of N. We want to select N such that the square risk, defined by

$$R(\hat{f}_N,f_\eta) = \mathbb{E}(\hat{f}_N(X) - f_\eta(X))^2,$$

is minimum. The principle of the method follows Cavalier and Golubev [39]. Given a training set D_n, we focus on the square loss, defined as:

$$r(\theta,N) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_N(X_i) - f_\eta(X_i))^2.$$

From Lemma 4.1, this loss can be written in the sequence space (4.12) as follows:

$$r(\theta,N) = \sum_{k=1}^{n}(\hat{\theta}_k(N)^2 - \theta_k)^2 = \sum_{k>N}^{n}\theta_k^2 + \epsilon^2\sum_{k=1}^{N}\xi_k^2. \tag{4.13}$$

In (4.13), the two terms are random: $\sum_{k>N}\theta_k^2$ depends on the design points X_i, $i = 1,\dots,n$. The second term $\epsilon^2\sum_{k=1}^{N}\xi_k^2$ is also random. It depends on the design and the noise σ_i, $i = 1,\dots,n$. To apply the RHM heuristic to classification, we fix the design and concentrate into the randomness of the noise.
Considering (4.13), we are looking at a quantity $V(N)$ such that:

$$\mathbb{E}\sup_{N\geq 1}\left[\epsilon^2\sum_{k=1}^{N}\xi_k^2 - V(N)|X_1,\dots,X_n\right] \leq 0.$$

Here $V(N)$ depends on the random variables X_i. As a result it is random. However, it controls the stochastic variations of (4.13) produced by the noise. For any data-driven choice $\widehat{N}$, we have:

$$\mathbb{E}(r(\theta,\widehat{N})|X_1,\dots X_n) \leq \mathbb{E}\left(\sum_{k>\widehat{N}}^{n}\theta_k^2 + V(\widehat{N})|X_1,\dots X_n\right),$$

and integrating with respect to the design, one gets:

$$R(\hat{f}_N,f_\eta) \leq \mathbb{E}\left(\sum_{k>\widehat{N}}^{n}\theta_k^2 + V(\widehat{N})\right).$$

The quantity $\sum_{k>N}^{n}\theta_k^2 + V(N)$ will be called a risk hull.
To get a risk hull in classification, we study the random process:

$$\zeta(N) = \epsilon^2\sum_{k=1}^{N}(\xi_k^2 - 1), N \geq 1, \tag{4.14}$$

where ϵ and ξ_k are defined in (4.12). From a probabilistic viewpoint, the present paper can be viewed as a generalization of Cavalier and Golubev [39]. When dealing with the classical

Gaussian setting, variables ξ_k in (4.1) are i.i.d. $\mathcal{N}(0,1)$. The random process η_N is a Wiener process. In our context, the presence of non-gaussian random noise σ_i in (4.9) leads to non i.i.d. random variables ξ_k in (4.11). Hence, we want the same kind of results but with more general random process. Following Cao and Golubev [33], Golubev [63] or more recently Marteau [96], we introduce a simple notion of ordered process $\zeta(t)$ characterized by the property $\mathbb{E}\zeta(t_1)\zeta(t_2) \geq \min(\mathbb{E}\zeta(t_1)^2, \mathbb{E}\zeta(t_2)^2)$. Such a process ζ is such that, for any $p > 1$,

$$\mathbb{E} \sup_{t \geq 0} [\zeta(t) - \mu\sigma(t)^p] \leq \frac{C(p)}{\mu},$$

where $\sigma(t)$ is the standard deviation of $\zeta(t)$.

The rest of the paper is organized as follows. In Section 3, we propose a risk hull and state the oracle efficiency of the method. This is the main result of the chapter. This chapter is a work in progress and offers several open problems discussed in Section 4. Section 5 is dedicated to the proofs of the main results whereas in Section 6 is presented the theory of ordered processes.

4.3 Oracle efficiency of the method

The essence of the risk hull method is rather simple. It consists of two steps: (1) the construction of a risk hull; (2) the minimization of the risk hull based on the observations. A risk hull is proposed in Lemma 4.2 thanks to the theory of ordered processes. Theorem 4.1 is an oracle inequality. It quantifies the price to pay for (2) comparing the square risk of the data-driven estimator with the best among the family.

4.3.1 A risk hull

Lemma 4.2 *Suppose there exists a constant $\kappa > 0$ such that:*

$$\sup_{N \geq 1} \mathbb{E} \exp \left(\kappa \frac{\sum_{k=1}^N (\xi_k^2 - 1)}{\sqrt{\mathbb{E}\left(\sum_{k=1}^N (\xi_k^2 - 1)\right)^2}} \right) < \infty. \tag{4.15}$$

Moreover, suppose (4.10) *holds for $h \geq \frac{1}{\sqrt{3}}$. Denote $\Sigma(N) = \sqrt{var(\sum_{k=1}^N \xi_k^2)}$. Then for any $p > 1$, there exists a constant $C > 0$ such that for any $\alpha > 0$,*

$$l_\alpha(\theta, N) = \sum_{k > N} \theta_k^2 + \epsilon^2 N + \alpha\epsilon^2 \Sigma(N)^p + \frac{C\epsilon^2}{\alpha}$$

is a risk hull, i.e.

$$\mathbb{E} \sup_{N \geq 1} [r(\theta, N) - l_\alpha(\theta, N)] \leq 0.$$

The proof is presented in Section 5. It is a direct consequence of Lemma 4.5 applied to the ordered process (4.14).

Remark 4.1 *Hypothesis* (4.15) *appears to control the stochastic ordered process* (4.14). *If the variables ξ_k are i.i.d. $\mathcal{N}(0,1)$,* (4.15) *holds. This is the case in the Gaussian white noise*

model. Here ξ_k *are neither Gaussian nor independent. However,* $\xi_k = \sum_{i=1}^{n} V_k^i \sigma_i'$ *where* σ_i' *are zero mean with variance 1 and* $(V_k)_{k=1,\dots,n}$ *are the eigenvectors of the kernel matrix. There is nice hope that* (4.15) *holds.*

Remark 4.2 *The condition* $h \geq \frac{1}{\sqrt{3}}$ *is related to the law of* σ_i *under assumption* (4.10)*. In our classification toy model, it ensures the kurtosis of* σ_i *to be greater than* 3 *(see Remark 4.8 in the appendix for a definition). We propose in the Appendix a generalization of this assumption in regression with non gaussian random noise.*

To complete the minimization of $l_\alpha(\theta,N)$, we use the principle of the URE method. We replace each θ_k^2 by its unbiased estimates $y_k^2 - \epsilon^2$. We arrive at the following choice of N:

$$N_{rhm} = \arg\min_{N\geq 1} \left[-\sum_{k=1}^{N} y_k^2 + 2\epsilon^2 N + \alpha\epsilon^2 \Sigma(N)^p \right] := \arg\min_{N\geq 1} \overline{R}_{rhm}(y,N). \tag{4.16}$$

Similarly to the gaussian case, the rish hull method can be embedded into the class of penalized URE method. Here the penalty depends on the standard deviation of the process. It ensures the stability of the procedure. The URE method for classification consists here in minimizing an unbiaised estimate of the mean square risk. This method is studied both from theoretical and practical viewpoint in the Appendix of this thesis.

4.3.2 Oracle inequality

Theorem 4.1 *Let* $\mathcal{H}_K$ *a RKHS of kernel* K*. We consider the family of projection estimators* (4.8)*, and the data-driven estimator* $\hat{f}_{N_{rhm}}$ *defined in* (4.16) *for* $\alpha > 0$ *and* $p > 1$*. Suppose* (4.15) *holds and* (4.10) *holds for* $h \geq \frac{1}{\sqrt{3}}$*. Then there exist constants* B*,* C*,* $D > 0$ *and* $\gamma_1 > 0$ *independent of* n *such that for any* $0 < \gamma < \gamma_1$*,*

$$\mathbb{E}(\hat{f}_{N_{rhm}}(X) - f_\eta(X))^2 \leq (1 + B\gamma) \inf_{N\geq 1} R_{rhm}(f_\eta,N) + \frac{(1+h)(1-h)}{n}\left(\frac{BC}{\alpha} + \frac{BD}{2\gamma}\right),$$

where $R_{rhm}(f_\eta,N) = \mathbb{E}(\hat{f}_N(X) - f_\eta(X))^2 + \alpha\frac{(1+h)(1-h)}{n}\Sigma(N)^p$*.*

Proof is presented in Section 5. It is based on the theory of ordered processes summarized in Section 6.

Remark 4.3 *The statistical sense of Theorem 4.1 is rather clear. Estimator* $\hat{f}_{N_{rhm}}$ *performs as well as the best KPM over the family, up to a residual term. This term depends on the amount of noise in the model and the number of observations. This is the price to pay for minimizing an estimate of the hull of Lemma 4.2.*

Remark 4.4 *Theorem 4.1 is an oracle inequality. It only compares the classifier* $\hat{f}_{N_{rhm}}$ *with the best KPM. As a result the quality of our procedure strongly depends on the quality of the best projection estimator. An interesting future direction will be to state learning rates to the Bayes. It depends on the behavior of the approximation error* $\inf_{N\geq 1} R_{rhm}(f_\eta,N)$*.*

Remark 4.5 *This oracle inequality controls the mean square risk of $\hat{f}_{N_{rhm}}$. It comes from the use of the least square loss $l(y,f(x)) = (y - f(x))^2$ in the minimization. This loss gives us many technical advantages. However in classification we usually want to control the excess risk of a classifier, defined as:*

$$\epsilon(\hat{f},f^*) = P(\hat{f}(X) \neq Y) - P(f^*(X) \neq Y).$$

Using for instance Bartlett et al. [14], it is possible to bound the excess classification risk in terms of the excess square risk. With Theorem 4.1, this leads to a control of the excess risk of our procedure.
However it must be better to consider a more direct approach to control the excess risk, considering directly the classification loss or the hinge loss in the empirical minimization procedure. This is a future direction.

4.4 Conclusion

This chapter tries to link the Risk Hull Minimization method with the context of classification. We consider a family of Kernel Projection Machines $\{\hat{f}_N, N \geq 1\}$. We write these estimators as a linear combination of the observations:

$$y_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_k(X_i),$$

where φ_k depends on the kernel generating the RKHS $\mathcal{H}_K$ in the procedure. This allows us to work in a sequence space and to study the stochastic variations of the square loss. It involves a random process which is shown to be ordered. It results in an oracle inequality for the procedure. This result has to be considered as preliminary into the application of RHM viewpoint in classification. Below we list several open problems. We invite the interested reader to go to the end of this thesis for a more general discussion about classification and risk hull.

In this chapter, we focus on regularization by finite dimensional projections. It will be interesting to consider also the least-square SVM minimization:

$$\arg\min_{f\in\mathcal{H}_K} \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \alpha\|f\|_K^2 \right). \tag{4.17}$$

In this case, RHM should give a way of choosing the smoothing parameter α in (4.17). It is easy to see that the solution of (4.17) can be written, as $n \to \infty$:

$$\hat{f}_\alpha = \sum_{k=1}^{n} h_k(\alpha) y_k \phi_k \text{ where } h_k(\alpha) = \frac{1}{1+\alpha/\lambda_k},$$

for $(\lambda_k)_{k=1,\dots,n}$ the eigenvalues of the kernel matrix. The family $\{\hat{f}_\alpha, \alpha \in \mathbb{R}^+\}$ can be identify to the family of coefficients $\{h_k(\alpha), \alpha \in \mathbb{R}^+\}$. The square loss of $\hat{f}_\alpha$ is given by:

$$r(\theta,\alpha) = \sum_{k=1}^{n}(1 - h_k(\alpha))^2\theta_k^2 + \epsilon^2\sum_{k=1}^{n} h_k(\alpha)^2\xi_k^2.$$

As a result, to find a hull, we have to consider the random process:

$$\eta(\alpha) = \epsilon^2 \sum_{k=1}^{n} h_k(\alpha)^2(\xi_k^2 - 1).$$

The control of the stochastic variations of $\eta(\alpha)$ may lead to an oracle inequality.

Moreover in this chapter we make a strong hypothesis. It concerns the distribution of the observations. We suppose in (4.9) that the σ_i are random variables with constant variance:

$$\mathbb{E}\sigma_i^2 = (1+h)(1-h).$$

In this case, the probability function $\eta(x) = \mathbb{P}(Y = 1|X = x) = (1 \pm h)/2$ for any $x \in \mathcal{X}$. It corresponds to a constant noise level in the couple (X,Y) of law P. This assumption can be relaxed. If we make no assumption on η, we have an heteroscedastic regression problem

$$Y_i = f_\eta(X_i) + \kappa_i \sigma_i,$$

where $\kappa_i = \eta(X_i)(1 - \eta(X_i))$. In this case the level of noise depends on the design points X_i. As a result, to state an oracle inequality for the family of KPM, we have to study the random process:

$$\zeta(N) = \epsilon^2 \sum_{k=1}^{N} (\xi_k^2 - \delta_k^2),$$

where $\delta_k^2 = \mathbb{E}\xi_k^2 \neq 1$.

Finally in this chapter we only consider the least square loss. It gives many technical advantages and gives a first step into the application of the risk hull minimization in classification. However an interesting future direction will be to consider more standard losses such as the classification loss or the hinge loss. It would be possible to consider more classical regularization schemes such as the SVM algorithm using the hinge loss.

4.5 Proofs

In the sequel, we write $\mathbb{E}_n = \mathbb{E}(\cdot|X_1, \dots X_n)$ for the expectation conditioning to the $X_i, i = 1, \dots, n$.

4.5.1 Proof of Lemma 4.2

Lemma 4.2 immediately follows from this result, gathering with Lemma 4.5.

Lemma 4.3 *The processes $\zeta(N)$ defined by:*

$$\zeta(N) = \epsilon^2 \sum_{k=1}^{N} (\xi_k^2 - 1)$$

is ordered for $h \geq \frac{1}{\sqrt{3}}$.

Proof. It is clear that $\mathbb{E}\zeta(N) = 0$. Then

$$\begin{aligned} \text{var}\zeta(N) &= \mathbb{E}\zeta(N)^2 \\ &= \epsilon^4 \sum_{k=1}^{N} \left(\mathbb{E}(\xi_k^2 - 1)^2 + \sum_{k' \neq k} cov(\xi_k^2, \xi_{k'}^2) \right). \end{aligned}$$

To show that $\text{var}\zeta(N) \leq \text{var}\zeta(M)$ for $N \leq M$, it remains to show that $cov(\xi_k^2, \xi_{k'}^2) \geq 0$, i.e. $\mathbb{E}\xi_k^2\xi_{k'}^2 \geq 1$ since

$$cov(\xi_k^2, \xi_{k'}^2) = \mathbb{E}\xi_k^2\xi_{k'}^2 - \mathbb{E}\xi_k^2\mathbb{E}\xi_{k'}^2 = \mathbb{E}\xi_k^2\xi_{k'}^2 - 1.$$

We have, conditioning with respect to the X_i:

$$\mathbb{E}_n(\xi_k^2\xi_{k'}^2) = \frac{6h^2 - 2}{(1+h)(1-h)} \sum_{i=1}^{n} V_{ki}^2 V_{k'i}^2 + 2\delta_{kk'} + 1,$$

where $\delta_{kk'} = 1$ if $k = k'$ and zero otherwise. As a result, a sufficient condition to have $\mathbb{E}(\xi_k^2\xi_{k'}^2) \geq 1$ for $k \neq k'$ is:

$$\frac{6h^2 - 2}{(1+h)(1-h)} \geq 0 \Leftrightarrow h \geq \frac{1}{\sqrt{3}}. \tag{4.18}$$

Integrating with respect to the X_i, we hence have $cov(\xi_k^2, \xi_{k'}^2) \geq 0$ and then $\text{var}\zeta(N) \leq \text{var}\zeta(M)$ for $N \leq M$.
Finally, notice that, for any $N \leq M$,

$$\begin{aligned} \mathbb{E}\zeta(N)^2 &= \epsilon^4 \mathbb{E} \left(\sum_{k=1}^{n} \mathbb{1}(k \leq N)\ \mathbb{1}(k \leq M)(\xi_k^2 - 1) \right)^2 \\ &= \mathbb{E}\zeta(N)\zeta(M) + \epsilon^4 \mathbb{E} \sum_{k=1}^{n} \left(\sum_{k' \neq k} \mathbb{1}(k \leq N)(\ \mathbb{1}(k' \leq N) - \ \mathbb{1}(k' \leq M)) cov(\xi_k^2, \xi_{k'}^2) \right) \\ &\leq \mathbb{E}\zeta(N)\zeta(M) \end{aligned}$$

where last line uses again $cov(\xi_k^2, \xi_{k'}^2) \geq 0$. As a result, (4.19) holds and the random process $\zeta(N)$ is ordered. □

4.5.2 Proof of Theorem 4.1

In view of Lemma 4.2, for any $\mu > 0$, we have for $\tilde{N} = N_{rhm}$:

$$\mathbb{E}r(\theta, \tilde{N}) \leq \mathbb{E}l_\mu(\theta, \tilde{N}).$$

On the other hand, since $\tilde{N}$ minimizes $\bar{R}_{rhm}(y, N)$, for any $N \geq 1$:

$$\mathbb{E}\bar{R}_{rhm}(y, \tilde{N}) \leq \mathbb{E}\bar{R}_{rhm}(y, N) = R_{rhm}(\theta, N) - \mathbb{E}\sum_{k=1}^{n} \theta_k^2,$$

where $R_{rhm}(\theta,N)$ is defined in Theorem 4.1. Now it is easy to see that

$$\bar{R}_{rhm}(y,\tilde{N}) + \sum_{k=1}^{n} \theta_k^2 + \frac{C(p)\epsilon^2}{\mu} = l_\mu(\theta,\tilde{N}) - 2\epsilon \sum_{k=1}^{\tilde{N}} \theta_k \xi_k - \epsilon^2 \sum_{k=1}^{\tilde{N}} (\xi_k^2 - 1) + (\alpha - \mu)\epsilon^2 \Sigma(\tilde{N})^p.$$

Therefore, using this equation with previous ones, we obtain for any integer N:

$$\begin{aligned}
\mathbb{E}r(\theta,\tilde{N}) &\leq \mathbb{E}\bar{R}_{rhm}(y,\tilde{N}) + \mathbb{E}\sum_{k=1}^{n} \theta_k^2 + \frac{C(p)\epsilon^2}{\mu} \\
&+ 2\epsilon\mathbb{E}\sum_{k=1}^{\tilde{N}} \theta_k \xi_k + \mathbb{E}\left(\epsilon^2 \sum_{k=1}^{\tilde{N}} (\xi_k^2 - 1) - (\alpha - \mu)\epsilon^2 \Sigma(\tilde{N})^p \right) \\
&\leq R_{rhm}(f_\eta,N) + \frac{C(p)\epsilon^2}{\mu} + 2\epsilon\mathbb{E}\sum_{k=1}^{\tilde{N}} \theta_k \xi_k + \epsilon^2 \mathbb{E}\left(\sum_{k=1}^{\tilde{N}} (\xi_k^2 - 1) - (\alpha - \mu)\Sigma(\tilde{N})^p \right).
\end{aligned}$$

Next step is to control the last two terms in the previous bound. To bound the first term, consider the ordered process $\rho(N) = \epsilon \sum_{k=1}^{N} \theta_k \xi_k$. Conditionning with respect to the $X_i, i = 1,\ldots,n$, we have for any $N_0 \geq 1$:

$$\begin{aligned}
\mathbb{E}_n(\rho(\tilde{N})) &= \mathbb{E}_n\left(\epsilon \sum_{k=1}^{n} \Big(\mathbb{1}(k \leq \tilde{N}) - \mathbb{1}(k \leq N_0) \Big) \theta_k \xi_k \right) \\
&= \mathbb{E}_n(\rho'(\tilde{N})).
\end{aligned}$$

Note that $\rho'(N)$ is ordered on $[N_0,\infty[$ and $\rho''(t) = \rho'(\lfloor t^{-1} \rfloor)$ is also ordered on $[N_0^{-1},\infty[$. Hence using Lemma 4.6, we have:

$$\mathbb{E}_n\left(\left| \rho(\tilde{N}) \right| \right) = \mathbb{E}_n\left(\left| \rho'(\tilde{N}) \right| \right) \leq \log^2(K)\mathbb{E}_n\left(\sqrt{ \epsilon^2 \sum_{k=1}^{n} \Big(\mathbb{1}(k \leq \tilde{N}) - \mathbb{1}(k \leq N_0) \Big)^2 \theta_k^2 } \right)$$
$$+ \frac{C}{K} \sqrt{ \mathbb{E}_n\left(\epsilon^2 \sum_{k=1}^{n} \Big(\mathbb{1}(k \leq \tilde{N}) - \mathbb{1}(k \leq N_0) \Big)^2 \theta_k^2 \right) }.$$

Hence we obtain, for any $N_0 \geq 1$:

$$\mathbb{E}_n\left(\rho(\tilde{N}) \right) \leq D(K) \sqrt{ \mathbb{E}_n\left(\left[\epsilon^2 \sum_{k=N_0+1}^{n} \theta_k^2 + \epsilon^2 \sum_{k=\tilde{N}+1}^{n} \theta_k^2 \right] \right) },$$

where $D(K) = \log^2(K) + \frac{C}{K}$ for any $K > 1$. Now using Young's inequality, one gets, for any $\gamma > 0$:

$$\mathbb{E}_n\left(\rho(\tilde{N}) \right) \leq D(K) \left[\gamma \sum_{k=N_0+1}^{n} \theta_k^2 + \gamma \mathbb{E}_n\left(\sum_{k=\tilde{N}+1}^{n} \theta_k^2 \right) + \frac{\epsilon^2}{\gamma} \right].$$

Integrating with respect to the X_i, we arrive at:

$$\mathbb{E}\left(\rho(\tilde{N})\right) \leq D(K)\left[\gamma R(\theta,N_0) + \gamma \mathbb{E}r(\theta,\tilde{N}) - \gamma\epsilon^2\mathbb{E}\sum_{k=1}^{\tilde{N}}(\xi_k^2-1) + \frac{\epsilon^2}{\gamma}\right].$$

As a result, for any $N_0 \geq 1$:

$$\begin{aligned}(1-D\gamma)\mathbb{E}r(\theta,\tilde{N}) &\leq (1+D\gamma)R_{rhm}(f_\eta,N_0) + \frac{C(p)\epsilon^2}{\mu} + \frac{D\epsilon^2}{\gamma} \\ &+ (1-D\gamma)\epsilon^2\mathbb{E}\left[\sum_{k=1}^{\tilde{N}}(\xi_k^2-1) - \frac{\alpha-\mu}{1-D\gamma}\Sigma(\tilde{N})^p\right].\end{aligned}$$

Finally, using Lemma 4.5, we arrive at:

$$(1-D\gamma)\mathbb{E}r(\theta,\tilde{N}) \leq (1+D\gamma)R_{rhm}(f_\eta,N_0) + \frac{C(p)\epsilon^2}{\mu} + \frac{D\epsilon^2}{\gamma} + \frac{(1-D\gamma)^2C(p)\epsilon^2}{\alpha-\mu}.$$

Now choosing $\gamma < \gamma_1$ for some γ_1 we have:

$$\begin{aligned}\mathbb{E}r(\theta,\tilde{N}) &\leq \frac{(1+D\gamma)}{(1-D\gamma)}R_{rhm}(f_\eta,N_0) + \frac{C(p)\epsilon^2}{\mu(1-D\gamma)} + \frac{D\epsilon^2}{\gamma(1-D\gamma)} + \frac{(1-D\gamma)C(p)\epsilon^2}{\alpha-\mu} \\ &\leq (1+B\gamma)R_{rhm}(f_\eta,N_0) + \frac{C(p)\epsilon^2}{\mu(1-D\gamma)} + \frac{D\epsilon^2}{\gamma(1-D\gamma)} + \frac{(1-D\gamma)C(p)\epsilon^2}{\alpha-\mu}.\end{aligned}$$

Choosing μ such that:

$$\frac{1}{\mu(1-D\gamma)} = \frac{(1-D\gamma)}{\alpha-\mu},$$

concludes the proof.

4.6 Ordered processes

4.6.1 Definition and main property

Our method to derive Theorem 4.1 is related to the control of processes $\zeta(N)$ and $\zeta'(N)$ defined above. These processes are embedded into a special class of random processes called ordered.

Definition 4.1 *Let $\zeta(t)$, $t \geq 0$ a separable random process with $\mathbb{E}\zeta(t) = 0$ and finite variance $\sigma(t)^2$ such that $\sigma(t_1)^2 \leq \sigma(t_2)^2$ for $t_1 \leq t_2$. It is called ordered if for all $t_1 \leq t_2$, we have:*

$$\mathbb{E}(\zeta(t_2)-\zeta(t_1))^2 \leq \sigma(t_2)^2 - \sigma(t_1)^2.$$

This condition can be rewritten as

$$\mathbb{E}\zeta(t_2)\zeta(t_1) \geq \min(\mathbb{E}\zeta(t_1)^2, \mathbb{E}\zeta(t_2)^2). \tag{4.19}$$

Ordered processes is a class of random processes rather vast. It can be viewed as a generalization of the Wiener process $W(t)$ for which $\mathbb{E}W(t_2)W(t_1) = \min(\mathbb{E}W(t_1)^2, \mathbb{E}W(t_2)^2)$. The simplest example of ordered process is $\zeta(t) = \xi t$ where ξ is a zero mean random variable with a finite exponential moment.

Here we propose to enumerate the principal properties of ordered processes. It almost immediately results in Theorem 4.1. The main characteristic of ordered processes is presented in the following lemma.

Lemma 4.4 *Let $\zeta(t)$, $t \geq 0$ an ordered process and suppose that there exists $\kappa > 0$ such that*

$$\varphi(\kappa) := \sup_{t_1,t_2} \mathbb{E} \exp\left(\kappa \frac{\zeta(t_1) - \zeta(t_2)}{\sqrt{\mathbb{E}(\zeta(t_1) - \zeta(t_2))^2}} \right) < \infty. \tag{4.20}$$

Then there exists a constant C depending on κ such that for any $T > 0$, and all $p \geq 1$,

$$\left[\mathbb{E} \sup_{t,s \in [0,T]} |\zeta(t) - \zeta(s)|^p \right]^{\frac{1}{p}} \leq Cp\sigma(T).$$

This lemma is rather important. In particular, the ordered processes considered in the proof of Theorem 4.1 have to satisfy (4.20). The proof is based on a chaining argument. It has been done in Cao and Golubev [34]. An extension is presented in Marteau [97]. We write it for reader convenience.

Proof. Denote for brievity:

$$\Delta_\zeta(t_1,t_2) = \frac{\zeta(t_1) - \zeta(t_2)}{\sqrt{\mathbb{E}(\zeta(t_1) - \zeta(t_2))^2}}.$$

For $p \geq 1$, consider the function $L(x) = \log^p(x + e^{p-1})$. We have, for any $x > 0$:

$$L''(x) = \frac{p \log^{p-2}(x + e^{p-1})}{(x + e^{p-1})^2} [p - 1 - \log(x - e^{p-1})] \leq 0,$$

which implies that the function L is concave.
For a given integer $s \geq 0$, define the point t_k^s on $[0,T]$ by

$$\sigma^2(t_k^s) = 2^{-s} k \sigma^2(T), k = 0, \ldots, 2^s - 1,$$

and denote by $\mathcal{T}^s$ the set of these points. Let $u \in \mathcal{T}^s$. Then we can find a chain $\tau_k(u) \in \mathcal{T}^k, k = 0, \ldots s$ such that

(a) $\tau_0(u) = 0$ and $\tau_s(u) = u$,

(b) $|\sigma^2(\tau_k(u)) - \sigma^2(\tau_{k-1}(u))| \leq 2^{-k+1}\sigma^2(T)$.

Moreover we can write

$$u = \sum_{k=0}^{s-1} (\tau_{k+1}(u) - \tau_k(u)).$$

Therefore since the process is ordered we have:

$$\left[\mathbb{E}\sup_{u,v\in\mathcal{T}^s}|\zeta(u)-\zeta(v)|^p\right]^{\frac{1}{p}} \leq 2\sum_{k=0}^{s-1}\left[\mathbb{E}\sup_{u\in\mathcal{T}^{k+1}}|\zeta(\tau_{k+1}(u))-\zeta(\tau_k(u))|^p\right]^{\frac{1}{p}}$$

$$= 2\sum_{k=0}^{s-1}\left[\mathbb{E}\sup_{u\in\mathcal{T}^{k+1}}|\Delta_\zeta(\tau_{k+1}(u),\tau_k(u))|^p[\sigma^2(\tau_{k+1}(u))-\sigma^2(\tau_k(u))]^{p/2}\right]^{\frac{1}{p}}$$

$$\leq 2\sigma(T)\sum_{k=0}^{s-1}2^{-k/2}\left[\mathbb{E}\sup_{u\in\mathcal{T}^{k+1}}|\Delta_\zeta(\tau_{k+1}(u),\tau_k(u))|^p\right]^{\frac{1}{p}}. \tag{4.21}$$

Next, using the concavity of L, and (4.20), we obtain:

$$\left[\mathbb{E}\sup_{u\in\mathcal{T}^{k+1}}|\Delta_\zeta(\tau_{k+1}(u),\tau_k(u))|^p\right]^{\frac{1}{p}} \leq \frac{1}{\kappa}\log\left[\sum_{u\in\mathcal{T}^{k+1}}\mathbb{E}\exp(\kappa|\Delta_\zeta(\tau_{k+1}(u),\tau_k(u))|)+e^{p-1}\right]$$

$$\leq \frac{1}{\kappa}\log\left[2^{k+2}\varphi(\kappa)+e^{p-1}\right] = \frac{(k+2)\log 2+p-1}{\kappa}+\frac{\log\varphi(\kappa)}{\kappa}.$$

Substituing into (4.21), we finally get:

$$\left[\mathbb{E}\sup_{u,v\in\mathcal{T}^s}|\zeta(u)-\zeta(v)|^p\right]^{\frac{1}{p}} \leq 2\sigma(T)\sum_{k=0}^{s-1}2^{-k/2}\left[\frac{(k+2)\log 2+p-1}{\kappa}+\frac{\log\varphi(\kappa)}{\kappa}\right]$$

$$\leq 2\sigma(T)\sum_{k=0}^{s-1}2^{-k/2}\frac{Ck+Cp}{\kappa} \leq Cp\sigma(T),$$

which proves the lemma by separability of $\zeta(t)$.□

4.6.2 Technical lemmas

Previous lemma results in the following facts. It will be useful to derive Lemma 4.2 and Theorem 4.1.

Lemma 4.5 *Let $\zeta(t)$, $t\geq 0$ an ordered process satisfying (4.20) and $\zeta(0)=0$. Then for all $p>1$, there exists a constant C depending on κ and p such that for any γ,*

$$\mathbb{E}\sup_{t\geq 0}[\zeta(t)-\gamma\sigma(t)^p]_+ \leq \frac{C}{\gamma}.$$

The proof is presented in Marteau [97]. We recall it for the sack of completeness.

Proof. The proof is based on a chaining argument. We will use the following form of the Markov inequality:

$$\mathbb{E}\eta^p\ \mathbb{1}(\eta>x) \leq \frac{\mathbb{E}|\eta|^{p+q}}{x^q} \tag{4.22}$$

which immediately results from the banal inequality,

$$\eta^p\ \mathbb{1}(\eta>x) \leq |\eta|^p\left|\frac{\eta}{x}\right|^q.$$

Let $\gamma > 0$. We can assume without loss of generality that $\sigma^2(t)$ is continuous such that $\sigma^2(t) \to \infty$ as $t \to \infty$. Then there exists $(t_k(\gamma))_{k\in\mathbb{N}}$ such that:

$$\sigma(t_k(\gamma)) = \frac{k}{\gamma}, k \in \mathbb{N}.$$

The function $x \mapsto x^p \mathbb{1}(x > x_0)$ is monotone on $\mathbb{R}^+$. Hence we have:

$$\begin{aligned}
&\mathbb{E}\sup_{t\geq 0}[\zeta(t) - \gamma\sigma(t)^p]_+ \leq \sum_{k=0}^{+\infty}\mathbb{E}\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}[\zeta(t) - \gamma\sigma(t)^p]_+ \\
\leq &\sum_{k=0}^{+\infty}\mathbb{E}\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t)\ \mathbb{1}(\zeta(t) > \gamma\sigma^p(t)) \\
\leq &\sum_{k=0}^{+\infty}\mathbb{E}\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t)\ \mathbb{1}(\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t) > \gamma\sigma^p(t_k(\gamma))) \\
\leq &\mathbb{E}\sup_{0\leq t\leq t_1(\gamma)}|\zeta(t)| \\
&+\sum_{k=1}^{+\infty}\mathbb{E}\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t)\ \mathbb{1}(\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t) > \gamma\sigma^p(t_k(\gamma)))
\end{aligned}$$

By Lemma 4.4, the first term of the above inequality is bounded as follows:

$$\mathbb{E}\sup_{0\leq t\leq t_1(\gamma)}|\zeta(t)| \leq C\sigma(t_1(\gamma)) = \frac{C}{\gamma}. \tag{4.23}$$

In view of (4.22), the second one is controled by:

$$\begin{aligned}
&\sum_{k=1}^{+\infty}\mathbb{E}\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t)\ \mathbb{1}(\sup_{t\in[t_k(\gamma),t_{k+1}(\gamma)[}\zeta(t) > \gamma\sigma^p(t_k(\gamma))) \\
&\leq \sum_{k=1}^{+\infty}\frac{\mathbb{E}\sup_{0\leq t<t_{k+1}(\gamma)}|\zeta(t)|^{q+1}}{(\gamma\sigma^p(t_k(\gamma)))^q} \leq C\sum_{k=1}^{+\infty}\frac{\sigma^{q+1}(t_{k+1}(\gamma))}{\gamma^q\sigma^{pq}(t_k(\gamma))}
\end{aligned}$$

Gathering (4.23) with previous inequality, we arrive at:

$$\mathbb{E}\sup_{t\geq 0}[\zeta(t) - \gamma\sigma(t)^p]_+ \leq \frac{C}{\gamma} + C\sum_{k=1}^{+\infty}\frac{(k+1)^{q+1}\gamma^q}{\gamma^{2q+1}k^{pq}} \leq \frac{C}{\gamma} + \frac{C}{\gamma^{q+1}}\sum_{k=1}^{+\infty}\frac{1}{k^{pq-q-1}}.$$

Setting $q > \frac{2}{p-1}$ proves the lemma. □

Lemma 4.6 *Let $\zeta(t)$, $t \geq 0$ an ordered process satisfying (4.20) such that $\zeta(0) = 0$ and t^* measurable with respect to ζ. Then there exists a positive constant C depending on κ such that for all $K > 1$:*

$$\mathbb{E}|\zeta(t^*)| \leq \log^2(K)\mathbb{E}\sigma(t^*) + \frac{C}{K}\sqrt{\mathbb{E}\sigma^2(t^*)}.$$

Proof. First note that:

$$\begin{aligned}\mathbb{E}\zeta(t^*) &= \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) \leq \log^2(K)\sigma(t^*)) + \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*)) \\ &\leq \log^2(K)\mathbb{E}\sigma(t^*) + \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*)).\end{aligned}$$

Once again we can assume without loss of generality that $\sigma^2(t)$ is continuous such that $\sigma^2(t) \to \infty$ as $t \to \infty$. Then we can find a real sequence $(t_k)_{k\in\mathbb{N}}$ where:

$$\sigma(t_k) = \frac{k^d\sqrt{\mathbb{E}\sigma^2(t^*)}}{K}, \tag{4.24}$$

where $d > 0$ will be chosen later. Using Lemma 4.4, one gets:

$$\begin{aligned}&\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*)) = \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* \leq t_1) \\ +\ &\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\ \leq\ &\mathbb{E}\sup_{t<t_1}|\zeta(t)| + \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\ =\ &\frac{\mathbb{E}\sigma^2(t^*)}{K} + \mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1)\end{aligned}$$

Let $1 < p < 2$. Using a Hölder inequality:

$$\begin{aligned}&\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\ =\ &\sum_{k=1}^{+\infty}\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) \geq \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* \in]t_k, t_{k+1}]) \\ =\ &\sum_{k=1}^{+\infty}\mathbb{E}\sigma^p(t^*)\frac{\zeta(t^*)}{\sigma^p(t^*)}\ \mathbb{1}(\zeta(t^*) \geq \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* \in]t_k, t_{k+1}]) \\ \leq\ &\sum_{k=1}^{+\infty}(\mathbb{E}\sigma^{pr}(t^*))^{\frac{1}{r}}\left(\mathbb{E}\frac{|\zeta(t^*)|^s}{\sigma^{ps}(t^*)}\ \mathbb{1}(\zeta(t^*) \geq \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* \in]t_k, t_{k+1}])\right)^{\frac{1}{s}},\end{aligned}$$

where $\frac{1}{r} + \frac{1}{s} = 1$. In the following, we choose $r = \frac{2}{p}$ to get:

$$\begin{aligned}&\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\ \leq\ &\left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p}{2}}\sum_{k=1}^{+\infty}\left(\frac{\mathbb{E}\sup_{t\in[t_k,t_{k+1}]}|\zeta(t)|^s}{\sigma^{ps}(t_k)}\ \mathbb{1}(\sup_{t\in[t_k,t_{k+1}]}|\zeta(t)| \geq \log^2(K)\sigma(t_k))\right)^{\frac{1}{s}}.\end{aligned}$$

Let $q > 0$ will be chosen later. Using again (4.22) and Lemma 4.4, we arrive at:

$$\begin{aligned}&\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\ \leq\ &\left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p}{2}}\sum_{k=1}^{+\infty}\left(\frac{\mathbb{E}\sup_{t\in[t_k,t_{k+1}]}|\zeta(t)|^{s+q}}{\sigma^{ps+q}(t_k)}\frac{1}{(\log(K))^{2q}}\right)^{\frac{1}{s}} \\ \leq\ &\left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p}{2}}\sum_{k=1}^{+\infty}\left(\frac{C(s+q)^{s+q}\sigma^{s+p}(t_{k+1})}{\sigma^{ps+q}(t_k)}\frac{1}{(\log(K))^{2q}}\right)^{\frac{1}{s}}.\end{aligned}$$

Using (4.24), we finally have:

$$\begin{aligned}
&\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \\
\leq\ & \left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p}{2}} \sum_{k=1}^{+\infty} \left(\frac{C(s+q)^{s+q}(k+1)^{d(s+q)}\sqrt{\mathbb{E}\sigma^2(t^*)}^{s+p}K^{-(s+q)}}{k^{d(ps+q)}\sqrt{\mathbb{E}\sigma^2(t^*)}^{ps+q}K^{-(ps+q)}} \frac{1}{(\log(K))^{2q}} \right)^{\frac{1}{s}} \\
\leq\ & \frac{\left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p}{2}}}{\left(\mathbb{E}\sigma^2(t^*)\right)^{\frac{p-1}{2}}} \times \frac{1}{K} \times \sum_{k=1}^{+\infty} \frac{1}{k^{d(p-1)}} \left(\frac{C(s+q)^{s+q}2^{d(s+p)}K^{sp}}{(\log K)^{2q}} \right)^{\frac{1}{s}}.
\end{aligned}$$

Setting for instance $d = \frac{2}{p-1}$, we obtain:

$$\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \leq C\frac{\sqrt{\mathbb{E}\sigma^2(t^*)}}{K} \times \frac{q^{q/s}2^{dq}K^p}{(\log K)^{2q/s}}.$$

Setting $q = s\log K$, one haves:

$$\mathbb{E}\zeta(t^*)\ \mathbb{1}(\zeta(t^*) > \log^2(K)\sigma(t^*))\ \mathbb{1}(t^* > t_1) \leq C\frac{\sqrt{\mathbb{E}\sigma^2(t^*)}}{K},$$

for a constant $C > 0$ independent of K. This concludes the proof.□

4.7 Appendix

4.7.1 SVD of the sampling operator

Vito et al. [143] gives an interpretation of learning from examples as inverse problem. From (4.9), we can write, for $f_\eta \in \mathcal{H}_K$,

$$Y = A_X f_\eta + \sigma, \tag{4.25}$$

where we write $Y = (Y_1, \dots Y_n)^T$, $\sigma = (\sigma_1, \dots \sigma_n)^T$ and $A_X : \mathcal{H}_K \to \mathbb{R}^n$ the linear compact operator given by $A_X f = (f(X_1), \dots f(X_n))^T$. A_X is often called the sampling operator. Here we want to underline that transformation (4.11) using Lemma 4.1 can be seen as the direct consequence of the singular value decomposition (SVD) of A_X.
It is easy to see that $A_X^* : \mathbb{R}^n \to \mathcal{H}_K$ is defined by, for $u \in \mathbb{R}^n$:

$$A_X^* u = \frac{1}{n}\sum_{i=1}^n u_i K(X_i, \cdot),$$

where we use the scalar product $< u,v >= \frac{1}{n}\sum u_i v_i$ in $\mathbb{R}^n$. Then we have $A_X^* A_X = C_n$ is the empirical covariance operator. We know that the eigenvalues of the kernel matrix $K_n = (K(X_i,X_j))_{i,j=1,\dots,n}$ correspond to the eigenvalues $\{\lambda_k, k = 1, \dots, n\}$ of C_n. Its eigenfunctions $\{\phi_k, k = 1, \dots, n\}$ are given by:

$$\phi_k(\cdot) = \frac{1}{\sqrt{n\lambda_k}}\sum_{i=1}^n V_k^i K(X_i, \cdot), \tag{4.26}$$

where $\{V_k, k = 1, \dots, n\}$ comes from the proof of Lemma 4.1. We hence have the following SVD for A_X:

$$\begin{cases} A_X \phi_k = \sqrt{\lambda_k}\psi_k \\ A_X^* \psi_k = \sqrt{\lambda_k}\phi_k, \end{cases}$$

where ϕ_k is defined in (4.26) and $\psi_k = \sqrt{n}(V_k^1, \dots V_k^n)^T$.
Then we can project the observations on the basis $\{\psi_k, k = 1, \dots, n\}$ to obtain:

$$< Y, \psi_k >_n = < A_X f, \psi_k >_n + < \sigma, \psi_k >_n .$$

As a consequence, since $\psi_k^i = \varphi_k(X_i)$:

$$< A_X f, \psi_k >_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)\psi_k^i = \theta_k,$$

and for the same reason:

$$< \sigma, \psi_k >_n = \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi_k^i = \epsilon \xi_k.$$

As a result, the linear transformation of Section 4.2.2 using $\{\varphi_k(X_i) = \frac{\phi(X_i)}{\sqrt{\lambda_k}}, k = 1, \dots, n\}$ as basis corresponds to the projection of the observations into the SVD of the sampling operator A_X. If we follow the inverse problem setting, this SVD decomposition lead to the following model:

$$y_k = b_k \Theta_k + \epsilon \xi_k, k = 1, \dots, n. \tag{4.27}$$

where here $\Theta_k = < f_\eta, \phi_k >_K$ and $b_k = \sqrt{\lambda_k}$ are the singular values.

Remark 4.6 *Let compare this model with the Gaussian sequence space model* (4.1)*. Both come from the projection of the observations into the orthonormal basis made of eigenfunctions of A^*A. However in* (4.27) *the singular values are random. It comes from the definition of A_X which depends on the design points. Moreover, from statistical viewpoint, the main difference is contained into the random variables ξ_k. The projection of a Gaussian white noise into a orthonormal basis gives i.i.d. $\mathcal{N}(0,1)$ random variables. Here variables ξ_k are linear combinations of variables σ_i, $i = 1, \dots, n$.*

Remark 4.7 *In inverse problem, the aim is to reconstruct a signal f from indirect observation Af. In a sequence space model such as* (4.27)*, we have to estimate Θ_k from observations $b_k\Theta_k$. As a result, the square risk of a spectral cut-off $\hat{f}_N$ is given by:*

$$\mathbb{E}\|\hat{f}_N - f\|^2 = \sum_{k>N} \Theta_k^2 + \epsilon^2 \sum_{k=1}^{N} b_k^{-2}.$$

In this work, we are interested in the square risk of $A_X\hat{f}_N$, defined by:

$$\mathbb{E}(\hat{f}_N(X) - f_\eta(X))^2 = \mathbb{E}\|A_X\hat{f}_N - A_X f_\eta\|_n^2 = \sum_{k>N} \mathbb{E} b_k^2 \Theta_k^2 + \epsilon^2 N.$$

As a result, we prefer to consider the model (4.12) *where we write:*

$$y_k = \theta_k + \epsilon \xi_k, k = 1, \dots, n,$$

for $\theta_k = b_k\Theta_k = \frac{1}{n}\sum_{k=1}^{n} f_\eta(X_i)\varphi_k(X_i)$. Classification appears in this case closer to regression than to inverse problems.

4.7.2 Generalization of Lemma 4.3 to regression

Condition to have an ordered process in Lemma 4.3 can be generalized to regression with random noise. In this case, an assumption over the moment of order 4 of the noise leads to an ordered process.

Consider the regression model

$$Y_i = f(X_i) + \sigma\epsilon_i, i = 1 \dots n, \tag{4.28}$$

where ϵ_i are i.i.d. random variables with variance 1. To estimate the unknown function f, we make the linear transformation (4.11). For a given kernel K, we consider:

$$y_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{\phi_k(X_i)}{\lambda_k},$$

where the sequence (λ_k) are the eigenvalues of the kernel matrix $(K(X_i,X_j))_{i,j}$ and (ϕ_k) is an orthonormal basis of $\mathcal{H}_K$. As a result, the regression model (4.9) is equivalent to:

$$y_k = \theta_k + \epsilon\xi_k, \tag{4.29}$$

where

$$\theta_k = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{\phi_k(X_i)}{\sqrt{\lambda_k}},\ \epsilon = \frac{\sigma}{\sqrt{n}},$$

and

$$\xi_k = \sum_{i=1}^{n} V_k^i \epsilon_i,$$

where (V_k) is an orthonormal basis of $\mathbb{R}^n$ of eigenvectors of $(K(X_i,X_j))_{i,j}$.
We hence have the following result:

Lemma 4.7 *The random process $\eta(N) = \epsilon^2\sum_{k=1}^{N}(\xi_k^2 - 1)$ is ordered if and only if the random variable ϵ_i are such that*

$$\mathbb{E}\epsilon_i^4 \geq 3. \tag{4.30}$$

Proof. $\Leftarrow$) Proof of Lemma 1 could be used. It only remains to note that with the previous notations:

$$h \geq \frac{1}{\sqrt{3}} \Leftrightarrow \mathbb{E}\epsilon_i^4 \geq 3.$$

$\Rightarrow$) If $\eta(N)$ is ordered, then, for any $N \geq 1$:

$$\mathbb{E}(\zeta(N+1) - \zeta(N))^2 \leq \sigma^2(N+1) - \sigma^2(N), \tag{4.31}$$

where $\sigma^2(N) = \text{var}\zeta(N)$. Moreover, since (V_k) is a o.n.b. of $\mathbb{R}^n$, we have:

$$\mathbb{E}(\zeta(N+1) - \zeta(N))^2 = \mathbb{E}\xi_{N+1}^4 - 1 = (\mathbb{E}\epsilon_i^4 - 3)\sum_{i=1}^{n}(V_{N+1}^i)^4 + 2,$$

and for the same reasons:

$$\sigma^2(N+1) - \sigma^2(N) = (\mathbb{E}\epsilon_i^4 - 3)\left(\sum_{i=1}^{n}(V_{N+1}^i)^4 + \sum_{k=1}^{N}\sum_{i=1}^{n}(V_k^i)^2(V_{N+1}^i)^2\right) + 2.$$

Hence we have with (4.31):

$$(\mathbb{E}\epsilon_i^4 - 3)\sum_{i=1}^{n}(V_{N+1}^i)^4 \leq (\mathbb{E}\epsilon_i^4 - 3)\left(\sum_{i=1}^{n}(V_{N+1}^i)^4 + \sum_{k=1}^{N}\sum_{i=1}^{n}(V_k^i)^2(V_{N+1}^i)^2\right).$$

As a result, we have coarselly $\mathbb{E}\epsilon_i^4 \geq 3$.□

Remark 4.8 *Assumption* (4.30) *is related to the kurtosis of the noise. Kurtosis is a measure of whether a distribution is peaked or flat relative to a normal distribution. The kurtosis of a random variable X is defined by:*

$$\beta_2 = \frac{\mathbb{E}X^4}{(varX)^2}.$$

The kurtosis of a normal distribution is 3. Then if ϵ_i are i.i.d. $\mathcal{N}(0,1)$, an equality holds in (4.30).

Conclusion et perspectives

Cette thèse apporte un point de vue théorique sur les performances des méthodes à noyaux. Les problèmes soulevés sont nombreux. Dans cette conclusion, on propose un survol des principaux axes de ce manuscrit en précisant les futures directions possibles.

Vitesses de convergence

Les Chapitres 2 et 3 proposent des vitesses de convergence pour des minimisations de type SVM. Sous une hypothèse de régularité sur la fonction à estimer, la solution atteint une certaine vitesse de convergence. Cette vitesse augmente avec la régularité de la fonction à estimer. Ce phénomène est classique en statistique non-paramétrique. Pour compléter ces résultats, on peut étudier l'optimalité de ces vitesses au sens minimax.
Notons $r(\hat{f}_n,\mathcal{B})$ le risque maximal du classifieur $\hat{f}_n$ défini par :

$$r(\hat{f}_n,\mathcal{B}) = \max_{f^* \in \mathcal{B}} \mathbb{E}R(\hat{f}_n,f^*),$$

où $\mathcal{B}$ est un espace fonctionnel avec une certaine régularité et $R(\hat{f}_n,f^*)$ est l'excès de risque de $\hat{f}_n$. On cherche à obtenir une borne inférieure de la forme :

$$\min_{\hat{f}_n} \max_{f^* \in \mathcal{B}} \mathbb{E}R(\hat{f}_n,f^*) \geq c_0 \psi_n. \tag{4.32}$$

Dans ce cas si $\hat{f}_n$ atteint la vitesse de convergence ψ_n, il sera dit optimal au sens minimax.

Il serait intéressant de relier l'hypothèse de marge à ce type d'approche. On peut en effet écrire l'hypothèse de marge comme une condition sur la régularité d'une transformée de la probabilité conditionelle $\eta(x) = P(Y = 1|X = x)$. Une loi P a un paramètre de marge q si $(2\eta - 1)^{-1} \in L_{q,\infty}(P_X)$, où $L_{q,\infty}$ est un espace de Lorentz (voir DeVore and Sharpley [49]). On pourrait obtenir des vitesses de convergence qui ne dépendraient que du paramètre de marge.

Pour espérer obtenir des résultats optimaux, il reste à mieux contrôler l'erreur d'approximation. Cette thèse propose un cadre fonctionnel où les espaces à noyaux approchent la règle de Bayes. Le Chapitre 2 montre que l'utilisation des EHNR n'est pas idéale pour converger rapidement vers f^*. Le Chapitre 3 propose d'utiliser des espaces plus vastes : les espaces de Besov. Cependant, l'utilisation d'espaces fonctionnels classiques et de leurs propriétés d'approximation semble limiter les performances des méthodes à noyaux étudiées.

On peut espérer atteindre de meilleures vitesses en construisant ses propres espaces d'hypothèses. Ces espaces seraient constitués de fonctions à valeurs discrètes et approcheraient plus facilement la règle de Bayes. Un premier pas dans cette direction est présenté dans Lecué [82]. A l'aide d'une suite de partitions dyadiques de $[0,1]^d$, Lecué [82] construit un espace d'hypothèses de fonctions $f : [0,1]^d \to \{-1, +1\}$. Ces fonctions sont les combinaisons linéaires d'un système d'indicatrices proche de la base de Haar. Pour obtenir une classe de fonctions à valeurs dans $\{-1,1\}$, les coefficients sont -1, 0 ou 1 seulement. Cet ensemble de règles de décision peut être représenté par un arbre de décision dyadique, où chaque feuille est un coefficient. Sous une hypothèse portant sur la parcimonie de la représentation de f^* dans le système, on peut contrôler l'erreur d'approximation. Un algorithme de type LASSO (pénalité l^1 sur les coefficients) atteint des vitesses de convergence rapides satisfaisant (4.32).

Perspectives expérimentales

Le Chapitre 2 propose une application directe des résultats théoriques discutés ci-dessus. On propose un classifieur "parameter free" qui résout des problèmes réels de classification. La principale caractéristique de ce classifieur est sa justification théorique. L'implémentation de notre agrégat suit pas à pas la construction de la Section 2.3. Cela entraîne des avantages mais aussi des inconvénients. Cette approche permet d'illustrer et de confronter les résultats théoriques de C. Scovel et I. Steinwart à ceux du Chapitre 2 de manière expérimentale.

La principale faiblesse des résultats théoriques du Chapitre 2 sont leur caractère essentiellement asymptotique. D'une part, le choix du paramètre de régularisation est de l'ordre d'une puissance de n. D'autre part, la constante devant les vitesses de convergence n'est pas explicite. Il serait intéressant de se tourner vers une étude plus précise des constantes dans les bornes supérieures. Ces précisions permettraient peut-être d'améliorer les résultats expérimentaux lorsque le nombre d'observations n est petit. Ce raffinement permettrait aussi de mieux traiter les données en grandes dimensions. Un regard détaillé sur les preuves du Chapitre 2 montre que lorsque d est grand, les constantes apparaissant dans les bornes supérieures augmentent, affaiblissant les performances statistiques établies.

Plus généralement, la méthode d'agrégation avec poids exponentiels pour l'adaptation s'apparente à de la validation croisée. En effet, l'échantillon est divisé en deux : la première partie permet de construire une famille de SVM. La deuxième partie permet de tester leurs performances (partie test ou validation). Dans la combinaison finale, le calcul du poids de chaque classifieur est fonction croissante de ses performances sur cette deuxième partie. Cependant, l'atout de cette construction réside dans le choix de la grille. Ce choix est dicté par la partie théorique et s'adapte à chaque problème spécifique. Cela permet d'éviter le problème de temps de calcul de la validation croisée classique.
On peut ajouter un paramètre supplémentaire à notre procédure d'agrégation : le paramètre

de température. Cette quantité intervient dans le calcul des poids, qui devient :

$$\omega_\alpha^T = \frac{\exp\left(T\sum_{i=n_1+1}^{n} Y_i \hat{f}_{n_1}^{\alpha}(X_i)\right)}{\sum_{\alpha'\in\mathcal{G}(n_2)} \exp\left(T\sum_{i=n_1+1}^{n} Y_i \hat{f}_{n_1}^{\alpha'}(X_i)\right)},$$

où $T > 0$ est appelée la température. Ce paramètre influence l'écart entre les différentes valeurs des poids, c'est-à-dire l'influence de la partie validation dans notre procédure. Si la température T est proche de 0, les poids sont tous égaux et la combinaison convexe revient à moyenner les classifieurs de la famille. On ne tient alors pas compte de la partie test. Au contraire, si la température est élevée, seul le meilleur SVM sur l'échantillon test sera utilisé. C'est la méthode de minimisation du risque empirique. Une étude de l'influence de T dans le "single index model" est présentée dans Lecué [78]. L'optimisation de ce paramètre peut permettre un gain dans la généralisation. On pourrait s'intéresser à l'influence de la température en classification avec les SVM, notamment dans le cas gaussien où les résultats de généralisation dans la famille sont assez dispersés.

Vers un noyau Besov

On a vu dans le Chapitre 1 que les méthodes à noyaux sont liées au formalisme des Espaces de Hilbert à noyaux reproduisants. On a démontré dans le Chapitre 3 qu'un espace plus général (un espace de Besov) permettait d'obtenir des performances statistiques comparables. La théorie des ondelettes se substitue à la machinerie des noyaux définis positifs. Pour concrétiser ce résultat et obtenir un algorithme implémentable, un moyen est de relier les espaces de Besov à la théorie des noyaux. Pour cela, il faut élargir le concept de noyau et d'EHNR à un cadre non-hilbertien.

Considérons deux espaces vectoriels normés $\mathcal{E}$ et $\mathcal{F}$. On dit que $(\mathcal{E},\mathcal{F})$ est une dualité s'il existe une forme bilinéaire $\mathcal{L}$ qui sépare $\mathcal{E}$ et $\mathcal{F}$, c'est-à-dire :

$$\forall\varphi\neq 0\in\mathcal{F},\exists f\in\mathcal{E}:\ \mathcal{L}(\varphi,f)\neq 0 \text{ et } \forall f\neq 0\in\mathcal{E},\exists\varphi\in\mathcal{F}:\ \mathcal{L}(\varphi,f)\neq 0.$$

L'exemple fondamental est le couple $(\mathcal{E},\mathcal{E}^*)$ où $\mathcal{E}^*$ est le dual algébrique de $\mathcal{E}$. Ce couple est mis en dualité par la forme canonique $\mathcal{L}:\mathcal{E}^*\times\mathcal{E}\to\mathbb{R}$ définie par $\mathcal{L}(\varphi,f)=\varphi(f)$.
Notons $\mathbb{R}^{\mathcal{X}}$ l'espace des fonctions $f:\mathcal{X}\to\mathbb{R}$ définies point par point. Alors $(\mathbb{R}^{\mathcal{X}},\mathbb{R}^{[\mathcal{X}]})$ est une dualité, où $\phi\in\mathbb{R}^{[\mathcal{X}]}\Leftrightarrow\exists(\alpha_i)_{i\in I}:\phi(\cdot)=\sum\alpha_i\ \mathbb{1}_{x_i}(\cdot)$. L'application qui sépare $\mathbb{R}^{\mathcal{X}}$ et $\mathbb{R}^{[\mathcal{X}]}$ est définie par :

$$\forall\phi\in\mathbb{R}^{[\mathcal{X}]},\forall f\in\mathbb{R}^{\mathcal{X}},\ <\phi,f>_{\mathcal{X}}=<\sum\alpha_i\ \mathbb{1}_{x_i},f>_{\mathcal{X}}=\sum_{i\in I}\alpha_i f(x_i).$$

Pour généraliser la notion d'EHNR, Mary [98] s'intéresse à une classe particulière de dualités, constituée d'espaces vectoriels contenus dans $\mathbb{R}^{\mathcal{X}}$. Une dualité $(\mathcal{E},\mathcal{F})$ est une sous-dualité d'évaluation si et seulement si :
(i) $\mathcal{E},\mathcal{F}\subset\mathbb{R}^{\mathcal{X}}$.
(ii) $\forall\phi\in\mathbb{R}^{[\mathcal{X}]},\exists\varphi\in\mathcal{F}:\forall f\in\mathcal{E},\mathcal{L}(\varphi,f)=<\phi,f>_{\mathcal{X}}$.

(ii)' $\forall \phi \in \mathbb{R}^{[\mathcal{X}]}, \exists f \in \mathcal{E} : \forall \varphi \in \mathcal{F}, \mathcal{L}(\varphi,f) =< \phi,\varphi >_{\mathcal{X}}$.
Dans ce cas les injections $i : \mathcal{E} \to \mathbb{R}^{\mathcal{X}}$ et $j : \mathcal{F} \to \mathbb{R}^{\mathcal{X}}$ sont faiblement continues.
Dans le cas particulier des EHNR, $\mathcal{H}_K$ est un espace de Hilbert, alors $\mathcal{H}_K = \mathcal{H}_K^*$ est mis en dualité par son produit scalaire. On peut vérifier aisément que cette dualité est une sous-dualité d'évaluation. *(i)* est satisfait par définition. De plus, pour $\phi(\cdot) = \sum \alpha_i \ \mathbb{1}_{x_i}(\cdot) \in \mathbb{R}^{[\mathcal{X}]}$, et $g(\cdot) = \sum \beta_j K(\cdot,y_j) \in \mathcal{H}_K$:

$$\begin{aligned} < \phi,g >_{\mathcal{X}} = \sum_{i=1}^{n} \alpha_i g(x_i) &= \sum_{i,j=1}^{n} \alpha_i \beta_j K(x_i,y_j) \\ &= < \sum \alpha_i K(x_i,\cdot), \sum \beta_j K(\cdot,y_j) >_K \\ &= < f,g >_K , \end{aligned}$$

où $f(\cdot) = \sum \alpha_i K(x_i,\cdot) \in \mathcal{H}_K$. Cela assure que $\mathcal{H}_K$ est un cas particulier de sous-dualité d'évaluation.
De manière analogue aux EHNR, on peut associer à toute sous-dualité d'évaluation un unique noyau $\kappa : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$ défini par :

$$\kappa(\phi) = i \circ j^*(\phi),$$

où j^* est la transposée[1] de j pour la forme bilinéaire $\mathcal{L}$. Ce noyau κ définit de manière unique l'application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ par :

$$K(x,y) = \mathcal{L}(\kappa^*(\delta_x),\kappa(\delta_y)).$$

K est appelé noyau reproduisant la sous-dualité $(\mathcal{E},\mathcal{F})$. On peut vérifier (Mary [98]) que les propriétés reproduisantes usuelles des EHNR sont vérifiées dans ce cadre :

- $\forall y \in \mathcal{X}, \forall f \in \mathcal{E}, f(y) = \mathcal{L}(K(\cdot,y),f)$.
- $\forall x \in \mathcal{X}, \forall \varphi \in \mathcal{F}, \varphi(x) = \mathcal{L}(\varphi,K(x,\cdot))$.

Ces propriétés permettent d'étendre le concept de noyau de la Définition 1.10 à un cadre non-hilbertien. Dans ce cas, le noyau K n'est plus symétrique ni défini-positif.

Canu et al. [32] propose une méthode pour construire de manière explicite une sous-dualité d'évaluation à partir d'une dualité (E,F) et de son produit dual $\mathcal{L}$. On dispose de $\{\Gamma_x, x \in \mathcal{X}\}$, une famille totale de E et $\{\Lambda_y, y \in \mathcal{X}\}$ une famille totale de F. On définit $S : E \to \mathbb{R}^{\mathcal{X}}$ l'opérateur (injectif) par $S(g)(y) = \mathcal{L}(\Lambda_y,g)$ et $T : F \to \mathbb{R}^{\mathcal{X}}$ par $T(f)(x) = \mathcal{L}(f,\Gamma_x)$. On peut montrer que $(S(E),T(F))$ est une sous-dualité d'évaluation de noyau reproduisant K telle que :

$$K(x,y) = \mathcal{L}(\Lambda_y,\Gamma_x).$$

En reliant ces résultats topologiques à la dualité dans les espaces de Besov (Meyer [105] par exemple), on peut espérer obtenir un "noyau Besov".

1. Si $j : \mathcal{F} \to \mathbb{R}^{\mathcal{X}}$ est faiblement continue, $j^* : \mathbb{R}^{[\mathcal{X}]} \to \mathcal{F}$ est définie de manière unique par : $\forall \phi \in \mathbb{R}^{[\mathcal{X}]}, \forall f \in \mathcal{E}, < \phi,j(f) >_{\mathcal{X}} = \mathcal{L}(j^*(\phi),f)$.

Enveloppe du risque

L'enveloppe du risque propose un nouveau point de vue en sélection de modèles. Initié par Cavalier and Golubev [39], cette méthode apporte des résultats très pertinents dans le modèle statistique des problèmes inverses, dans le cadre d'un bruit blanc gaussien. Son application en classification est un vaste chantier et les problèmes soulevés sont nombreux. Ce sont autant de futures directions.

Le Chapitre 4 doit être considéré comme un premier pas dans cette direction. La représentation spectrale du problème de classification proposée dans ce chapitre permet de faire un pont avec des modèles statistiques classiques. Le bruit non-gaussien de classification engendre des processus stochastiques de la forme :

$$\zeta(t) = \epsilon^2 \sum_{k=1}^{n} h_k(t)(\xi_k^2 - 1), \tag{4.33}$$

où $h_k(t), k = 1, \ldots, n$ est appelé "filtre", et t représente le paramètre de régularisation. La famille des KPM correspond à $h_k(t) = \mathbb{1}(k \leq t)$, où dans ce cas $t \in \mathbb{N}^*$. Les processus du type (4.33) appartiennent à une classe particulière de processus : les processus ordonnés. Grâce à la théorie des processus ordonnés, on peut étudier la variabilité du modèle et obtenir une enveloppe du risque. Cependant, une hypothèse sur les moments du processus est nécessaire dans ce cadre. Pour aborder cette difficulté, on peut se pencher sur la forme particulière des variables ξ_k (somme de variables aléatoires indépendantes et bornées) intervenant dans (4.33). On peut aussi imaginer étendre l'étude du Chapitre 4 à d'autres familles d'estimateurs comme les LS-SVM (least-square SVM), où dans ce cas $h_k(t) := h_k(\frac{1}{\alpha}) = \frac{1}{1+\alpha/\lambda_k}$ dépend du paramètre de régularisation α de la procédure. Il serait aussi intéressant de considérer le cas hétéroscédastique où le bruit de classification n'est plus supposé constant sur l'espace d'entrée $\mathcal{X}$. Cela engendre des difficultés techniques dans la forme des processus à contrôler.

Enfin, la principale limitation des résultats du Chapitre 4 se trouve dans la perte utilisée. Pour se ramener au cadre de Cavalier and Golubev [39], on utilise la perte des moindres carrés $l(y,f(x)) = (y - f(x))^2$. Ce critère est habituellement utilisé en régression lorsque la réponse Y du couple (X,Y) est à valeur réelle. En classification on préfère utiliser la perte dite dure définie par $l(y,f(x)) = \mathbb{1}(y \neq f(x))$ ou des pertes basées sur la marge, comme la perte douce. Dans ce cas $l(y,f(x)) = (1 - yf(x))_+$. Un problème ouvert est d'utiliser la méthode de l'enveloppe du risque avec la perte dure ou la perte charnière. Cela permettrait d'étudier le choix du paramètre de régularisation dans l'algorithme SVM et d'obtenir des inégalités oracles pour l'excès de risque.

Afin de parvenir à cet objectif, on peut se pencher sur la famille d'estimateurs de type LASSO considérée dans Lecué [82]. L'écriture relativement simple de la perte de ces classifieurs permet d'extraire des processus stochastiques particuliers.
On suppose que l'espace d'entrée $\mathcal{X} = [0,1]^d$. On considère une suite de partitions de $\mathcal{X}$ définie par :

$$\forall j \in \mathbb{N}^*, I_k^j = E_{k_1}^j \times \cdots \times E_{k_d}^j,$$

où $k = (k_1, \ldots, k_d) \in I_d(j) = \{0, \ldots, 2^j - 1\}^d$ et $\forall i \in \{1, \ldots, d\}$:

$$E_{k_i}^j = \begin{cases} [\frac{k_i}{2^j}, \frac{k_i+1}{2^j}) \text{ si } k_i = 0, \ldots 2^j - 2 \\ [\frac{2^j-1}{2^j}, 1] \text{ si } k_i = 2^j - 1. \end{cases}$$

On considère le système $\mathcal{S} = \{\phi_k^j, j \in \mathbb{N}^*, k \in I_d(j)\}$ où

$$\phi_k^j = \mathbb{1}_{I_k^j}.$$

Le paramètre j est appelé niveau de résolution de la partition I_k^j.
Etant donné le système $\mathcal{S}$, on considère une classe de règle de décision $\mathcal{F}^d$ telle que :

$$f \in \mathcal{F}^d \Leftrightarrow f(x) = \sum_{j=1}^{\infty} \sum_{k \in I_d(j)} a_k^j \phi_k^j, \ \lambda_d\text{-p.s., où } a_k^j \in \{-1,0,1\} \ . \tag{4.34}$$

L'écriture de $f \in \mathcal{F}^d$ n'est pas unique dans le système $\mathcal{S}$ mais on peut utiliser une convention pour lever cette ambiguïté. On peut aussi vérifier que sous une hypothèse assez faible, toute règle de décision peut s'écrire sous la forme (4.34). Pour estimer f^*, Lecué [82] considère la famille d'estimateurs $\{\hat{f}_J, J \geq 1\}$ définie par :

$$\hat{f}_J = \sum_{k \in I_d(J)} \hat{A}_k^J \phi_k^J,$$

où

$$\hat{A}_k^J = \begin{cases} 1 \text{ si } \exists X_i \in I_k^J \text{ et } \text{card}\{X_i \in I_k^J : Y_i = 1\} > \ \text{card}\{X_i \in I_k^J : Y_i = -1\}, \\ -1 \text{ sinon.} \end{cases}$$

Ces estimateurs sont appelés estimateurs par vote majoritaire. Ils consistent, pour un niveau de résolution donné, à répondre 1 sur I_k^J si la classe majoritaire dans cet hypercube est 1, et -1 sinon. Il est clair que pour J trop grand, $\hat{f}_J$ sur-apprend alors que pour J petit, il sera trop grossier. Le problème de sélection de modèle est donc le choix optimal de J.

Pour comprendre l'influence de J dans la qualité de l'estimation, on peut calculer l'excès de risque de $\hat{f}_J$. On peut montrer que sous une hypothèse de marge forte (du type (1.13) du Chapitre 1), et si P_X est absolument continue par rapport à λ_d,

$$\frac{ah}{2} \|\hat{f}_J - f^*\|_{L^1(\lambda_d)} \leq R(\hat{f}_J, f^*) \leq \frac{A}{2} \|\hat{f}_J - f^*\|_{L^1(\lambda_d)},$$

où h est la marge. Alors on peut étudier les variations de $\|\hat{f}_J - f^*\|_{L^1(\lambda_d)}$ par rapport à $J \geq 1$. On peut écrire :

$$\|\hat{f}_J - f^*\|_{L^1(\lambda_d)} = \|\hat{f}_J - f_J^*\|_{L^1(\lambda_d)} + \|f_J^* - f^*\|_{L^1(\lambda_d)}, \tag{4.35}$$

où f_J^* est le meilleur classifieur au niveau de résolution J. Il consiste à répondre 1 sur I_k^J si $P(Y = 1 | X \in I_k^J) > 1/2$ et -1 sinon. Il est clair que dans la décomposition (4.35), le deuxième terme n'est pas aléatoire. Il vérifie $\|f_J^* - f^*\|_{L^1(\lambda_d)} \to 0$ lorsque $J \to \infty$. Pour

proposer une enveloppe, on s'intéresse à la partie stochastique de (4.35). Par définition de $\hat{f}_J$, on a :

$$\begin{aligned} \|\hat{f}_J - f_J^*\|_{L^1(\lambda_d)} &= \int_{[0,1]^d} |\hat{f}_J(x) - f_J^*(x)| d\lambda_d(x) \\ &= \sum_{k \in I_d(J)} \int_{I_k^J} |\hat{A}_k^J - A_k^{J,*}| d\lambda_d(x), \end{aligned}$$

où $A_k^{J,*}$ est défini par :

$$A_k^{J,*} = \begin{cases} 1 \text{ si } P(Y=1|X \in I_k^J) \geq \frac{1}{2}, \\ -1 \text{ sinon.} \end{cases}$$

Le terme stochastique de la perte s'écrit :

$$\|\hat{f}_J - f_J^*\|_{L^1(\lambda_d)} = \frac{1}{2^{dJ-1}} \sum_{k \in I_d(J)} Z_k,$$

où $Z_k^J \sim \mathcal{B}(p_k^J)$ est une variable aléatoire de Bernoulli de paramètre $p_k^J = P(|\hat{A}_k^J - A_k^{J,*}| = 2)$. La recherche d'une enveloppe dans ce cas revient à l'étude du processus stochastique suivant :

$$\zeta(J) = \frac{1}{2^{dJ-1}} \sum_{k \in I_d(J)} (Z_k^J - p_k^J).$$

Finalement, les thèmes abordés dans cette thèse mettent en lumière de nombreuses futures directions, aussi bien d'un point de vue théorique que dans un aspect plus pratique. Certains problèmes ouverts sont exprimés de manière relativement précise alors que d'autres constituent des pistes de recherches plus vastes, à explorer plus profondément. Toutes ces problématiques s'inscrivent dans un objectif commun : mieux comprendre les algorithmes de classification et proposer de nouveaux outils plus performants.

Appendix

On the performances of Unbiased Risk Estimation in classification

Abstract

This appendix proposes to study the model selection method of Unbiased Risk Estimation (URE) in classification. We consider the family of KPM classifiers $\{\hat{f}_N, N \geq 1\}$ using the least square loss. We are interested in both theoretical and practical performances of URE to select the bandwidth parameter N. We state an oracle inequality for the mean square risk and illustrate this result numerically, thanks to simulated data sets. We finally compare the adaptive KPM classifier using URE with the agregate of SVM classifiers using Sobolev spaces.

A-1 Model and framework

We observe a training set $D_n = \{(X_i,Y_i), i = 1 \dots n\}$ of i.i.d. random pairs (X_i,Y_i) of law P. For $i = 1, \dots n$, $X_i \in \mathcal{X}$ is an input variable with associated class $Y_i \in \{-1, +1\}$. Given these observations, the goal is to mimic the best decision rule, given by:

$$f^*(x) = \text{sign}(2\eta(x) - 1) = \text{sign}(f_\eta(x)),$$

where $\eta(x) = \mathbb{P}(Y = 1|X = x)$ is the conditional probability function. To avoid some technical details (see Chapter 4), we make the following assumption on η:

(A-1) $$\exists 0 < h < 1 : \eta(x) = (1 \pm h)/2, \text{ for any } x \in \mathcal{X}.$$

This appendix proposes to estimate the regression function $E(Y|X = x) = f_\eta(x) = 2\eta(x) - 1$ thanks to a regularized empirical risk minimization using the least square loss. We study the family of Kernel Projection Machines (KPM) $\{\hat{f}_N, N \geq 1\}$. Given a symmetric and positive definite kernel K generating a RKHS $\mathcal{H}_K$, we associated to each dimension $N \geq 1$ the following estimator:

(A-2) $$\hat{f}_N = \arg \min_{f \in <\phi_1,\dots,\phi_N>} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

where $(\phi_k)_{k\geq 1}$ is an orthonormal basis of $\mathcal{H}_K$. We are interested in the model selection problem of choosing the dimension N into this family of KPM.

From Chapter 4, this problem can be formulated in the spectral domain as follows. If we choose $(\phi_k)_{k\geq 1}$ in (A-2) to be the eigenfunctions of the empirical covariance operator, with associated eigenvalues $(\lambda_k)_{k\geq 1}$, observations D_n can be written as follows:

$$y_k = \theta_k + \epsilon \xi_k, \tag{A-3}$$

where

$$y_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{\phi_k(X_i)}{\sqrt{\lambda_k}}, \theta_k = \frac{1}{n}\sum_{i=1}^{n} f_\eta(X_i)\frac{\phi_k(X_i)}{\sqrt{\lambda_k}}, \epsilon = \frac{\sqrt{(1+h)(1-h)}}{\sqrt{n}},$$

and

$$\xi_k = \frac{1}{\sqrt{n(1+h)(1-h)}}\sum_{i=1}^{n}(Y_i - f_\eta(X_i))\frac{\phi_k(X_i)}{\sqrt{\lambda_k}}.$$

The estimation of f_η corresponds to the estimation of the sequence $\theta_k, k = 1\dots n$ in (A-3). Moreover, with Lemma 4.1, we can identify the family of KPM classifiers $\{\hat{f}_N, N \geq 1\}$ with the family of spectral cut-off $\{\hat{\theta}_N, N \geq 1\}$ where:

$$\hat{\theta}_k(N) = y_k \ \mathbb{1}(k \leq N).$$

It allows us to consider the model selection problem in the sequence space model (A-3). Table A.1 summarizes the correspondence between the classification model and the spectral sequence space model (A-3).

Classification model	Sequence space model
Training set $D_n = (X_i,Y_i), i = 1\dots n$	observations y_k, $k = 1\dots n$
Unknown $f_\eta(X_i)$, $i = 1\dots n$	Unknown θ_k, $k = 1\dots n$
Probability function $\eta(x) = \frac{1\pm h}{2}$	Level of noise $\epsilon = \sqrt{\frac{(1+h)(1-h)}{2}}$
KPM classifiers $\{\hat{f}_N, N \geq 1\}$	Spectral cut-off $\{\hat{\theta}_N, N \geq 1\}$
Mean square risk $\mathbb{E}(f_\eta(X) - \hat{f}_N(X))^2$	$\mathbb{E}\|\theta - \hat{\theta}(N)\|^2$

Table A.1: Relationship between Classification and Sequence space model

In this appendix, we study the performances from both theoretical and practical viewpoint of a widely used approach in non-parametric statistics: the Unbiased Risk Estimation (URE). We begin with the oracle efficiency of the method. We state an oracle inequality for the mean square risk and illustrate numerically the performances of the method in classification. Finally, we propose generalization efficiency of the KPM classifier using URE method. We compare it classification errors with the data-driven SVM of Chapter 2.

A-2 Theoretical result

A-2.1 Description of the method

In this section we work in the sequence space (A-3). To describe the method, we consider the variables X_i, $i = 1\dots n$ to be fixed. As a result, the sequence θ_k is not random. We

are looking at an adaptive estimator $\tilde{\theta} = \hat{\theta}(\tilde{N})$ where $\tilde{N}$ is a data-driven choice of N. A well-known idea of choosing N is based on the unbiased estimation of the risk. The mean square risk of a spectral cut-off $\hat{\theta}(N)$ can be written:

$$R(\theta,N) = \sum_{k>N} \theta_k^2 + \epsilon^2 N.$$

Then, for any $N \geq 1$, the function

$$\bar{R}(y,N) = \sum_{k>N} (y_k^2 - \epsilon^2) + \epsilon^2 N,$$

satisfies:

$$\mathbb{E}\bar{R}(y,N) = R(\theta,N).$$

In other words, the function $\bar{R}(y,N)$ is an unbiased estimator of the risk $R(\theta,N)$. To mimic the best estimator over the family, the URE method proposes to minimize $\bar{R}(y,N)$ over $N \geq 1$. We arrive at the following data-driven choice for N:

$$N_{URE} = \arg\min_{N\geq 1} \left(-\sum_{k=1}^{N} y_k^2 + 2\epsilon^2 N \right). \tag{A-4}$$

It is clear that N_{URE} depends on the level of noise ϵ in (A-3). As a result, $\hat{\theta}(N_{URE})$ is not adaptive with respect to ϵ. In this Appendix, we do not consider the problem of unknown variance whereas it is an interesting problem in practice. Fortunately, there is a simple trick to overcome this difficulty. It consists in estimating the noise level and plug-in into (A-4) (see Cao and Golubev [33, 34] or Green and Silverman [65] for details).

A-2.2 Oracle inequality

The main result is an oracle inequality. It compares the mean square risk of the data-driven estimator $\hat{\theta}(N_{ERM})$ with the best among the family, $\hat{\theta}(N^*)$ where:

$$N^* = \arg\min_{N\geq 1} R(\theta,N).$$

The estimator $\hat{\theta}(N^*)$ is called the oracle. To get this oracle inequality, we apply the theory of ordered processes (see Section 4.6 in Chapter 4) to the random process:

$$\zeta(N) = \sum_{k=1}^{N} (\xi_k^2 - 1). \tag{A-5}$$

We write empirical minimization (A-4) as the minimization of a particular risk hull. Then we follow the proof of Theorem 4.1 in Chapter 4.

Proposition A-1 *Suppose* (A-1) *holds for* $h \geq \frac{1}{\sqrt{3}}$. *Moreover suppose the sequence of random variables* ξ_k, $k = 1 \dots n$ *is such that:*

$$\sup_{N\geq 1} \mathbb{E} \exp\left(\kappa \frac{\sum_{k=1}^{N} (\xi_k^2 - 1)}{\mathbb{E}\left[\sum_{k=1}^{N} (\xi_k^2 - 1)\right]^2} \right) < \infty, \tag{A-6}$$

for a constant $\kappa > 0$.
Then for any $\gamma > 0$, we have:

$$\mathbb{E}\|\hat{\theta}(N_{ERM}) - \theta\|^2 \leq (1+\gamma) \min_{N \geq 1} R(\theta,N) + \frac{\epsilon^2 C}{\gamma}. \tag{A-7}$$

Remark A-1 (A-7) *controls the mean square risk of a spectral cut-off. Equivalently, integrating with respect to the design, we have using Table A.1:*

$$\mathbb{E}(\hat{f}_{N_{ERM}}(X) - f_\eta(X))^2 \leq (1+\gamma) \min_{N \geq 1} \mathbb{E}(\hat{f}_N(X) - f_\eta(X))^2 + C\frac{(1+h)(1-h)}{n\gamma}.$$

We hence have a control of the mean square risk of $\hat{f}_{N_{URE}}$ with respect to f_η. Using for instance Bartlett et al. [14], we can bound the excess risk of $\hat{f}_{N_{URE}}$ with (A-7)*. In the sequel, we propose to estimate the generalization error of this classifier with simulated data. It gives rather good performances.*

Remark A-2 *Assumption* (A-6) *is related to the theory of ordered process. To control the variations of the process* (A-5)*, we use Lemma 4.5. It results almost immediately in* (A-7)*. Moreover the condition $h \geq \frac{1}{\sqrt{3}}$ ensures the process* (A-5) *to be ordered (see Lemma 4.3 in Chapter 4).*

Proof
The proof follows the proof of Theorem 4.1 in Chapter 4. First step is to find a hull for the URE method.

Lemma A-1 *Under the previous assumption, there exists constants B,$C > 0$ such that for any $\alpha > 0$:*

$$l_{URE}(\theta,N) = (1+\alpha)\left(\sum_{k>N} \theta_k^2 + \epsilon^2 N\right) + \frac{\epsilon^2 CB}{\alpha}$$

is a risk hull, i.e.:

$$\mathbb{E} \sup_{N \geq 1} \left[\sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^N \xi_k^2 - l_\alpha(\theta,N)\right] \leq 0.$$

Proof. We have:

$$\begin{aligned}
& \mathbb{E}\left(\sum_{k>N} \theta_k^2 + \epsilon^2 \sum_{k=1}^N \xi_k^2 - l_{URE}(\theta,N)\right) \\
\leq\ & \mathbb{E}\left(\epsilon^2 \sum_{k=1}^N (\xi_k^2 - 1) - \alpha\epsilon^2 N\right) - \frac{CB\epsilon^2}{\alpha} \\
\leq\ & \mathbb{E}\left(\epsilon^2 \sum_{k=1}^N (\xi_k^2 - 1) - \frac{\alpha}{B}\epsilon^2 \mathrm{var}\left[\sum_{k=1}^N \xi_k^2\right]^{p/2}\right) - \frac{CB\epsilon^2}{\alpha},
\end{aligned}$$

where last line uses the fact that:

$$\begin{aligned} \text{var}\left[\sum_{k=1}^{N}\xi_k^2\right] &= \frac{6h^2-2}{(1+h)(1-h)}\sum_{k=1}^{N}\sum_{k'=1}^{N} < V_k^2, V_{k'}^2 > +2N \\ &\leq \left(\frac{6h^2-2}{(1+h)(1-h)}+2\right)N := BN. \end{aligned} \tag{A-8}$$

We know from Chapter 4 that $\zeta(N) = \epsilon^2 \sum_{k=1}^{N}(\xi_k^2 - 1)$ is an ordered process. Hence with (A-6), gathering with Lemma 4.5, one gets:

$$\mathbb{E}\left(\zeta(N) - \frac{\alpha}{B}\epsilon^2 \text{var}\zeta(N)^{p/2}\right) \leq \frac{\epsilon^2 CB}{\alpha},$$

which concludes the proof.□
$l_{URE}(\theta,N)$ is a risk hull. Then for $\tilde{N} = N_{URE}$:

$$\mathbb{E}r(\theta,\tilde{N}) \leq \mathbb{E}l_{URE}(\theta,\tilde{N}) = (1+\alpha)\mathbb{E}R(\theta,\tilde{N}) + \frac{\epsilon^2 CB}{\alpha}. \tag{A-9}$$

Moreover, we have for $\tilde{N}$:

$$\begin{aligned} R(\theta,\tilde{N}) &= \sum_{k=1}^{n}\theta_k^2 - \sum_{k=1}^{\tilde{N}} y_k^2 + \epsilon^2\tilde{N} - \epsilon^2\sum_{k=1}^{\tilde{N}}\xi_k^2 + 2\epsilon\sum_{k=1}^{\tilde{N}} y_k\xi_k \\ &= \sum_{k=1}^{n}\theta_k^2 - \sum_{k=1}^{\tilde{N}} y_k^2 + 2\epsilon^2\tilde{N} - \epsilon^2\sum_{k=1}^{\tilde{N}}(\xi_k^2-1) + 2\epsilon\sum_{k=1}^{\tilde{N}}\theta_k\xi_k. \end{aligned}$$

From (A-4), we have for any $N \geq 1$:

$$-\sum_{k=1}^{\tilde{N}} y_k^2 + 2\epsilon^2\tilde{N} \leq -\sum_{k=1}^{N} y_k^2 + 2\epsilon^2 N$$

Hence gathering with the previous identity, we obtain for any $N \geq 1$:

$$\mathbb{E}R(\theta,\tilde{N}) \leq R(\theta,N) + \mathbb{E}\epsilon^2\sum_{k=1}^{\tilde{N}}(\xi_k^2-1) + \mathbb{E}2\epsilon\sum_{k=1}^{\tilde{N}}\theta_k\xi_k.$$

Using the proof of Theorem 4.1 in Chapter 4, we can write, for any $\gamma > 0$:

$$\mathbb{E}2\epsilon\sum_{k=1}^{\tilde{N}}\theta_k\xi_k \leq 2D(\kappa)\left[\gamma R(\theta,N) + \gamma\mathbb{E}R(\theta,\tilde{N}) - \gamma\mathbb{E}\epsilon^2\tilde{N} + \frac{\epsilon^2}{\gamma}\right].$$

Using again Lemma 4.5, for any $\mu > 0$, we have:

$$\mathbb{E}\epsilon^2\sum_{k=1}^{\tilde{N}}(\xi_k^2-1) \leq \epsilon^2\mu\mathbb{E}\text{var}\zeta(\tilde{N})^p + \frac{\epsilon^2 C}{\mu}.$$

Gathering with previous inequality, we arrive at, for any $\gamma < \frac{1}{D}$:

$$\mathbb{E}R(\theta,\tilde{N}) \leq \frac{1+D\gamma}{1-D\gamma}R(\theta,N) + \frac{\epsilon^2\mu\mathbb{E}\mathrm{var}\zeta(\tilde{N})^p - D\gamma\epsilon^2\mathbb{E}\tilde{N}}{1-D\gamma} + \frac{\epsilon^2 D}{\gamma(1-D\gamma)} + \frac{\epsilon^2 C}{\mu(1-D\gamma)}.$$

Using (A-8) and putting $\mu = \gamma D$ we obtain:

$$\mathbb{E}R(\theta,\tilde{N}) \leq \frac{1+D\gamma}{1-D\gamma}R(\theta,N) + \frac{\epsilon^2 D}{\gamma(1-D\gamma)} + \frac{\epsilon^2 C}{\gamma D(1-D\gamma)}.$$

From (A-9), one gets:

$$\mathbb{E}r(\theta,\tilde{N}) \leq (1+\alpha)\left(\frac{1+D\gamma}{1-D\gamma}R(\theta,N) + \frac{\epsilon^2 D}{\gamma(1-D\gamma)} + \frac{\epsilon^2 C}{\gamma D(1-D\gamma)}\right) + \frac{\epsilon^2 CB}{\alpha},$$

which concludes the proof by choosing α properly.

A-3 Simulations

A-3.1 Description of the data

We use R to simulate training sets D_n satisfying assumption (A-1). We propose to illustrate the behavior of the URE method for different simulated data. These datasets are made of 500 observations (X_i,Y_i) where $X_i \in [-1,1]^2$ are i.i.d. with uniform law in $[-1,1]^2$. These training sets depend on two parameters (h,α) such that:

- h is the classification ability in assumption (A-1). It represents the inverse level of noise in the labels.
- α describes the complexity of the Bayes decision rule. We choose a Bayes rule f^*_α such that

$$f^*_\alpha(x) = \begin{cases} 2\ \mathbb{1}_{\{x=(x_1,x_2)\in\mathcal{X}:x_2-\cos(\alpha x_1)>0\}}(x) - 1 \text{ if } \alpha > 0, \\ 2\ \mathbb{1}_{\{x=(x_1,x_2)\in\mathcal{X}:x_2>0\}} \text{ for } \alpha = 0. \end{cases}$$

It gives a large choice of training sets to generate. In Figure A.1-A.4, we plot four different datasets varying the level of noise h and the complexity α.

A-3.2 Oracle efficiency

The behavior of the URE method is studied for differents amont of noise h. We generate 1000 realizations of data sets $D_n = \{(X_i,Y_i), i = 1\dots n\}$. Each realization D_n^p corresponds to a vector θ^p to estimate. We construct the family of spectral cut-off $\{\hat{\theta}^p(N), 1 \leq N \leq n\}$ using the Laplace kernel defined by:

$$K(x,y) = \exp(-5\|x-y\|).$$

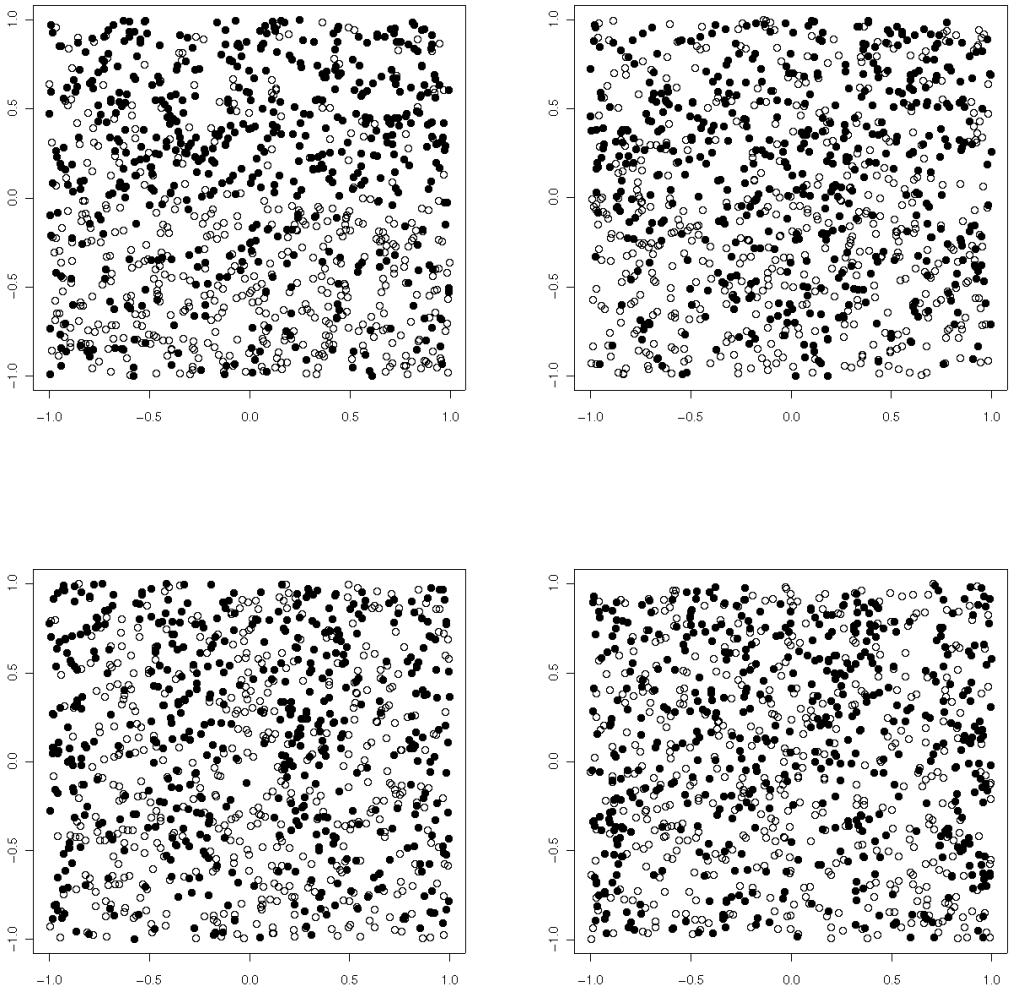

Figure A.1-A.4: Representations of Datasets[2] *of size* $n = 500$.

2.

- top left: $(h,\alpha) = (0.5,0)$.
- top right: $(h,\alpha) = (0.3,0)$.
- bottom left: $(h,\alpha) = (0.5,10)$.
- bottom right: $(h,\alpha) = (0.3,10)$.

We calculate N_{URE}^p with URE method. We want to compare the risk $R(\theta,N_{URE})$ of the adaptive estimator $\hat{\theta}(N_{URE})$ with the risk $R(\theta,N^*)$ of the oracle. The quality of the estimator is measured by its oracle efficiency over the 1000 realizations:

$$r_h(N_{URE},N^*) = \frac{1}{1000}\sum_{p=1}^{1000} \frac{R(\theta^p,N_{URE}^p)}{R(\theta^p,N^{*,p})}.$$

We compare the mean bandwidth choice of the URE method with the oracle thanks to the quantities:

$$mean(N_{URE}) = \frac{1}{1000}\sum_{p=1}^{1000} N_{URE}^p \text{ and } mean(N^*) = \frac{1}{1000}\sum_{p=1}^{1000} N^{*,p},$$

and the associated standard deviations. The oracle efficiency of the URE method is summarized in Table A.2 for $\alpha = 0$.

Classification ability	mean(N^*)	mean(N_{URE})	$r_h(N_{URE},N^*)$
$h = 0.5$	8.71±2.29	12.84±8.57	0.92
$h = 0.3$	4.65±1.92	6.24±4.66	0.91
$h = 0.1$	2.65±0.48	3.09±2.92	0.80

Table A.2: Behavior of URE method for $\alpha = 0$.

The performances of URE are rather good. The ratio is close to one, even in the worst case ($h = 0.1$). It corresponds to a good oracle efficiency. The risk of the adaptive estimator $\hat{\theta}(N_{URE})$ is close to the risk of the oracle. However the choice of the bandwidth N_{URE} is instable. Moreover the method seems to propose large choice of N, compared to the oracle. This table illustrates the instability of the method in this case.

To take a closer look at this phenomena, we plot for each realization $p = 1,\dots 1000$ the loss and the associated bandwidth N_{URE}^p as a stem diagram (Figure A.5). A minority of N_{URE}^p are chosen inadequately. It results in a larger square loss but does not affect substancially the performances of the method.
This instability could be explain as follows. The principle of URE is to minimize a criterion based on the unbiased estimation of the true risk:

$$\text{(A-10)} \qquad U(y,N) = -\sum_{k=1}^{N} y_k^2 + 2\epsilon^2 N \xrightarrow{\text{estimates}} U(\theta,N) = -\sum_{k=1}^{N} \theta_k^2 + \epsilon^2 N.$$

This quantity contains a deterministic positive term and a stochastic negative term. The variance of the stochastic term depends on the noise level. It could affects the behavior of the data-driven choice N_{URE}. This could explain why URE allows large N in some cases. Figure A.6 illustrates numerically this phenomena. We generate a design $X_i, i = 1\dots n$ and plot the corresponding $U(\theta,N)$ (solid line) defined in (A-10). We compare the true risk with 3 realizations of $U(y,N)$ (dashed lines) for different realizations of labels $Y_i, i = 1\dots n$ for two level of noise $h = 0.5$ and $h = 0.3$.

The same performances are reached for different complexity parameters α. This parameter does not affect the oracle efficiency of the method. However, it reduces the performances of the oracle.

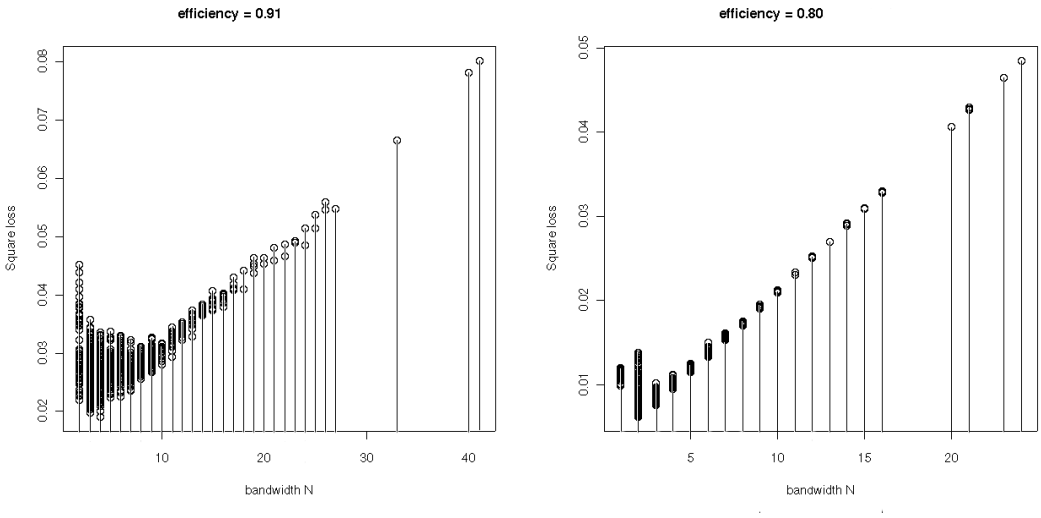

Figure A.5: Behavior of the URE method for $h = 0.3$ and $h = 0.1$ when $\alpha = 0$.

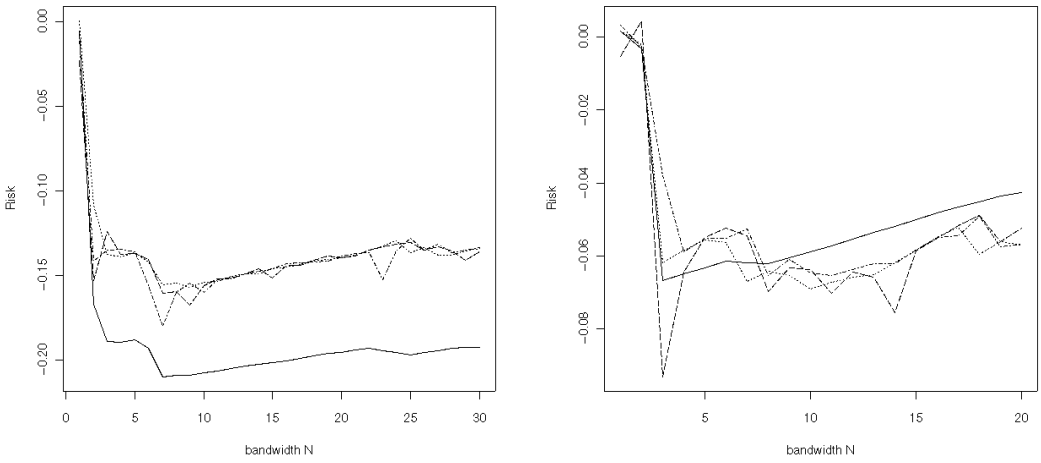

Figure A.6: Behavior of criterion $U(y,N)$ for $h = 0.5$ and $h = 0.3$.

A-3.3 KPM using URE VS SVM using Aggregation

We illustrate the generalization performances of the adaptive classifier $\hat{f}_{N_{URE}}$ over simulated data sets. We compare the classification errors of KPM using URE method with the aggregate $\tilde{f}_n$ of Chapter 2. Recall this classifier is given by:

$$\tilde{f}_n = \sum_{\lambda \in \mathcal{G}} \omega_\lambda \hat{f}_\lambda,$$

where $\hat{f}_\lambda$ is a SVM classifier with smoothing parameter λ. $\mathcal{G}$ is a finite grid and ω_λ are weights such that $\sum \omega_\lambda = 1$.
We generate 100 training and test sets of size $n = 500$ for different parameters (h,α). For each realization of training and test set (D_n^p, T_n^p), we act in two steps:

- We train the KPM using URE method $\hat{f}^p_{N_{URE}}$ and the aggregate of SVM $\tilde{f}^p_n$.
- We test the performances of the two classifiers over the test set $T_n^p = (X_i^p, Y_i^p)$.

For each $p = 1, \dots 100$, the winner is given one point. For completeness, we also mention the mean classification error of $\hat{f}_{N_{URE}}$ and $\tilde{f}_n$ as follows:

$$\text{risk}(\hat{f}) = \frac{1}{100} \sum_{p=1}^{100} \frac{1}{500} \sum_{i=1}^{500} \mathbb{1}(\hat{f}^p(X_i^p) \neq Y_i^p), \text{ for } \hat{f} = \tilde{f}_n^p \text{ and } \hat{f} = \hat{f}^p_{N_{URE}}.$$

Table A.3 summarizes the numerical results.

parameters (h,α)	SVM	KPM	risk$(\tilde{f}^p_n)$	risk$(\hat{f}^p_{N_{URE}})$
(0.5,0)	67	25	26.72	26.25
(0.3,0)	45	49	37.23	37.08
(0.5,10)	2	97	38.87	34.42
(0.3,10)	59	37	44.79	45.36

Table A.3: KPM using URE VS SVM using Aggregation.

In this particular toy model, $\hat{f}_{N_{URE}}$ reaches the same generalization performances as the agregate of SVM $\tilde{f}_n$. However, the KPM is not adaptive with respect to the level of noise in the data. On the contrary, $\tilde{f}_n$ is data-dependent and can be used to solve real-world problems. Moreover the KPM is constructed in the spectral domain. It requires the diagonalization of the kernel matrix $K_n = (K(X_i, X_j))_{i,j=1\dots n}$. It gives numerical problems when the number of observations grows.

Finally in these numerical experiments, we do not treat the influence of the kernel in the performances. We focus on the Laplace kernel with fixed variance $\sigma = 5$. An interesting question is the influence of the kernel in both the representation of the observations y_k and the construction of the adaptive KPM.

Bibliographie

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] U. Amato, A. Antoniadis, and M. Pensky. Wavelet kernel penalized estimation for non-equispaced design regression. *Stat. Comput.*, 16 (1):37–55, 2006.

[4] S. Arlot. *Rééchantillonage et sélection de modèles*. PhD thesis, Université Paris-Sud, 2007.

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

[6] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54 (2):317–331, 1997.

[7] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sc. Paris*, 296:1021–1024, 1983.

[8] J.Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35 (2):608–633, 2007.

[9] Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

[10] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[11] P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44 (2):525–536, 1998.

[12] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[13] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

[14] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101 (473):138–156, 2006.

[15] P.L. Bartlett, S.R. Kulkarni, and S.E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43 (5):1721–1724, 1997.

[16] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[17] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.

[18] P.L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In *Proc. 17th Annu. Conference on Comput. Learning Theory*, volume 3120, pages 270–284, 2004.

[19] P.L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.

[20] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, 1988.

[21] G. Biau, F. Bunea, and M.H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[22] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

[23] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3 (3):203–268, 2001.

[24] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36 (2), 2008.

[25] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66 (2-3):259–294, 2007.

[26] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, 2003.

[27] G. Blanchard and L. Zwald. Finite dimensional projection for classification andstatistical learning. submitted to IEEE transactions on Information Theory, 2005.

[28] B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.

[29] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[30] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sc. Paris, Ser. I*, 334:495–500, 2002.

[31] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Machine Learning Summer School 2003*, 3176 of LNAI:208–240, 2004.

[32] S. Canu, X. Mary, and A. Rakotomamonjy. Functional learning through kernel. *Advances in Learning Theory: Methods, Models and Applications*, 190:89–110, 2003.

[33] Y. Cao and Y. Golubev. On oracle inequalities related to a polynomial fitting. *Math. Methods Statist.*, 6 (4):431–450, 2005.

[34] Y. Cao and Y. Golubev. On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15 (4):398–414, 2006.

[35] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4 (4): 377–408, 2006.

[36] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29:131–160, 1997.

[37] L. Cavalier. Inverse problems with non-compact operators. *Journal of Statistical Planning and Inference*, 136 (2):390–400, 2006.

[38] L. Cavalier, G.K. Golubev, D. Picard, and A.B. Tsybakov. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30 (3):843–874, 2002.

[39] L. Cavalier and Y. Golubev. Risk hull method and regularization by projections of ill-posed inverse problems. *The Annals of Statistics*, 34 (4):1653–1677, 2006.

[40] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19 (5):1155–1178, 2007.

[41] D.R. Chen, Q. Wu, Y. Ying, and D.X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

[42] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

[43] N. Cristianini and H. Shawe-Taylor. *Introduction to Support Vector Machines, and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.

[44] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39 (1):1–49, 2002.

[45] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41 (7):909–996, 1988.

[46] R.A. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Approximation methods for supervised learning. *Foundations of Computational Mathematics*, 6 (1): 3–58, 2006.

[47] R.A. DeVore and B.J. Lucier. Wavelets. *Acta Numerica*, pages 1–56, 1992.

[48] R.A. DeVore and V.A. Popov. Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, 305 (1):397–414, 1988.

[49] R.A. DeVore and R.C. Sharpley. Besov spaces on domains in $\mathcal{R}^d$. *Transactions of the American Mathematical Society*, 335 (2):843–864, 1993.

[50] L. Devroye. Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Vew. Gebiete*, 61 (4): 467–481, 1982.

[51] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[52] D.L. Donoho and I.M. Johnstone. Wavelet shrinkage: asymptotia? *J. Roy. Statist. Soc. Ser. B*, 57 (2):301–369, 1995.

[53] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24 (2):508–539, 1996.

[54] W.H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, 2000.

[55] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121 (2):256–285, 1995.

[56] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[57] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66 (2-3):165–207, 2007.

[58] J.B. Gao, C.J. Harris, and S.R. Gunn. On a class of support vector kernels based on frames in function Hilbert spaces. *Neural Computation*, 13:1975–1994, 2001.

[59] G.K. Golubev. How does the method of model selection explode? Manuscript, 2002.

[60] G.K. Golubev. The method of risk envelope in estimation of linear functionals. *Problem of Information Transmission*, 40 (1):53–65, 2004.

[61] G.K. Golubev. On a method of empirical risk minimization. *Problems of Information Transmission*, 40 (3):195–204, 2004.

[62] G.K. Golubev. The principle of penalized empirical risk in severely ill-posed problems. *Probability Theory and Related Fields*, 130 (1):18–38, 2004.

[63] Y. Golubev. On oracle inequalities related to high linear models. In *MA Proceedings. Topics in stochastic analysis and nonparametric estimation*, volume Springer-Verlag, 2007.

[64] Y. Golubev and B. Levit. An oracle approach to adaptive estimation of linear functionals in a gaussian model. *Math. Methods Statist.*, 13 (4):392–408, 2004.

[65] P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman and Hall, 1994.

[66] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48 (1):239–250, 2002.

[67] P.R. Halmos. What does the spectral theorem say? *Amer. Math. Monthly*, 70:241–247, 1963.

[68] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*. Lecture Notes in Statistics, 1997.

[69] W. Härdle, B.U. Park, and A.B. Tsybakov. Estimation of non-sharp support boundaries. *Journal of Multivariate Analysis*, 55 (2):205–218, 1995.

[70] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2002.

[71] H. Hochstadt. *Integral Equations*. Wiley, New-York, 1973.

[72] S. Jaffard. Décompositions en ondelettes. *Developments of Mathematics 1950-2000*, pages 609–634, 2000.

[73] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28 (3):681–712, 2000.

[74] A. Karatzoglou, A. Smola, and K. Hornik. An S4 package for kernel methods in R. Reference manual, 2007.

[75] G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statistics and Probability Letters*, 13 (1):15–24, 1992.

[76] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47 (5):1902–1914, 2001.

[77] T. Kühn, H.G. Leopold, W. Sickel, and L. Skrzypczak. Entropy numbers of embeddings of weighted Besov spaces. *Constr. Approx.*, 23 (1):61–77, 2006.

[78] G. Lecué. *Méthodes d'agrégation : optimalité et vitesses rapides*. PhD thesis, Université Paris VI, 2007.

[79] G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Proc. 19th Annu. Conference on Comput. Learning Theory*, volume 4005, pages 364–378, 2006.

[80] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13 (4):1000–1022, 2007.

[81] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *The Annals of Statistics*, 35 (4):1698–1721, 2007.

[82] G. Lecué. Classification with minimax fast rates for classes of Bayes rules with sparse representation. *Electronic Journal of Statistics*, 2:741–773, 2008.

[83] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

[84] O.V. Lepskii. Asymptotically minimax adaptative estimation I: Upper bounds. Optimally adaptative estimates. *Theory Probab Appl.*, 36 (4):682–697, 1991.

[85] Y. Lin. Support vector machines and the Bayes rule in classification. *Data Min. Knowl. Discov.*, 6 (3):259–275, 2002.

[86] Y. Lin and L.D. Brown. Statistical properties of the method of regularization with periodic Gaussian reproducing kernel. *The Annals of Statistics*, 32 (4):1723–1743, 2004.

[87] S. Loustau. Aggregation of SVM classifiers using Sobolev spaces. *Journal of Machine Learning Research*, 9:1559–1582, 2008.

[88] S. Loustau. Penalized empirical risk minimization over Besov spaces. Submitted, 2008.

[89] S. Loustau. Risk hull minimization and kernel projection machines in classification. Manuscript, 2008.

[90] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32 (4):1679–1697, 2004.

[91] S. Mallat. *Une exploration des signaux en ondelettes*. Ellipses, 2000.

[92] P. Malliavin. *Analyse de Fourier-Analyses spectrales*. Ecole Polytechnique, 1974.

[93] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[94] E. Mammen and A.B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *The Annals of Statistics*, 23 (2):502–524, 1995.

[95] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

[96] C. Marteau. *Recherche d'inégalités oracles pour des problèmes inverses*. PhD thesis, Université de Provence, 2007.

[97] C. Marteau. Risk hull method for general families of estimators. Manuscript, 2007.

[98] X. Mary. *Sous-espaces hilbertiens, sous-dualités et applications*. PhD thesis, INSA de Rouen, 2003.

[99] X. Mary, D. De Brucq, and S. Canu. Sous-dualités et noyaux (reproduisants) associés. *C. R. Acad. Sc. Paris*, 336 (1):949–954, 2003.

[100] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9 (2):245–303, 2000.

[101] P. Massart. Concentration inequalities and model selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer, 2007.

[102] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006.

[103] M. Matache and V. Matache. Hilbert spaces induced by Toeplitz covariance kernels. *Lecture notes in Control and Information Sciences*, 280:319–334, 2002.

[104] S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[105] Y. Meyer. *Ondelettes et Opérateurs 1 : Ondelettes*. Hermann, 1990.

[106] F. Natterer. Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis*, 18 (1-2):29–37, 1984.

[107] A. Nemirovski. *Topics in Nonparametric Statistics*. Ecole d'été de Saint-Flour XX-VIII, Springer, N.Y., 1998.

[108] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. In *Proc. 21st International Conference on Machine Learning*, volume 69, pages 639–646, 2004.

[109] J. Peetre. *New thoughts on Besov spaces*. Mathematics Department, Duke University, Durham, N.C., 1976.

[110] A. Rakotomamonjy and S. Canu. Frame, reproducing kernel, regularization and learning. *Journal of Machine Learning Research*, 6:1485–1515, 2005.

[111] D. Rätsch, T. Onoda, and K.R. Müller. Soft margin for adaboost. Esprit Working Group in Neural and Computational Learning II, 1998.

[112] R. Rifkin, G. Yeo, and T. Poggio. *Advances in Learning Theory : Methods, Model and Applications*. Nato Sciences Serie III, Vol. 190, Chapter 7, 131-154, 2003.

[113] E. Rio. *Théorie asymptotique des processus faiblement dépendants*. Springer-Verlag, 2000.

[114] E. Rio. A Bennet type inequality for maxima of empirical processes. *Ann. I. H. Poincaré*, 38 (6):1053–1057, 2002.

[115] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Cornell Aeronautical Laboratory, Psychological Review*, 65 (6):386–408, 1958.

[116] H.P. Rosenthal. On the span in l_p of sequences of independent random variables. *Israël J. Math.*, 8:273–303, 1972.

[117] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[118] S. Smale and D.X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1 (1):17–41, 2003.

[119] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9 (6):1135–1151, 1981.

[120] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[121] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[122] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51 (1):128–142, 2005.

[123] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52 (10):4635–4643, 2006.

[124] I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. In *Proc. 19th Annu. Conference on Comput. Learning Theory*, volume 4005, pages 79–93, 2006.

[125] I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. *Neural Information Processing Systems*, 19:1321–1328, 2007.

[126] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35 (2):575–607, 2007.

[127] H. Sun. Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21 (3):337–349, 2005.

[128] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.*, 58 (1):267–288, 1996.

[129] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K.Knight. Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B*, 67 (1):91–108, 2005.

[130] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston, Washington, 1977.

[131] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Company, 1978.

[132] H. Triebel. *Theory of Functions Spaces II*. Birkhauser, 1992.

[133] H. Triebel. *The Structure of Functions*. Birkhauser, 2001.

[134] A.B. Tsybakov. Optimal rates of aggregation. *Learning Theory and Kernel Machines*, 2777:303–313, 2003.

[135] A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer-Verlag, 2004.

[136] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004.

[137] A.B. Tsybakov and S.A. Van De Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33 (3):1203–1224, 2005.

[138] V. Vapnik. *Estimation of Dependances Based on Empirical Data*. Springer Verlag, 1982.

[139] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[140] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 (2):264–280, 1971.

[141] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

[142] R. Vert and J.P. Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

[143] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6: 883–904, 2005.

[144] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. 59. Philadelphia, SIAM, 1990.

[145] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47 (6):2516–2532, 2001.

[146] Q. Wu, Y. Ying, and D.X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23 (1):108–134, 2007.

[147] Q. Wu and D.X. Zhou. Analysis of support vector machine classification. *J. Comput. Anal. Appl.*, 8 (2):99–119, 2006.

[148] Y. Yang. Minimax nonparametric classification-Part I: rates of convergence. *IEEE Transactions on Information Theory*, 45 (7):2271–2284, 1999.

[149] Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28 (1): 75–87, 2000.

[150] G. Zeng and R. Zhao. Image denoising using least squares wavelet support vector machines. *Chinese Optics Letters*, 5 (11):632–635, 2007.

[151] L. Zhang, W. Zhou, and L. Jiao. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 34 (1):34–39, 2004.

[152] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32 (1):56–85, 2004.

[153] D.X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49 (7):1743–1752, 2003.

[154] L. Zwald. *Performances statistiques d'algorithmes d'apprentissage : Kernel Projection Machine et analyse en composantes principales à noyau.* PhD thesis, Université Paris-Sud, 2005.

Résumé

Cette thèse se concentre sur le modèle de classification binaire. Etant donné n couples de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) (X_i,Y_i), $i = 1,\ldots,n$ de loi P, on cherche à prédire la classe $Y \in \{-1,+1\}$ d'une nouvelle entrée X où (X,Y) est de loi P. La règle de Bayes, notée f^*, minimise l'erreur de généralisation $R(f) = P(f(X) \neq Y)$. Un algorithme de classification doit s'approcher de la règle de Bayes. Cette thèse suit deux axes : établir des vitesses de convergence vers la règle de Bayes et proposer des procédures adaptatives.

Les méthodes de régularisation ont montrées leurs intérêts pour résoudre des problèmes de classification. L'algorithme des Machines à Vecteurs de Support (SVM) est aujourd'hui le représentant le plus populaire. Dans un premier temps, cette thèse étudie les performances statistiques de cet algorithme, et considère le problème d'adaptation à la marge et à la complexité. On étend ces résultats à une nouvelle procédure de minimisation de risque empirique pénalisée sur les espaces de Besov. Enfin la dernière partie se concentre sur une nouvelle procédure de sélection de modèles : la minimisation de l'enveloppe du risque (RHM). Introduite par L.Cavalier et Y.Golubev dans le cadre des problèmes inverses, on cherche à l'appliquer au contexte de la classification.

Mots Clés : Apprentissage statistique, Classification, Vitesses de convergence, Adaptation, Sélection de modèles.

Abstract

This manuscript studies the statistical performances of kernel methods to solve the binary classification problem. We observe an independent and identically distributed (i.i.d.) sequence of random pairs (X_i,Y_i), $i = 1,\ldots,n$ of unknown probability distribution P over $\mathcal{X} \times \{-1,+1\}$. The aim is to learn, from a new observation X, the corresponding class $Y \in \{-1,+1\}$ where $(X,Y) \sim P$. A well-known minimizer of the generalization error $R(f) = P(f(X) \neq Y)$ is called the Bayes rule. The present work proposes to give rates of convergence to the Bayes for SVM-type minimization. We also study a new procedure of model selection called the Risk Hull Minimization.

Kernel methods are based on the minimization of an empirical risk and a regularizer. Support Vector Machines is one of the most classical representant. In a first time, we propose learning rates for this algorithm and consider the problem of adaptation to the margin and to the complexity. We also provide a new procedure of penalized empirical risk minimization using Besov spaces as hypothesis spaces. Finally we study a new method of model selection called the Risk Hull Minimization. We use this procedure in binary classification to select the bandwidth of a KPM classifier and give statistical performances in terms of oracle inequality.

Keywords : Statistical Learning, Classification, Learning rates, Adaptation, Model selection.

AMS 2000 subject classification: 62H30, 62G05, 68Q32.