ORSAY
N° D'ORDRE : ????

UNIVERSITÉ PARIS XI
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

SPÉCIALITÉ : MATHÉMATIQUES

par

Laurent Zwald

Sujet: **PERFORMANCES STATISTIQUES D'ALGORITHMES D'APPRENTISSAGE : "KERNEL PROJECTION MACHINE" ET ANALYSE EN COMPOSANTES PRINCIPALES À NOYAU**

Rapporteurs : M. Peter L. BARTLETT
M. Jean-Philippe VERT

Soutenue le 23 novembre 2005 devant le jury composé de :

M.	Gilles BLANCHARD	Examinateur
M.	Stéphane BOUCHERON	Président
M.	Stéphane ROBIN	Examinateur
M.	Pascal MASSART	Directeur de Thèse
M.	Jean-Philippe VERT	Rapporteur

Acknowledgements - Remerciements

À Pascal et Stéphane pour m'avoir proposé un sujet passionnant. C'est aussi grâce à vous que de fructueuses collaborations ont pu voir le jour. Votre confiance et votre optimisme ont constitué un soutien inestimable pendant ces trois années.

À Olivier Bousquet : ton enthousiasme et ta disponibilité lors de nos nombreux échanges d'e-mails ont été précieux. De plus, la semaine passée au "Max Planck Institute" a été riche d'enseignements.

Un merci tout particulier à Gilles Blanchard pour s'être autant intéressé à mon travail. L'attention et la patience dont tu as fais preuve ont été une source constante de motivation. Sans la perspicacité de tes remarques, ce travail n'aurait jamais pu aboutir.

À Régis Vert pour m'avoir initié à la programmation en Matlab. Ta modestie et ta ténacité ont permis de rendre notre travail agréable et efficace.

I should acknowledge the researchers of the departement "Empirical Inference for Machine Learning and Perceptron" of the Max Planck Institute for Biological Cybernetics (Tübingen) and those of the "Intelligent Data Analysis Group" of the Fraunhofer Institute (Berlin) for their warm welcome and the interesting discussions. Special thanks to Bernhard Schölkopf, Mikhail Belkin, Olivier Chapelle, Matthias Hein, Alexander Zien, Arthur Gretton, Ulrike von Luxburg for the Max Planck Institute and Klaus-Robert Müller, Mikio L. Braun, Christin Schäfer for the Fraunhofer Institute.

Thanks to my referees, Jean-Philippe Vert and Peter L. Bartlett.

À tous les membres du jury pour leur présence : Gilles, Stéphane, Pascal, Jean-Philippe et Stéphane Robin.

À toutes les personnes avec qui j'ai eu l'occasion de parler de mathématiques : Jean-Michel Poggi, Émilie Lebarbier, Jean-Philippe Vert, Michèle Sebag, Gilles Celeux, Yves Grandvalet, Laurent Cavalier, Jean-Pierre Kahane et Bernard Helffer.

À toute l'équipe de Probabilité et Statistiques de l'université d'Orsay pour leur sympathie. En particulier, à Vincent Rivoirard, Marie-Luce Taupin, Jean-Michel Loubes. Merci aussi à Yves Misiti et Patrick Jakubowic pour m'avoir rajouté de la mémoire et pour leur courtoisie.

À Valérie Lavigne pour son efficacité.

À tous les (ex)-doctorants d'Orsay pour leur soutien : Christine, pour m'avoir aidé durant tout notre cursus universitaire commun, Ismaël B., pour ses connaissances mathématiques et sa jovialité, Laurent T. , dont les plaisanteries sur les belettes me font encore rire, Nicolas, qui apprécie la compagnie des chiens, Aurélien, pour ses connaissances informatiques et pour l'organisation des nombreuses sorties, Sylvain et Jonas, pour les pause-thé relaxantes, Marie, pour sa gentillesse, Sophie, pour sa bonne humeur, Bouthaina, une des survivantes du mois d'août, Graham, infatigable relecteur des textes en anglais, Neil, Christian, Karine, Guillemette, Mina, Stefano, Ismaël C., Marion, Cédric, imbattable pour les jeux de mots et à la boursicotte, Mathieu, Magalie, Estelle, Béatrice, Réda, Thomas, Violaine, Antoine, Céline et Gilles S., relecteur attentif d'introduction.

Aux amis et, en particulier, au joyeux groupe d'Igny pour les soirées de détente : Yohann, Carole, Yann, Sophie, Stéphanie, Jean-Patrick, Monsieur Thomas, Madame Stéphanie, Denis, Xavier, Pauline, Virginie.

À Nadine, pour sa présence réconfortante qui m'a permis de surmonter les périodes de doutes.

À mes parents

Contents

Chapter 1

Présentation générale

Ce chapitre introductif poursuit un double but. D'une part, il vise à expliquer comment les algorithmes d'apprentissage peuvent bénéficier des méthodes de statistique théorique. Des liens étroits entre les domaines du **machine learning** et de la **statistique** classique seront exhibés. D'autre part, il résume les contributions de la présente thèse :

1. Études statistiques d'un algorithme de réduction de la dimension : l'analyse en composantes principales à noyau (chapitres 3 et 4).

2. Conception d'un nouvel algorithme de classification : la Kernel Projection Machine. Évaluation de ses performances et analyses statistiques (chapitres 5 et 6).

Contents

1.1 Apprentissage et statistique.

1.1.1 Apprentissage.

Un problème d'*apprentissage*, aussi appelé apprentissage à partir d'exemples, se formule de façon formelle comme suit : un objet X est observé et le but est de lui associer une *sortie* Y. Le qualificatif de "sortie" provient de l'anglais "output" utilisé par la communauté informatique. On suppose pouvoir effectuer des mesures sur l'objet. Dans le cas le plus simple de la *classification binaire*, *l'étiquette* Y peut prendre deux valeurs notées arbitrairement -1 et $+1$. Ce cadre abstrait de l'apprentissage englobe de nombreuses applications. Par exemple (cette liste est loin d'être exhaustive),

- l'aide au diagnostic médical,
- la classification des e-mails,
- la catégorisation de textes,
- la bio-informatique : analyse de données biologiques, compréhension de la fonction de certaines parties de séquences d'ADN.

Dans le cas de l'aide au diagnostic médical, si on cherche à étudier une certaine maladie, X représente un patient sur lequel on fait des relevés médicaux (examens sanguins, taille, poids,...) et le fait qu'il ait contracté la maladie est représenté par l'étiquette : elle vaut $+1$ s'il est sain et -1 s'il est malade. De façon générale, la qualité et le type de mesures prises sur X sont importantes : leur pertinence favorise une détermination fiable de Y. Dans ce travail, nous ne traiterons pas directement de ce problème. Toute l'information dont nous disposons sera toujours sous la forme d'un nombre *fini* d'exemples $X_1, \ldots, X_n$ représentés par leurs différentes mesures ainsi que les réels $Y_1, \ldots, Y_n$ qui leur sont associés. Les n couples $(X_i, Y_i), i = 1 \ldots n$ forment *l'échantillon d'apprentissage* noté $\mathcal{L}_n$. Le but est alors de pouvoir prédire la sortie Y associée à un objet X ne faisant pas nécessairement partie de l'échantillon d'apprentissage : pour cela, les méthodes doivent *apprendre* la structure sous-jacente entre objets et sorties en utilisant uniquement le nombre fini de données fournies par l'échantillon d'apprentissage. De plus, afin de pouvoir utiliser ces méthodes pour des problèmes concrets, elles doivent être réalisables de façon automatique, c'est-à-dire programmables sur une machine sous forme d'un *algorithme d'apprentissage*. Le principe d'un tel algorithme est donc le suivant. L'entrée est uniquement composée des données d'apprentissage et la sortie est une fonction $\widehat{f}$ dont l'évaluation en X donne une prédiction de la sortie qui lui est associée. Dans le cas où on connaît aussi la valeur de la sortie en question, on dit que l'algorithme a *appris* (ou qu'il *généralise*) si sa prédiction est suffisamment proche de la véritable valeur. La qualité d'un algorithme se mesure donc uniquement par la qualité de ses prédictions sur des données qu'il ne connaissait pas : les *données de test*. Ces données sont une façon de reproduire la structure qui lie les objets aux sorties Y et permettent de savoir si l'algorithme a appris à la reconnaître. Le fait que l'algorithme retrouve les bonnes sorties de l'échantillon d'apprentissage ne garantie pas de bonnes performances sur les données de test. En effet, les données d'apprentissage ne sont que des exemples : elles comportent leur part d'incertitude. Si l'algorithme la prend en compte, il accorde trop de confiance aux données d'apprentissage et généralise moins bien : c'est l'*overfitting*. Cependant, s'il n'est pas assez fidèle aux données (son unique source d'information), ses capacités de généralisation en seront aussi affectés : c'est l'*underfitting*. L'algorithme d'apprentissage doit donc rester fidèle à l'échantillon d'apprentissage sans donner trop d'importance à chaque donnée prise individuellement.

Il est intuitif que les performances de l'algorithme seront meilleures si on lui présente de "bonnes" données. Leur mise en forme est un aspect à la fois essentiel et difficilement quantifiable mathématiquement : de nombreuses expériences ont montré qu'un algorithme d'apprentissage est plus efficace s'il est précédé d'une étape de compression des données.

Les performances d'un algorithme seront comparées à celles d'un cas idéal où on posséderait une infinité de données et où on pourrait explorer toutes les fonctions possibles. Une des difficultés réside dans le fait qu'on ne connaît pas la fonction de sortie de l'algorithme tant qu'il n'a pas fonctionné sur les données : les études générales de ses performances nécessitent donc des contrôles valables *uniformément* sur l'ensemble des fonctions possibles considérées par l'algorithme en question. Il est important de respecter le principe de fonctionnement de l'algorithme et, en particulier, de faire le moins possible d'hypothèses pour obtenir ces contrôles.

La partie suivante vise à introduire le cadre statistique général de cette thèse tout en montrant qu'il est adapté aux spécificités des problèmes d'apprentissage.

1.1.2 Cadre mathématique.

Le cadre statistique classique d'un problème d'apprentissage est le suivant : les données d'apprentissage sont modélisées comme étant n couples de variables aléatoires indépendantes et identiquement distribuées (X_i, Y_i) , $i = 1 \dots n$. C'est la base même des statistiques : on suppose pouvoir tirer infiniment de données suivant une loi fixée. Les données X_i sont les *variables de position* (ou variables d'entrée) : elles appartiennent à *l'espace des entrées* $\mathcal{X}$ et sont toutes distribuées suivant la loi P. Y_i désigne la sortie associée à chaque X_i. Afin d'évaluer la performance de l'algorithme, on introduit une variable aléatoire générique (X, Y) indépendante et de même loi Q que les données d'apprentissage. Elle représente une donnée de test quelconque. Ainsi, les données de test et d'apprentissage sont des tirages aléatoires d'une même loi de probabilité.

Le contrôle de la qualité de l'algorithme se traduit mathématiquement par l'évaluation la plus fine possible de son *erreur de généralisation* $\mathbb{E}\left[\gamma(\widehat{f}, (X,Y))\right]$. Elle est définie relativement à un contraste γ mesurant l'écart entre la prédiction $\widehat{f}(X)$ donné par l'algorithme et la véritable sortie Y. Sa forme dépend du type de problème d'apprentissage considéré. Elle représente la moyenne des erreurs de l'algorithme sur toutes les données de test possibles. Les contrastes utilisés ici seront toujours choisis positifs et nuls si la sortie de l'algorithme prédit la bonne valeur. On quantifie mathématiquement le fait que l'algorithme soit capable d'apprendre par le fait que sa fonction de sortie ait une faible erreur de généralisation.

La loi sous-jacente des données étant inconnue, on ne peut qu'approcher cette erreur. Dans les expériences numériques, il suffit d'avoir préservé des données de test : elles serviront à quantifier la performance de l'algorithme en approchant l'intégrale définissant son erreur de généralisation par la moyenne empirique prise sur ces données de test. D'un point de vue théorique, on peut obtenir des *bornes* supérieures sur l'erreur de généralisation : idéalement, si ces bornes sont optimales (c'est-à-dire qu'on possède aussi des bornes inférieures du même ordre), elles doivent permettre de mettre en lumière des quantité-clés dans le contrôle de la performance.

C'est pour ce type de raison que les travaux de Vapnik (entre autres) visent à obtenir des *bornes en généralisation*, c'est-à-dire des bornes supérieures sur l'erreur de généralisation.

La vision statistique est légèrement différente : de ce point de vue, la sortie de l'algorithme est un *estimateur* de la meilleure sortie possible notée s. Elle est définie comme étant celle qui possède la plus petite erreur de généralisation parmi un très gros ensemble de fonctions S :

$$s = \arg\min_{f \in S} \mathbb{E}\left[\gamma(f,(X,Y))\right] . \tag{1.1}$$

Ainsi, les problèmes d'apprentissage suivent la problématique générale de la statistique qui consiste à obtenir des informations sur une distribution inconnue à partir de l'observation d'un nombre fini d'exemples. L'idée fondamentale de l'apprentissage consiste à ne pas chercher à expliquer entièrement le phénomène sous-jacent : on souhaite simplement être capable de faire de bonnes prédictions sans chercher à estimer la distribution elle-même. La *perte* statistique naturellement associée à ce problème,

$$L(\widehat{f},s) = \mathbb{E}_{X,Y}\left[\gamma(\widehat{f},(X,Y))\right] - \mathbb{E}_{X,Y}\left[\gamma(s,(X,Y))\right] , \tag{1.2}$$

traduit bien cette préoccupation en ne considérant que les erreurs de prédiction commises par la sortie $\widehat{f}$ de l'algorithme. $\mathbb{E}_{X,Y}$ désigne l'espérance par rapport à la variable aléatoire (X,Y)

Le *risque* est défini comme étant l'espérance de la perte L par rapport à l'échantillon d'apprentissage (notée $\mathbb{E}_{\mathcal{L}_n}$). Si S_m désigne un sous-ensemble de S, les bornes de risque obtenues sont de la forme :

$$\mathbb{E}_{\mathcal{L}_n}\left[L(\widehat{f},s)\right] \leq cL(f_m,s) + C(S_m,n) ,$$

où f_m désigne un élément particulier de S_m et $C(S_m,n)$ est une quantité reflétant la taille du sous-ensemble S_m. Si la constante c est égale à 1, cette inégalité est une borne en généralisation comme celles obtenues par Vapnik. Cependant, certaines bornes de cette thèse seront obtenues avec $c > 1$.

Introduisons maintenant la notion de risque minmax : c'est une mesure de la meilleure performance possible dans le pire cas.

$$\mathcal{R}(A) = \inf_{\widehat{t}} \sup_{Q} \mathbb{E}_{\mathcal{L}_n}\left[L(\widehat{t},s)\right] ,$$

où l'infimum est pris sur tous les estimateurs possibles $\widehat{t}$ construit à partir de l'observation. Quant au supremum, il est considéré sur toutes les distributions Q du couple (X,Y) pour lesquelles s appartient à A. On note

$$\mathcal{R}(\widehat{f},A) = \sup_{Q} \mathbb{E}_{\mathcal{L}_n}\left[L(\widehat{f},s)\right] ,$$

la plus mauvaise performance d'un estimateur $\widehat{f}$ pour une cible s appartenant à l'espace A. Le supremum est pris sur le même ensemble de distribution que précédemment. En utilisant la terminologie de [BBM99], l'estimateur $\widehat{f}$ est *approximativement minmax* (de façon non-asymptotique) par rapport à A s'il existe une constante universelle $\square$ telle que $\mathcal{R}(\widehat{f},A) \leq \square\,\mathcal{R}(A)$.

Quelques situations englobées par ce cas général sont maintenant décrites. Dans le cadre de la **régression**, le contraste est celui des moindres carrés $\gamma(f,(X,Y)) = (Y - f(X))^2$ où

$\mathcal{Y} = \mathbb{R}$. S est l'ensemble des fonctions mesurables de carré intégrable. La cible s est la *fonction de régression* $s(x) = \mathbb{E}[Y|X=x]$ et, de par le Théorème de Pythagore, la perte vérifie

$$\mathbb{E}\left[(Y-t(X))^2\right] - \mathbb{E}\left[(Y-s(X))^2\right] = \mathbb{E}\left[(t(X)-s(X))^2\right] .$$

Dans le cadre de la **classification** binaire supervisée, $\mathcal{Y} = \{-1,+1\}$, S est l'ensemble des fonctions mesurables de $\mathcal{X}$ à valeurs dans $\mathcal{Y}$, c'est-à-dire l'ensemble de toutes les *fonctions de classification* possibles. Le contraste est la fonction de *perte dure* (ou "hard loss") $\gamma(f,(X,Y)) = \mathbb{1}_{Y\neq f(X)}$ qui est la fonction de perte naturelle du problème. La cible s est alors la fonction de classification (appelée *de Bayes*) minimisant la proportion d'erreurs de prédiction. Une notation spéciale f^* lui est réservée. On montre facilement qu'il s'agit de la fonction

$$f^*(x) = \begin{cases} 1 & si\, \eta(x) \geq \frac{1}{2}\,, \\ -1 & si\, \eta(x) < \frac{1}{2}\,, \end{cases}$$

où $\eta(x) = \mathbb{P}[Y=1|X=x]$ joue un rôle analogue à celui de la fonction de régression du problème. On peut aussi remarquer que si f est à valeurs dans $\{-1,+1\}$, on a $\mathbb{E}\left[(Y-f(X))^2\right] = 4P(Y \neq f(X))$.

L'*erreur de classification* est la perte associée. Afin de ne pas perdre de vue que c'est la quantité à contrôler dans un problème de classification, une notation spéciale ℓ lui est consacrée et sera conservée tout au long de cette présentation. Elle vérifie ([DGL96]),

$$\begin{aligned} \ell(f,f^*) = \mathbb{P}[Y\neq f(X)] - \mathbb{P}[Y \neq f^*(X)] &= \mathbb{E}\left[|2\eta(X)-1|\mathbb{1}_{f(X)\neq f^*(X)}\right] \\ &= \frac{1}{2}\mathbb{E}\left[|2\eta(X)-1||f(X)-f^*(X)|\right] . \end{aligned} \tag{1.3}$$

Dans la suite, on dira qu'on est sous l'hypothèse de *marge* s'il existe $h > 0$ tel que

$$\forall x \in \mathcal{X}\,, |2\eta(x)-1| \geq h\,.$$

C'est un cas particulier de la *condition de marge de Tsybakov* qui permet de quantifier la difficulté d'un problème de classification et d'obtenir des vitesses rapides de convergence ([MT95]).

Pour des raisons explicitées dans la partie 1.5, l'algorithme des Support Vector Machines (SVM) exploite la *convexification du risque*. Dans ce cas, le contraste se met sous la forme $\gamma(g,(X,Y)) = \phi(-Yg(X))$ où $\phi(x) = (1+x)_+$. La fonction ϕ est appelée *hinge-loss*. La cible reste la même que pour la perte dure puisque f^* vérifie toujours ([Lin99])

$$f^* = \arg \min_{g \text{ mesurable}} \mathbb{E}\left[\phi(-Yg(X))\right] ,$$

De plus, on a $L(g,f^*) \geq \ell(\text{signe}(g), f^*)$ et

$$L(g,f^*) \geq \mathbb{E}\left[|2\eta(X)-1||\phi(-Yg(X)) - \phi(-Yf^*(X))|\right] . \tag{1.4}$$

1.2 Analyse en composantes principales à noyau.

1.2.1 Objectifs.

Comme évoqué dans la partie 1.1.1, la mise en forme des données fournies à l'algorithme est un élément essentiel de l'apprentissage. Imaginons que l'on dispose de "beaucoup" de données dans le sens où de nombreuses mesures ont pu être prises sur les objets. Chaque variable de position X contient alors une quantité importante d'information sur l'objet. Cependant, les différentes mesures sont probablement liées entre elles et l'information qu'elles apportent est alors globalement redondante. Elles contiennent aussi du "bruit" provenant, par exemple, d'imprécisions dans les mesures. Cette redondance et ce bruit masquent la structure qui relient les objets à leur sortie. Afin de permettre un bon apprentissage, il faut donc extraire "l'information essentielle" contenue dans les données. Cela peut aussi être une nécessité liée à la place que celles-ci occupent dans la mémoire de la machine. Les algorithmes d'apprentissage sont donc souvent précédés d'un *pre-process* visant à compresser l'information contenue dans les variables de position. Un des pre-process les plus utilisés est l'analyse en composantes principales à noyau (KPCA, en abrégé). La version d'origine de cette algorithme est l'analyse en composantes principales (PCA, en abrégé) qui fût beaucoup utilisée comme outil d'analyse descriptive des données en grande dimension. La KPCA permet de dépasser les limites dues à la linéarité de la PCA. La partie suivante décrit les critères que la PCA cherche à optimiser afin de donner un sens mathématique à ce que signifie le terme "information essentielle". L'algorithme de la PCA sera ensuite décrit . On en déduit aisément celui de la KPCA. Finalement, la dernière partie concerne les contributions originales de la thèse pour la KPCA.

1.2.2 Critères.

L'analyse en composantes principales (PCA) a été introduite par Pearson ([Pea01]) et développée indépendamment par Hotelling ([Hot33]). Bien que simple, son principe est utilisé dans des domaines extrêmement divers (météorologie, économie, ...). Elle considère uniquement les objets X. Pour le moment, ils sont supposés appartenir à un espace vectoriel de dimension finie $\mathbb{R}^p$. Une *variable* représente une mesure effectuée sur l'objet : p désigne le nombre de ces variables. La PCA vise à réduire le nombre de variables, c'est-à-dire la dimension du jeu de données, en exploitant la corrélation potentielle entre les variables initiales. Pour cela, elle propose de nouvelles variables *décorrélées* et ordonnées de telle façon que les premières retiennent le plus possible la variation présente dans les données initiales. Si les variables de départ sont très corrélées entre elles, l'information qu'elles donnent est très redondante et il suffira de considérer peu de nouvelles variables pour prendre en compte la plus grande partie de cette information. Formalisons maintenant ces idées avec un point de vue mathématique. Sauf mention explicite du contraire, cette partie utilise la norme et le produit scalaire euclidien.

Pour commencer, on suppose connaître la loi de la variable aléatoire X ; son espérance est notée $\mu = \mathbb{E}[X]$ et sa matrice de covariance est $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)']$. Il est ici naturel de considérer la matrice de covariance puisque on veut étudier les *liens* entre les variables: en effet, le coefficient $\Sigma_{i,j}$ est la covariance $\mathrm{Cov}(X^i, X^j) = \mathbb{E}\left[(X^i - \mathbb{E}\left[X^i\right])(X^j - \mathbb{E}\left[X^j\right])\right]$ entre la i^{e} et la j^{e} variable (notées X^i et X^j). Pour rendre l'explication suivante plus simple, on suppose que toutes les valeurs propres de Σ sont strictement positives. Afin de considérer simultanément toutes les variables, la PCA étudie les *combinaisons linéaires* $(X-\mu)'a$ où $\|a\| = 1$. D'un point de vue géométrique, sa valeur absolue représente la longueur de la projection de la variable aléatoire recentrée sur la droite vectorielle engendrée par a. Le vecteur a maximisant $\mathbb{E}\left[\langle a, X-\mu\rangle^2\right]$ sous la contrainte $\|a\| = 1$ correspond donc à la direction de plus grand allongement des données. D'un point de vue statistique, cela revient à chercher la combinaison linéaire ayant une variance maximale afin de rendre compte de la variation des données. L'expression de cette variance est :

$$\mathrm{Var}(\langle a, X-\mu\rangle) = a'\Sigma a\,.$$

Algébriquement, il faut donc maximiser la forme quadratique induite par la matrice de covariance sur la sphère euclidienne. La diagonalisation de Σ implique que le maximum soit atteint pour $a = \gamma_1$ un vecteur propre de Σ associé à sa plus grande valeur propre λ_1. La nouvelle variable $z_1 = (X-\mu)'\gamma_1$ est la première *composante principale* : elle vérifie $\mathrm{Var}(z_1) = \lambda_1$. La deuxième composante principale est de la forme $z_2 = (X-\mu)'b$. La corrélation linéaire entre deux variables aléatoires réelles W_1 et W_2 est mathématiquement définie par :

$$\mathrm{Corr}(W_1, W_2) = \frac{\mathrm{Cov}(W_1, W_2)}{\sqrt{\mathrm{Var}(W_1)\mathrm{Var}(W_2)}}\,. \tag{1.5}$$

La covariance entre z_1 et z_2 vérifie

$$\mathrm{Cov}(z_1, z_2) = \mathbb{E}\left[b'(X-\mu)(X-\mu)\gamma_1\right] = b'\Sigma\gamma_1 = \lambda_1 b'\gamma_1\,.$$

La PCA cherchant des variables décorrélées, z_2 doit donc être décorrélée de z_1. Ainsi, elle est choisie de variance maximale sous les contraintes $\|b\| = 1$ et $b'\gamma_1 = 0$. La diagonalisation de Σ fournit que la maximisation de $b'\Sigma b$ sous les contraintes précédentes est obtenue pour $b = \gamma_2$ un vecteur propre de Σ associé à sa deuxième plus grande valeur propre λ_2. La deuxième composante principale $z_2 = (X-\mu)'\gamma_2$ satisfait $\mathrm{Var}(z_2) = \lambda_2$.

En itérant ce procédé, les vecteurs propres de la matrice de covariance fournissent une base orthonormée de $\mathbb{R}^p$ adaptée aux données. Elle est obtenue par rotation du repère d'origine et représente les directions de plus grandes variances. La quantité de variance représentée par chacun des vecteurs de base est mesurée par les valeurs propres de la matrice de covariance. Cette base est ordonnée pour que les vecteurs prenant en compte le plus de variance soient les premiers de la base.

1.2.3 Algorithme de la PCA.

En pratique, la loi de X est inconnue. On ne dispose que d'un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et toutes distribuées suivant la loi de X. En analyse multidimensionnelle, chaque X_j est un *individu* et leur ensemble forme le *nuage des individus*. La i^{e} coordonnée de X_j est notée X_j^i : elle représente la mesure de la i^{e} variable sur le j^{e} individu. Par analogie avec la partie précédente, la i^{e} variable est le vecteur $X^i = (X_1^i, \ldots, X_n^i)$. Par abus, un vecteur de $\mathbb{R}^n$ sera appelé variable. L'espérance μ est alors estimée par $g = 1/n\sum_{j=1}^n X_j$ et la matrice de covariance Σ par la covariance empirique,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(X_j - g)(X_j - g)'\,.$$

Le coefficient $\widehat{\Sigma}_{i,j}$ est la covariance empirique entre la i^{e} et la j^{e} variable :

$$\widehat{\Sigma}_{i,j} = \mathrm{Cov}_n(X^i, X^j) = \frac{1}{n}\sum_{\ell=1}^{n}(X_\ell^i - g_i)(X_\ell^j - g_j)\,.$$

Dans la suite, les données seront regroupées dans la matrice T à p lignes et n colonnes. La j^e colonne est constituée des coordonnées de l'individu X_j. Elle est donc définie par

$$T_{i,j} = X_j^i \text{ pour } 1 \leq i \leq p \text{ et } 1 \leq j \leq n \,.$$

Le nuage recentré est constitué des nouveaux individus $\widetilde{X}_j = X_j - g \,, j = 1 \ldots n$. Son centre de gravité est l'origine puisque $1/n \sum_{i=1}^n \widetilde{X}_j = 0$. La matrice des données correspondante est $\widetilde{T}$: c'est une matrice à p lignes et n colonnes où la j^e colonne est constituée des coordonnées de l'individu $\widetilde{X}_j$. Les deux matrices T et $\widetilde{T}$ sont liées par la relation

$$\widetilde{T} = T(I_n - \frac{1}{n} 1\!\!1_{n,n}) \,, \tag{1.6}$$

où I_n et $1\!\!1_{n,n}$ sont deux matrices carrées de taille n : I_n est la matrice identité et tous les coefficients de $1\!\!1_{n,n}$ valent 1. De plus, la matrice de covariance satisfait :

$$\widehat{\Sigma} = \frac{1}{n} \widetilde{T} \widetilde{T}' \,. \tag{1.7}$$

Pour la PCA, les vecteurs propres de la matrice de covariance Σ sont approchés par les vecteurs propres de la matrice de covariance empirique $\widehat{\Sigma}$. Ces derniers sont notés $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$ et appelés *directions principales*. Ils sont associés aux valeurs propres $\widehat{\lambda}_1 \geq \widehat{\lambda}_2, \ldots \geq \widehat{\lambda}_p$ de $\widehat{\Sigma}$ ordonnées par valeurs décroissantes. Comme expliqué dans la partie précédente, ils représentent les directions de plus grande variation du nuage des individus. Les nouvelles variables proposées par la PCA sont les composantes principales : la i^e composante principale est maintenant le vecteur des projections de chaque individu $\widetilde{X}_j$ sur la droite engendrée par $\widehat{\gamma}_i$. D'après le paragraphe précédent, la i^e composante principale est de variance empirique $\widehat{\lambda}_i$ maximale sous contrainte de décorrélation empirique avec les autres composantes principales. On retrouve aussi l'interprétation géométrique : la projection de chaque individu $\widetilde{X}_j$ sur la droite engendrée par a avec $\|a\| = 1$ est le vecteur

$$(\widetilde{X}_1' a, \ldots, \widetilde{X}_n' a) \,.$$

La variance empirique de cette variable est

$$a' \widehat{\Sigma} a = \frac{1}{n} \sum_{j=1}^n \langle a, \widetilde{X}_j \rangle^2 \,.$$

La maximisation de la variance empirique $a'\widehat{\Sigma}a$ revient donc à chercher la direction de $\mathbb{R}^p$ où la somme des carrés des longueurs des projections de chaque individu est maximale.

L'espace choisi par la PCA est l'espace vectoriel engendré par les directions principales :

$$\widehat{V}_d = \langle \widehat{\gamma}_1, \ldots, \widehat{\gamma}_d \rangle \,.$$

L'erreur quadratique de reconstruction associée à un sous-espace V est la quantité

$$\frac{1}{n} \sum_{j=1}^n \|\widetilde{X}_j - \Pi_V(\widetilde{X}_j)\|^2 \,,$$

où Π_V désigne la projection orthogonale sur V. Elle représente l'erreur commise en projetant le nuage des individus sur l'espace V. On peut montrer (voir chapitre 3) que l'espace de dimension d choisi par la PCA minimise l'erreur quadratique moyenne parmi tous les espaces vectoriels de dimension d :

$$\widehat{V}_d = \arg \min_{V, \dim(V)=d} \frac{1}{n} \sum_{j=1}^{n} \|\widetilde{X}_j - \Pi_V(\widetilde{X}_j)\|^2 . \tag{1.8}$$

De plus, l'erreur de reconstruction commise par ce sous-espace vectoriel est donnée par la somme des petites valeurs propres de la matrice de covariance :

$$\frac{1}{n} \sum_{j=1}^{n} \|\widetilde{X}_j - \Pi_{\widehat{V}_d}(\widetilde{X}_j)\|^2 = \sum_{\ell=d+1}^{p} \widehat{\lambda}_\ell . \tag{1.9}$$

La quantité de variance d'un nuage d'individus est définie comme étant la somme, normalisée par $1/n$, des carrés des normes de chacun de ces individus. Elle représente la variation totale du nuage. La relation de Pythagore implique que

$$\sum_{j=1}^{n} \|\Pi_V(\widetilde{X}_j)\|^2 = \frac{1}{n} \sum_{j=1}^{n} \|\widetilde{X}_j\|^2 - \frac{1}{n} \sum_{j=1}^{n} \|\widetilde{X}_j - \Pi_V(\widetilde{X}_j)\|^2 .$$

La relation (1.8) montre donc que l'espace choisi par la PCA contient un maximum de la variation initiale des données. En effet, la quantité de variance du nuage projeté sur cet espace est maximale parmi celles obtenues par projection sur les espaces vectoriels de dimension d. La variation initiale (c'est-à-dire obtenue sans projection) des données est $\frac{1}{n} \sum_{j=1}^{n} \|\widetilde{X}_j\|^2 = \operatorname{tr} \widehat{\Sigma} = \sum_{i=1}^{p} \widehat{\lambda}_i$. Celle prise en compte par $\widehat{V}_d$ est la somme des d plus grandes valeurs propres de la matrice de covariance. En effet, de par les deux relations précédentes et l'égalité (1.9),

$$\sum_{j=1}^{n} \|\Pi_{\widehat{V}_d}(\widetilde{X}_j)\|^2 = \sum_{\ell=1}^{d} \widehat{\lambda}_\ell . \tag{1.10}$$

La somme cumulée des grandes valeurs propres de la matrice de covariance détermine donc la quantité de variance prise en compte. En pratique, la qualité de $\widehat{V}_d$ est souvent mesurée par cette quantité relativement à la variance totale : des critères purement empiriques de choix d'une dimension d'arrêt pour la PCA exploitent le pourcentage $100 \sum_{\ell=1}^{d} \widehat{\lambda}_\ell / \sum_{\ell=1}^{p} \widehat{\lambda}_\ell$ de variation conservée.

> **Question ouverte :** sous quel critère la variation non-conservée peut-elle être considérée comme négligeable ?

Cette question n'est pas résolue par les critères fondés sur le pourcentage de variation conservée puisque plus d est grand et plus cette proportion est élevée. En pratique, on est donc amené à fixer arbitrairement un seuil pour choisir la dimension de la PCA.

L'étude de l'erreur de généralisation de l'espace choisi par la PCA repose sur son erreur quadratique moyenne $\mathbb{E}\left[\|\Pi_{\widehat{V}_d^\perp}(X-\mu)\|^2\right]$. Elle représente l'erreur commise en projetant une données de test sur l'espace choisi par la PCA. Elle est étudiée dans le chapitre 3 mais ne permet pas de donner directement un critère pour le problème du choix de dimension en PCA puisque c'est une fonction décroissante de d (les espaces $\widehat{V}_d$ sont emboîtés).

Soit V_d l'espace ayant la plus petite erreur quadratique moyenne parmi tous les sous-espaces de dimension d :

$$V_d = \arg \min_{\dim(V)=d} \mathbb{E}\left[\|\Pi_{V^\perp}(X-\mu)\|^2\right] . \tag{1.11}$$

La qualité de l'espace $\widehat{V}_d$ choisi par la PCA peut se mesurer par rapport à celle de V_d. Ce dernier est engendré par les vecteurs propres de la matrice de covariance associés aux d plus grandes valeurs propres de Σ :

$$V_d = \langle \gamma_1, \dots, \gamma_d \rangle .$$

La PCA n'est pas invariante par changement d'échelle : si on change l'unité dans laquelle est exprimée une variable, cela modifie artificiellement sa variance. C'est pour cette raison qu'en pratique, les PCA ne se basent pas sur la diagonalisation de la matrice de covariance mais sur la diagonalisation de la matrice des corrélations linéaires (matrice carrée de taille p dont le $(i,j)^{\mathrm{e}}$ coefficient est $\mathrm{Corr}(X_i, X_j)$ définie par l'équation (1.5)). Elle permet de considérer uniquement des variables de variance 1. Ce problème ne sera pas abordé dans cette thèse.

1.2.4 Algorithme de la KPCA.

L'intérêt porté à la PCA pour résoudre des problèmes d'apprentissage a été récemment relancé par l'obtention d'une version non-linéaire de cet algorithme ([SSM99]) : la KPCA. En effet, tel qu'il a été présenté ici, il ne s'appuie que sur des dépendances *linéaires* entre les variables et cela restreint considérablement son efficacité. La KPCA permet d'exploiter des relations potentiellement non-linéaires entre les variables. On décrit maintenant comment la non-linéarité est introduite.

Les objets ne sont plus supposés de nature vectorielle : ils appartiennent à l'espace des entrées $\mathcal{X}$ qui est quelconque. La version non-linéaire de la PCA repose sur un principe général permettant d'obtenir des versions non-linéaires d'algorithmes : *l'astuce du noyau*. Ce principe est aussi utilisé pour l'algorithme des SVM (voir partie 1.5 ci-dessous). Il consiste à envoyer préalablement les données X_j par une application $\phi : \mathcal{X} \mapsto \mathcal{H}$ (appelée *feature map*) dans un espace linéaire de grande dimension $\mathcal{H}$ muni d'un produit scalaire. Les nouveaux individus sont alors les vecteurs $\phi(X_j)$, $j = 1 \dots n$. La KPCA agit sur les $\phi(X_j)$ de la même façon que la PCA agissait sur les X_j. Ainsi, la KPCA correspond à une PCA dans un espace de grande dimension $\mathcal{H}$.

Les colonnes de la matrice T sont maintenant constituées des coordonnées des $\phi(X_j)$ et celles de la matrice $\widetilde{T}$ des coordonnées des $\widetilde{\phi(X_j)} = \phi(X_j) - \frac{1}{n}\sum_{j=1}^n \phi(X_j)$. La matrice de covariance à diagonaliser est

$$\widehat{\Sigma} = \frac{1}{n}\sum_{j=1}^n (\phi(X_j) - \overline{\phi})(\phi(X_j) - \overline{\phi})' , \tag{1.12}$$

où $\overline{\phi} = \frac{1}{n}\sum_{j=1}^n \phi(X_j)$ est le centre de gravité du nuage des individus envoyés dans $\mathcal{H}$. Par souci de clarté, l'espace $\mathcal{H}$ est supposé de dimension finie mais le principe reste le même s'il est de dimension infinie. Il suffit de considérer des opérateurs à la place des matrices.

La "feature map" ϕ n'est pas connue explicitement : on ne la connaît que par ses produits scalaires

$$\langle \phi(x_1), \phi(x_2) \rangle = k(x_1, x_2) , \tag{1.13}$$

où k est une fonction connue à l'avance et appelée *noyau*. Afin d'obtenir l'algorithme de la KPCA, il faut donc pouvoir calculer les vecteurs propres de $\widehat{\Sigma}$. Par extension du cas linéaire, ces vecteurs propres sont encore appelés *directions principales* : il est important de noter que ce sont des vecteurs de $\mathcal{H}$. Cette diagonalisation engendre deux difficultés d'ordre algorithmique. D'une part, $\widehat{\Sigma}$ est une matrice de grande dimension et sa diagonalisation peut donc être difficile. D'autre part, la matrice $\widehat{\Sigma}$ n'est pas connue explicitement puisqu'elle fait intervenir la feature map. Ces problèmes sont résolus en introduisant la *matrice noyau*.

Soit $\overline{K} = \frac{1}{n}\widetilde{T}'\widetilde{T}$. Un résultat classique d'algèbre linéaire stipule que $\overline{K}$ et $\widehat{\Sigma} = \frac{1}{n}\widetilde{T}\widetilde{T}'$ ont les mêmes valeurs propres non-nulles. De plus, si $\widehat{w}_i$ est un vecteur propre de $\overline{K}$, $\widehat{\gamma}_i = \widetilde{T}\widehat{w}_i$ est un vecteur propre de $\widehat{\Sigma}$. Si on considère des vecteurs propres orthogonaux de K, les vecteurs propres correspondants de $\widehat{\Sigma}$ sont eux aussi orthogonaux. Finalement, pour diagonaliser $\widehat{\Sigma}$, il suffit de diagonaliser la matrice symétrique $\overline{K}$ de taille n. La relation (1.6) implique $\overline{K} = \frac{1}{n}(I_n - \frac{1}{n}1\!\!1_{n,n})T'T(I_n - \frac{1}{n}1\!\!1_{n,n})$ où le $(i,j)^{\text{e}}$ coefficient de $K = T'T$ est

$$K_{i,j} = \langle \phi(X_i), \phi(X_j) \rangle = k(X_i, X_j)\,. \tag{1.14}$$

La *matrice noyau* (ou *matrice de Gram*) K est une matrice carrée de taille n accessible uniquement à partir du noyau et des données dont la diagonalisation permet accéder aux vecteurs propres de $\widehat{\Sigma}$. Précisément, on a :

$$\widehat{\gamma}_i = \sum_{j=1}^{n} \widehat{w}_i^j \widetilde{\phi(X_j)}\,, \tag{1.15}$$

où $\widehat{w}_i^j$ désigne la j^{e} coordonnée du vecteur $\widehat{w}_i$ et $\widehat{\gamma}_i$ est un vecteur propre de $\widehat{\Sigma}$ associé à la i^{e} plus grande valeur propre non-nulle de $\widehat{\Sigma}$. L'expression (1.15) fait encore intervenir la feature map. Cela ne pose pas de difficultés puisque qu'on peut exprimer la projection de n'importe quel élément de la forme $\phi(x) - \overline{\phi}$ sur la droite engendrée par $\widehat{\gamma}_i$ uniquement en fonction des valeurs de la fonction k grâce à la relation (1.13). Soit $f_i(x) = \langle \phi(x) - \overline{\phi}, \widehat{\gamma}_i \rangle$ la projection de l'image d'un point x quelconque de l'espace $\mathcal{X}$ sur $\widehat{\gamma}_i$. Contrairement à la PCA, les courbes de niveau de f_i dépendent non-linéairement de x.

La KPCA nécessite la donnée à priori d'une fonction k liée à une application ϕ par la relation (1.13). La section 1.5 donne une condition nécessaire et suffisante sur k assurant l'existence de la feature map ϕ.

Formellement, les propriétés de la PCA se généralisent à la KPCA en remplaçant les individus $\widetilde{X}_i$ par $\widetilde{\phi(X_i)}$. L'espace choisi par la KPCA est celui engendré par les d premières directions principales :

$$\widehat{V}_d = \langle \widehat{\gamma}_1, \ldots, \widehat{\gamma}_d \rangle\,, \tag{1.16}$$

où $\widehat{\gamma}_i$ désigne un vecteur propre associé à la i^{e} plus grande valeur propre non-nulle de l'opérateur $\widehat{\Sigma}$ défini par l'équation (1.12). La somme des grandes (resp petites) valeurs propres de la matrice noyau recentrée $\overline{K}$ permet de quantifier la quantité de variance prise en compte par l'espace $\widehat{V}_d$ (resp. l'erreur quadratique de reconstruction). En effet, les relations (1.10) et (1.9) se traduisent ici par:

$$\sum_{j=1}^{n} \|\Pi_{\widehat{V}_d}(\widetilde{\phi(X_j)})\|^2 = \sum_{\ell=1}^{d} \widehat{\lambda}_\ell\,, \tag{1.17}$$

et

$$\frac{1}{n}\sum_{j=1}^{n}\|\widetilde{\phi(X_j)} - \Pi_{\widehat{V}_d}(\widetilde{\phi(X_j)})\|^2 = \sum_{\ell=d+1}^{n}\widehat{\lambda}_\ell\,, \tag{1.18}$$

où $(\widehat{\lambda}_j)_{j\geq 1}$ désignent les valeurs propres de $\overline{K}$ rangées par ordre décroissant.

D'autre part, l'erreur quadratique de reconstruction associée à un espace vectoriel V est maintenant $\overline{R}_n(V) = \frac{1}{n}\sum_{j=1}^{n}\|\widetilde{\phi(X_j)} - \Pi_V(\widetilde{\phi(X_j)})\|^2$. Son erreur quadratique moyenne est $\overline{R}(V) = \mathbb{E}\left[\|\Pi_{V^\perp}(\phi(X) - \mathbb{E}[\phi(X)])\|^2\right]$. Les équations (1.8) et (1.11) impliquent que

$$\widehat{V}_d = \arg\min_{V\in\mathcal{V}_d}\overline{R}_n(V)\,,$$

et

$$V_d = \arg\min_{V\in\mathcal{V}_d}\overline{R}(V)\,,$$

où V_d est l'espace vectoriel engendré par les vecteurs propres associés aux d plus grande valeurs propres de la matrice $\Sigma = \mathbb{E}\left[(\phi(X) - \mathbb{E}[\phi(X)])(\phi(X) - \mathbb{E}[\phi(X)])'\right]$ et $\mathcal{V}_d$ désigne l'ensemble des sous espaces de dimension d de $\mathcal{H}$. Les études statistiques menées dans les chapitres 3 et 4 consistent en partie à étudier le comportement de l'espace $\widehat{V}_d$ choisi par la KPCA vis-à-vis de V_d.

1.3 Contributions concernant l'analyse en composantes principales à noyau.

Le chapitre 3 donne des contrôles non-asymptotiques sur les sommes des grandes et des petites valeurs propres des matrices de Gram K (dont chaque coefficient est défini par la relation (1.14)) et $\overline{K}$. Les études précédentes s'étaient focalisées sur les valeurs propres de K alors que l'algorithme utilisé en pratique repose sur $\overline{K}$. Cela revient à négliger le centrage, c'est-à-dire le fait que, afin d'étudier la variance, le nuage des individus soit préalablement recentré. Concernant les valeurs propres de K, des résultats avaient déjà été obtenus par [KG00] dans un cadre asymptotique. Les résultats non-asymptotiques obtenus dans le chapitre 3 améliorent l'étude de [STWCK05]. Comme indiqué par les relations (1.17) et (1.18), la somme des grandes (resp. des petites) valeurs propres de cette matrice s'interprète en KPCA comme étant la quantité de variance prise en compte (resp. l'erreur quadratique de reconstruction) de l'espace de dimension d choisi par la KPCA. Ainsi, ces résultats traitent de la stabilisation de l'erreur commise par la KPCA. Ils pourraient également servir à analyser d'autres algorithmes obtenus grâce à l'astuce du noyau comme par exemple les Support Vector Machines décrit dans la partie 1.5.

Les propriétés de généralisation de la KPCA sont explorées dans le chapitre 3 sous forme de bornes sur l'erreur quadratique moyenne. Dans le cas où on ne tient pas compte du centrage effectué en pratique, ce chapitre améliore les résultats de [STWCK05] en fournissant des vitesses de convergence rapides de l'erreur quadratique moyenne de la KPCA. Précisément, il montre que cette vitesse de convergence dépend de la vitesse de décroissance des valeurs propres et est typiquement meilleure que $1/\sqrt{n}$. De plus, il utilise un cadre fonctionnel permettant de traiter rigoureusement le cas où l'espace $\mathcal{H}$ est de dimension infinie. Il donne aussi des résultats originaux de convergence de l'erreur quadratique moyenne en tenant compte du centrage. Ainsi, le Théorème 3.4.4 montre que sous certaines hypothèses peu restrictives, pour tout $\xi \geq 1$, avec probabilité au moins $1 - e^{-\xi}$,

$$\overline{R}(\widehat{V}_d) - \overline{R}(V_d) \leq \square \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\xi}{n}} \right) ,$$

où $\square$ désigne une constante universelle. Ce résultat s'interprète en terme de stabilité de l'erreur commise par la KPCA : il assure la convergence de l'erreur quadratique moyenne à vitesse $1/\sqrt{n}$.

Le chapitre 4 fournit d'autres contributions originales concernant les propriétés statistiques de la KPCA en les étudiant avec un point de vue différent de celui du chapitre précédent. Au lieu de considérer l'erreur quadratique moyenne comme critère, il étudie directement l'erreur d'approximation entre les espaces propres $\widehat{V}_d$ et V_d. Par souci de clarté, supposons que les valeurs propres de la matrice Σ soient simples. Dans ce cas, le Théorème 4.3.2 stipule que l'erreur d'approximation ne dépend de d que via l'écart δ_d entre la d^{e} et la $(d+1)^{\text{e}}$valeur propre de Σ. Plus précisément, pour n assez grand et pour tout $\xi > 0$, avec probabilité au moins $1 - e^{-\xi}$,

$$\|\Pi_{\widehat{V}_d} - \Pi_{V_d}\| \leq \square \left(\frac{1}{\delta_d \sqrt{n}} + \frac{1}{\delta_d} \sqrt{\frac{\xi}{n}} \right) .$$

Ce résultat a une conséquence géométrique : pour tout vecteur de $\widehat{V}_d$, la tangente de son angle avec sa projection sur V_d est majorée par $\frac{c}{\delta_d}\sqrt{\frac{\xi}{n}}$. Cela peut s'interpréter comme une propriété de stabilité.

1.4 Sélection de modèle type Birgé et Massart.

Revenons maintenant au cadre mathématique général de la partie 1.1.2. On souhaite estimer la caractéristique s (définie par la relation (1.1)) de P qui se trouve dans un grand modèle S. La communauté informatique sait depuis longtemps qu'un algorithme qui explorerait tout l'ensemble S ne peut être performant à cause du phénomène d'overtitting. La taille de l'ensemble des fonctions considérées par l'algorithme doit donc être restreinte. Il existe beaucoup de notions pour mesurer cette taille : ce sont des mesures de *complexité* (ou *capacité*). D'un point de vue statistique, cela revient à considérer un *modèle* en argumentant que l'estimation est trop difficile à faire dans S. Dans les deux cas, la raison fondamentale réside dans la structure même du problème et, plus particulièrement, dans le fait qu'on ne possède qu'un nombre fini de données. On dispose donc de trop peu de données pour pouvoir travailler dans des modèles de taille importante.

Afin de se donner plus de liberté, on considère généralement une *collection* $(S_m)_{m \in \mathcal{M}}$ de modèles (des sous-ensembles de S) telle que la complexité de chaque S_m soit contrôlée. Le but de la sélection de modèle est de choisir le "meilleur" d'entre eux puis d'y faire l'estimation. Usuellement, chacun est représenté par un estimateur $\widehat{f}_m$ et le problème revient donc à choisir le "meilleur" estimateur à partir des données limitées dont on dispose. Naturellement, on souhaiterait sélectionner l'estimateur $\widehat{f}_{m^*}$ ayant le plus petit risque possible. La loi sous-jacente des données étant inconnue, cette quantité est inaccessible et on va chercher à définir une procédure de sélection de modèle ne dépendant que des données. Cette *procédure adaptative* de sélection doit fournir un modèle $\widehat{m}$ qui pourra se substituer efficacement au modèle idéal m^*. L'estimateur sélectionné $\widehat{f}_{\widehat{m}}$ devrait donc avoir un risque comparable à celui du modèle idéal. Cette condition est assurée s'il vérifie une *inégalité oracle* :

$$\mathbb{E}\left[L(\widehat{f}_{\widehat{m}}, s)\right] \leq c \inf_{m \in \mathcal{M}} \mathbb{E}\left[L(\widehat{f}_m, s)\right] . \tag{1.19}$$

De par la définition de la perte par l'équation (1.2), le risque de chaque estimateur se décompose sous la forme suivante :

$$\mathbb{E}\left[L(\widehat{f}_m, s)\right] = L(f_m, s) + \mathbb{E}\left[L(\widehat{f}_m, f_m)\right] .$$

où $f_m = \arg\min_{f \in S_m} \mathbb{E}\left[\gamma(f, (X,Y))\right]$ est la meilleure approximation de s dans le modèle S_m.

Le choix d'un modèle comporte donc deux sources d'erreur : la première est *l'erreur d'approximation* de s par f_m – $L(f_m, s)$ – et la seconde à *l'erreur d'estimation* de f_m par $\widehat{f}_m$ dans le modèle S_m – $\mathbb{E}\left[L(\widehat{f}_m, f_m)\right]$ –. L'erreur d'approximation est une quantité déterministe reflétant la distance du modèle S_m à la cible s. On peut aussi remarquer qu'elle dépend explicitement de s : son traitement nécessite donc généralement des hypothèses sur la cible. L'erreur d'estimation reflète quant à elle la difficulté à estimer dans le modèle S_m : elle dépend essentiellement de la complexité de ce modèle et de la quantité de données disponibles. Finalement, la procédure de sélection de modèle vise à réaliser le meilleur *compromis* entre ces deux sortes d'erreur : typiquement, plus le modèle choisi sera de complexité importante, plus l'erreur d'approximation sera petite mais plus l'erreur d'estimation sera importante. En particulier, même si on sait a priori que la cible s appartient à un modèle S^*, on peut avoir intérêt à en choisir un qui aura de bonnes propriétés d'approximation tout en ayant une complexité plus petite.

Le compromis idéal étant inconnu, il faut définir une procédure de sélection uniquement à partir des données et fournissant un modèle aux propriétés proches de celui réalisant le meilleur compromis entre les erreurs d'approximation et d'estimation.

Au regard de la propriété (1.1) de la cible s que l'on cherche à estimer, on considère maintenant le cas où l'estimateur représentant le modèle est obtenu par la méthode de *minimisation empirique du risque* :

$$\widehat{f}_m = \arg\min_{f \in S_m} \gamma_n(f) ,$$

où $\gamma_n(f) = \frac{1}{n}\sum_{i=1}^{n} \gamma(f, (X_i, Y_i))$. La procédure de *minimisation pénalisée de la perte empirique* consiste alors à choisir le modèle par le critère

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} (\gamma_n(\widehat{f}_m) + \mathrm{pen}(m)) . \tag{1.20}$$

où *pen* est une fonction à choisir convenablement. Elle est appelée *pénalité* . L'estimateur final de s est $\widehat{f}_{\widehat{m}}$. Les propriétés statistiques générales de ce type d'estimateur ont été étudiées par Birgé et Massart dans [BM98, BBM99, BM97]. Le premier terme est connu et ne dépend que de l'observation: la procédure de sélection de modèle est donc maintenant entièrement déterminée par la fonction de pénalité *pen*. Celle-ci contrôle la complexité du modèle choisi: d'un point de vue apprentissage, elle évite l'overfitting c'est-à-dire l'obtention d'un estimateur s'ajustant trop sur les données d'apprentissage au détriment de ses performances en généralisation.

Afin d'obtenir un bon estimateur final, le représentant $\widehat{f}_m$ doit être approximativement minmax sur le modèle S_m, c'est-à-dire être optimal si la cible s appartient au modèle S_m. Dans ce cas, la pénalité $pen(m)$ doit être de l'ordre du risque minmax sur S_m.

La problématique de la classification est introduite à la page 9. La partie suivante décrit un algorithme de classification renommé : les Support Vector Machines (SVM). Cet algorithme sera ensuite décrit comme une procédure de *régularisation* qui s'interprète dans le cadre de la sélection de modèle de Birgé et Massart.

1.5 SVM et régularisation.

La classification est introduite à la page 9. D'un point de vue statistique, elle vise à d'obtenir une fonction de classification qui soit la meilleure *estimation* possible de la fonction de classification de Bayes à partir de l'échantillon d'apprentissage $((X_1, Y_1), \dots, (X_n, Y_n)) \in (\mathcal{X} \times \{-1, +1\})^n$. L'évaluation en un point x appartenant à $\mathcal{X}$ de cette fonction de classification donne une prédiction de l'étiquette de x. Obtenir une telle fonction revient donc à scinder l'espace $\mathcal{X}$ des entrées en deux régions : l'une pour les variables d'entrées associées à l'étiquette $+1$ et l'autre pour celles associées à l'étiquette -1. Supposons pour le moment que $\mathcal{X} = \mathbb{R}^p$. Il est naturel de considérer des régions *linéairement* séparées. Ainsi, la fonction de classification est cherchée sous la forme $f_{w,b}(x) = \text{signe}(\langle w, x\rangle + b)$ où $\text{sign}(a) = 2\mathbb{1}_{a \geq 0} - 1$. Les régions de $\mathcal{X}$ sont définies par le signe de $\langle w, x\rangle + b$ et elles sont séparées par l'hyperplan $\langle w, x\rangle + b = 0$.

Il est maintenant naturel de déterminer le vecteur normal w et l'ordonnée à l'origine b de l'hyperplan séparateur, en minimisant le nombre d'erreurs commises sur l'échantillon d'entraînement :

$$(\widehat{w}, \widehat{b}) = \arg \min_{\|w\|=1, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\text{signe}(\langle w, X_i\rangle + b) \neq Y_i} . \tag{1.21}$$

La fonction de classification finale sera $\widehat{f}(x) = \text{signe}(\langle \widehat{w}, x\rangle + \widehat{b})$.

Cependant, cette méthode d'obtention d'une fonction de classification n'est pas satisfaisante pour plusieurs raisons. Tout d'abord, une solution de ce problème d'optimisation n'est pas nécessairement unique. De plus, sauf situation particulière, elle n'est pas accessible en temps polynomial. Ce problème d'optimisation fait partie d'une classe de problèmes connus en informatique : il est NP-complet. La partie suivante explique brièvement une solution, proposée en partie par Vapnik. L'algorithme des Support Vector Machines (SVM) correspondant permet un accès automatisé à une fonction de classification.

1.5.1 Algorithme SVM.

La théorie de Vapnik a connu un grand succès grâce aux bonnes performances des algorithmes d'apprentissage qui en découlent et en particulier les SVM (présentés dans [CV95] et [Vap95]). Cependant, même si ces algorithmes justifient en partie la pertinence de l'approche Vapnik pour l'apprentissage, ils ne constituent pas une validation empirique de ses résultats fondamentaux.

Ce paragraphe est très inspiré de la présentation des SVM de [CST00]. Initialement, la solution proposée par Vapnik pour échapper aux problèmes inhérents de la minimisation empirique du risque associé à la fonction de perte dure (équation (1.21)) repose sur le concept de *marge* : la marge d'un hyperplan est définie comme étant sa distance à la donnée X_i la plus proche. Le fait de choisir un hyperplan qui ait une *marge maximale* permet de régler le problème d'unicité tout en accédant à une frontière de classification robuste : intuitivement, on peut penser que la fonction de classification correspondante aura de bonnes propriétés de généralisation. De plus, en considérant une normalisation adéquate du vecteur normal de l'hyperplan, on constate que la marge d'un hyperplan est inversement proportionnelle à la norme euclidienne de son vecteur normal.

L'idée de base de l'algorithme SVM repose donc sur une intuition géométrique et propose de choisir l'hyperplan correspondant à $(\widehat{w}, \widehat{b})$ obtenu en résolvant le problème d'optimisation quadratique suivant :

$$\begin{array}{cc} \text{minimiser en } w,\, b & \|w\|^2\,, \\ \text{sous} & Y_i(\langle w, X_i\rangle + b) \geq 1\,,\ i = 1\ldots n\,. \end{array}$$

Il signifie que l'on cherche l'hyperplan de marge maximale séparant correctement les données. Dans le but d'étendre cette idée à des problèmes de classification non linéairement séparables, les contraintes peuvent être relâchées de façon douce en introduisant des "slack variables" $\xi_i\,,\ i = 1\ldots n$ qui permettent aux contraintes de marge d'être violées. On obtient alors le problème d'optimisation suivant:

$$\begin{array}{rc} \text{minimiser en } w,\, b,\, \xi_i & \|w\|^2 + \lambda\sum_{i=1}^n \xi_i\,, \\ \text{sous} & Y_i(\langle w, X_i\rangle + b) \geq 1 - \xi_i\,,\ i = 1\ldots n\,, \\ & \xi_i \geq 0\,. \end{array} \tag{1.22}$$

Afin de pouvoir résoudre ce problème d'optimisation sous contraintes, on l'écrit sous sa forme duale où les multiplicateurs de Lagrange sont notés $\alpha_i\,, i = 1\ldots n$:

$$\begin{array}{cc} \text{maximiser en } \alpha_i & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n Y_iY_j\alpha_i\alpha_j\langle X_i, X_j\rangle\,, \\ \text{sous} & \lambda \geq \alpha_i \geq 0\,,\ i = 1\ldots n\,, \\ & \sum_{i=1}^n \alpha_i Y_i = 0\,. \end{array}$$

La fonction de classification finale sera $\widehat{f}(x) = \text{signe}(\sum_{i=1}^n Y_i\widehat{\alpha}_i\langle X_i, x\rangle + \widehat{b})$ où $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n$ sont les solutions du problème d'optimisation précédent et $\widehat{b}$ est choisi séparément.

Pour pouvoir considérer des frontières de classification non-linéaires, l'astuce du noyau (explicitée dans la partie 1.2.4) est maintenant utilisable puisque les données n'interviennent que par l'intermédiaire d'un produit scalaire (voir équation (1.13)). On obtient alors la formulation de l'algorithme des *SVM soft margin* :

$$\begin{array}{cc} \text{maximiser en } \alpha_i & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n Y_iY_j\alpha_i\alpha_j k(X_i, X_j)\,, \\ \text{sous} & \lambda \geq \alpha_i \geq 0\,,\ i = 1\ldots n\,, \\ & \sum_{i=1}^n \alpha_i Y_i = 0\,. \end{array}$$

La fonction de classification finale sera

$$\widehat{f}(x) = \text{signe}\left(\sum_{i=1}^n Y_i\alpha_i^* k(X_i, x) + b^*\right)\,. \tag{1.23}$$

Seuls les X_i pour lesquels le multiplicateur de Lagrange α_i est non-nul participent à la solution: ce sont les *vecteurs supports* qui donnent leur nom aux Support Vector Machines.

Cette algorithme comporte deux paramètres libres (à ajuster par l'utilisateur) : la constante λ et le noyau k. Leurs rôles respectifs seront éclairés par la formulation des SVM comme une procédure de régularisation donnée à la fin de ce paragraphe.

Comme expliqué dans la partie 1.2.4, le fait d'utiliser un noyau s'interprète de la façon suivante : les données sont envoyées dans un espace de grande dimension $\mathcal{H}$ par une *feature map* ϕ. Ensuite, une SVM linéaire est appliquée sur ces fonctions. La séparation linéaire dans l'espace de grande dimension correspond à une séparation non-linéaire dans l'espace de départ $\mathcal{X}$.

Finalement, l'idée proposée par Vapnik repose sur le fait de chercher une fonction à valeurs réelles adaptée aux données puis d'en prendre le signe pour obtenir une fonction de classification. La suite de cette partie vise à interpréter l'algorithme des SVM comme une procédure de régularisation. Pour ce faire, il est important de connaître l'utilité d'une telle procédure. Le moyen le plus simple pour expliquer son principe de fonctionnement est de revenir à ses origines : l'utilisation des splines par [Wah90] pour des problèmes de régression que nous décrirons ici uniquement en dimension 1. Ce procédé sera ensuite étendu en dimension plus grande grâce à la régularisation type Tikhonov.

1.5.2 Régularisation.

La considération des splines vient de l'analyse unidimensionnelle. Soit W_m l'espace de Sobolev d'ordre m sur $[0,1]$: il est constitué des fonctions $m-1$ fois continûment différentiables et de dérivée m^e de carré intégrable. Schoenberg ([Sch64b],[Sch64a]) a montré que le minimiseur naturel de $\int_0^1 (f^{(m)}(x))^2 dx$ parmi les fonctions de W_m satisfaisant (i) $f(X_i) = Y_i\,, i = 1 \ldots n$ est, si $n \geq m$, l'unique spline de W_m satisfaisant la contrainte (i). Un spline est un polynôme par morceaux satisfaisant certaines propriétés de régularité. Les statisticiens ne cherchent pas à interpoler les données : celles-ci peuvent contenir du bruit et ils veulent obtenir une fonction qui possède de bonnes propriétés de généralisation. Ils cherchent donc une fonction *régulière* qui s'ajuste le plus possible aux données. C'est pour cette raison que le fructueux travail de Wahba [Wah90] s'intéresse au problème de trouver f dans W_m qui minimise:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + C\int_0^1 (f^{(m)}(x))^2 dx\,. \tag{1.24}$$

On voit clairement sur cet exemple que l'ajout du terme $\int_0^1 (f^{(m)}(x))^2 dx$ assure la régularité du minimiseur : la constante de régularisation C contrôle le compromis entre la fidélité aux données et la régularité de la fonction obtenue. Schoenberg a montré que la solution de ce problème était encore un spline. Les splines sont aussi très utilisés en analyse numérique car ils possèdent de bonnes propriétés d'approximation et sont facilement utilisables par des ordinateurs.

Afin de généraliser cette approche en grandes dimensions, on définit la notion d'espace auto-reproduisant sur laquelle repose la régularisation type Tikhonov.

Définition 1. *Un espace de Hilbert auto-reproduisant $\mathcal{H}$ (RKHS, en abrégé) est un espace de Hilbert de fonctions ($\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$) pour lequel il existe une fonction $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ (appelée noyau auto-reproduisant) satisfaisant*

- *$\mathcal{H}$ contient toutes les fonctions $k(x,\cdot)$ pour $x \in \mathcal{X}$,*

- *la propriété de reproduction est satisfaite,*

$$\forall f \in \mathcal{H},\, \forall x \in \mathcal{X},\, \langle f, k(x,\cdot)\rangle_{\mathcal{H}} = f(x)\,. \tag{1.25}$$

Une fonction $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ est un *noyau défini positif* si elle est symétrique ($k(x,x') = k(x',x)\,,\, \forall x, x' \in \mathcal{X}$) et si toutes les matrices de Gram qu'elle engendre sont positives. Précisément, cela signifie que

$$\forall N \geq 1,\, \forall (x_1, \dots, x_N) \in \mathcal{X}^N,\, \forall (a_1, \dots, a_N) \in \mathbb{R}^N\,,\, \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0\,.$$

Un noyau auto-reproduisant est clairement défini positif et un résultat de [Aro50] stipule que cette condition est suffisante : à tout noyau défini positif correspond un *unique* RKHS. Ce résultat est frappant puisqu'il ne suppose *aucune* condition sur l'espace $\mathcal{X}$. En particulier, l'utilisation de l'astuce du noyau qui a mené aux algorithme SVM et de la KPCA peut s'interpréter comme une procédure d'envoi des données dans un RKHS si et seulement si elle utilise un noyau défini positif.

D'un point de vue statistique, le choix du noyau peut s'interpréter comme un a priori qu'on met sur les données : un des problèmes cruciaux de l'apprentissage réside dans la façon de le choisir.

Dans la preuve du résultat de [Aro50], le RKHS est construit comme une complétion de l'espace vectoriel $\mathcal{H}_0$ engendré par l'ensemble des fonctions $\{k(x,.)\,,\, x \in \mathcal{X}\}$. La norme considérée sur l'espace $\mathcal{H}_0$ est

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

On obtient alors la représentation sous forme *primale* du RKHS où la forme typique des fonctions est une combinaison linéaire de noyaux centrés en différents points.

Afin de se former une intuition sur ce que peut représenter un RKHS, quelques exemples plus explicites sont maintenant donnés. Le cas le plus connu est celui correspondant aux noyaux de *Mercer* : $\mathcal{X}$ est supposé être un espace compact et le noyau k est un noyau défini positif continu en tant que fonction définie sur $\mathcal{X} \times \mathcal{X}$. On rappelle que P désigne la loi de la variable d'entrée X. Dans ce cas, l'opérateur intégral à noyau,

$$\begin{aligned} T_k :\; L_2(P) &\mapsto\; L_2(P) \\ f \;&\mapsto\; \int_{\mathcal{X}} k(x,\cdot) f(x) dP(x)\,, \end{aligned} \tag{1.26}$$

définit un opérateur linéaire auto-adjoint compact. Il est diagonalisable dans une base orthonormée notée $(\phi_i)_{i \geq 1}$. Les valeurs propres associées sont notées $(\lambda_i)_{i \geq 1}$ et sont supposées être rangées par ordre décroissant. Dans ce cas, le Théorème de Mercer stipule que le noyau peut se décomposer sous la forme suivante :

$$k(x,x') = \sum_{i \geq 1} \lambda_i \phi_i(x) \phi_i(x')\,,$$

où la série est uniformément convergente sur $\mathcal{X} \times \mathcal{X}$. De plus, le RKHS associé au noyau k se met sous la forme

$$\mathcal{H} = \left\{ f \in L_2(P),\ f = \sum_{i \geq 1} \langle f, \phi_i \rangle \phi_i\,,\ \|f\|_{\mathcal{H}}^2 = \sum_{i \geq 1} \langle f, \phi_i \rangle^2 / \lambda_i < \infty \right\} . \tag{1.27}$$

On obtient la forme *duale* du RKHS : elle permet de constater que, dans ce cas, c'est un sous-espace de $L_2(P)$ constitué des fonctions ayant des coefficients – dans la base des fonctions propres – à décroissance suffisamment rapide par rapport aux valeurs propres. Afin de voir clairement le lien avec les problèmes de régularité formulés plus haut, nous traitons un exemple spécifique en dimension 1. Soit $\mathcal{X} = [0, 1]$ et k le noyau de Mercer suivant :

$$k(x, y) = 1 + 2 \sum_{j \geq 1} \lambda_j \left(\cos(2\pi j x) \cos(2\pi j y) + \sin(2\pi j x) \sin(2\pi j y) \right) ,$$

où $\lambda_j = (2\pi j)^{-2m}$, $m \in \mathbb{N}^*$. La base de $L_2([0, 1])$ considérée ici étant la base trigonométrique, les coefficients d'une fonction dans cette base sont donnés par ses coefficients de Fourier notés $f_j^c = \sqrt{2} \int_0^1 f(t) \cos(2\pi j t) dt$ et $f_j^s = \sqrt{2} \int_0^1 f(t) \sin(2\pi j t) dt$. Le RKHS associé à ce noyau est l'espace de Sobolev périodisé d'ordre m, c'est-à-dire l'ensemble des fonctions de $L_2([0, 1])$ qui sont $m - 1$ fois continûment différentiables, de dérivée m^{e}appartenant à $L_2([0, 1])$ et vérifiant les conditions au bord $f(0) = f(1), \ldots, f^{(m-1)}(0) = f^{(m-1)}(0)$. Cela peut se voir facilement puisque la norme du RKHS associée à k définie dans (1.27) satisfait

$$\|f\|_{\mathcal{H}}^2 = \sum_{j \geq 1} (2\pi j)^{2m} \left((f_j^c)^2 + (f_j^s)^2 \right) + \left(\int_0^1 f(t) dt \right)^2 = \int_0^1 |f^{(m)}(t)|^2 dt + \left(\int_0^1 f(t) dt \right)^2 .$$

où la deuxième égalité vient de l'identité de Parseval. La dernière quantité est par définition le carré de la norme $\|f\|_{\text{Sob}}$ dans l'espace de Sobolev.

L'inconvénient majeur du Théorème de Mercer est qu'il restreint l'espace $\mathcal{X}$ à être compact. Dans le cas où c'est un espace vectoriel de dimension finie ($\mathcal{X} = \mathbb{R}^p$), un des noyaux les plus utilisés est le *noyau gaussien* :

$$k(x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} = K(x - y) .$$

avec $K(x) = e^{\frac{-\|x\|^2}{2\sigma^2}}$. Dans ce cas, le RKHS associé se caractérise par une décroissance rapide de la transformée de Fourier $\widehat{f}$ de f :

$$\mathcal{H}_\sigma = \left\{ f \in C_0(\mathbb{R}^d)\ :\ f \in L_1(\mathbb{R}^d)\ et\ \int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{\widehat{K}(\omega)} d\omega < \infty \right\} . \tag{1.28}$$

avec $\widehat{K}(\omega) = (2\pi\sigma^2)^{d/2} e^{\frac{-\sigma^2 \|\omega\|^2}{2}}$ et $C_0(\mathbb{R}^d)$ est l'espace des fonctions continues nulles à l'infini.

Cette représentation intégrale du RKHS associé au noyau gaussien éclaire le rôle de régularisation joué par σ : plus il est élevé, plus les fonctions du RKHS sont régulières. Précisément, $\sigma < \sigma'$ implique $\mathcal{H}_{\sigma'} \subset \mathcal{H}_\sigma$.

En conclusion, la norme RKHS d'une fonction peut être vue comme une mesure de sa régularité.

1.5.3 Approche statistique pour la SVM.

[SS98] et [EPP00] ont remarqué que le problème d'optimisation définissant $\widehat{f}$ dans (1.23) pouvait se réécrire comme une procédure de régularisation type Tikhonov. En effet, grâce à la vision fonctionnelle du RKHS précédemment décrite, on a $\widehat{f} = \text{signe}(\widehat{g})$ où $\text{sign}(a) = 2\mathbb{1}_{a\geq 0} - 1$ et

$$\widehat{g} = \arg\min_{f \in \mathcal{H}^b} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + C_n \|f\|_{\mathcal{H}}^2 \right) , \tag{1.29}$$

où $\mathcal{H}^b = \{f(x) + b,\ f \in \mathcal{H},\ b \in \mathbb{R}\}$, $\mathcal{H}$ est le RKHS associé au noyau k et $C_n = \frac{1}{n\lambda}$ (λ apparaît dans la partie 1.5.1). La régularisation type Tikhonov se caractérise par l'ajout du carré de la norme dans le RKHS. Cette formulation n'est pas celle utilisée en pratique pour programmer la SVM mais elle permet de prendre conscience du procédé de régularisation utilisé dans cet algorithme. Précisément, il est de la même forme que (1.24) : simplement, le risque empirique associé à la perte des moindres carrés est remplacé par celui associé à la perte dite *hinge loss* $\gamma(f,(x,y)) = (1 - yf(x))_+$. Cette dernière est un majorant convexe de la fonction de perte dite *hard loss* $(x, y) \mapsto \mathbb{1}_{f(x)\neq y}$. Le terme de régularisation est plus général dans le sens où $\mathcal{X}$ peut être un espace de grande dimension et pas uniquement $[0, 1]$. Le calibrage de la constante de régularisation C_n contrôle le compromis entre la fidélité aux données et la régularité de la fonction obtenue.

De plus, cette formulation permet d'interpréter l'algorithme SVM comme une procédure de pénalisation du *risque convexifié* (voir [BJM03] pour une étude générale). Il entre ainsi dans le cadre de la sélection de modèles de Birgé et Massart de la section 1.4. En effet, si on considère les boules de $\mathcal{H}^b$,

$$\mathcal{E}(R) = \{g \in \mathcal{H}^b,\ \|g\| \leq R\},$$

$\widehat{g}$ défini par l'équation (1.29) satisfait clairement $\widehat{g} = \widehat{g}_{\widehat{R}}$ où

$$\widehat{g}_R = \arg\min_{g \in \mathcal{E}(R)} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ , \tag{1.30}$$

et

$$\widehat{R} = \arg\min_{R \geq 0} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widehat{g}_R(X_i))_+ + C_n R^2 \right) .$$

Ce choix de R par minimisation du risque empirique pénalisé est de la forme de l'équation (1.20) où la fonction de pénalité *pen* est

$$\text{pen}(R) = C_n R^2 . \tag{1.31}$$

Cette pénalité représente la régularisation : elle impose un choix de boule de complexité contrôlée. De plus, si k est un noyau de Mercer, les boules du RKHS sont des ellipsoïdes de $L_2(P)$ homothétiques les uns des autres par la relation (1.27) :

$$\{g, \|g\|_{\mathcal{H}} \leq R\} = \left\{ \sum_{i\geq 1} a_i \phi_i \, ; \, \sum_{i \geq 1} \frac{a_i^2}{\lambda_i} \leq R^2 \right\} .$$

Ils ont tous comme directions principales les fonctions propres de l'opérateur intégral T_k défini par l'équation (1.26). Ainsi, les SVM constituent une procédure de sélection d'ellipsoïde.

On peut se demander si l'ordre de grandeur de la pénalité en R^2 est statistiquement convenable. En effet, il est choisi pour des raisons purement algorithmiques. [BBM04] a montré que, d'un point de vue statistique, sous certaines conditions, le carré de la norme est trop important et qu'il faudrait plutôt ajouter la norme elle-même.

La contribution principale de la présente thèse en terme de classification repose sur l'utilisation d'un autre procédé de régularisation : la projection fini-dimensionnelle. Ce procédé est étudié avec un point de vue théorique à travers des résultats de sélection de modèle et avec un point de vue pratique par la conception d'un nouvel algorithme de classification: la Kernel Projection Machine.

1.6 Contributions concernant la classification.

On rappelle que la fonction que l'on cherche à estimer est la fonction de classification de Bayes f^* définie par

$$f^*(x) = \begin{cases} 1 & si\, \eta(x) \geq \frac{1}{2}\,, \\ -1 & si\, \eta(x) < \frac{1}{2}\,, \end{cases}$$

où $\eta(x) = \mathbb{P}[Y = 1|X = x]$. On dit qu'on est sous l'hypothèse de marge s'il existe $h > 0$ tel que

$$\forall x \in \mathcal{X}\,, |2\eta(x) - 1| \geq h\,.$$

1.6.1 Projection fini-dimensionnelle.

Au regard des travaux de statistique de Birgé et Massart [BM97, BBM99, BM01], on peut penser à la projection fini-dimensionnelle comme étant une alternative à la régularisation de Tikhonov. Au lieu d'être minimisé sur des ellipsoïdes comme c'est le cas pour les SVM (voir équation (1.30)), le risque empirique est minimisé sur des espaces vectoriels

$$S_D = \langle \phi_1, \dots, \phi_D \rangle\,, \tag{1.32}$$

où $\phi_1, \dots, \phi_D$ sont les vecteurs propres de l'opérateur intégral T_k (défini par l'équation (1.26)) associés à ses D plus grandes valeurs propres. Ainsi, à chaque dimension, on associe l'estimateur

$$\widehat{g}_D = \arg\min_{g \in S_D} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+\,. \tag{1.33}$$

La régularité de $\widehat{g}_D$ dépend de son nombre de coefficients non-nuls dans la base $(\phi_i)_{i\geq 1}$. On considère maintenant la collection de modèles $\{S_D\}_{D\geq 1}$. Suivant le critère général (1.20), la dimension est choisie par minimisation pénalisée de la perte empirique :

$$\widehat{D} = \arg\min_{D\geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widehat{g}_D(X_i))_+ + \mathrm{pen}(D) \right)\,.$$

La fonction de pénalité *pen* assure la régularité de la fonction $\widehat{g}_{\widehat{D}}$ en contrôlant la dimension de l'espace vectoriel choisi. Contrairement à la SVM, l'ordre de grandeur de la pénalité sera choisi par des critères statistiques. La fonction de classification finale sera $\widehat{f} = \mathrm{signe}(\widehat{g}_{\widehat{D}})$.

La détermination de l'ordre de grandeur de la pénalité dans ce cadre comporte des difficultés liées à la non-compacité des espaces vectoriels. Ce problème est abordé dans le chapitre 5. Dans le chapitre 6, l'estimateur associé à chaque dimension est

$$\widetilde{g}_D(x) = \begin{cases} 1 & \text{si } \widehat{g}_D(x) \geq 1\,, \\ \widehat{g}_D(x) & \text{si } -1 < \widehat{g}_D(x) < 1\,, \\ -1 & \text{si } \widehat{g}_D(x) \leq -1\,. \end{cases}$$

Ce procédé n'induit pas de perte de généralité dans le sens où $\widetilde{g}_D$ et $\widehat{g}_D$ fournissent la même fonction de classification (puisq'ils ont le même signe). La dimension est maintenant choisie par minimisation pénalisée de la perte empirique tronquée :

$$\widetilde{D} = \arg\min_{D \geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{g}_D(X_i))_+ + \mathrm{pen}(D) \right) .$$

La fonction de classification finale est $\widehat{f} = \mathrm{signe}(\widetilde{g}_{\widetilde{D}})$.

Sous l'hypothèse de marge, le Théorème 6.2.1 stipule que, à un facteur logarithmique près, si on choisit une pénalité linéaire en la dimension

$$\mathrm{pen}(D) = \square \frac{D}{nh}\,, \tag{1.34}$$

alors la fonction de classification correspondante satisfait l'inégalité de type oracle suivante

$$\mathbb{E}L(\widetilde{g}_{\widetilde{D}}, f^*) \leq \square \inf_{D \geq 1} \big(\inf_{g \in S_D} L(g, f^*) + \frac{D}{nh} \big)\,, \tag{1.35}$$

où L est l'excès de risque associé à la hinge loss $L(g, f^*) = \mathbb{E}\left[(1 - Yg(X))_+ - (1 - Yf^*(X))_+\right]$ et $\square$ désigne une constante universelle. Ce résultat signifie qu'une pénalité linéaire en la dimension est statistiquement convenable.

L'algorithme de la Kernel Projection Machine (KPM) permet de tester concrètement la pertinence de la régularisation par projection fini-dimensionnelle. Ses performances seront comparées à celles de la SVM utilisant la régularisation type Tikhonov dans le chapitre 5. La partie suivante décrit les idées utilisées dans l'algorithme de la KPM ainsi que ses liens avec la KPCA.

1.6.2 Un nouvel algorithme de classification : la Kernel Projection Machine.

Puisque l'opérateur T_k dépend de la loi sous-jacente des données, ses fonctions propres sont inconnues et le minimiseur empirique $\widehat{g}_D$ donné par l'équation (1.33) n'est pas directement calculable. En pratique, l'algorithme de la KPM exploitera donc une version approchée définie à partir de la matrice de Gram introduite par l'équation (1.14). K_n désigne la matrice de Gram normalisée. C'est une matrice carrée de taille n ayant pour coefficient :

$$(K_n)_{i,j} = \frac{1}{n} k(X_i, X_j)\,.$$

L'*approximation de Nyström* stipule que les vecteurs propres de la matrice de Gram normalisée approchent les vecteurs propres de l'opérateur intégral T_k ([Bak77]). [Kol98] exploite le fait que la matrice K_n soit la partie empirique de l'opérateur intégral T_k et étudie la convergence (en un certain sens) des vecteurs propres de K_n vers les vecteurs propres de T_k. Le fait que la matrice de Gram soit la version empirique de l'opérateur intégral T_k peut se voir en quelques lignes. En effet, si $f \in L_2(P)$ est une fonction propre de T_k, elle satisfait

$$\forall x' \in \mathcal{X}\,, \int_{\mathcal{X}} f(x)k(x,x')dP(x) = \lambda f(x')\,.$$

Si on remplace formellement la probabilité P par la mesure empirique $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, on obtient

$$\frac{1}{n}\sum_{i=1}^n f(X_i)k(X_i,x') \sim \widehat{\lambda} f(x')\,.$$

Puis, en choisissant successivement $X_1, \ldots, X_n$ pour valeurs de x', on conclut que la matrice de Gram normalisée admet $\widetilde{f} = (f(X_1), \ldots, f(X_n))$ comme vecteur propre :

$$K_n \widetilde{f} \sim \widehat{\lambda}\widetilde{f}\,,$$

où $\widehat{\lambda}$ est une valeur propre de K_n.

De par la formulation (1.33) du minimiseur empirique, on constate que les fonctions de S_D n'interviennent que par les valeurs qu'elles prennent sur les points $X_1, \ldots, X_n$ de l'échantillon d'apprentissage. L'approximation précédemment décrite permet donc d'obtenir une version approchée $\widehat{\widehat{g}}_D$ de $\widehat{g}_D$. D'un point de vue algorithmique, $\widehat{\widehat{g}}_D$ est accessible en résolvant un problème de programmation linéaire. La dernière étape de l'algorithme est une phase de sélection de modèle: la dimension peut être sélectionnée par minimisation pénalisée de la perte empirique tronquée :

$$\widetilde{\widetilde{D}} = \arg\min_{D \geq 1} \left(\frac{1}{n}\sum_{i=1}^n (1 - Y_i \widetilde{\widetilde{g}}_D(X_i))_+ + \mu D \right)\,, \tag{1.36}$$

où μ est une constante de pénalisation à choisir et

$$\widetilde{\widetilde{g}}_D(x) = \begin{cases} 1 & \text{si } \widehat{\widehat{g}}_D(x) \geq 1\,, \\ \widehat{\widehat{g}}_D(x) & \text{si } -1 < \widehat{\widehat{g}}_D(x) < 1\,, \\ -1 & \text{si } \widehat{\widehat{g}}_D(x) \leq -1\,. \end{cases}$$

Le choix d'une pénalité μD linéaire en la dimension est guidé par le Théorème 6.2.1 rappelé au travers des inégalités (1.34) et (1.35). La fonction de classification finale sera $\widehat{f}(x) = \text{signe}(\widetilde{\widetilde{g}}_{\widetilde{\widetilde{D}}}(x))$.

Le chapitre 6 montrera clairement le lien entre la KPM et la KPCA. D'après la partie 1.2.4, pour chaque dimension D, l'espace choisi par la KPCA est $\widehat{V}_D$ (défini par l'équation (1.16)). C'est le sous-espace du RKHS engendré par les D premières fonctions propres de l'opérateur de covariance $\widehat{\Sigma}$ (défini par l'équation (1.12)). En utilisant la représentation fonctionnelle du RKHS donnée dans la partie 1.5.2, on montre aisément que l'estimateur $\widehat{\widehat{g}}_D$ correspond au minimiseur empirique sur $\widehat{V}_D$:

$$\widehat{\widehat{g}}_D = \arg\min_{g \in \widehat{V}_D} \frac{1}{n}\sum_{i=1}^n (1 - Y_i g(X_i))_+\,.$$

Le chapitre 4 étudie la stabilisation de l'espace aléatoire $\widehat{V}_D$ autour de l'espace déterministe S_D (défini par l'équation 1.32). Les résultats y sont interprétés dans le cadre de la KPCA. Ils sont aussi motivés par l'étude du comportement statistique de $\widehat{\widehat{g}}_D$. Cela demeure cependant un problème ouvert.

Finalement, l'algorithme de la KPM repose sur le fait suivant.

> *La régularisation peut être obtenue par la sélection d'un espace vectoriel via une méthode de réduction de la dimension telle que la KPCA.*

Comme précisé dans le problème ouvert à la page 13 de la partie 1.2, l'obtention d'un critère d'arrêt en KPCA ayant un sens statistique et ne prenant en compte que les variables d'entrées reste un problème ouvert. Cependant, la dimension choisie par la KPM peut s'interpréter comme une dimension d'arrêt pour la KPCA en tenant compte des étiquettes.

1.7 Limites et perspectives des résultats de classification.

Cette partie propose des prolongements aux contributions obtenues en classification en vue d'améliorer les liens entre la théorie et la pratique. D'une part, le Théorème 6.2.1, rappelé au travers des équations (1.34) et (1.35), ne justifie que *partiellement* la pénalité linéaire utilisée dans l'algorithme de la KPM. D'autre part, l'étude de l'influence du calibrage des constantes sur la performance des algorithmes reste à mener.

1.7.1 Bornes de risque.

La borne de type oracle (1.35) représente un premier pas vers l'analyse des performances de la KPM. Cependant, cette approche théorique ne prend pas pleinement en compte la spécificité des modèles utilisés pour l'algorithme de la KPM. De plus, la pénalité obtenue dépend du paramètre de marge h qui est inconnu.

- Les espaces vectoriels possèdent-ils de meilleures capacités d'approximation que les ellipsoïdes dans le cadre de la classification ? Afin de répondre à cette question, une des possibilités consiste à contrôler le terme d'erreur d'approximation $\inf_{g \in S_D} L(g, f^*)$ de l'inégalité (1.35). Puis, à comparer la vitesse de convergence de la fonction de classification obtenue par projection fini-dimensionnelle à celle obtenue par régularisation type Tikhonov. De façon générale, la théorie de l'approximation pour l'apprentissage n'en est qu'à ses débuts. Les travaux de [SS03] et [VV05] sont, en ce sens, précurseurs. La qualité d'approximation de la base des fonctions propres de l'opérateur intégral (déterminé par le noyau utilisé) est un élément déterminant dans le contrôle de l'erreur d'approximation mais la littérature comporte très peu de résultats sur ces fonctions propres.

- Puisque la pénalité (1.34) dépend de la marge h, la procédure de pénalisation proposée par l'analyse théorique n'est pas adaptative à la marge. À ce jour, une telle procédure de classification théoriquement justifiée n'existe pas. En pratique, cela pose d'importants problèmes pour le calibrage des constantes. Nous reviendrons sur ce point dans la partie suivante.

- Les modèles S_D considérés pour justifier la pénalité linéaire sont déterministes alors que ceux ($\widehat{V}_D$) exploités par l'algorithme sont aléatoires. Afin d'analyser complètement l'algorithme de la KPM, il faudrait obtenir un résultat théorique tenant compte de cet aléa.

1.7.2 Calibrage des constantes.

Cette partie repose sur des considérations d'ordre plus pratique. La détermination des paramètres libres des algorithmes est un élément clef dans les applications numériques qui n'est, en général, pas contrôlé par des arguments théoriques. Par exemple, quelle constante μ (équation (1.36)) doit-on choisir pour optimiser les performances de la KPM ? De façon générale, les constantes fournies par des arguments théoriques du type Théorème 6.2.1 (équation (1.34)) sont trop mauvaises pour être utilisables. En pratique, elles sont donc calibrées pour s'adapter aux données. Les remarques suivantes concernent les constantes de pénalisation que sont C_n pour la SVM (équation (1.31)) et μ pour la KPM. Dans les deux cas, l'ordre de grandeur de la pénalité est fixé et la procédure de sélection de modèle est déterminée par la constante de pénalisation. On peut noter que le calibrage de ces constantes doit se faire de façon adaptative à la marge.

- Il est important de se poser la question de l'importance de l'ordre de grandeur de la pénalité en pratique. Dans le cas de la SVM, C_n est souvent choisie par validation croisée. Cette méthode de calibrage compense l'éventuel défaut de l'ordre de grandeur de la pénalité. Il se produit le même phénomène pour la KPM. En effet, les résultats numériques des chapitres 5 et 6 montrent que les performances de la KPM sont sensiblement les mêmes si la pénalité est choisie linéaire ou quadratique en la dimension et que la constante μ est obtenue par validation croisée.

- Dans le chapitre 6, un autre procédé de calibrage de la constante μ est adopté : *l'heuristique de pente*. Contrairement à la validation croisée, il profite pleinement de l'ordre de grandeur linéaire de la pénalité. On verra les gains que cela peut apporter, au moins en termes de temps de calcul.

- La compréhension de l'effet de la validation croisée sur les performances d'un algorithme nécessite une étude théorique. À part le travail de Zhang [Zha93] dans le cadre de la régression gaussienne, la littérature est particulièrement pauvre sur ce sujet. Les considérations sur l'heuristique de pente sont elles aussi purement empiriques et méritent d'être approfondies.

Le chapitre 7 donnera des propositions d'utilisation de la projection fini-dimensionnelle pour d'autres problèmes d'apprentissage.

Le chapitre suivant présente, en français, les techniques utilisées dans la majeure partie de cette thèse. Les chapitres 3 à 6 correspondent essentiellement à quatre articles publiés ou soumis. Ils sont rédigés en anglais.

Chapter 2

Introduction mathématique

Ce chapitre ne donne pas de nouveaux résultats. Il met simplement en parallèle différentes méthodes afin d'expliciter les liens entre les chapitres de la présente thèse.

La première partie décrit *l'analyse globale* de la minimisation empirique du risque. Elle permet de retrouver les résultats théoriques fondamentaux d'apprentissage de Vapnik. La deuxième partie se place dans un modèle très utilisé en statistique : le *bruit blanc gaussien*. Dans ce cadre, elle compare le comportement des régularisations type Tikhonov et fini-dimensionnelle dont il est déjà question dans la présentation générale. Enfin, la troisième partie décrit les principales étapes de *l'analyse locale* de la minimisation empirique du risque. Cette analyse raffine celle de Vapnik en s'inspirant du travail dans le cas gaussien de la deuxième partie.

L'analyse locale sera utilisée pour étudier le comportement statistique de la KPCA, des valeurs propres de la matrice de Gram (chapitre 3) et pour analyser les propriétés statistiques de la projection fini-dimensionnelle en classification (chapitres 5 et 6).

Contents

L'échantillon d'apprentissage est constitué de n variables aléatoires indépendantes et identiquement distribuées (X_i, Y_i), $i = 1 \dots n$ de même loi que (X, Y). Les notations Pf (resp. $P_n f$) désignent les quantités $\mathbb{E}[f(X,Y)]$ (resp. $\frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$). Soit $\mathcal{F}$ une classe de fonctions. Un *processus empirique* indexé par $\mathcal{F}$ désigne la collection de variables aléatoires $\{(P_n - P)(f),\, f \in \mathcal{F}\}$. Une variable aléatoire Z est *sous-gaussienne* si elle vérifie :

$$\forall t > 0\,,\, \mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq e^{-t^2/(2\mathrm{Var}(Z))}\,.$$

Par exemple, si Z suit une loi gaussienne centrée, elle vérifie la précédente inégalité.

Dans sa théorie concernant l'apprentissage statistique, V. N. Vapnik a proposé une façon d'analyser statistiquement la minimisation empirique du risque : le but de cette partie est d'expliquer en détails cette approche “classique” (dite *globale*) afin d'en extraire les idées intéressantes et d'en souligner les lacunes.

2.1 Approche globale : premières bornes de risque.

Cette partie est inspirée de la formulation de l'approche de Vapnik par [Mas00b].

2.1.1 Analyse de Vapnik.

La théorie d'origine qui nous sert de référence est celle de Vapnik. Elle a introduit des considérations statistiques pour des méthodes d'apprentissage et porte le nom de *statistical learning* ([Vap95],[Vap98]). Elle est basée sur des principes issus des contraintes imposées par la communauté du machine learning. La sélection de modèles de Birgé et Massart est aussi régie par ces principes. En effet, dans le cadre de sélection de modèle décrite dans la partie 1.4, le but est d'obtenir des bornes non-asymptotiques avec le moins possible d'hypothèses. Les bornes doivent ensuite être effectivement calculables pour pouvoir réellement évaluer la qualité des algorithmes. Les inégalités de concentration dont il sera question plus tard sont des outils intéressants pour atteindre ce but de façon élégante.

En considérant la propriété (1.1) satisfaite par la cible s, la minimisation empirique du risque (ERM) étudiée, entre autres, par Vapnik consiste à associer à chaque modèle S_m l'estimateur $\widehat{f}_m$ minimisant le risque empirique :

$$\widehat{f}_m = \arg \min_{f \in S_m} \frac{1}{n} \sum_{i=1}^{n} \gamma(f, (X_i, Y_i))\,, \tag{2.1}$$

où $\frac{1}{n}\sum_{i=1}^n \gamma(f, (X_i, Y_i))$ désigne le risque empirique et γ est un contraste adapté au type de problème d'apprentissage considéré. En considérant comme critère uniquement la qualité de prédiction (mesurée à l'aide du contraste), cette formulation du problème d'apprentissage satisfait le principe de Vapnik qui consiste à chercher directement une fonction qui ait de bonnes propriétés de prédiction. Une fois le modèle fixé, elle permet de réduire le problème d'apprentissage à un problème d'optimisation. Les choix du critère γ et du modèle S_m sont aussi guidés par des contraintes de calcul : on doit pouvoir effectivement résoudre le problème d'optimisation correspondant.

Si le minimum n'est pas atteint ou s'il est difficilement calculable de façon exacte, on peut considérer un minimiseur approximatif (en ajoutant un petit paramètre à la fonction à minimiser) : les méthodes développées ici seront robustes à ce type de modification.

L'analyse de Vapnik a été construite pour travailler avec des contrastes peu réguliers tels que la perte dure en classification et fournit des inégalités valables sans hypothèse sur la loi sous-jacente des données : elles doivent être considérées comme des bornes de référence et toute méthode de majoration plus subtile doit donner des bornes au moins aussi bonnes.

D'après la définition (2.1) de l'estimateur $\widehat{f}_m$ comme étant un minimiseur empirique, on a $P_n\gamma(\widehat{f}_m) \leq P_n\gamma(f_m)$ d'où

$$L(\widehat{f}_m, f_m) \leq (P - P_n)(\gamma(\widehat{f}_m)) - (P - P_n)(\gamma(f_m)), \tag{2.2}$$

où $L(f, g) = P\gamma(f) - P\gamma(g)$ désigne l'excès de risque associé au contraste γ et f_m est un élément quelconque de S_m. Par la suite, il sera choisi pour optimiser la borne de risque obtenue et doit dès maintenant être considéré comme représentant l'erreur d'approximation du modèle S_m :

$$f_m = \arg \min_{f \in S_m} P\gamma(f).$$

Le manque de régularité du contraste est géré en faisant la majoration suivante :

$$(P - P_n)(\gamma(\widehat{f}_m)) - (P - P_n)(\gamma(f_m)) \leq 2 \sup_{f \in S_m} |(P - P_n)(\gamma(f))|. \tag{2.3}$$

Cette inégalité, caractéristique de l'approche de Vapnik, est brutale pour plusieurs raisons : tout d'abord, elle gère séparément les termes $(P - P_n)(\gamma(\widehat{f}_m))$ et $(P - P_n)(\gamma(f_m))$. Si S_m est un "bon"modèle, $\widehat{f}_m$ sera "proche" de f_m : ces deux termes peuvent donc se compenser et la majoration devient sous-optimale. Il serait plus précis de considérer le module de continuité associé au processus empirique $\{(P - P_n)(\gamma(f)) - (P - P_n)(\gamma(f_m)), f \in S_m\}$ pour quantifier un gain potentiel. Ensuite, la fluctuation due à la quantité aléatoire $\widehat{f}_m$ est contrôlée par la fluctuation maximale du contraste sur le modèle, ce qui peut aussi s'avérer être sous-optimal. Ces observations seront à la base des raffinements obtenus par l'approche locale considérée dans la partie 2.3.

Une vertu de cette inégalité est de ramener l'étude de la borne de risque à un contrôle d'un supremum de processus empirique indexé par le modèle. Ce type de quantité se contrôle en utilisant des *inégalités de concentration* : elles quantifient la déviation d'une variable aléatoire par rapport à son espérance. Dans les bons cas (considérer une variable aléatoire bornée suffit), cette déviation décroît exponentiellement vite. En utilisant ce type d'inégalités, on obtient des bornes valables avec grande probabilité. Ces contrôles sont aussi utilisés dans la communauté informatique sous le nom de "bornes-PAC" (probably approximately correct) dont la paternité en revient à [Val84]. On peut retrouver les résultats de Vapnik en utilisant l'inégalité de concentration suivante attribuée à McDiarmid.

Théorème 2 ([McD98]). *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées et soit $Z = f(X_1, ..., X_n)$ avec f telle que:*

$$\sup_{x_1,...,x_n,x'_i} |f(x_1, ..., x_n) - f(x_1, ., x'_i, ., x_n)| \leq c_i \;, \forall\, 1 \leq i \leq n\,,$$

alors

$$\mathbb{P}[Z - \mathbb{E}[Z] \geq x] \leq e^{-2x^2/(c_1^2+...+c_n^2)}.$$

Initialement, cette inégalité est issue de la théorie des martingales (inégalité d'Hoeffding-Azuma). Elle est simple à formuler et à utiliser et semble donner un comportement sous-gaussien de la variable aléatoire. Cependant, il faut faire attention au comportement de la variance de Z : elle peut être beaucoup plus petite que $\sum_{i=1}^n c_i^2$ et, dans ce cas, l'inégalité obtenue est loin d'être sous-gaussienne. L'approche raffinée développée dans la section 2.3 utilisera une inégalité de concentration prenant en compte la variance : l'inégalité de concentration des processus empiriques de Talagrand.

En utilisant le Théorème 2, on obtient que pour tout $\xi > 0$, avec probabilité au moins $1 - e^{-\xi}$,

$$\sup_{f \in S_m} |(P - P_n)(\gamma(f))| \leq \mathbb{E}\left[\sup_{f \in S_m} |(P - P_n)(\gamma(f))|\right] + b\sqrt{\frac{\xi}{2n}}. \tag{2.4}$$

où b est une borne sur le contraste : $\sup_{f \in S, x \in \mathcal{X}, y \in \mathcal{Y}} \gamma(f, (x, y)) \leq b$. Grâce à l'inégalité de concentration, l'étude du supremum de processus empirique est ramenée à celle de son espérance. Finalement, en combinant la dernière inégalité avec (2.2) et (2.3), on obtient qu'avec probabilité plus grande que $1 - e^{-\xi}$,

$$L(\widehat{f}_m, f_m) \leq 2\mathbb{E}\left[\sup_{f \in S_m} |(P - P_n)(\gamma(f))|\right] + b\sqrt{\frac{2\xi}{n}}. \tag{2.5}$$

L'inégalité-clé permettant de contrôler une espérance de supremum de processus empirique est *l'inégalité de symétrisation* (Lemme 2.3.1 p. 108 de [vdVW96]) couramment attribuée à J.P. Kahane. Via un raisonnement très simple basé sur la considération d'une copie indépendante de l'échantillon et sur l'inégalité de Jensen, elle assure que

$$\mathbb{E}\left[\sup_{f \in S_m} |(P - P_n)(\gamma(f))|\right] \leq 2\mathbb{E}\left[\sup_{f \in S_m} \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_i \gamma(f, (X_i, Y_i))\right|\right]. \tag{2.6}$$

où $\varepsilon_1, \ldots, \varepsilon_n$ désignent n variables aléatoires indépendantes suivant une loi de Rademacher et indépendantes de l'échantillon ($\mathbb{P}[\varepsilon_1 = 1] = \mathbb{P}[\varepsilon_1 = -1] = 1/2$).

Cette inégalité permet de contrôler l'espérance d'un supremum de processus empirique par l'espérance d'un supremum de processus sous-gaussien (le processus de Rademacher). Pour ce dernier, nous disposons d'outils récents issus de la théorie des processus gaussiens (nous y reviendrons plus précisément dans la partie 2.3.3).

Afin de faciliter la comparaison avec l'approche locale, on écrit l'inégalité finale en introduisant la cible s : avec probabilité plus grande que $1 - e^{-\xi}$,

$$L(\widehat{f}_m, s) \leq L(f_m, s) + 4\mathbb{E}\left[\sup_{f \in S_m} \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_i \gamma(f, (X_i, Y_i))\right|\right] + b\sqrt{\frac{2\xi}{n}}. \tag{2.7}$$

Cette méthode est dite globale puisque la complexité du modèle S_m est déterminée par un supremum indexé par *toutes* les fonctions de ce modèle : l'espérance apparaissant dans le membre de droite est appelée *moyenne de Rademacher*.

Cette approche a été utilisée sous cette forme par [STWCK05] pour contrôler l'erreur quadratique moyenne de la KPCA et donner des inégalités de concentration des valeurs propres de la matrice de Gram (toutes ces notions sont définies dans la partie 1.2). En travaillant avec le bon contraste γ, compte tenu de l'inégalité précédente, il suffit de contrôler la moyenne de Rademacher associée au problème. Avec les notations de la partie 1.2, [STWCK05] montre que, sous certaines conditions peu restrictives,

$$\mathbb{E}\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\left|\sum_{j=1}^{n}\varepsilon_j\|\Pi_{V^\perp}(k(X_j,.))\|^2\right|\right]\leq\Box\sqrt{\frac{d}{n}}\,, \tag{2.8}$$

où $\Box$ désigne une constante universelle. Le raisonnement précédent leur permet d'obtenir une vitesse de convergence en $1/\sqrt{n}$ pour l'erreur de généralisation de la KPCA. Les résultats du chapitre 3 consistent, en partie, à affiner les résultats de [STWCK05] en utilisant l'approche locale de la partie 2.3.

2.1.2 Classification.

D'un point de vue statistique, le problème de classification consiste à estimer la fonction de classification de Bayes f^* :

$$f^*(x)=\begin{cases}1 & si\ \eta(x)\geq\frac{1}{2}\,,\\ -1 & si\ \eta(x)<\frac{1}{2}\,,\end{cases}$$

où $\eta(x)=\mathbb{P}[Y=1|X=x]$ à l'aide de l'échantillon d'apprentissage $((X_1,Y_1),\cdots,(X_n,Y_n))\in(\mathcal{X}\times\{-1,+1\})^n$. Il est naturel de chercher des estimateurs sous la forme $2\mathbb{1}_A-1$ où l'ensemble A est suffisamment proche de $\{x\in\mathcal{X},\ \eta(x)\geq 1/2\}$. La formulation en terme d'ensembles du problème a été utilisée, par exemple dans [TvdG05] mais, dans un souci d'homogénéisation, des notations fonctionnelles seront utilisées.

La minimisation empirique du risque respecte le principe de Vapnik qui stipule que, pour éviter d'avoir à traiter un problème potentiellement plus difficile, il vaut mieux directement chercher une fonction de classification qui ait de bonnes performances en terme de prédictions sans passer par l'estimation de la fonction η.

Dans le cadre de l'ERM de Vapnik, le modèle considéré est $S_m=\{2\mathbb{1}_A-1, A\in\mathcal{C}_m\}$ où $\mathcal{C}_m$ est une classe d'ensembles de complexité contrôlée. Le contraste est la fonction de perte dure $\gamma(f,(x,y))=\mathbb{1}_{f(x)\neq y}$: dans ce cas, le minimiseur $\widehat{f}_m$ du risque empirique

$$\widehat{f}_m=\arg\min_{f\in S_m}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i\neq f(X_i)}\,,$$

est la fonction du modèle minimisant le nombre de mauvaises prédictions.

La notion de complexité utilisée dans l'ERM est la *dimension de Vapnik-Chervonenkis* (VC-dimension). Pour une classe d'ensembles $\mathcal{A}$, elle se définit comme étant le plus grand entier n tel qu'il existe un ensemble de n points dont on puisse retrouver toutes les sous-parties via des intersections avec les ensembles de la classe $\mathcal{A}$:

$$VC(\mathcal{A})=\sup\left\{n\geq 1\,,\max_{z_1,\dots,z_n}|\{A\cap\{z_1,\dots,z_n\},\ A\in\mathcal{A}\}|=2^n\right\}\,.$$

On dit que $\mathcal{A}$ est une classe de Vapnik-Chervonenkis si $VC(\mathcal{A})<\infty$.

La VC-dimension permet donc de rendre compte de la complexité (*capacité*) d'une classe d'ensembles au travers de sa richesse combinatoire. A l'origine, elle a été introduite pour l'étude des suprema de processus empiriques indexés par des classes d'ensembles. En effet, comme rappelé dans [EPP99], elle permet de caractériser les classes d'ensembles satisfaisant une loi des grands nombres uniformes. On a

$$\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}1\!\!1_{X_i\in A}-\mathbb{P}[X\in A]\right|\xrightarrow{\mathbb{P}}0\,,$$

pour tout $\mathbb{P}$ si et seulement si $VC(\mathcal{A})<\infty$. Ce résultat généralise celui de Glivenko-Cantelli à des classes d'ensembles plus générales que les intervalles du type $]-\infty,t]$.

Afin de pouvoir évaluer le risque, il faut contrôler la moyenne de Rademacher apparaissant dans l'inégalité (2.7) par des quantités connues et, en particulier, indépendantes de la loi sous-jacente des données.

La façon dont Vapnik a procédé est la base de raisonnements plus fins. Il est nécessaire de revenir à son raisonnement désormais "classique" en théorie de l'apprentissage : les techniques utilisées sont fondamentales et l'approche locale de la partie 2.3 les exploitera de façon plus fines. De plus, le raisonnement suivant souligne comment la VC-dimension permet de contrôler une moyenne de Rademacher : de nombreuses approches cherchant à pallier aux défauts de la théorie de Vapnik évitent cette étape et travaillent directement avec les moyennes de Rademacher. Dans la suite, on suppose que la classe d'ensembles $\mathcal{C}_m$ est une classe de Vapnik-Chervonenkis de VC-dimension notée V_m.

Le point-clé est le théorème suivant, attribué à Pisier, qui permet de contrôler l'espérance du supremum de variables aléatoires sous-gaussiennes.

Théorème 3. *Soient $Y_1,\dots,Y_N$ des variables aléatoires avec $N\geq 2$ telles que $\forall\,\lambda>0$, $\forall\, i=1\dots N$*

$$\mathbb{E}\left[e^{\lambda Y_i}\right]\leq e^{\lambda^2R^2/2}\,,$$

et

$$\mathbb{E}\left[e^{-\lambda Y_i}\right]\leq e^{\lambda^2R^2/2}\,,$$

où $R>0$. On a alors

$$\mathbb{E}\left[\max_{i=1\dots N}|Y_i|\right]\leq R\sqrt{2\log 2N}\,.$$

Ce théorème permet d'exploiter le fait que les variables de Rademacher soient sous-gaussiennes. En effet, $\forall\,a,\lambda\in\mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda\varepsilon_1 a}\right]\leq e^{\lambda^2a^2/2}\,. \tag{2.9}$$

On peut donc appliquer le Théorème 3 conditionnellement aux variables (X_i,Y_i), $i=1\dots n$ avec $R=\frac{1}{n}\sqrt{\sup_{f\in S_m}\sum_{i=1}^{n}1\!\!1_{Y_i\neq f(X_i)}}$. On obtient

$$\mathbb{E}_\epsilon\left[\sup_{f\in S_m}\frac{1}{n}\left|\sum_{i=1}^{n}\varepsilon_i 1\!\!1_{f(X_i)\neq Y_i}\right|\right]\leq R\sqrt{2H_m}\,,$$

où $H_m=\log 2|\{C\cap\{X_1,\dots,X_n\},C\in\mathcal{C}_m\}|$ et $\mathbb{E}_\varepsilon$ signifie que l'on intègre uniquement part rapport aux ε : les variables (X_i,Y_i), $i=1\dots n$ sont "fixées". Le lemme de Sauer ([Vap82]) permet de contrôler H_m et il est clair que $R \leq \frac{1}{\sqrt{n}}$. Donc, en supposant $n > V_m$, on obtient un majorant de la moyenne de Rademacher indépendant de la loi sous-jacente des données,

$$\mathbb{E}\left[\sup_{f \in S_m} \frac{1}{n}\left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right] \leq \sqrt{\frac{2V_m}{n} \log \frac{2en}{V_m}}.$$

Le facteur logarithmique est en fait superflu : grâce à un argument de chaînage (qui consiste à utiliser le Théorème 3 de façon locale pour être plus précis), [Lug02] montre que

$$\mathbb{E}\left[\sup_{f \in S_m} \frac{1}{n}\left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right|\right] \leq \square\sqrt{\frac{V_m}{n}}, \tag{2.10}$$

où $\square$ désigne une constante universelle qu'on ne cherchera pas à expliciter. Ce résultat illustre le contrôle des moyennes de Rademacher par la VC-dimension.

En passant à l'espérance dans l'inégalité (2.2), puis en utilisant l'inégalité (2.6) et le précédent contrôle de moyenne de Rademacher, on obtient

$$\mathbb{E}\left[\ell(\widehat{f}_m, f_m)\right] \leq \square\sqrt{\frac{V_m}{n}},$$

où $\ell(f, f^*) = \mathbb{P}[f(X) \neq Y] - \mathbb{P}[f^*(X) \neq Y]$. Finalement

$$\sup_Q \mathbb{E}\left[\ell(\widehat{f}_m, f^*)\right] \leq \square\sqrt{\frac{V_m}{n}}, \tag{2.11}$$

où le supremum est pris sur toutes les distributions Q de (X, Y) pour lesquelles f^* appartient à S_m. Selon la théorie de Vapnik, le problème de la complexité en classification se résume donc à l'évaluation de la VC-dimension. De plus, cette inégalité est optimale dans le sens où elle permet de conclure que l'estimateur de minimum de contraste $\widehat{f}_m$ est approximativement minmax (notion définie à la page 8) sur S_m. En effet, on a:

$$R(\widehat{f}_m, S_m) = \inf_{\widetilde{s}_m} \sup_Q \mathbb{E}\left[\ell(\widetilde{s}_m, f^*)\right] \geq \square\sqrt{\frac{V_m}{n}}, \tag{2.12}$$

où l'infimum est pris sur tous les estimateurs possibles à partir de $(X_1, Y_1), \ldots, (X_n, Y_n)$ et le supremum sur toutes les distributions Q de (X, Y) pour lesquelles f^* appartient à S_m.

Toutefois, cette approche ne résout pas toutes les difficultés du problème de classification. Tout d'abord, d'un point de vue pratique, la VC-dimension peut être difficilement calculable. De plus, ce paramètre évalue de façon pessimiste les performances de l'algorithme puisqu'il ne s'adapte pas aux données. La théorie de Vapnik est connue comme étant celle du "pire cas" : cela se voit sur le risque minmax (2.12) puisque, dans la preuve de cette inégalité, on constate que les distributions qui contribuent à la borne inférieure sont celles qui correspondent à des problèmes de classification très difficiles. La théorie Vapnik ne reflète donc pas toujours bien la réalité.

Plusieurs idées ont été émises pour obtenir une théorie moins conservative (voir par exemple [Vay00]). Par exemple, [Kol01] et [BBL02] utilisent un argument de concentration de [BLM00] pour contrôler la moyenne de Rademacher apparaissant dans la borne (2.7) par la *moyenne de Rademacher conditionnée* correspondante $\mathbb{E}_\varepsilon\left[\sup_{f \in S_m} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)\neq Y_i}\right]$, où la notation $\mathbb{E}_\varepsilon$ signifie que l'on intègre uniquement part rapport aux ε : les variables X_i, $i = 1 \dots n$ sont "fixées". Cette approche permet d'obtenir un contrôle de la complexité qui s'adapte aux données. Bien qu'elle considère des moyennes de Rademacher globales, cette méthode est proche, dans l'esprit, de l'approche locale détaillée dans la partie 2.3. En effet, elle souligne que l'utilisation de résultats de concentration adaptés permet d'obtenir des termes de complexité plus adéquats. De plus, elle met en valeur l'importance des moyennes de Rademacher tant d'un point de vue pratique que théorique : ces quantités sont plus souples à utiliser que la VC-dimension et seront à la base du développement de la méthode de localisation.

La partie suivante se place dans un autre contexte en considérant un modèle statistique classique de régression gaussienne : le *bruit blanc gaussien.* Elle utilisera un vocabulaire plutôt statistique. Ainsi, les termes de *biais* et de *variance* utilisés par la suite désignent l'erreur d'approximation et l'erreur d'estimation utilisées dans le cadre de l'apprentissage. Tout comme l'approche globale de Vapnik, le travail mené dans ce cadre a inspiré l'élaboration de la méthode de localisation de la partie 2.3.

Le but est maintenant de comparer les propriétés d'adaptativité de la régularisation type Tikhonov à celles de la régularisation fini-dimensionnelle dans le cadre du bruit blanc gaussien. Comme expliqué dans la présentation générale, la régularisation fini-dimensionnelle a été étudiée dans les chapitres 5 et 6 de la présente thèse dans le cadre de la classification tandis que l'algorithme des SVM repose sur la régularisation de Tikhonov. Le cadre gaussien permet de calculer facilement les quantités d'intérêts et servira à guider notre intuition.

2.2 Intuition gaussienne pour la régularisation.

Cette section présente des résultats de [Mas04] obtenus dans le cadre du modèle gaussien généralisé (introduit dans [BM01]) : afin de simplifier la présentation, ils sont ici énoncés dans le cadre particulier du *bruit blanc gaussien.* C'est le modèle le plus simple de statistique non-paramétrique : on suppose observer un signal bruité modélisé sous la forme suivante

$$\begin{cases} dY_n(x) = s(x)dx + \frac{1}{\sqrt{n}}dB(x) \quad , x \in [0,1] \, , \\ Y_n(0) = 0 \, , \end{cases}$$

où B est un mouvement brownien sur [0,1] et s le signal à estimer supposé appartenir à $L_2([0,1])$. Par la suite, la fonction de perte utilisée est la norme naturelle de l'espace $L_2([0,1])$: elle est notée $\|\cdot\|$. La notation $\mathbb{E}_s$ signifie qu'on intègre par rapport à la loi de Y_n, définie par l'équation précédente.

Pour un sous-ensemble A de $L_2([0,1])$ et un estimateur $\widehat{s}$, on note

$$R(\widehat{s}, A) = \sup_{\mathcal{P}(A)} \mathbb{E}_s \|\widehat{s} - s\|^2 \, ,$$

où $\mathcal{P}(A)$ désigne l'ensemble des distributions de Y_n pour lesquelles s appartient à A. De plus, le risque minmax sur A est

$$\mathcal{R}(A) = \inf_{\widehat{s}} \sup_{\mathcal{P}(A)} \mathbb{E}_s \|\widehat{s} - s\|^2 \, ,$$

où l'infimum est pris sur tous les estimateurs $\widehat{s}$ basés sur l'observation du processus Y_n. Si $(A_\theta)_{\theta\in\Theta}$ désigne une collection de sous-ensembles de $L_2([0,1])$, un estimateur $\widehat{s}$ est simultanément minmax sur $(A_\theta)_{\theta\in\Theta}$ (de façon non-asymptotique) si $\forall\theta \in \Theta$, $\mathcal{R}(\widehat{s}, A_\theta) \leq c\mathcal{R}(A_\theta)$ où la constante c ne dépend ni de n ni de θ : on dit alors qu'il est *adaptatif* par rapport à $(A_\theta)_{\theta\in\Theta}$. La notion d'adaptativité définie dans [BBM99] autorise la constante c à dépendre de θ mais nous utiliserons la définition précédente.

Pour un noyau de Mercer k (défini à la page 22 : ici, cela signifie simplement que la fonction k soit continue comme fonction de deux variables), on considère la base orthonormée $(\phi_i)_{i\geq 1}$ de $L_2([0,1])$ diagonalisant l'opérateur T_k défini sur $L_2([0,1])$ par

$$T_k : f \mapsto \int_0^1 k(x,.)f(x)dx \,.$$

Il est important de constater que cet opérateur est un cas particulier de celui défini par l'équation (1.26) où l'espace $\mathcal{X}$ est le segment $[0,1]$ et la loi P est la loi uniforme sur $[0,1]$. Les valeurs propres associées sont notées $(\lambda_i)_{i\geq 1}$ et sont rangées par ordre décroissant. Le produit scalaire de $L_2([0,1])$ est noté sans indice $\langle\cdot,\cdot\rangle$ et la norme associée $\|\cdot\|$.

La fonction à estimer satisfait $s = \sum_{i\geq 1}\langle s,\phi_i\rangle\phi_i = \sum_{i\geq 1}\left(\int_0^1 \phi_i(x)s(x)dx\right)\phi_i$. De plus, les coefficients $\beta_i = \langle s,\phi_i\rangle$ sont naturellement estimés par $\widehat{\beta}_i = \int_0^1 \phi_i(x)dY_n(x)$. Cependant, l'estimateur $\sum_{i\geq 1}\widehat{\beta}_i\phi_i$ possède de très mauvaises qualités de généralisation. Un procédé de régularisation modifie les coefficients de cet estimateur pour le rendre, comme son nom l'indique, plus régulier dans le but d'obtenir un estimateur plus performant. Dans cette partie, nous comparons les performances de deux estimateurs : l'un est régularisé par une procédure type Tikhonov et l'autre par une procédure de projection fini-dimensionnelle.

Étant donné un modèle S_m, la meilleure approximation s_m de s appartenant à S_m est le minimiseur, pour g parcourant S_m, de $\|s-g\|$ c'est-à-dire de $\|g\|^2 - 2\langle s,g\rangle$. L'estimateur associé au modèle S_m sera donc le minimiseur du critère des moindres carrés

$$\gamma_n(g) = \|g\|^2 - 2\int_0^1 g(x)dY_n(x)\,,$$

avec g variant dans S_m.

Ce cadre semble a priori assez éloigné de celui de l'apprentissage. En effet, l'observation est constituée d'un processus entier $\{Y_n(x), x\in[0,1]\}$ de réels et non d'un nombre fini d'observations en dimension potentiellement élevée. De plus, les erreurs sont supposées gaussiennes et le critère des moindres carrés est très régulier. Néanmoins, il permet de donner un cadre fixant les objectifs qu'on peut atteindre en classification.

2.2.1 Régularisation type Tikhonov.

L'estimateur que nous considérons dans cette partie est le suivant :

$$\widehat{g} = \arg\min_{g\in\mathcal{H}} \left(\gamma_n(g) + \mathrm{pen}(\|g\|_{\mathcal{H}})\right)\,, \tag{2.13}$$

où $\mathcal{H}$ désigne le RKHS (définition 1 de la page 21) associé au noyau k. Si la fonction de pénalité *pen* est croissante, cet estimateur s'interprète dans le cadre de la sélection de modèles définie dans la partie 1.1.2. En effet, on montre aisément que $\widehat{g} = \widehat{g}_{\widehat{R}}$ où

$$\widehat{g}_R = \arg\min_{g\in\mathcal{H}, \|g\|_{\mathcal{H}}\leq R} \gamma_n(g)\,,$$

et

$$\widehat{R} = \arg\min_{R>0}\left(\gamma_n(\widehat{g}_R) + \mathrm{pen}(R)\right)\,. \tag{2.14}$$

La dernière égalité correspond à la minimisation pénalisée de la perte empirique de l'équation (1.20). Dans ce cas, la collection de modèles est donc constituée des boules de l'espace auto-reproduisant $B_{\mathcal{H}}(0,R) = \{g \in \mathcal{H}, \|g\|_{\mathcal{H}} \leq R\}$ où R varie dans $\mathbb{R}^+$ et la perte est le critère des moindres carrés.

L'estimateur $\widehat{g}$ introduit ci-dessus a déjà été beaucoup étudié tant d'un point de vue pratique que théorique dans le cas où $\mathrm{pen}(R) = C_n R^2$ (C_n désigne un réel dépendant de n). Cette pénalité joue un rôle particulier car elle facilite les calculs : $\widehat{g}$ a alors une expression explicite

$$\widehat{g} = \sum_{i\geq 1} \omega_i^{Tik} \widehat{\beta}_i \phi_i \,, \tag{2.15}$$

où $\omega_i^{Tik} = \frac{1}{1+C_n\lambda_i^{-1}}$. $\widehat{g}$ ressemble donc à un estimateur linéaire. Cependant, la valeur des coefficients dépend de la valeur propre associée et de la constante de régularisation C_n.

Afin de bien comprendre la portée des résultats d'adaptativité (ou de non-adaptativité) présentés ici, il est impératif que le lecteur ait bien à l'esprit que la constante de régularisation C_n est supposée *déterministe* : en particulier, les résultats présentés ici sont de nature différente de ceux de [CT01]).

Par définition de la norme dans l'espace auto-reproduisant $\mathcal{H}$, les boules de cet espace sont des ellipsoïdes de $L_2([0,1])$. La régularisation type Tikhonov revient donc à considérer comme collection de modèles des ellipsoïdes de $L_2([0,1])$:

$$B_{\mathcal{H}}(0,R) = \left\{ g \in L_2([0,1])\,, \sum_{i\geq 1} \frac{\langle g,\phi_i\rangle^2}{\lambda_i} \leq R^2 \right\} .$$

Soit $\mathcal{E}(c) = \{g \in L_2([0,1])\,, \sum_{i\geq 1}\langle g,\phi_i\rangle^2/c_i^2 \leq 1\}$ l'ellipsoïde de $L_2([0,1])$ d'axes principaux les fonctions propres de l'opérateur intégral et de demi-axes associés la suite des $(c_j)_{j\geq 1}$ supposée décroissante. Le risque minmax sur les ellipsoïdes est connu dans le cadre gaussien ([DLM90]) : en effet, si $\mathcal{E}(c)$ est un ellipsoïde non-dégénéré, c'est-à-dire que $c_1 \geq \frac{1}{\sqrt{n}}$, alors

$$\mathcal{R}(\mathcal{E}(c)) \geq \square \inf_{D\geq 1} \left(c_{D+1}^2 + \frac{D}{n} \right) , \tag{2.16}$$

où $\square$ désigne une constante universelle.

Dans la suite, nous nous placerons dans le cas dit "Sobolev" où $\lambda_j \sim j^{-2r}$ avec $r > 1/2$. La boule $B_{\mathcal{H}}(0,R)$ est alors une boule de Besov et sera notée $\mathcal{B}_2(r,R)$. Dans ce cas, (2.16) donne

$$\mathcal{R}(\mathcal{B}_2(r,R)) \geq \square R^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}} . \tag{2.17}$$

Soit $\widehat{g}$ l'estimateur obtenu par la relation (2.13) où $\mathrm{pen}(\|g\|_{\mathcal{H}}) = C_n\|g\|_{\mathcal{H}}^2$. Le point-clé des minorations obtenues par [Mas04] réside dans le Lemme 66 : grâce à l'expression explicite (2.15) de $\widehat{g}$, on a

$$R(\widehat{g}, \mathcal{B}_2(r,R)) \geq C(r) \left\{ C_n R^2 + \frac{1}{n} C_n^{-\frac{1}{2r}} \right\} , \tag{2.18}$$

où la constante $C(r)$ ne dépend que de r. De part cette inégalité, si on veut que l'estimateur $\widehat{g}$ soit approximativement minmax (notion définie à la page 8) sur $\mathcal{B}_2(r,1)$, c'est-à-dire (au regard de (2.17)) que

$$R(\widehat{g}, \mathcal{B}_2(r,1)) \leq \square n^{-\frac{2r}{2r+1}} ,$$

la constante de régularisation doit satisfaire

$$C_1(r)n^{-\frac{2r}{2r+1}} \leq C_n \leq C_2(r)n^{-\frac{2r}{2r+1}},$$

où $C_1(r)$ et $C_2(r)$ sont des constantes adéquates ne dépendant que de r. L'ordre de grandeur de R étant imposé, une fois qu'on considère $C_n \sim C(r)n^{-\frac{2r}{2r+1}}$ la pénalité est déterminée et on a, en réutilisant l'inégalité (2.18),

$$R(\widehat{g}, \mathcal{B}_2(r,R)) \geq C(r)R^2 n^{-\frac{2r}{2r+1}},$$

où $\widehat{g}$ est l'estimateur obtenu en prenant comme pénalité $\mathrm{pen}(R) = \square n^{-\frac{2r}{2r+1}}R^2$. Au moins quand R est grand, l'estimateur obtenu par la régularisation type Tikhonov n'est donc pas approximativement minmax sur les ensembles de la collection de modèles $\mathcal{B}_2(r,R)$: on ne retrouve pas le facteur $R^{\frac{2}{2r+1}}$ du risque minmax donné par l'équation (2.17). L'approche non-asymptotique permet de détecter le défaut d'adaptativité. La vitesse $n^{-\frac{2r}{2r+1}}$ reste cependant correcte et on pourrait penser ne perdre que sur les constantes mais ce n'est pas le cas. L'estimateur $\widehat{g}$ préalablement défini satisfait : $\forall \alpha \in]0,r]$:

$$R(\widehat{g}, \mathcal{B}_2(\alpha,1)) \geq C(r)n^{-\frac{2\alpha}{2r+1}}. \tag{2.19}$$

Cette minoration assure que la vitesse de l'estimateur obtenu par la régularisation type Tikhonov est moins bonne que la vitesse minmax sur les $\mathcal{B}_2(\alpha,1)$ avec $\alpha < r$.

Pour conclure, en terme de vitesses, la procédure de régularisation type Tikhonov fournit un estimateur approximativement minmax vis-à-vis de la collection de modèles mais présentant un défaut d'adaptativité sur des ellipsoïdes de taille plus importante.

Un moyen naturel de remédier à ce problème consisterait à changer l'ordre de grandeur de R dans la pénalité. Soit $\widetilde{g}$ l'estimateur obtenu en prenant $\mathrm{pen}(R) = \square R^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}}$ dans l'équation (2.14) : la Proposition 65 de [Mas04] stipule que $\widetilde{g}$ est approximativement minmax, non seulement sur $\mathcal{B}_2(r,R)$, mais aussi sur tous les $\mathcal{B}_2(\alpha,R)$ avec $\alpha \leq r$ et $R \geq 1/\sqrt{n}$.

En conclusion, le fait d'adapter uniquement la constante de régularisation à la décroissance des valeurs propres peut être sous-optimal : adapter la puissance de R est aussi profitable. Le problème est maintenant d'ordre algorithmique : l'estimateur obtenu n'a plus d'expression explicite.

La projection fini-dimensionnelle présentée dans la partie suivante utilise uniquement la base pour régulariser et elle pallie aux défauts d'adaptativité de la régularisation type Tikhonov.

2.2.2 Projection fini-dimensionnelle.

Le principe de la projection fini-dimensionnelle consiste à considérer des espaces linéaires comme modèles : les ellipsoïdes intrinsèquement liés à la régularisation type Tikhonov sont remplacés par des espaces vectoriels de dimension finie engendrés par les fonctions d'une base de $L_2([0,1])$. Le comportement des estimateurs de minimum de contraste correspondants ont été étudiés dans un cadre général dans [BBM99].

La collection de modèles considérée dans l'approche fini-dimensionnelle est constituée des espaces vectoriels $S_D = \langle \phi_1, \ldots, \phi_D \rangle$ où D varie ($D \geq 1$). Les estimateurs de minimum de contraste correspondants sont :

$$\widehat{f}_D = \arg\min_{f \in S_D} \left(\|f\|^2 - 2\int_0^1 f(x)dY_n(x) \right) = \sum_{i \geq 1} \omega_i^{f.d.} \widehat{\beta}_i \phi_i, \tag{2.20}$$

où $\omega_i^{f.d.} = \mathbb{1}_{i \leq D}$. Comme on considère un contraste L_2 dans un cadre hilbertien, ces estimateurs sont des projections orthogonales sur un espace vectoriel de dimension finie d'où le nom de projection fini-dimensionnelle. Comme le souligne les notations des poids ω_i^{Tik} et $\omega_i^{f.d.}$ dans les équations (2.15) et (2.20), contrairement à l'estimateur $\widehat{g}$ le seuillage des coefficients opéré par ce procédé de régularisation se fait *indépendamment* des valeurs propres. Suivant le procédé général de sélection de modèles (partie 1.4), on choisit l'estimateur final, c'est-à-dire la dimension, par la procédure de minimisation pénalisée de la perte empirique correponsdante à l'équation (1.20) :

$$\widehat{D} = \arg\min_{D \geq 1} \left(\gamma_n(\widehat{f}_D) + \text{pen}(D) \right) , \tag{2.21}$$

où on rappelle que $\gamma_n(f) = \|f\|^2 - 2\int_0^1 f(x)dY_n(x)$. Il faut maintenant choisir convenablement la fonction de pénalité *pen*.

Une façon simple de comprendre les arguments statistiques qui mène à sa détermination consiste à remonter aux origines de la sélection de modèle en décrivant l'heuristique de Mallows. Cette heuristique, d'abord justifiée par des arguments asymptotiques, a inspiré les travaux de sélection de modèles non-asymptotiques de Birgé et Massart (par exemple, [BBM99, BM01]) où un rôle prépondérant est accordé à l'utilisation de bases. Elle est aussi à l'origine du cadre général du problème de sélection de modèle présenté dans la section 1.4.

Heuristique de Mallows.

Le choix de la pénalité est guidé par le fait que la dimension $\widehat{D}$ correspondante doit être la plus proche possible de la dimension optimale D_{oracle} définie suivant l'objectif donné par l'équation (1.19). Cette dernière correspond à l'estimateur ayant la meilleure performance en généralisation :

$$D_{oracle} = \arg\min_{D \geq 1} \mathbb{E}_s \|\widehat{f}_D - s\|^2 .$$

D'après le théorème de Pythagore, le risque se décompose en un terme de biais et un terme de variance :

$$\mathbb{E}_s \|s - \widehat{f}_D\|^2 = \|s - f_D\|^2 + \mathbb{E}_s \|f_D - \widehat{f}_D\|^2 , \tag{2.22}$$

avec $\|s - f_D\|^2 = \|s\|^2 - \|f_D\|^2$ et $f_D = \arg\min_{f \in S_D} \|f - s\| = \sum_{j=1}^D \beta_j \phi_j$ est la projection orthogonale de s sur S_D.

Le terme de variance se calcule ici explicitement puisque, par construction de l'intégrale stochastique,

$$\mathbb{E}_s \|f_D - \widehat{f}_D\|^2 = \sum_{i=1}^D \mathbb{E}_s \left[\left(\int_0^1 \phi_j(x) \frac{1}{\sqrt{n}} dB(x) \right)^2 \right] = \frac{D}{n} . \tag{2.23}$$

Finalement,

$$\mathbb{E}_s \|s - \widehat{f}_D\|^2 = \|s\|^2 - \|f_D\|^2 + \frac{D}{n} ,$$

d'où

$$D_{oracle} = \arg\min_{D \geq 1} \left(-\|f_D\|^2 + \frac{D}{n} \right) . \tag{2.24}$$

La fonction cible s étant inconnue, il faut estimer le carré de la norme de sa projection $\|f_D\|^2$. Il est clair que

$$\|\widehat{f}_D\|^2 = \|\widehat{f}_D - f_D\|^2 + \|f_D\|^2 + 2\langle \widehat{f}_D - f_D, f_D\rangle\,.$$

En utilisant que $\widehat{f}_D$ est un estimateur sans biais de f_D et en réutilisant le calcul (2.23) de variance, on constate que $\|\widehat{f}_D\|^2$ est un estimateur biaisé de $\|f_D\|^2$:

$$\mathbb{E}\|\widehat{f}_D\|^2 = \frac{D}{n} + \|f_D\|^2\,.$$

En remplaçant $\|f_D\|^2$ par son estimateur sans biais $\|\widehat{f}_D\|^2 - \frac{D}{n}$ dans l'équation (2.24), on constate qu'il faudrait choisir la dimension par le critère suivant :

$$\widehat{D} = \arg\min_{D\geq 1}\left(-\|\widehat{f}_D\|^2 + \frac{2D}{n}\right)\,.$$

Si maintenant, on note que $\gamma_n(\widehat{f}_D) = -\|\widehat{f}_D\|^2$, l'heuristique de Mallows préconise de choisir

$$\mathrm{pen}(D) = \frac{2D}{n}\,,$$

dans le critère (2.21). L'ordre de grandeur de cette pénalité a été justifiée dans notre cas puisque la collection de modèles comporte un seul modèle par dimension. La constante 2 apparaissant dans la pénalité n'est pas miraculeuse : en pratique, il faut calibrer cette constante à partir des données pour avoir une procédure de sélection de modèles efficace.

Résultats. [BBM99] établit que l'estimateur $\widehat{f}_{\widehat{D}}$, où $\widehat{D}$ est choisi par le critère (2.21) avec $\mathrm{pen}(D) = \frac{KD}{n}$ $(K > 1)$, satisfait l'inégalité oracle (1.19) :

$$\mathbb{E}\|f - \widehat{f}_{\widehat{D}}\|^2 \leq \Box \inf_{D\geq 1} \mathbb{E}\|s - \widehat{f}_D\|^2\,. \tag{2.25}$$

On obtient donc qu'une pénalité linéaire en la dimension est statistiquement convenable.

Comparaison avec la régularisation type Tikhonov. Considérons un ellipsoïde quelconque $\mathcal{E}(c)$. Le risque minmax de $\widehat{f}_{\widehat{D}}$ sur cet ellipsoïde est maintenant facilement contrôlable. D'une part, l'inégalité (2.25) implique

$$\mathcal{R}(\widehat{f}_{\widehat{D}}, \mathcal{E}(c)) \leq \Box \sup_{s\in\mathcal{E}(c)} \inf_{D\geq 1} \mathbb{E}\|s - \widehat{f}_D\|^2\,,$$

puis, via la décomposition biais-variance (inégalité (2.22)) et le calcul de la variance (équation (2.23)), on a:

$$\mathcal{R}(\widehat{f}_{\widehat{D}}, \mathcal{E}(c)) \leq c \sup_{s\in\mathcal{E}(c)} \inf_{D\geq 1}\left(\sum_{j\geq D+1}\langle s, \phi_j\rangle^2 + \frac{D}{n}\right)\,.$$

Finalement, en utilisant la décroissance de la suite $(c_j)_{j\geq 1}$:

$$\mathcal{R}(\widehat{f}_{\widehat{D}}, \mathcal{E}(c)) \leq \Box \inf_{D\geq 1}\left(c_{D+1}^2 + \frac{D}{n}\right)\,.$$

L'équation (2.16) implique que l'estimateur par projection $\widehat{f}_{\widehat{D}}$ est adaptatif par rapport à *tous* les ellipsoïdes non-dégénérés ($c_1 \geq \frac{1}{\sqrt{n}}$) d'axes principaux les fonctions propres de l'opérateur intégral. En particulier, il est adaptatif par rapport à R et n sur la collection de modèles et sur les boules de Besov $\mathcal{B}_2(\alpha, 1)$. Ce n'était pas le cas pour l'estimateur $\widehat{g}_{\widehat{R}}$ obtenu par régularisation type Tikhonov. En conclusion, dans ce contexte, les espaces vectoriels ont de meilleures propriétés d'adaptation au sens du minmax que les ellipsoïdes.

On peut aussi noter qu'une approche similaire est menée pour des problèmes inverses (type de modèles utilisé, par exemple, en tomographie). Cela peut paraître naturel puisque la formulation d'un problème inverse est formellement assez proche de celle du bruit blanc gaussien. En effet, un problème inverse linéaire se formule de la façon suivante :

$$dY_n(x) = Af(x)dx + \frac{1}{\sqrt{n}}dB(x) \ , x \in [0, 1] \, ,$$

où A est un opérateur compact et f la fonction à estimer. Dans ce cadre, les propriétés de la régularisation type Tikhonov sont connues et [CG05] applique une approche finidimensionnelle. Ce type de problèmes comporte bien sûr ses propres difficultés. En particulier, il faut maintenant tenir compte des valeurs propres de l'opérateur A.

La partie suivante revient sur l'étude de la minimisation empirique du risque en apprentissage. Elle fournit les principales étapes de l'approche *locale* qui vise à raffiner l'approche globale de Vapnik. Cette approche repose à la fois sur les techniques utilisées dans l'approche globale (partie 2.1), sur un travail dans le cadre gaussien (inégalité (2.25) de la partie 2.2) et sur des progrès récents de l'étude des processus empiriques.

2.3 Approche locale.

Les travaux précurseurs de l'approche locale ont été menés dans un cadre de statistique gaussienne. Les différentes étapes de cette analyse sont données dans le paragraphe suivant : l'approche locale suivra le même déroulement tout en s'adaptant aux difficultés des problèmes d'apprentissage. [Bou03] traite de l'approche locale avec un point de vue plus proche des problèmes d'apprentissage.

2.3.1 Pénalisation type Birgé et Massart.

La preuve de l'inégalité (2.25) présentée dans [BM01] se divise en plusieurs grandes étapes. Pour commencer, en utilisant les notations de la partie 2.2.2 et les définitions des différents minimiseurs, on a

$$\left\| \widehat{f}_{\widehat{D}} - s \right\|^2 \leq \|s - f_D\|^2 + \frac{2}{\sqrt{n}} \left[W(\widehat{f}_{\widehat{D}}) - W(f_D) \right] - \mathrm{pen}(\widehat{D}) + \mathrm{pen}(D) \, , \tag{2.26}$$

où $W(f) = \int_0^1 f(x)dB(x)$.

1. On considère $V_{D'}$ un supremum de processus gaussien normalisé associé à la famille $\mathcal{G}_{D'} = \{W(f) - W(f_D), \, f \in S_{D'}\}$:

$$V_{D'} = \sup_{f \in S_{D'}} \left[\frac{W(f) - W(f_D)}{\omega_{D'}(f)} \right] \, ,$$

où le poids $\omega_{D'}(f)$ assure que $\sup_{f \in S_{D'}} \mathrm{Var}\left[\frac{W(f) - W(f_D)}{\omega_{D'}(f)}\right]$ soit contrôlée.

2. Par construction de l'intégrale stochastique, W définit un processus gaussien (isonormal) et on peut donc appliquer l'inégalité de concentration des suprema de processus gaussien de Cirelson, Ibragimov et Sudakov ([BCS76]) au processus normalisé $V_{D'}$: il suffit maintenant de contrôler $\mathbb{E}[V_{D'}]$ pour contrôler $V_{D'}$.

3. Le contrôle de la complexité de l'espace vectoriel $S_{D'}$ de dimension D' intervient par le fait que

$$\mathbb{E}\left[\sup_{f\in S_{D'}}\left[\frac{W(f)-W(f_{D'})}{\|f-f_{D'}\|^2+r}\right]\right]\le \mathbb{E}\left[\sup_{f\in S_{D'}}\left[\frac{W(f)-W(f_{D'})}{\sqrt{r}\,\|f-f_{D'}\|}\right]\right]\le \frac{1}{2}\sqrt{\frac{D'}{r}}\,.$$

Combiné avec un choix convenable du poids $\omega_{D'}(f)$ on obtient qu'avec grande probabilité,

$$V_{D'}\le \frac{\sqrt{n}}{K}\,.$$

4. En utilisant une disjonction d'évènements et en injectant l'inégalité précédente dans l'équation (2.26), on obtient:

$$\|\widehat{f}_{\widehat{D}}-s\|^2\le \|s-f_D\|^2+2K^{-1}\omega_{\widehat{D}}(\widehat{f}_{\widehat{D}})-\mathrm{pen}(\widehat{D})+\mathrm{pen}(D)\,.$$

Le poids $\omega_{D'}(f)$ étant choisi de la forme $\|s-f_D\|^2+\|s-f\|^2+c_{D'}$, on obtient une inégalité du type :

$$\|\widehat{f}_{\widehat{D}}-s\|^2\le \|s-f_D\|^2+2K^{-1}(\|s-f_D\|^2+\|s-\widehat{f}_{\widehat{D}}\|^2+c_{\widehat{D}})-\mathrm{pen}(\widehat{D})+\mathrm{pen}(D)\,.$$

5. L'étape finale consiste à regrouper les termes en $\|\widehat{f}_{\widehat{D}}-s\|^2$ qui apparaissent de chaque côté de l'inégalité. Par un choix adapté de c_D et en utilisant le fait que $\mathrm{pen}(D)\ge KD$ avec $K>1$, on obtient alors l'inégalité (2.25) par intégration.

2.3.2 Approche locale.

L'approche locale s'appuie sur une extension du cadre gaussien : les idées sont les mêmes que ci-dessus mais elles sont utilisées dans un cadre différent ([Mas00b]).

Par analogie avec la famille $\mathcal{G}_{D'}$ de l'étape 1 de la preuve de l'inégalité (2.25), l'inégalité (2.2) incite à considérer la *famille translatée de fonctions*,

$$\mathcal{G}_m=\{(P-P_n)(\gamma(f))-(P-P_n)(\gamma(f_m)),f\in S_m\}\,,$$

où

$$f_m=\arg\min_{f\in S_m}P\gamma(f)\,.$$

Elle vise à profiter d'éventuelles compensations dans le cas où $\widehat{f}_m$ est proche de f_m. Contrairement à l'approche de Vapnik décrite dans la section 1.1, le terme f_m représentant l'erreur d'approximation du modèle S_m est donc utilisé pour contrôler la complexité.

Idéalement, par analogie avec le Théorème "Central Limite", on devrait contrôler les déviations de

$$\widetilde{Z}=\sup_{f\in S_m}\frac{(P-P_n)(\gamma(f)-\gamma(f_m))}{\sqrt{\mathrm{Var}(\gamma(f)-\gamma(f_m))}}\,.$$

La normalisation du processus assure que la variance de chaque incrément du supremum vaut 1 uniformisant ainsi les déviations de $(P-P_n)(\gamma(f)-\gamma(f_m))$. Cependant, on doit considérer des poids un peu plus importants que la variance elle-même et introduire un *paramètre de localisation* r. Dans cette thèse, deux types de poids seront considérés : tout d'abord le poids "découplé" dans le chapitre 6, correspondant à

$$Z_r = \sup_{f \in S_m} \frac{(P-P_n)(\gamma(f)-\gamma(f_m))}{\mathrm{Var}(\gamma(f)-\gamma(f_m))+r}, \tag{2.27}$$

mais aussi le poids "non-découplé" dans le chapitre 5 amenant à considérer

$$Z'_r = \sup_{f \in S_m} \frac{(P-P_n)(\gamma(f)-\gamma(f_m))}{\sqrt{r\mathrm{Var}(\gamma(f)-\gamma(f_m))}}. \tag{2.28}$$

Ils sont liés par le fait que

$$\mathrm{Var}(\gamma(f)-\gamma(f_m))+r \geq 2\sqrt{r\mathrm{Var}(\gamma(f)-\gamma(f_m))}.$$

Le choix du poids à adopter dans la méthode de localisation est à la fois subtil et important. Il doit être choisi de façon à contrôler la variance des incréments du processus empirique mais de façon assez fine : le choix est aussi guidé par la nature de la communication biais-variance (notion qui sera expliquée par la suite).

L'étape suivante consiste (comme dans l'étape 2 de la preuve de l'inégalité (2.25)) à contrôler uniformément les déviations du processus re-normalisé. Contrairement à l'inégalité de Mc-Diarmid (Théorème 2) qui ne prend en compte que la norme infinie de la variable aléatoire, le résultat de concentration des processus empiriques utilisé ici prend en compte le comportement de sa variance afin d'obtenir une inégalité de concentration plus précise. L'outil fondamental de l'approche locale est l'inégalité de concentration de Talagrand : elle vient de travaux récents ([Tal96]) sur les suprema de processus empiriques et a fait l'objet d'un travail approfondi pour en améliorer les constantes et obtenir de bonnes inégalités de déviation vers le bas successivement par [Led96], [Mas00a], [Rio01] et [Kle02]. La version utilisée dans cette thèse est celle de [Bou02b] où les constantes sont optimales en un certain sens. Elle permet, sous des conditions de bornitude, d'obtenir que pour tout $\xi > 0$, avec probabilité au moins $1-e^{-\xi}$,

$$Z_r \leq 2\mathbb{E}[Z_r] + \square\sqrt{\frac{\sigma^2\xi}{n}} + c_1\frac{\xi}{n}, \tag{2.29}$$

où $\square$ est une constante universelle qu'on ne cherche pas à déterminer dans le raisonnement suivant et c_1 fait intervenir r et la norme infinie de la classe $\mathcal{F}_m$. La variance

$$\sigma^2 = \sup_{f \in S_m} \mathrm{Var}\left(\frac{\gamma(f)-\gamma(f_m)}{\mathrm{Var}(\gamma(f)-\gamma(f_m))+r}\right),$$

est contrôlée par $1/(4r)$ grâce à la normalisation du processus.

Finalement, afin de contrôler les déviations du supremum de processus empirique re-normalisé Z, on est donc amené à majorer son espérance

$$\mathbb{E}[Z_r] = \mathbb{E}\left[\sup_{f \in S_m} \frac{(P-P_n)(\gamma(f)-\gamma(f_m))}{\mathrm{Var}(\gamma(f)-\gamma(f_m))+r}\right].$$

Utiliser un processus re-normalisé revient moralement à éviter d'avoir à considérer la variance maximale sur le modèle S_m. Précisément, en découpant le modèle en couronnes (selon les valeurs de la variance) et en considérant le maximum de variance sur chaque couronne, le *peeling device* permet de faire le lien entre $\mathbb{E}[Z_r]$ et des quantités où le supremum est localisé. De fait, le Lemme 5.1 de [Mas00b] montre que si

$$\mathbb{E}\left[\sup_{f\in S_m,\,\mathrm{Var}(\gamma(f)-\gamma(f_m))\le r} |(P-P_n)(\gamma(f)-\gamma(f_m))|\right] \le \phi_m(r)\,, \tag{2.30}$$

avec ϕ_m une fonction croissante et telle que $\phi_m(r)/\sqrt{r}$ est décroissante, alors pour tout $r \ge r_m^*$,

$$\mathbb{E}[Z_r] \le 4\sqrt{\frac{r_m^*}{r}}\,,$$

où r_m^* est le point fixe de ϕ_m (il existe toujours).

On voit ici clairement apparaître le module de continuité : plutôt que d'avoir à contrôler un supremum de processus empirique sur tout le modèle, il suffit de le faire *localement* c'est-à-dire pour les fonctions de faible variance. Cela revient à contrôler le processus empirique uniquement pour de petits incréments. Du point de vue de l'analyse statistique d'algorithmes, cela correspond à tenir compte du faible risque empirique de l'estimateur produit par l'algorithme.

Par l'inégalité (2.29), on a donc qu'avec probabilité plus grande que $1-e^{-\xi}$,

$$Z_r \le 8\sqrt{\frac{r_m^*}{r}} + \square\sqrt{\frac{\xi}{rn}} + c_1\frac{\xi}{n}\,.$$

Soit $K>1$. Via un choix adapté de r sous la forme $r_0 = c(K)(r_m^* + \frac{\xi}{n})$, où $c(K)$ ne dépend que de K, on a

$$Z_{r_0} \le \frac{1}{K}\,.$$

Par définition de Z_r (équation (2.27)) avec probabilité plus grande que $1-e^{-\xi}$, $\forall f\in S_m$

$$(P-P_n)(\gamma(f)-\gamma(f_m)) \le \frac{1}{K}\mathrm{Var}(\gamma(f)-\gamma(f_m)) + \square K r_m^* + \square\frac{K\xi}{n}\,.$$

En choisissant l'élément aléatoire $f=\widehat{f}_m$, on obtient que si ϕ_m satisfait la condition (2.30), alors avec probabilité plus grande que $1-e^{-\xi}$,

$$(P-P_n)(\gamma(\widehat{f}_m)-\gamma(f_m)) \le \frac{1}{K}\mathrm{Var}(\gamma(\widehat{f}_m)-\gamma(f_m)) + \square K r_m^* + \square\frac{K\xi}{n}\,,$$

L'inégalité (2.2) conduit à

$$L(\widehat{f}_m, f_m) \le \frac{1}{K}\mathrm{Var}(\gamma(\widehat{f}_m)-\gamma(f_m)) + \square K r_m^* + \square\frac{K\xi}{n}\,. \tag{2.31}$$

Dans le cas gaussien, la preuve s'arrête ici (étape 5 de la preuve de l'inégalité 2.25) car la fonction de perte L est la perte des moindres carrés qui correspond à la variance. Cependant, dans un cadre d'apprentissage, L représente la perte pour un *autre* contraste: afin d'obtenir une borne de risque, on utilise donc une *communication biais-variance*, c'est-à-dire un contrôle de la variance de l'erreur relative par l'espérance de la perte relative L. De façon générale, rien n'assure que cette communication biais-variance ait lieu et il faut le vérifier dans chaque cas. De façon intuitive, elle a lieu pour les contrastes réguliers. Elle est très importante dans la méthode de localisation et peut prendre plusieurs formes. Le cas le plus agréable est celui où la variance est directement dominée par le risque. Dans ce contexte, deux cas de figure sont a priori possibles.

- La communication idéale se fait autour du meilleur représentant du modèle S_m c'est-à-dire

$$\forall f \in S_m \, , \, \mathbb{E}\left[(\gamma(f) - \gamma(f_m))^2\right] \le c_m L(f, f_m) \, . \tag{2.32}$$

- Le cas de figure le plus crédible est celui où la communication a lieu autour de la cible s :

$$\forall f \in S_m \, , \, \mathbb{E}\left[(\gamma(f) - \gamma(s))^2\right] \le \tau L(f, s) \, . \tag{2.33}$$

  Dans certains cas, la constante τ peut dépendre de m.

Dans le cadre du chapitre 5, la non-compacité des espaces vectoriels nous oblige à travailler avec une communication biais-variance plus faible que celles considérées précédemment. En effet, par homogénéité, il est impossible d'avoir l'inégalité (2.33) avec S_m un espace vectoriel et γ la fonction de perte dite hinge-loss (il suffit d'utiliser plusieurs fois l'inégalité de Jensen). C'est pourquoi le chapitre 5 utilisera le poids de Z_r' (défini dans l'équation (2.28)) : il est adapté à une communication biais-variance "faible". Par souci de clarté, ce cas ne sera pas abordé dans la présente introduction. Grâce à un seuillage des estimateurs, le chapitre 6 se ramène sans perte de généralité à un cas borné où une communication biais-variance de la forme (2.33) a lieu sous certaines conditions.

Muni de toutes ces notions, on peut maintenant contrôler le *risque sur un modèle*. Dans le cas d'une communication biais-variance autour du meilleur représentant du modèle, les inégalités (2.32) et (2.31) impliquent qu'avec probabilité au moins $1 - e^{-\xi}$,

$$L(\widehat{f}_m, f_m) \le \frac{c_m}{K} L(\widehat{f}_m, f_m) + \square K r_m^* + \square \frac{K\xi}{n} \, ,$$

donc, pour $K > c_m$,

$$L(\widehat{f}_m, f_m) \le \square \frac{K^2}{K - c_m} r_m^* + \square \frac{K^2 \xi}{(K - c_m) n} \, , \tag{2.34}$$

c'est-à-dire

$$L(\widehat{f}_m, s) \le L(f_m, s) + \square \frac{K^2}{K - c_m} r_m^* + \square \frac{K^2 \xi}{(K - c_m) n} \, . \tag{2.35}$$

Cette inégalité est à rapprocher de la borne de risque (2.7) obtenue par l'approche globale. On constate que, sous l'hypothèse où les inégalités (2.32) et (2.30) sont vérifiées, la complexité globale est remplacée par une complexité locale : le point fixe r_m^* est calculé uniquement en contrôlant des supremum sur de petites boules autour de la fonction d'intérêt. Sauf cas pathologique, le terme de complexité r_m^* donné par l'approche locale sera donc nettement plus petit que celui donné par l'approche globale qui est $\mathbb{E}\left[\sup_{f \in S_m} (P - P_n)(\gamma(f))\right]$. Cependant, les constantes de l'inégalité (2.35) sont plus grandes que celles de (2.7) et la complexité dépend aussi de la communication biais-variance via c_m : meilleure est la communication biais-variance et meilleure sera la borne de risque obtenue. Si la communication biais-variance n'est pas bonne, l'approche locale peut devenir moins fine que celle de Vapnik.

Malgré sa formulation non-asymptotique, cette borne assure donc, dans les bons cas, une meilleure vitesse de convergence. Cependant, à cause des constantes, il n'est pas clair qu'elle fournisse une meilleure borne pour un échantillon de petite taille.

Le cadre de communication biais-variance plus général est celui où elle a lieu autour de s (inégalité (2.33)) : pour contrôler la variance, on introduit donc la cible s en utilisant que $(a+b)^2 \leq 2(a^2+b^2)$:

$$\mathbb{E}\left[(\gamma(\widehat{f}_m)-\gamma(f_m))^2\right] \leq 2\left(\mathbb{E}\left[(\gamma(\widehat{f}_m)-\gamma(s))^2\right]+\mathbb{E}\left[(\gamma(s)-\gamma(f_m))^2\right]\right),$$

puis par l'équation (2.33) :

$$\mathbb{E}\left[(\gamma(\widehat{f}_m)-\gamma(f_m))^2\right] \leq 2\tau(L(\widehat{f}_m,s)+L(f_m,s)).$$

Finalement, en utilisant l'inégalité (2.31), avec probabilité au moins $1-e^{-\xi}$,

$$L(\widehat{f}_m,f_m) \leq \frac{2\tau}{K}(L(\widehat{f}_m,s)+L(f_m,s))+\square K r_m^* + \square\frac{K\xi}{n}.$$

Afin d'obtenir une borne de risque, on remarque que la définition (1.2) de la perte L implique $L(\widehat{f}_m,f_m)=L(\widehat{f}_m,s)-L(f_m,s)$. Ainsi, on obtient qu'avec probabilité au moins $1-e^{-\xi}$,

$$L(\widehat{f}_m,s) \leq \frac{K+2\tau}{K-2\tau}L(f_m,s)+\square\frac{K^2}{K-2\tau}r_m^*+\square\frac{K^2}{K-2\tau}\frac{\xi}{n}, \tag{2.36}$$

si $K>2\tau$.

Comme dans le cas précédent, on garde la complexité reflétant la structure métrique locale des modèles : r_m^* ne dépend que du contrôle du module de continuité du processus empirique sur le modèle. Pour cette raison, r_m^* est appelé *complexity radius*. La différence majeure avec l'inégalité précédente et la borne (2.7) réside dans le terme d'approximation : il apparaît maintenant avec une constante $(K+2\tau)/(K-2\tau)$ strictement plus grande que 1. Afin d'illustrer les problèmes engendrés par ce terme d'approximation plus important, on peut noter que l'approche globale (inégalités (2.7) et (2.10)) assure que $L(\widehat{f}_m,f_m)$ tende vers 0 quand n tend vers l'infini alors que (2.34) n'assure que

$$\limsup_{n\to\infty} L(\widehat{f}_m,f_m) \leq \square L(f_m,s),$$

si r_m^* tend vers 0 quand n tend vers l'infini. Si l'erreur d'approximation $L(f_m,s)$ est non-nul, contrairement à l'approche de Vapnik, l'approche locale n'implique donc pas que $L(\widehat{f}_m,f_m)$ tende vers 0 quand n tend vers l'infini.

2.3.3 Moyenne de Rademacher localisée.

Comme le souligne la partie précédente, dans l'approche locale, les problèmes d'estimation sont gérés par des quantités ressemblant à des modules de continuité de processus empiriques,

$$\mathbb{E}\left[\sup_{f\in S_m,\,\mathrm{Var}(\gamma(f)-\gamma(f_m))\leq r}|(P-P_n)(\gamma(f)-\gamma(f_m))|\right].$$

Elles contrôlent la complexité du modèle. Le moyen le plus général et le plus pratique pour les borner est d'utiliser une technique de théorie des processus empiriques qui apparaît déjà dans l'approche globale : la symétrisation. Cette fois, elle est utilisée pour une famille localisée de fonctions et permet de majorer le module de continuité précédent par

$$2\mathbb{E}\left[\sup_{f\in S_m,\,\mathrm{Var}(\gamma(f)-\gamma(f_m))\leq r}\frac{1}{n}\Big|\sum_{i=1}^{n}\varepsilon_i(\gamma(f,(X_i,Y_i))-\gamma(f_m,(X_i,Y_i)))\Big|\right],$$

où $\varepsilon_1,\ldots,\varepsilon_n$ désignent n variables aléatoires indépendantes suivant une loi de Rademacher et indépendantes de l'échantillon ($\mathbb{P}[\varepsilon_1=1]=\mathbb{P}[\varepsilon_1=-1]=1/2$). Cette quantité est appelée *moyenne de Rademacher localisée* puisque le supremum est restreint aux fonctions de petite variance. En notant $\mathcal{F}'_m=\{\gamma(f)-\gamma(f_m),\ f\in S_m\}$, elle est de la forme

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}'_m,\,\mathrm{Var}(f)\leq r}\frac{1}{n}\Big|\sum_{i=1}^{n}\varepsilon_i f(Z_i)\Big|\right],$$

où $Z_i=(X_i,Y_i)$. Lorsque le contraste utilisé est une fonction lipschitzienne, [Mas00b] a remarqué qu'on pouvait localiser par rapport à la distance $P(f-f_m)^2$ et, par un argument de contraction ([LT91]), on peut se ramener à travailler avec la classe de fonctions $\mathcal{F}_m=\{f-f_m, f\in S_m\}$.

Par des arguments de concentration type [BLM00] (voir aussi [BBM03] pour notre situation), sous des conditions de bornitude, il suffit de contrôler *la moyenne de Rademacher conditionnée localisée empiriquement*

$$R_n(\mathcal{F}_m,r)=\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}_m,\,P_nf^2\leq r}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Z_i)\right],$$

où la notation $\mathbb{E}_\varepsilon$ signifie que l'on intègre uniquement par rapport aux ε : les variables Z_i, $i=1\ldots n$ sont "fixées". Un processus de Rademacher est de la forme $\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i t_i\right)_{t\in T}$ où T est une sous-partie de $\mathbb{R}^n$. L'argument de symétrisation permet donc de se ramener à l'étude du module de continuité d'un processus de Rademacher sur le modèle. Précisément, il est associé à $T=\{(f(Z_1),\ldots,f(Z_n)),\ f\in\mathcal{F}_m,\ P_nf^2\leq r\}$. Les moyennes de Rademacher localisées ont déjà été utilisées dans [LW03], [BBM03], [BMP04] et [Kol04] pour obtenir de meilleures performances et constituent un élément-clé pour la mise en oeuvre de l'approche locale.

Les arguments de *chaînage* permettent de contrôler des quantités du type $\mathbb{E}[\sup_{t\in T}X_t]$ où $(X_t)_{t\in T}$ est un processus dont les queues vérifient une certaine condition de décroissance. Initialement, ces quantités ont été étudiées pour obtenir des critères de continuité uniforme (p.s.) d'une version d'un processus gaussien ([Dud67]). Dans notre cas, le processus d'intérêt est celui de Rademacher (appelé aussi parfois de Bernoulli dans un cadre plus probabiliste). C'est un processus sous-gaussien puisqu'il vérifie l'inégalité (2.9) : les techniques de chaînage sont donc utilisables. L'argument de chaînage usuel consiste à considérer une *chaîne* d'approximations et à utiliser le Théorème 3 sur chacun des maillons. Ainsi, [Dud84] établit que

$$R_n(\mathcal{F}_m,r)\leq\square\frac{1}{n}\int_0^{\sqrt{r}}\sqrt{\log\mathcal{N}(u,\mathcal{F}_m,L_2(P_n))}du\,. \tag{2.37}$$

Cette inégalité fait intervenir l'intégrale de Koltchinskii-Pollard ([DL01]). Elle reflète que l'espérance du processus ne dépend essentiellement que de la structure métrique de l'espace des indices. La notion de taille d'un espace métrique utilisée dans ce résultat est l'entropie métrique $\log \mathcal{N}(u, \mathcal{F}_m, L_2(P_n))$ où $\mathcal{N}(u, \mathcal{F}_m, L_2(P_n))$ désigne le nombre de recouvrement de $\mathcal{F}_m$ pour la métrique $L_2(P_n)$: c'est le nombre minimal de boules pour la métrique $L_2(P_n)$ nécessaires au recouvrement de $\mathcal{F}_m$. Par comparaison avec l'approche globale de Vapnik, on peut noter que la notion de complexité utilisée ici est de nature métrique et non combinatoire. On constate aussi que les moyennes de Rademacher se prêtent bien à la localisation : l'argument de chaînage est facilement utilisable. Cette remarque est très largement utilisée pour contrôler la complexité d'algorithmes d'apprentissage (à titre d'exemple, on peut citer [Men02] pour le cas général et, plus spécifiquement [BLV03] pour le boosting). Il sera aussi utilisé pour l'analyse de la KPM du chapitre 6.

Il a été remarqué par Fermique et Talagrand qu'on pouvait rendre compte plus précisément de la complexité du modèle en utilisant des partitions au lieu des nombres de recouvrement : cette idée conduit au concept de mesure majorante et à une condition nécessaire et suffisante pour l'existence d'une version p.s. uniformément continue d'un processus gaussien ([Fer75], [Tal87]). Soit $r > 0$ et $(\mathcal{A}_j)_{j\geq 1}$ une suite de partitions de $\mathcal{F}(m,r) = \{f \in \mathcal{F}_m,\ P_n f^2 \leq r\}$ de diamètre r^{-j} par rapport à la distance $L_2(P_n)$ telle que $\mathcal{A}_{j+1}$ raffine $\mathcal{A}_j$ alors

$$R_n(\mathcal{F}_m, r) \leq \square \inf_{\pi\in\mathcal{M}_+^1(\mathcal{F}(m,r))} \sup_{f\in\mathcal{F}(m,r)} \sum_{j\geq 1} r^{-j}\sqrt{\log 1/\pi A_j(f)}\,,$$

où $\mathcal{M}_+^1(\mathcal{F}(m,r))$ désigne l'ensemble des mesures de probabilité sur $\mathcal{F}(m,r)$ et $A_j(f)$ est l'ensemble de la partition $\mathcal{A}_j$ contenant la fonction f. Cette inégalité est plus fine que (2.37) puisqu'en considérant une partition induite par un recouvrement minimal de rayon r^{-j}, on retrouve l'inégalité (2.37) aux constantes près. Cette nouvelle forme de chaînage porte le nom de *chaînage générique* ([Tal05]). [AB04] l'utilise dans un cadre d'apprentissage mais on ne connaît pas encore de cas d'analyse d'algorithme où le chaînage serait insuffisant et où il faudrait recourir au chaînage générique.

Quelques Exemples. Les moyennes de Rademacher localisées seront utilisées tout au long de cette thèse pour contrôler la complexité des modèles intervenant dans différents algorithmes d'apprentissage. Par exemple, le contrôle de Rademacher localisée sur lequel repose le chapitre 3 concernant la KPCA (voir partie 1.2) est le suivant. Définissons

$$\widetilde{\mathcal{F}}_d = \left\{ x \mapsto \|\Pi_{V^\perp}(k(x,.))\|^2 - \|\Pi_{V_d^\perp}(k(x,.))\|^2, V \in \mathcal{V}_d \right\}\ ,$$

où k est un noyau, $\Pi_{V^\perp}$ est la projection orthogonale sur $V^\perp$, $\mathcal{V}_d$ est l'ensemble des sous-espaces de dimension d du RKHS associé à k et V_d est l'espace de dimension d engendré par les fonctions propres de l'opérateur intégral T_k (défini par l'équation (1.26)) associées aux d plus grandes valeurs propres. Alors

$$\mathbb{E}\left[\sup_{f\in\widetilde{\mathcal{F}}_d, Pf^2\leq r} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right] \leq \frac{1}{\sqrt{n}} \inf_{h\geq 0}\left\{ \sqrt{rh} + 2\sqrt{d\sum_{j>h}\lambda_j(K_2)} \right\}\ ,$$

où $(\lambda_j(K_2))_{j\geq 1}$ désignent les valeurs propres rangées par ordre décroissant de l'opérateur intégral T_{k^2} défini par l'équation (1.26) associé au noyau k^2.

Les moyennes de Rademacher localisées ont aussi été étudiées pour analyser l'algorithme des SVM (voir [SS03] et [BBM04]). Le Théorème 41 de [Men02] stipule que, sous certaines conditions peu restrictives,

$$\mathbb{E}\left[\sup_{f\in B_{\mathcal{H}}(0,1),Pf^2\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]\leq \square\frac{1}{\sqrt{n}}\left(\sum_{i\geq 1}\lambda_i\wedge r\right)^{1/2}, \tag{2.38}$$

où $(\lambda_i)_{i\geq 1}$ désignent les valeurs propres de l'opérateur intégral T_k défini par l'équation (1.26) et $\mathcal{H}$ est le RKHS associé au noyau k. De plus, la référence précédemment citée montre que cette majoration est optimale. Ce résultat met en évidence le contrôle de la complexité des modèles utilisés par la SVM par les valeurs propres de l'opérateur intégral. Ce type de résultat motive l'étude des propriétés de concentration des valeurs propres de la matrice de Gram autour de celles de l'opérateur intégral du chapitre 3.

Dans le cas de la projection fini-dimensionnelle, le chapitre 5 repose sur le Théorème 5.8.5 qui stipule que

$$\mathbb{E}\left[\sup_{g\in S_D,Pg^2\leq r}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i)\right|\right]\leq\sqrt{\frac{rD}{n}},$$

où S_D désigne un espace vectoriel de dimension D. Contrairement au cas de la SVM, la complexité des modèles est indépendante des valeurs propres de l'opérateur intégral. On peut aussi remarquer une autre différence : dans le cas des SVM, sous des conditions peu restrictives sur le noyau, les modèles considérés sont compacts et on peut contrôler les moyennes de Rademacher globales associées (il suffit de faire tendre r vers l'infini dans l'équation (2.38)). Dans le cas de la projection fini-dimensionnelle, la situation est différente : l'approche globale n'a aucune chance d'aboutir puisque les espaces vectoriels ne sont pas compacts. Précisément, on peut montrer facilement que, sauf situation triviale, $\mathbb{E}\left[\sup_{g\in S_D}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i)\right]=\infty$. L'approche locale est donc, en un certain sens, la seule analyse possible. Les moyennes de Rademacher sont alors considérées sur des boules qui sont compactes puisque les espaces vectoriels considérés sont de dimension finie (et donc localement compacts). C'est l'analyse que mène le chapitre 5 avec un succès discutable. Le chapitre 6 montrera qu'on peut se ramener sans perte de généralité à un cadre plus favorable et la moyenne de Rademacher localisée considérée sera contrôlée explicitement grâce à l'inégalité de chaînage (2.37).

L'approche locale permet aussi d'affiner les propriétés d'optimalité minmax du minimiseur empirique du risque associé au contraste dur de la classification. Les vitesses données dans la partie suivante serviront de point de repère pour celles obtenues par les algorithmes de classification.

2.3.4 Classification.

Revenons au cadre de la classification. S_m est le modèle constitué des fonctions $\{2\mathbb{1}_A-1, A\in\mathcal{C}_m\}$ où $\mathcal{C}_m$ est une classe d'ensembles de dimension de Vapnik-Chervonenkis V_m. On rappelle aussi que $\ell(f,f^*)=\mathbb{P}[Y\neq f(X)]-\mathbb{P}[Y\neq f^*(X)]$ et f^* est la fonction de classification de Bayes défini par

$$f^*(x)=\begin{cases}1 & si\ \eta(x)\geq\frac{1}{2}\,,\\ -1 & si\ \eta(x)<\frac{1}{2}\,,\end{cases}$$

où $\eta(x) = \mathbb{P}[Y = 1|X = x]$. Dans cette partie, f^* est estimé par

$$\widehat{f}_m = \arg\min_{f \in S_m} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq f(X_i)} .$$

Q désigne la loi du couple (X, Y).

Le Théorème 3.1 de [NM03] montre que l'ordre de grandeur du risque minmax change de façon drastique en présence de marge : si $n \geq V_m$,

$$\inf_{\widetilde{f} \in S_m} \sup_{Q \in \mathcal{P}(h, S_m)} \mathbb{E}\left[\ell(f^*, \widetilde{f})\right] \geq \square \left[\left(\frac{V_m}{nh}\right) \wedge \sqrt{\frac{V_m}{n}}\right] . \tag{2.39}$$

où l'infimum est considéré sur tous les estimateurs $\widetilde{f}$ construit à partir de l'échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_n.Y_n)$ et

$$\mathcal{P}(h, S_m) = \{Q, |2\eta(x) - 1| \geq h\,, \forall x \in \mathcal{X}\; et\; f^* \in S_m\} .$$

Ce résultat quantifie la déformation entre le risque minmax en $\frac{V_m}{n}$ du cas zero-error ($h = 1$) et celui en $\sqrt{\frac{V_m}{n}}$ du pire cas (qui correspond à l'approche globale de Vapnik développée précédemment) en fonction du paramètre de marge h. Si la quantité de marge n'est pas suffisante ($h \leq \sqrt{\frac{V_m}{n}}$), elle n'influe pas sur l'ordre de grandeur du risque minmax et on retrouve le risque en $\sqrt{\frac{V_m}{n}}$ de l'équation (2.12).

En s'inspirant des travaux de [MT95] et [Tsy04], [NM03] utilise l'approche locale pour montrer que $\widehat{f}_m$ est approximativement minmax sur $\mathcal{P}(h, S_m)$.

Soit d la distance $d(f, g) = P(f - g)^2$. Elle vérifie

$$\mathbb{E}\left[(\mathbb{1}_{Y \neq f(X)} - \mathbb{1}_{Y \neq g(X)})^2\right] = \frac{1}{4} d^2(f, g) .$$

De plus, en utilisant l'égalité (1.3), sous l'hypothèse de marge (c'est-à-dire qu'on suppose que $\forall x \in \mathcal{X}\,, |2\eta(x) - 1| \geq h$), on a

$$d^2(f, f^*) \leq \frac{4}{h} \ell(f, f^*) .$$

En classification, l'hypothèse de marge permet donc d'avoir une communication biais-variance autour de la fonction de classification de Bayes (du type de l'inégalité (2.33) avec $\tau = \frac{4}{h}$). Plus la marge est importante et meilleure est la communication biais-variance : la marge quantifie donc le degré de difficulté du problème de classification en mesurant sa distance avec un problème gaussien (dans lequel la communication biais-variance est une égalité).

Par définition du risque minmax, le classifieur de Bayes est supposé *a priori* être dans le modèle S_m et on est donc dans le cadre d'une communication biais-variance autour du meilleur du modèle (du type de l'inégalité (2.32)). L'approche locale utilisée dans [NM03] permet d'obtenir que, si $h > \sqrt{\frac{V_m}{n}}$,

$$\sup_{Q \in \mathcal{P}(h, S_m)} \mathbb{E}\left[\ell(\widehat{f}_m, f^*)\right] \leq \square \frac{V_m(1 + \log(nh^2))}{nh} . \tag{2.40}$$

L'inégalité (2.11) se reformule ici : si $h \leq \sqrt{\frac{V_m}{n}}$,

$$\sup_{Q \in \mathcal{P}(h, S_m)} \mathbb{E}\left[\ell(\widehat{f}_m, f^*)\right] \leq \square\sqrt{\frac{V_m}{n}}. \tag{2.41}$$

Finalement, l'approche locale permet de raffiner la borne de risque de Vapnik si la quantité de marge est suffisante. De par les inégalités (2.39), (2.40) et (2.41), le minimiseur empirique reste approximativement minmax (à un facteur logarithme près) si on tient compte de la marge, ce qui correspond à se placer dans des cas moins défavorables que ceux envisagés par Vapnik.

Part I

Kernel Principal Component Analysis and Kernel Matrix

Chapter 3

Statistical Properties of Kernel Principal Component Analysis

Most of this chapter is based on [BBZ04]. The extended work presented here is a submitted paper. This is a joint work with G. Blanchard and O. Bousquet.

Contents

Abstract

The main goal of this chapter is to prove non-asymptotic inequalities on the reconstruction error for Kernel Principal Component Analysis. Our contribution to this topic is two-fold: (1) we give bounds that explicitly take into account the empirical centering step in this algorithm, and (2) we show that a "localized" approach allows to show fast rates of convergence towards the minimum reconstruction error, more precisely we prove that the convergence rate is related to the decay of eigenvalues and can typically be faster than $n^{-1/2}$.

A secondary goal, for which we present similar contributions, is to obtain convergence bounds for the partial sums of the biggest or smallest eigenvalues of the Gram matrix towards eigenvalues of the corresponding kernel operator. These quantities are naturally linked to the KPCA procedure; furthermore these results can have applications to the study of various other kernel algorithms.

The results are presented in a functional analytic framework, which is suited to deal rigorously with reproducing kernel Hilbert spaces of infinite dimension.

3.1 Introduction

Due to their versatility, kernel methods are currently very popular as data-analysis tools. In such algorithms, the key object is the so-called kernel matrix (the Gram matrix built on the data sample) and it turns out that its spectrum can be related to the performance of the algorithm. This has been shown in particular in the case of Support Vector Machines ([WSTSS99]). Studying the behavior of eigenvalues of kernel matrices, their stability and how they relate to the eigenvalues of the corresponding kernel integral operator is thus crucial for understanding the statistical properties of kernel-based algorithms.

In the present work we focus on Principal Component Analysis (PCA), and its non-linear variant, kernel-PCA, which are widely used algorithms in data analysis. Their goal is to extract a basis adapted to the data, by looking for directions where the variance is maximized. Their applications are very diverse, ranging from dimensionality reduction to denoising. Applying PCA to a space of functions rather than a space of vectors was first proposed by [Bes79] (see also the survey of [RD91]). Kernel-PCA ([SSM99]) is an instance of such a method which has boosted the interest in PCA as it allows to overcome the limitations of linear PCA in a very elegant manner.

Despite being a relatively old and commonly used technique, little has been done on analyzing the statistical performance of PCA. Most of the previous work has focused on the asymptotic behavior of empirical covariance matrices of Gaussian vectors ([And63]). For the kernelized version, there is a tight connection between the covariance and the kernel matrix of the data. This is actually at the heart of the kernel-PCA algorithm itself, and also indicates that the properties of the kernel matrix, in particular its spectrum, play a crucial role in the properties of the kernel-PCA algorithm.

Recently [STWCK02, STWCK05] have undertaken an investigation of the properties of the eigenvalues of kernel matrices and related it to the statistical performance of kernel-PCA. Our goal in the present work is mainly to extend their results in several directions:

- In practice, for PCA or KPCA, an (empirical) recentering of the data is generally performed. This is because PCA is viewed as a technique to analyze the *variance* of the data; it is often desirable to treat the mean independently as a preliminary step (although, arguably it is also feasible to perform PCA on uncentered data). This centering was not considered in the cited previous work while we take this step into account explicitly and show that it leads to comparable convergence properties.

- to control the estimation error, [STWCK02, STWCK05] use what we would call a *global approach* which typically leads to convergence rates of order $n^{-1/2}$. Numerous recent theoretical works on M-estimation have shown that improved rates can be obtained by using a so-called *local approach*, which very coarsely speaking consists in taking the estimation variance precisely into account. We refer the reader to the works of [Mas00b], [BBM03], [BJM03] (among others). Here we show that this principle leads to improved convergence bounds.

Note that we consider these two types of extension *separately*, not simultaneously. While we believe it possible to combine these two extensions, in the framework of this chapter we choose to treat them independently to avoid additional technicalities and therefore leave this issue as an open problem.

To state and prove our results we have chosen to use a functional analysis formalism. Its main justification is that some of the most interesting positive definite kernels (e.g., the Gaussian RBF kernel) generate an infinite dimensional reproducing kernel Hilbert space (the "feature space" into which the data is mapped). This infinite dimensionality potentially raises a technical difficulty. In part of the literature on kernel methods a matrix formalism of finite-dimensional linear algebra is used for the feature space and it is generally assumed more or less explicitly that the results "carry over" to infinite dimension because (separable) Hilbert spaces have good regularity properties. In the present work we wanted to state rigorous results directly in an infinite-dimensional space using the corresponding formalism of Hilbert-Schmidt operators and of random variables in Hilbert spaces. We hope the necessary notational background which we introduce first will not tax the reader excessively and hope to convince her that it leads to a more rigorous and elegant analysis.

Finally, let us emphasize some open problems that will be discussed in more detail in the concluding part. We want to underline that in our results we consider the number of components d kept in the PCA procedure (or the number of eigenvalues) as a fixed constant. Our focus here is in the dependence of the bounds in the sample size n. As for the dependence in d for fixed n, unfortunately it is clear that our results do not capture the correct behavior: our bound on the reconstruction error eventually increases as a function of d while basic considerations show that the true reconstruction error is always decreasing in d. In other words, for fixed n there exists a certain dimension $d(n)$ such that the bound obtained for $d' > d(n)$ is actually *less informative* than the bound obtained for $d(n)$. The same issue surfaces in the work of [STWCK05] and as far as we know, this problem has not been solved. An indirectly linked issue is how to define a sensible criterion for what would be an optimal dimension choice in KPCA. Obviously the (true) reconstruction error alone is not enough since it is always a decreasing function of the dimension. We believe these two open issues to be most interesting for future research.

The chapter is organized as follows. Section 3.2 introduces the necessary background on functional analysis, the basic assumptions and some preliminary fundamental results. Section 3.3 concentrates on bounding the difference between sums of eigenvalues of the kernel matrix and of the associated kernel operator. Finally, Section 3.4 gives our main results, bounds on the reconstruction error of kernel-PCA. We conclude with an extended discussion on the open issues sketched above.

3.2 Preliminaries

The core of our results is concerned with estimating eigenvalues of certain operators in a reproducing Hilbert kernel space $\mathcal{H}_k$. The most convenient way to deal with these objects is to use formalisms from functional analysis, and in particular to introduce the space of Hilbert-Schmidt operators on $\mathcal{H}_k$ endowed with a suitable Hilbert structure. The present section is devoted to introducing the necessary notation and base properties that will be used repeatedly.

3.2.1 The Hilbert space of Hilbert-Schmidt operators

The provided material of this section can be found in [DS63]. Let $\mathcal{H}$ be a separable Hilbert space. A linear operator L from $\mathcal{H}$ to $\mathcal{H}$ is called Hilbert-Schmidt if

$$\sum_{i\geq 1} \|Le_i\|^2_{\mathcal{H}} = \sum_{i,j\geq 1} \langle Le_i, e_j\rangle^2 < \infty\,,$$

where $(e_i)_{i\geq 1}$ is an orthonormal basis of $\mathcal{H}$. This sum is independent of the chosen orthonormal basis and is the squared of the Hilbert-Schmidt norm of L when it is finite. The set of all Hilbert-Schmidt operators on $\mathcal{H}$ is denoted by $\mathrm{HS}(\mathcal{H})$. Endowed with the following inner product $\langle L, N\rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{i\geq 1} \langle Le_i, Ne_i\rangle = \sum_{i,j\geq 1} \langle Le_i, e_j\rangle \langle Ne_i, e_j\rangle$, it is a separable Hilbert space.

A Hilbert-Schmidt operator is compact, it has a countable spectrum and an eigenspace associated to a non-zero eigenvalue is of finite dimension. A compact, self-adjoint operator on a Hilbert space can be diagonalized, i.e., there exists an orthonormal basis of $\mathcal{H}$ made of eigenfunctions of this operator. If L is a compact, positive self-adjoint operator, we will denote $\lambda(L) = (\lambda_1(L) \geq \lambda_2(L) \geq \dots)$ the sequence of its *positive* eigenvalues sorted in non-increasing order, repeated according to their multiplicities; this sequence is well-defined and contains all nonzero eigenvalues since these are all nonnegative and the only possible limit point of the spectrum is zero. Note that $\lambda(L)$ may be a finite sequence. An operator L is called trace-class if $\sum_{i\geq 1} \langle e_i, Le_i\rangle$ is a convergent series. In fact, this series is independent of the chosen orthonormal basis and is called the trace of L, denoted by $\mathrm{tr}\, L$. Moreover, $\mathrm{tr}\, L = \sum_{i\geq 1} \lambda_i(L)$ for a self-adjoint operator L.

We will keep switching from $\mathcal{H}$ to $\mathrm{HS}(\mathcal{H})$ and treat their elements as vectors or as operators depending on the context. At times, for more clarity we will index norms and dot products by the space they are to be performed in, although this should always be clear from the objects involved. The following summarizes some notation and identities that will be used in the sequel.

Rank one operators. For $f, g \in \mathcal{H}\backslash\{0\}$ we denote by $f \otimes g^*$ the rank one operator defined as $f \otimes g^*(h) = \langle g, h\rangle f$. The following properties are straightforward from the above definitions:

$$\|f \otimes g^*\|_{\mathrm{HS}(\mathcal{H})} = \|f\|_{\mathcal{H}}\, \|g\|_{\mathcal{H}}\ ; \tag{3.1}$$

$$\mathrm{tr}\, f \otimes g^* = \langle f, g\rangle_{\mathcal{H}}\ ; \tag{3.2}$$

$$\langle f \otimes g^*, A\rangle_{\mathrm{HS}(\mathcal{H})} = \langle Ag, f\rangle_{\mathcal{H}} \quad \text{for any } A \in \mathrm{HS}(\mathcal{H})\,. \tag{3.3}$$

Orthogonal projectors. We recall that an orthogonal projector in $\mathcal{H}$ is an operator U such that $U^2 = U = U^*$ (hence positive). In particular one has

$$\|U(h)\|^2_{\mathcal{H}} = \langle h, Uh\rangle_{\mathcal{H}} \leq \|h\|^2_{\mathcal{H}} \;;$$
$$\langle f \otimes g^*, U\rangle_{\mathrm{HS}(\mathcal{H})} = \langle Uf, Ug\rangle_{\mathcal{H}} \;.$$

U has rank $d < \infty$ (i.e., it is a projection on a finite dimensional subspace), if and only if it is Hilbert-Schmidt with

$$\|U\|_{\mathrm{HS}(\mathcal{H})} = \sqrt{d}\,, \tag{3.4}$$
$$\operatorname{tr} U = d\,. \tag{3.5}$$

In that case it can be decomposed as $U = \sum_{i=1}^d \phi_i \otimes \phi_i^*$, where $(\phi_i)_{i=1}^d$ is an orthonormal basis of the image of U.

If V denotes a closed subspace of $\mathcal{H}$, we denote by Π_V the unique orthogonal projector such that $\operatorname{range}\Pi_V = V$ and $\ker \Pi_V = V^{\perp}$. When V is of finite dimension, $\Pi_{V^\perp}$ is not Hilbert-Schmidt, but we will denote (with some abuse of notation), for a trace-class operator A,

$$\langle \Pi_{V^\perp}, A\rangle := \operatorname{tr} A - \langle \Pi_V, A\rangle \;. \tag{3.6}$$

Eigenvalues formulas. We denote by $\mathcal{V}_d$ the set of subspaces of dimension d of $\mathcal{H}$. The following theorem sums up important formulas concerning the individual or summed eigenvalues of self-adjoint compact operators; the first one is due to Fan (see [Tor97] for a proof) while the second is the so-called Courant-Fischer-Weyl formula (see e.g.[DS63]).

Theorem 3.2.1. *Let C a compact self-adjoint operator on $\mathcal{H}$, then for all $d \geq 0$,*

$$\sum_{i=1}^d \lambda_i(C) = \max_{V \in \mathcal{V}_d} \langle \Pi_V, C\rangle_{\mathrm{HS}(\mathcal{H})}\,, \tag{3.7}$$

$$\text{and} \qquad \lambda_{d+1}(C) = \min_{V\in\mathcal{V}_d} \max_{f \perp V} \frac{\langle f, Cf\rangle}{\|f\|^2}\,, \tag{3.8}$$

where in both cases, the optimum is attained when V is the span of the first d eigenvectors of C.

3.2.2 Second order integral operators

We recall basic facts about random elements in Hilbert spaces. A random variable Z in a separable Hilbert space has an expectation $e \in \mathcal{H}$ when $\mathbb{E}\,\|Z\| < \infty$ and e is the unique vector satisfying $\langle e, f\rangle_{\mathcal{H}} = \mathbb{E}\,\langle Z, f\rangle_{\mathcal{H}}\,, \forall f \in \mathcal{H}$. We now introduce the (noncentered) covariance operator through this theorem and definition:

Theorem 3.2.2. *If $\mathbb{E}\,\|Z\|^2 < \infty$, there exists a unique operator $C : \mathcal{H} \to \mathcal{H}$ such that*

$$\langle f, Cg\rangle_{\mathcal{H}} = \mathbb{E}\left[\langle f, Z\rangle_{\mathcal{H}}\,\langle g, Z\rangle_{\mathcal{H}}\right], \forall f, g \in \mathcal{H}\,.$$

This operator is self-adjoint, positive, trace-class with $\operatorname{tr} C = \mathbb{E}\,\|Z\|^2$, and satisfies

$$C = \mathbb{E}\left[Z \otimes Z^*\right] \,.$$

This result is proved in the appendix. We call C the *noncentered covariance* operator of Z.

The core property of covariance operators that we will use is its intimate relationship with another integral operator summarized in the next theorem. This property was first used in a similar but more restrictive context (finite dimensional) by [STWCK02, STWCK05].

Theorem 3.2.3. *Let $(\mathcal{X}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space, X be a $\mathcal{X}$-valued random variable and Φ be a map from $\mathcal{X}$ to $\mathcal{H}$ such that for all $h \in \mathcal{H}$, $\langle h, \Phi(.)\rangle$ is measurable and $\mathbb{E}\left\|\Phi(X)\right\|^2 < \infty$. Let C_Φ be the covariance operator associated to $\Phi(X)$ and $K_\Phi : L_2(P) \to L_2(P)$ be the integral operator defined as*

$$(K_\Phi f)(t) = \mathbb{E}\left[f(X)\left\langle \Phi(X), \Phi(t)\right\rangle\right] = \int f(x)\left\langle \Phi(x), \Phi(t)\right\rangle dP(x)\,.$$

Then K is a Hilbert-Schmidt, positive self-adjoint operator, and

$$\lambda(K_\Phi) = \lambda(C_\Phi)\,.$$

In particular, K_Φ is a trace-class operator and $\operatorname{tr}(K_\Phi) = \mathbb{E}\left\|\Phi(X)\right\|^2 = \sum_{i\geq 1}\lambda_i(K_\Phi)$.

This result is proved in the appendix. If we denote $\langle\Phi(x), \Phi(y)\rangle = k(x, y)$, then K_Φ is called the integral operator with kernel k.

3.2.3 Main framework and assumptions

Let $\mathcal{X}$ denote the input space (an arbitrary measurable space) and P denote a distribution on $\mathcal{X}$ according to which the data is sampled i.i.d. We will denote by P_n the empirical measure associated to a sample $X_1, \ldots, X_n$ from P, i.e., $P_n = \frac{1}{n}\sum \delta_{X_i}$. With some abuse of notation, for a function $f : \mathcal{X} \to \mathbb{R}$, we may use the notation $Pf := \mathbb{E}\left[f(X)\right]$ and $P_n f := \frac{1}{n}\sum_{i=1}^n f(X_i)$.

Let k be a positive definite function on $\mathcal{X}$ and $\mathcal{H}_k$ the associated reproducing kernel Hilbert space (RKHS for short in the sequel). We recall (see ,e.g., [Aro50]) that $\mathcal{H}_k$ is a Hilbert space of real functions on $\mathcal{X}$, containing functions $k(x, \cdot)$ for all $x \in \mathcal{H}_k$ and such that the following *reproducing property* is satisfied:

$$\forall f \in \mathcal{H}_k \qquad \forall x \in \mathcal{X} \qquad \langle f, k(x, .)\rangle = f(x)\,, \tag{3.9}$$

and in particular

$$\forall x\,, y \in \mathcal{X} \qquad \langle k(x, \cdot), k(y, \cdot)\rangle = k(x, y)\,.$$

Finally, let $\mathcal{V}_d$ denote the set of all vector subspaces of dimension d of $\mathcal{H}_k$.

We will always work with the following assumptions which we will refer collectively to as "assumption **(A)**" in the sequel:

(A1) *$\mathcal{H}_k$ is separable.*

(A2) *For all $x \in \mathcal{X}$, $k(x, .)$ is P-measurable.*

(A3) *There exists $M > 0$ such that $k(X, X) \leq M$ P-almost surely.*

Note that assumptions **(A1)**-**(A2)** ensure the measurability of all functions in $\mathcal{H}_k$ since they are obtained by linear combinations and pointwise limits of functions $k(x, \cdot)$.

Notation for the noncentered case. For $x \in \mathcal{X}$, we denote

$$\varphi_x = k(x,.) \in \mathcal{H}_k\,,$$
$$C_x = \varphi_x \otimes \varphi_x^* \in \mathrm{HS}(\mathcal{H}_k)\,.$$

The following properties are then straightforward from the preceding sections:

$$\operatorname{tr} C_x = \|C_x\|_{\mathrm{HS}(\mathcal{H}_k)} = k(x,x)\,, \tag{3.10}$$
$$\langle C_x, C_y\rangle_{\mathrm{HS}(\mathcal{H}_k)} = k^2(x,y)\,, \tag{3.11}$$
$$\langle f, C_x g\rangle_{\mathcal{H}_k} = \langle C_x, f\otimes g^*\rangle_{\mathrm{HS}(\mathcal{H}_k)} = f(x)g(x)\,, \tag{3.12}$$

and for any orthogonal projector U,

$$\langle U, C_x\rangle_{\mathrm{HS}(\mathcal{H}_k)} = \|U\varphi_x\|^2_{\mathcal{H}_k}\,. \tag{3.13}$$

Note incidentally that (3.11) implies that $\mathrm{HS}(\mathcal{H}_k)$ is actually a natural representation of the RKHS with kernel $k^2(x,y)$. Namely to an operator $A \in \mathrm{HS}(\mathcal{H}_k)$ we can associate the function

$$f_A(x) = \langle A, C_x\rangle_{\mathrm{HS}(\mathcal{H}_k)} = \langle A\varphi_x, \varphi_x\rangle_{\mathcal{H}_k} = (A\varphi_x)(x)\,;$$

with this notation, we have $f_{C_x} = k^2(x,\cdot)$, and one can check that (3.9) is satisfied in $\mathrm{HS}(\mathcal{H}_k)$ with the kernel $k^2(x,y)$ when identifying an operator to its associated function. Also, paralleling the earlier remark about measurability of functions in $\mathcal{H}_k$, assumptions **(A1)-(A2)** ensure that f_A is measurable for any A.

Now, let us denote $C_1 : \mathcal{H}_k \to \mathcal{H}_k$, resp. $C_2 : \mathrm{HS}(\mathcal{H}_k) \to \mathrm{HS}(\mathcal{H}_k)$, the noncentered covariance operator associated to the random element φ_X in $\mathcal{H}_k$, resp. C_X in $\mathrm{HS}(\mathcal{H}_k)$; and $K_1, K_2 : L_2(P) \to L_2(P)$ the integral operators with kernel $k(x,y)$, resp. $k^2(x,y)$ (Note that all these operators are well-defined due to assumption **(A)**). We then have the following property:

Lemma 3.2.4. *Under assumption* **(A)** *the operators C_1, C_2, K_1, K_2 defined above satisfy the following :*

(i) C_1 *is the expectation in* $\mathrm{HS}(\mathcal{H}_k)$ *of* $C_X = \varphi_X \otimes \varphi_X^*$.

(ii) C_2 *is the expectation in* $\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))$ *of* $C_X \otimes C_X^*$.

(iii) $\lambda(C_1) = \lambda(K_1)$, *and* $\operatorname{tr} C_1 = \operatorname{tr} K_1 = \mathbb{E}\left[k(X,X)\right]$.

(iv) $\lambda(C_2) = \lambda(K_2)$, *and* $\operatorname{tr} C_2 = \operatorname{tr} K_2 = \mathbb{E}\left[k^2(X,X)\right]$.

This Lemma is a direct consequence of Theorems 3.2.2 and 3.2.3. (noting that the measurability conditions have been established in the preceding discussions).

Notation for the recentered case. We will be interested in the sequel in recentered versions of the above quantities (which appear for standard covariance operators and PCA techniques), which we now define accordingly. Let us define for all $x \in \mathcal{X}$

$$\mu = \mathbb{E}\left[\varphi_X\right]\,,$$
$$\overline{\varphi}_x = \varphi_x - \mu \in \mathcal{H}_k\,,$$
$$\overline{C}_x = \overline{\varphi}_x \otimes \overline{\varphi}_x^* \in \mathrm{HS}(\mathcal{H}_k)\,;$$

we then have $\|\mu\|^2 = \mathbb{E}k(X, X')$ and

$$\operatorname{tr} \overline{C}_x = \left\|\overline{C}_x\right\|_{\mathrm{HS}(\mathcal{H}_k)} = \|\varphi_x - \mu\|^2 = k(x,x) + \mathbb{E}k(X,X') - 2\mathbb{E}k(X,x)\,,$$

where X' denotes an independent copy of X.

Similarly, let us denote $\overline{C}_1$ the covariance operator associated to $\overline{\varphi}_X$, and $\overline{K}_1$ the integral operator with kernel $\overline{k}(x,y) = \langle \varphi_x - \mu, \varphi_y - \mu\rangle = k(x,y) - \mathbb{E}k(X,x) - \mathbb{E}k(X,y) + \mathbb{E}k(X,X')$; then the following holds:

Lemma 3.2.5. *Under assumption* **(A)** *the operators $\overline{C}_1, \overline{K}_1$ defined above satisfy the following :*

(i) $\overline{C}_1$ is the expectation in $\mathrm{HS}(\mathcal{H}_k)$ *of $\overline{C}_X = \overline{\varphi}_X \otimes \overline{\varphi}_X^*$; moreover one has*

$$\overline{C}_1 = C_1 - \mu \otimes \mu^* \tag{3.14}$$

(ii) $\lambda(\overline{C}_1) = \lambda(\overline{K}_1)$ *, and* $\operatorname{tr} \overline{C}_1 = \operatorname{tr} \overline{K}_1 = \mathbb{E}\left[k(X,X)\right] - \mathbb{E}\left[k(X,X')\right]$.

Again, this lemma is a direct consequence of Theorems 3.2.2 and 3.2.3 and of straightforward computations.

Notations for the empirical case. In the following we will study empirical counterparts of the above quantities. The generality of the above results implies that we can replace the distribution P by the empirical measure P_n associated to an i.i.d. sample $X_1, \ldots, X_n$ without any changes and we merely need to introduce adequate notation.

In the noncentered case, the corresponding operators are denoted by $K_{1,n}$ and $C_{1,n}$; we define $K_{2,n}$ and $C_{2,n}$ similarly. In particular, Lemma 3.2.4 implies that $\lambda(K_{1,n}) = \lambda(C_{1,n})\,, \lambda(K_{2,n}) = \lambda(C_{2,n})\,, \operatorname{tr} K_{1,n} = \operatorname{tr} C_{1,n} = \frac{1}{n}\sum_{i=1}^n k(X_i, X_i)$, and $\operatorname{tr} K_{2,n} = \operatorname{tr} C_{2,n} = \frac{1}{n}\sum_{i=1}^n k^2(X_i, X_i)$.

Note that $C_{1,n}$ is the empirical covariance operator, i.e., $\langle f, C_{1,n} g\rangle = \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$. An important point is that $K_{1,n}$ can be identified (as in [KG00]) with the normalized kernel matrix of size $n \times n$, $K_{1,n} \equiv (k(X_i, X_j)/n)_{i,j=1,\ldots,n}$. This comes from the fact that $L_2(P_n)$ is a finite-dimensional space so that any function $f \in L_2(P_n)$ can be identified to the n-uple $(f(X_i))_{i=1,\ldots,n}$; this way the Hilbert structure of $L_2(P_n)$ is isometrically mapped into $\mathbb{R}^n$ embedded with the standard Euclidian norm rescaled by n^{-1} (note that this mapping may not be *onto* in the case where two datapoints are identical, but this does not cause a problem).

For the centered case, note that the quantities $\overline{\varphi}_x, \overline{C}_x$ already depend on P through the centering, so that we will define the corresponding quantities for P_n with an index n:

$$\begin{aligned}
\overline{\varphi}_{x,n} &= \varphi_x - \frac{1}{n}\sum_{i=1}^n \varphi_{X_i}\,,\\
\overline{C}_{x,n} &= \overline{\varphi}_{x,n} \otimes \overline{\varphi}_{x,n}^*\,,\\
\overline{C}_{1,n} &= \frac{1}{n}\sum_{i=1}^n \overline{\varphi}_{X_i,n} \otimes \overline{\varphi}_{X_i,n}^* = \frac{1}{n}\sum_{i=1}^n \overline{C}_{X_i,n}\,.
\end{aligned}$$

The associated centered kernel operator is denoted $\overline{K}_{1,n}$ and identified with the following centered kernel matrix :

$$\overline{K}_{1,n} \equiv \left(\left\langle \overline{\varphi}_{X_i}, \overline{\varphi}_{X_j}\right\rangle_{\mathcal{H}_k}\right)_{1\le i,j\le n} = \left(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n'\right) K_{1,n} \left(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n'\right)\,.$$

where $\mathbb{1}_n = (1, \cdots, 1)' \in \mathbb{R}^n$. As a consequence of Lemma 3.2.5, we have $\lambda(\overline{C}_{1,n}) = \lambda(\overline{K}_{1,n})$. Finally, note that $\overline{C}_{1,n}$ is a biased estimator of $\overline{C}_1$, so we will additionally introduce

$$\widetilde{C}_{1,n} = \frac{n}{n-1}\overline{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi^*_{X_j} , \tag{3.15}$$

which satisfies $\mathbb{E}\left[\widetilde{C}_{1,n}\right] = \overline{C}_1$.

3.3 General Results on Eigenvalues of Gram Matrices

We are now able to proceed to our first goal, the estimation of sums of eigenvalues of kernel operators K_1 or $\overline{K}_1$ from eigenvalues of their empirical counterparts $K_{1,n}$ and $\overline{K}_{1,n}$. For this, we will make use of the preliminary results to relate these sums of eigenvalues to an empirical process on classes of functions of type $x \mapsto \langle \Pi_V, C_x \rangle$. In turn, this will allow us to introduce classical tools of empirical process theory to obtain our results.

Let us formulate precisely this stepping stone in the case of K_1 through the following corollary:

Corollary 3.3.1. *Under Assumption* **(A)**, *we have*

$$\sum_{k=1}^{d} \lambda_k(K_1) = \max_{V \in \mathcal{V}_d} \mathbb{E}\left[\langle \Pi_V, C_X \rangle\right] , \tag{3.16}$$

$$\sum_{k \geq d+1} \lambda_k(K_1) = \min_{V \in \mathcal{V}_d} \mathbb{E}\left[\langle \Pi_{V^\perp}, C_X \rangle\right] . \tag{3.17}$$

The first equality in this corollary is an immediate consequence of (3.7) and assertions (i), (iii) of Lemma 3.2.4. The second is a consequence of the first, of definition (3.6) and of the definition of the trace. Of course, corresponding results for the centered and empirical versions hold as well in a parallel fashion. As will be clear in section 3.4, these quantities play a crucial role in Kernel-PCA.

3.3.1 Noncentered Case

In this section we consider the easier case of the noncentered kernel operator.

Global approach

The first result consists in data-dependent upper and lower bounds for the sum of the d largest or smallest eigenvalues of the integral operator. It is essentially the same as the result obtained by [STWCK05], but we give a proof for completeness and to show how it fits in our framework.

Theorem 3.3.2 (Shawe-Taylor et al.). *Under Assumption* **(A)**, *with probability at least* $1 - 3e^{-\xi}$,

$$-M\sqrt{\frac{\xi}{2n}} \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}} . \tag{3.18}$$

Also, with probability at least $1-3e^{-\xi}$,

$$-M\sqrt{\frac{\xi}{2n}} \leq \sum_{i\geq d+1} \lambda_i(K_1) - \sum_{i\geq d+1} \lambda_i(K_{1,n}) \leq 2\sqrt{\frac{d}{n}\operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}}. \tag{3.19}$$

Proof. We start with the first statement. Each inequality is proved separately and we use a union-of-event bound to gather them. From equation (3.16) and its counterpart for $K_{1,n}$ we have

$$\sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) = \max_{V\in\mathcal{V}_d} P_n \left\langle \Pi_V, C_X \right\rangle - \max_{V\in\mathcal{V}_d} P \left\langle \Pi_V, C_X \right\rangle .$$

This gives, denoting by V_d the subspace attaining the second maximum,

$$(P_n - P)\left\langle \Pi_{V_d}, C_X \right\rangle \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) \leq \sup_{V\in\mathcal{V}_d} (P_n - P)\left\langle \Pi_V, C_X \right\rangle .$$

The lower bound above leads to the lower bound of the theorem by an application of Hoeffding's inequality for the empirical mean (Theorem 3.9.2 with $r=1$) of an i.i.d. sample of the bounded random variable $\left\langle \Pi_{V_d}, C_X \right\rangle$, namely

$$0 \leq \left\langle \Pi_{V_d}, C_X \right\rangle = \left\langle \Pi_{V_d}, \varphi_X \otimes \varphi_X^* \right\rangle = \left\| \Pi_{V_d}(\varphi_X) \right\|^2 \leq \left\| \varphi_X \right\|^2 \leq M\,, \tag{3.20}$$

where we have used the fact that Π_{V_d} is a projector.

For the upper bound, we use standard techniques of concentration and symmetrization. Since $\left\langle \Pi_{V_d}, C_X \right\rangle \in [0, M]$, we can apply the bounded difference concentration inequality (Theorem 3.9.1) to the variable $\sup_{V\in\mathcal{V}_d}(P_n - P)\left\langle \Pi_V, C_X \right\rangle$. Thus with probability $1-e^{-\xi}$,

$$\sup_{V\in\mathcal{V}_d} (P_n - P)\left\langle \Pi_V, C_X \right\rangle \leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d} (P_n - P)\left\langle \Pi_V, C_X \right\rangle\right] + M\sqrt{\frac{\xi}{2n}}. \tag{3.21}$$

By a standard symmetrization argument,

$$\mathbb{E}\left[\sup_{V\in\mathcal{V}_d} (P_n - P)\left\langle \Pi_V, C_X \right\rangle\right] \leq 2\mathbb{E}\mathbb{E}_\varepsilon \left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle\right] , \tag{3.22}$$

where $(\varepsilon_i)_{i=1,\ldots,n}$ is an i.i.d. family of Rademacher variables. We can apply the bounded difference inequality a second time to this quantity, so that with probability $1-e^{-\xi}$:

$$\mathbb{E}\mathbb{E}_\varepsilon \left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle\right] \leq \mathbb{E}_\varepsilon \left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle\right] + M\sqrt{\frac{\xi}{2n}}. \tag{3.23}$$

The expectation on the right-hand-side is then bounded by an application of Lemma 3.3.3 below, leading to the conclusion.

The second inequality of the Theorem follows from similar arguments. Equation (3.17) leads to

$$\sum_{i>d} \lambda_i(K_1) - \sum_{i>d} \lambda_i(K_{1,n}) = \min_{V\in\mathcal{V}_d} P \left\langle \Pi_{V^\perp}, C_X \right\rangle - \min_{V\in\mathcal{V}_d} P_n \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

and thus, denoting by $\widetilde{V}_d$ the subspace attaining the first minimum,

$$(P - P_n) \left\langle \Pi_{\widetilde{V}_d^\perp}, C_X \right\rangle \leq \sum_{i>d} \lambda_i(K_1) - \sum_{i>d} \lambda_i(K_{1,n}) \leq \sup_{V \in \mathcal{V}_d} (P - P_n) \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

The rest of the proof parallels exactly the proof of the first part. □

We have used the following Lemma in the completion of the proof:

Lemma 3.3.3.

$$\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_{V^\perp}, C_{X_j} \right\rangle \right] = \mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle \right] \leq \sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} .$$

and

$$\mathbb{E}\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_{V^\perp}, C_{X_j} \right\rangle \right] = \mathbb{E}\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle \right] \leq \sqrt{\frac{d}{n} \operatorname{tr} K_2}$$

Proof. First note that for the two statements, the first equality is straightforward from the definition and the symmetry of Rademacher variables. We then have

$$\begin{aligned} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V, C_{X_j} \right\rangle &= \left\langle \Pi_V, \sum_{j=1}^{n} \varepsilon_j \varphi_{X_j} \otimes \varphi_{X_j}^* \right\rangle_{\mathrm{HS}(\mathcal{H}_k)} \\ &\leq \sqrt{d} \left\| \sum_{j=1}^{n} \varepsilon_j \varphi_{X_j} \otimes \varphi_{X_j}^* \right\|_{\mathrm{HS}(\mathcal{H}_k)} = \sqrt{d \sum_{i,j=1}^{d} \varepsilon_i \varepsilon_j k^2(X_i, X_j)} , \end{aligned}$$

where the inequality is Cauchy-Schwarz. Finally, by Jensen's inequality,

$$\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \,\in\, \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_{V^\perp}, C_{X_j} \right\rangle \right] \leq \sqrt{\frac{d}{n}} \sqrt{\frac{\sum_{i=1}^{n} k^2(X_i, X_i)}{n}} .$$

This concludes the proof of the first statement. The second is obtained by a second application of Jensen's inequality. □

Remark. Notice that the upper and lower bounds in Theorem 3.3.2 are of a different nature. One way to explain this is to consider directly the expectation of the involved quantities: one has

$$0 \leq \mathbb{E} \left[\sum_{i=1}^{d} \lambda_i(K_{1,n}) \right] - \sum_{i=1}^{d} \lambda_i(K_1) \leq \mathbb{E} \left[\sup_{V \in \mathcal{V}_d} (P_n - P) \left\langle \Pi_V, C_X \right\rangle \right] \leq 2 \sqrt{\frac{d}{n} \operatorname{tr} K_2} ,$$

where the lower bound is a consequence of (3.16) and Jensen's inequality, and the upper bound follows from arguments similar to the above proof.

We see that the empirical eigenvalues are biased estimators of the population ones (although the above inequality only provides an upper bound on the bias); therefore the difference between upper and lower bound in (3.18) is to be interpreted as bias rather than estimation error. If we additionally apply McDiarmid's inequality twice to the above bound, on the one hand to the quantity $\sum_{i=1}^{d} \lambda_i(K_{1,n})$, and on the other hand to $\operatorname{tr} K_{2,n}$, then we are lead precisely to (3.18). This approach was followed by [STWCK02, STWCK05]. We have used the same arguments in the proof of Theorem 3.3.2, but in a different order, as this allows for further refinement (see next section).

Relative bounds

The *star-shaped hull* of a class of functions $\mathcal{F}$ is defined as

$$\operatorname{star}(\mathcal{F}) = \{\lambda f\,,\ f \in \mathcal{F}\,,\ \lambda \in [0,1]\}\,.$$

Also we need the following notation for Rademacher complexities:

$$R_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\,,$$

where (ε_i) are i.i.d. Rademacher.

We now use recent work based on Talagrand's inequality (see ,e.g., [Mas00b, BBM03]) to obtain improved concentration for the large eigenvalues of the Gram matrix. We obtain a better rate of convergence, but at the price of comparing the sums of eigenvalues up to a constant factor.

Theorem 3.3.4. *Under Assumption* **(A)**, *for all $K > 1$ and $\xi > 0$, with probability at least $1 - e^{-\xi}$, the following holds:*

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K+1}{K}\sum_{k=1}^{d} \lambda_k(K_1) \leq 6K \inf_{h\geq 0}\left\{\frac{Mh}{n} + 2\sqrt{\frac{d}{n}\sum_{j\geq h+1}\lambda_j(K_2)}\right\} + \frac{M\xi(11+5K)}{n}\,. \quad (3.24)$$

Also, for all $K > 1$ and $\xi > 0$, with probability at least $1 - 3e^{-\xi}$, we have

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K+1}{K}\sum_{k=1}^{d} \lambda_k(K_1) \leq 282K \inf_{h\geq 0}\left\{\frac{2hM}{n} + \sqrt{2}\sqrt{\frac{d}{n}\sum_{j\geq h+1}\lambda_j(K_{2,n})}\right\} + \frac{2620MK\xi}{n}. \quad (3.25)$$

Moreover, with probability at least $1 - e^{-\xi}$, for all $K > 1$,

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K-1}{K}\sum_{k=1}^{d} \lambda_k(K_1) \geq -\frac{5KM\xi}{6n}\,. \quad (3.26)$$

Comments. A superficial look at this result could lead to conclude that it is of the same form as Theorem 2 of [STWCK05] where an infimum operator also appears in the bound. However, the bounds are really of a different nature. In the latter reference the infimum operation comes from the observation that since obviously the partial sum S_d of the first d eigenvalues is increasing in d, we can lower bound S_d by $S_{d'}$ with $d' < d$; hence the empirical lower bound for $S_{d'}$ is *a fortiori* a lower bound for S_d. We could naturally also take advantage of this observation and introduce in the lower bound an additional maximum operation over $d' < d$ but opted against it for readability.

To illustrate the novelty introduced by our result, first notice that if we disregard the multiplicative constants, the complexity term obtained here is always better (or equal) in order than the one of (3.18) (take $h = 0$). As an example of how this bound differs from (3.18), assume that $\lambda_j(K_2) = O(j^{-\alpha})$ with $\alpha > 2$ ([BJ02] reports that this is the case for the gaussian kernel if the density of the law of X decay like $1/|x|^{\epsilon+\alpha}$ with $\epsilon > 0$), then (3.18) gives a bound of order $\sqrt{d/n}$, while Theorem 3.3.4 gives a bound of order $d^{1/(1+\alpha)}n^{-\alpha/(1+\alpha)}$ – hence a better rate. In the case of an exponential decay ($\lambda_j(K_2) = O(e^{-\gamma j})$ with $\gamma > 0$), the rate even drops to $\log(nd)/n$. If K_2 has a finite number k of non-zeros eigenvalues, the bound is of order $\frac{k}{n}$. Of course this improvement comes at the cost of an additional factor in front of the empirical sum, hence this bound is better understood as a *relative* performance bound.

Finally, Theorem 3.3.4 only covers the case of the sum of the *bigger* eigenvalues. Unfortunately, unlike in the global case, we were not able to use an identical reasoning for the smallest eigenvalues. It is actually possible to derive a result of a similar form, but with worse constants, as a consequence of our results for the generalization of kernel PCA. For this reason we postpone the statement of this result to section 3.4.

The proof of the Theorem uses a fundamental deviation inequality recalled in Appendix 3.7 and additional auxiliary results in Appendix 3.8.

Proof. As in the proof of Theorem 3.3.2, we have to consider the empirical process associated with $\langle \Pi_V, C_X \rangle$ for $V \in \mathcal{V}_d$. Let us define

$$\mathcal{F}_d = \{x \mapsto \langle \Pi_V, C_x \rangle \ , \ V \in \mathcal{V}_d\} \, .$$

In order to prove inequality (3.24), we will apply Theorem 3.7.1 (coming from [BBM03], and recalled in Appendix 3.7) to the class of functions $M^{-1}\mathcal{F}_d$. From equation (3.20), it holds that $\forall f \in M^{-1}\mathcal{F}_d \, , f(x) \in [0,1]$, and therefore $Pf^2 \leq Pf$, hence the first assumptions of the theorem are satisfied.

What we need is now to obtain upper bounds for localized Rademacher complexities where the localization is in terms of P or P_n. For this we will need some results about local Rademacher complexities on ellipsoids that are regrouped and shown in Appendix 3.8. Let us first denote the "localized" set

$$S_r = \left\{g \in \mathrm{star}(M^{-1}\mathcal{F}_d), \, Pg^2 \leq r\right\} = M^{-1}\left\{g \in \mathrm{star}(\mathcal{F}_d), \, Pg^2 \leq M^2 r\right\} . \tag{3.27}$$

Corollary 3.8.2 entails

$$\mathbb{E}R_n S_r \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) := \psi_d(r) \, .$$

We now need to upper-bound the fixed point r_d^* of $\psi_d(r)$. For this we use Lemma 3.8.4 with $c = 1\,, \alpha = M^{-1}$, leading to

$$r_d^* \leq \inf_{h \geq 0} \left\{ \frac{h}{n} + 2M^{-1} \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} . \tag{3.28}$$

Inequality (3.43) of Theorem 3.7.1 implies that with probability at least $1 - e^{-\xi}$, every $f \in \mathcal{F}_d$ satisfies

$$P_n f \leq \frac{K+1}{K} Pf + 6KMr_d^* + \frac{M\xi(11+5K)}{n} . \tag{3.29}$$

Substituting (3.28), taking the supremum over $f \in \mathcal{F}_d$ in both sides, and using (3.16), we obtain (3.24).

In order to prove inequality (3.25), we apply the second part of theorem 3.7.1, which gives us a confidence bound on r_d^* using the Rademacher complexity localized in terms of the empirical measure. For this we define $\widehat{S}_r$ like S_r in (3.27) but where P_n takes the role of P. Corollary 3.8.2 entails

$$\mathbb{E}_\varepsilon R_n \widehat{S}_r \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{k \geq h+1} \lambda_k(K_{2,n})} \right) := \widehat{\psi}_d(r) . \tag{3.30}$$

Then Theorem 3.7.1 tells us that with probability $1 - 2e^{-\xi}$, r_d^* is upper bounded by the fixed point of $20\widehat{\psi}_d(2r) + 31\xi/n$. To upper bound this quantity in turn, we first apply Lemma 3.8.4 with $c = 2, \alpha = M^{-1}$ as above to obtain a bound on the fixed point of $\widehat{\psi}_d(2r)$; then we apply Lemma 3.7.2 with $K = \frac{7}{6}$. Gathering these inequalities and after straightforward calculations, we finally get that with probability at least $1 - 3e^{-\xi}$, $\forall f \in \mathcal{F}_d$,

$$P_n f \leq \frac{K+1}{K} Pf + 282K \inf_{h \geq 0} \left\{ \frac{2hM}{n} + \sqrt{2} \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_{2,n})} \right\} + \frac{2620MK\xi}{n} ,$$

leading to (3.25).

Using Bernstein's inequality (3.53) (see Theorem 3.9.3) with $f(x) = \langle \Pi_{V_d}, C_x \rangle$, the proof of inequality (3.26) uses the same arguments (e.g., inequality (3.20)) that the proof of the lower bound of (3.18). □

3.3.2 Recentered Case

In the following result, we extend Theorem 3.3.2 to a more general case where the data is first recentered. Let us begin with a control of a supremum of random variables:

Theorem 3.3.5. *Under Assumption* **(A)**, *with probability at least* $1 - 3e^{-\xi}$,

$$\sup_{V \in \mathcal{V}_d} \left(\left\langle \Pi_V, \widetilde{C}_{1,n} \right\rangle - \left\langle \Pi_V, \overline{C}_1 \right\rangle \right) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + M \left(5\sqrt{\frac{\xi}{n}} + \frac{4}{\sqrt{n}} + \frac{6}{n-1} \right) ;$$

similarly, with probability at least $1 - 3e^{-\xi}$,

$$\sup_{V \in \mathcal{V}_d} \left(\left\langle \Pi_{V^\perp}, \overline{C}_1 \right\rangle - \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle \right) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + M \left(5\sqrt{\frac{\xi}{n}} + \frac{4}{\sqrt{n}} + \frac{6}{n-1} \right) ;$$

The proof of this Theorem follows the same structure as for Theorem 3.3.2, but some additional ingredients are needed to control U-processes arising from the recentering.

Proof. We prove the first statement of Theorem 3.3.5: the second one follows from the same arguments. First recall the following decomposition from equations (3.14) and (3.15):

$$\overline{C}_1 = C_1 - \mu \otimes \mu^* \quad \text{and} \quad \widetilde{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \varphi_{X_i} \otimes \varphi_{X_j}^* , \tag{3.31}$$

from which we obtain

$$\begin{aligned} \sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} - \overline{C}_1 \right\rangle &\leq \sup_{V \in \mathcal{V}_d} \langle \Pi_V, C_{1,n} - C_1 \rangle \\ &\quad + \sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j}^* \right\rangle . \end{aligned} \tag{3.32}$$

It was shown in the proof of Theorem 3.3.2 that the following holds with probability greater than $1 - 2e^{-\xi}$:

$$\sup_{V \in \mathcal{V}_d} \langle \Pi_V, C_{1,n} - C_1 \rangle \leq 2\sqrt{\frac{d}{n}} \sqrt{\operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}} ,$$

so we now concentrate on the second term of (3.32). If we denote $G(x_1, \ldots, x_n) = \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{x_i} \otimes \varphi_{x_i}^* \right\rangle$, then we have for any i_0:

$$\begin{aligned} &\left| G(x_1, \ldots, x_n) - G(x_1, \ldots, x_{i_0 - 1}, x'_{i_0}, x_{i_0+1}, \ldots, x_n) \right| \\ &\qquad \leq \frac{1}{n(n-1)} \left\| \sum_{j \neq i_0} \left((\varphi_{x_{i_0}} - \varphi_{x'_{i_0}}) \otimes \varphi_{x_j}^* + \varphi_{x_j} \otimes (\varphi_{x_{i_0}}^* - \varphi_{x'_{i_0}}^*) \right) \right\| \\ &\qquad \leq \frac{2}{n(n-1)} \sum_{j \neq i_0} \left\| \varphi_{x'_{i_0}} - \varphi_{x_{i_0}} \right\| \left\| \varphi_{x_j} \right\| \leq \frac{4M}{n} . \end{aligned}$$

Therefore we can apply the bounded difference inequality (Theorem 3.9.1) to G, so that with probability greater than $1 - e^{-\xi}$,

$$\begin{aligned} &\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j}^* \right\rangle \\ &\qquad \leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j}^* \right\rangle \right] + 4M\sqrt{\frac{\xi}{2n}} . \end{aligned}$$

To deal with the above expectation, we consider Hoeffding's decomposition (see [dlPnG99, p. 137]) for U-processes. To this end, we define the following quantities:

$$
\begin{aligned}
S_d &= \sup_{V \in \mathcal{V}_d} \left(\frac{2}{n} \sum_{j=1}^{n} \langle \Pi_V, \mu \otimes \mu^* \rangle - \left\langle \Pi_V(\varphi_{X_j}), \mu \right\rangle \right) \\
R_d &= \sup_{V \in \mathcal{V}_d} -\frac{1}{n(n-1)} \sum_{i \neq j} \Big(\left\langle \Pi_V, \varphi_{X_i} \otimes \varphi_{X_j}^* \right\rangle - \left\langle \Pi_V(\varphi_{X_j}), \mu \right\rangle \\
&\qquad - \langle \Pi_V(\varphi_{X_i}), \mu \rangle + \langle \Pi_V, \mu \otimes \mu^* \rangle \Big) .
\end{aligned}
$$

It can easily be seen that

$$
\mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j}^* \right\rangle \right] \leq \mathbb{E}\left[S_d\right] + \mathbb{E}\left[R_d\right] .
$$

Gathering the different inequalities up to now, we have with probability greater than $1 - 3e^{-\xi}$:

$$
\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} - \overline{C}_1 \right\rangle \leq 2\sqrt{\frac{d}{n}} \sqrt{\operatorname{tr} K_{2,n}} + \mathbb{E}[S_d] + \mathbb{E}[R_d] + 5M\sqrt{\frac{\xi}{2n}} . \tag{3.33}
$$

We now bound from above the expectation of S_d and R_d using Lemmas 3.3.6 and 3.3.7 below. This leads to the conclusion. $\square$

Lemma 3.3.6. *The following inequality holds:*

$$
\mathbb{E}\left[S_d\right] \leq 4 \frac{\mathbb{E}k(X,X)}{\sqrt{n}} .
$$

Proof. A standard symmetrization argument leads to

$$
\begin{aligned}
\mathbb{E}[S_d] &\leq \mathbb{E}\mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \frac{4}{n} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_V(\varphi_{X_j}), \mu \right\rangle \\
&\leq \frac{4}{n} \mathbb{E}\mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \Pi_V \left(\sum_{j=1}^{n} \varepsilon_j \varphi_{X_j} \right) \right\| \|\mu\| \\
&\leq \frac{4}{n} \mathbb{E}\mathbb{E}_\varepsilon \left\| \sum_{j=1}^{n} \varepsilon_j \varphi_{X_j} \right\| \|\mu\| \\
&\leq \frac{4}{\sqrt{n}} \mathbb{E}\sqrt{\operatorname{tr} K_{1,n}} \, \|\mu\| ,
\end{aligned}
$$

where we successively applied the Cauchy-Schwarz inequality, the contractivity of an orthogonal projector, and Jensen's inequality. Applying Jensen's inequality again, and the fact that $\|\mu\|^2 = \mathbb{E}k(X,X') \leq (\mathbb{E}k^{\frac{1}{2}}(X,X))^2$ yields the conclusion. $\square$

Lemma 3.3.7. *The following inequality holds:*

$$
\mathbb{E}\left[R_d\right] \leq \frac{6}{n-1} \mathbb{E}k(X,X) .
$$

Remark The proof uses techniques developed by [dlPnG99]. Actually, we could directly apply Theorems 3.5.3 and 3.5.1 of this reference, getting a factor 2560 instead of 6. We give here a self-contained proof tailored for our particular case for completeness and for the improved constant.

Proof. Since Π_V is a symmetric operator, using Jensen's inequality ,

$$\mathbb{E}\left[R_d\right] \leq \frac{1}{n(n-1)} \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i \neq j} f_V(X_i, X_i', X_j, X_j')\right] ,$$

where

$$f_V(X_i, X_i', X_j, X_j') = \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} - \varphi_{X_i'} \otimes \varphi^*_{X_j} - \varphi_{X_i} \otimes \varphi^*_{X_j'} + \varphi_{X_i'} \otimes \varphi^*_{X_j'} \right\rangle .$$

Since $f_V(X_i, X_i', X_j, X_j') = -f_V(X_i', X_i, X_j, X_j')$ and $f_V(X_i, X_i', X_j, X_j') = -f_V(X_i, X_i', X_j', X_j)$, following the proof of the standard symmetrization, we get:

$$\mathbb{E}\left[R_d\right] \leq \frac{1}{n(n-1)} \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i \neq j} \varepsilon_i \varepsilon_j f_V(X_i, X_i', X_j, X_j')\right]$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[R_d\right] \leq \frac{2}{n(n-1)} \Bigg(& \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i \neq j} \varepsilon_i \varepsilon_j \langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} \rangle\right] \\ & + \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} - \sum_{i \neq j} \varepsilon_i \varepsilon_j \langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j'} \rangle\right] \Bigg) = \frac{2}{n(n-1)} (A + B) \ ;\end{aligned}$$

for the first term above we have

$$A \leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i,j} \varepsilon_i \varepsilon_j \langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} \rangle\right] = C ,$$

while for the second we use

$$\begin{aligned} B &\leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} - \sum_{i,j} \varepsilon_i \varepsilon_j \langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j'} \rangle\right] + \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i} \langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_i'} \rangle\right] \\ &= D + E .\end{aligned}$$

We bound terms C, D, E by the following similar chains of inequalities where we successively use the Cauchy-Schwarz inequality, the contractivity of an orthogonal projector and a standard computation on sums of weighted Rademacher:

$$\begin{aligned} C &\leq \mathbb{E}_X \mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \Pi_V(\varphi_{X_j}) \right\| \leq \mathbb{E}_X \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\|^2 \\ &= n \mathbb{E} k(X, X) ;\end{aligned}$$

$$
\begin{aligned}
D &\le \mathbb{E}_{X,X'}\mathbb{E}_{\varepsilon} \sup_{V\in\mathcal{V}_d} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \Pi_V(\varphi_{X'_j}) \right\| \\
&\le \mathbb{E}_{X,X'}\mathbb{E}_{\varepsilon} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \varphi_{X'_j} \right\| \\
&\le \mathbb{E}_{X,X'} \sqrt{ \mathbb{E}_{\varepsilon} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\|^2 \mathbb{E}_{\varepsilon} \left\| \sum_j \varepsilon_j \varphi_{X'_j} \right\|^2 } \\
&\le \mathbb{E}_{X,X'} \sqrt{ \left(\sum_i k(X_i, X_i) \right) \left(\sum_i k(X'_i, X'_i) \right) } \le n\mathbb{E}k(X,X)\,;
\end{aligned}
$$

$$
\begin{aligned}
E \le \mathbb{E}_X \sup_{V\in\mathcal{V}_d} \sum_i \left\| \Pi_V(\varphi_{X'_i}) \right\| \|\varphi_{X_i}\| &\le \mathbb{E}_X \sum_i \left\| \varphi_{X'_i} \right\| \|\varphi_{X_i}\| \\
&\le \mathbb{E}_X \sum_i \sqrt{k(X'_i, X'_i) k(X_i, X_i)} \\
&= n\mathbb{E}k(X,X)\,.
\end{aligned}
$$

Gathering the previous inequalities, we obtain the conclusion. □

From Theorem 3.3.5 we deduce the following upper bounds:

Theorem 3.3.8. *Under Assumption* **(A)**, *for all* $\xi > 1$, *with probability greater than* $1 - 3e^{-\xi}$,

$$
-2M\sqrt{\frac{\xi}{n}} \le \frac{n}{n-1} \sum_{i=1}^{d} \lambda_i(\overline{K}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{K}_1) \le 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 15M\sqrt{\frac{\xi}{n}}\,;
$$

and with probability greater than $1 - 3e^{-\xi}$,

$$
-2M\sqrt{\frac{\xi}{n}} \le \sum_{i \ge d+1} \lambda_i(\overline{K}_1) - \frac{n}{n-1} \sum_{i \ge d+1} \lambda_i(\overline{K}_{1,n}) \le 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 15M\sqrt{\frac{\xi}{n}}\,.
$$

Proof. (Upper bound) Theorem 3.2.1 entails

$$
\frac{n}{n-1} \sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \le \sup_{V\in\mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} \right\rangle - \left\langle \Pi_V, \overline{C}_1 \right\rangle\,,
$$

and

$$
\sum_{i \ge d+1} \lambda_i(\overline{C}_1) - \frac{n}{n-1} \sum_{i \ge d+1} \lambda_i(\overline{C}_{1,n}) \le \sup_{V\in\mathcal{V}_d} \left\langle \Pi_{V^\perp}, \overline{C}_1 \right\rangle - \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle\,.
$$

Theorem 3.3.5 and Lemma 3.2.4 allow to conclude.

The lower bound follows from Hoeffding's inequality for U-statistics; details can be found in the Appendix 3.6.2. □

3.4 Application to Kernel-PCA

3.4.1 Uncentered Case

We first consider in this section the simpler case of "uncentered Kernel-PCA" where the goal is to reconstruct the signal using principal directions of the noncentered covariance operator.

Remember we assume that the number d of KPCA directions kept for projecting the observations has been fixed *a priori*. We wish to find the linear space of dimension d that conserves the maximal norm, i.e., which minimizes the error (measured with the RKHS norm) of approximating the data by their projections. The space $\widehat{V}_d$ minimizing the empirical error is given by

$$\widehat{V}_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \frac{1}{n} \sum_{j=1}^{n} \|\varphi_{X_j} - \Pi_V(\varphi_{X_j})\|^2 ; \tag{3.34}$$

$\widehat{V}_d$ is the vector space spanned by the first d eigenfunctions of $C_{1,n}$. Analogously, we denote by V_d the space spanned by the first d eigenfunctions of C_1. We will adopt the following notation for the true and empirical *reconstruction error*:

$$R_n(V) = \frac{1}{n} \sum_{j=1}^{n} \|\varphi_{X_j} - \Pi_V(\varphi_{X_j})\|^2 = P_n \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

$$R(V) = \mathbb{E}\left[\|\varphi_X - \Pi_V \varphi_X\|^2\right] = P \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

One has $R_n(\widehat{V}_d) = \sum_{i>d} \lambda_i(K_{1,n})$ and $R(V_d) = \sum_{i>d} \lambda_i(K_1)$.

Bound on the Reconstruction Error: global approach

We give a data dependent bound for the reconstruction error which is a simple consequence of Theorem 3.3.2.

Theorem 3.4.1. *Under Assumption* **(A)**, *with probability at least* $1 - 2e^{-\xi}$,

$$R(\widehat{V}_d) \leq \sum_{i=d+1}^{n} \lambda_i(K_{1,n}) + 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}} .$$

Also, with probability at least $1 - e^{-\xi}$,

$$R(\widehat{V}_d) - R(V_d) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_2} + 2M\sqrt{\frac{\xi}{2n}} .$$

Proof. We have

$$R(\widehat{V}_d) - R_n(\widehat{V}_d) = (P - P_n) \left\langle \Pi_{\widehat{V}_d^\perp}, C_X \right\rangle \leq \sup_{V \in \mathcal{V}_d} (P - P_n) \left\langle \Pi_{V^\perp}, C_X \right\rangle ; \tag{3.35}$$

we have already treated this quantity in the proof of Theorem 3.3.2, hence the first part is proved.

For the second part, the definition of $\widehat{V}_d$ implies that

$$R(\widehat{V}_d) - R(V_d) \leq \left(R(\widehat{V}_d) - R_n(\widehat{V}_d)\right) - \left(R(V_d) - R_n(V_d)\right).$$

The first term is controlled by gathering inequalities (3.21), (3.22) and (3.35). We obtain a lower bound for the second term using Hoeffding's inequality (again, exactly as in the proof of the lower bound in Theorem 3.3.2). This concludes the proof. □

Fast rates via localized approach

We now show that the excess error of the best empirical d-dimensional subspace with respect to the error of the best d-dimensional subspace can decay at a much faster rate than can be expected from Theorem 3.4.1. This however comes at the price of an additional factor related to the size of the gap between two successive distinct eigenvalues.

Here is the main result of the section:

Theorem 3.4.2. *Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues. Let $\widetilde{d}$ be such that $\lambda_d = \mu_{\widetilde{d}}$. Define*

$$\gamma_d = \begin{cases} \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1} & \text{if } \widetilde{d} = 1 \text{ or } \lambda_d > \lambda_{d+1}, \\ \min\left(\mu_{\widetilde{d}-1} - \mu_{\widetilde{d}}, \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1}\right) & \text{otherwise}; \end{cases} \tag{3.36}$$

and $B_d = 2\sqrt{\mathbb{E}k^4(X,X')}/\gamma_d$.

Then under Assumption **(A)**, *for all d, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$ the following holds:*

$$R(\widehat{V}_d) - R(V_d) \leq 7 \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4\sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{\xi(11M + 6B_d)}{n}. \tag{3.37}$$

Comments. Similarly to the remarks on Theorem 3.3.4, the complexity term obtained in Theorem 3.4.2 has a faster (or equal) decay rate, as a function of the sample size n, than the one of Theorem 3.4.1; this rate depends on the decay behavior of the eigenvalues.

We do not state a fully empirical version of the bound (using only empirical eigenvalues) to avoid additional burden. Let us sketch briefly how this could be obtained: in the proof of the Theorem, we can use the *empirically* localized Rademacher complexity at the price of worse constants (see the proof of Theorem 3.3.4 to see an example of how this plays out). This has the effect of replacing the true eigenvalues by the empirical ones in the sum appearing in (3.37). However the constant B_d still depends on the true eigenvalues. To deal with this, we can use a simple convergence result of the empirical eigenvalues to the true ones (as proved for example by [KG00]), so that for n big enough B_d is bounded by $2\widehat{B}_d$ (its empirical counterpart).

Proof. We will use here again Theorem 3.7.1. We define the following class of functions:

$$\widetilde{\mathcal{F}}_d = \left\{ f_V : x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x \right\rangle, V \in \mathcal{V}_d \right\},$$

where for each $V \in \mathcal{V}_d$, H_V is obtained via Lemma 3.6.1. We will apply Theorem 3.7.1 to the class $M^{-1}\widetilde{\mathcal{F}}_d$. For any $f \in M^{-1}\widetilde{\mathcal{F}}_d$, it holds that $f \in [-1,1]$; furthermore, Lemma 3.6.1 entails that $Pf^2 \leq M^{-1}B_d Pf$. To upper bound the local Rademacher complexities of this class we define

$$\widetilde{S}_r = \left\{ g \in \operatorname{star}(M^{-1}\widetilde{\mathcal{F}}_d),\ Pg^2 \leq r \right\} = M^{-1} \left\{ g \in \operatorname{star}(\widetilde{\mathcal{F}}_d),\ Pg^2 \leq M^2 r \right\} .$$

Corollary 3.8.3 entails

$$M^{-1}B_d \mathbb{E} R_n \widetilde{S}_r \leq \frac{M^{-1}B_d}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) := \widetilde{\psi}_d(r) .$$

Let $\widetilde{r}_d^*$ denote the solution of equation $\widetilde{\psi}_d(r) = r$. We apply Lemma 3.8.4 with the choice $c = M^{-1}B_d, \alpha = M^{-1}$ to obtain

$$\widetilde{r}^* \leq M^{-2} \inf_{h \geq 0} \left\{ \frac{B_d^2 h}{n} + 4B_d \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} .$$

We can now apply Theorem 3.7.1, obtaining that for any $K > 1$ and every $\xi > 0$, with probability at least $1 - e^{-\xi}$, $\forall\, V \in \mathcal{V}_d$,

$$Pf_V \leq \frac{K}{K-1} P_n f_V + 6K \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4 \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{\xi(11M + 5B_d K)}{n} \tag{3.38}$$

Choosing $V = \widehat{V}_d$, using that $R(H_{V_d}) = R(V_d)$ (by definition (3.41) of V_d) and noting that the definition (3.34) of $\widehat{V}_d$ yields $P_n f_{\widehat{V}_d} \leq 0$, this leads to the result. □

The techniques used to obtain the previous fast rates for reconstruction error of KPCA allows us to get improved bounds for the sum of the *smaller* eigenvalues.

Corollary 3.4.3. *Under Assumption* **(A)**,*with probability at least* $1 - e^{-\xi}$, *for all* $K > 1$,

$$\sum_{k \geq d} \lambda_k(K_1) - \frac{K}{K+1} \sum_{k \geq d} \lambda_k(K_{1,n}) \geq -\frac{5KM\xi}{6n} . \tag{3.39}$$

Moreover, if $\lambda_d > \lambda_{d+1}$, *for all* $K > 1$ *and* $\xi > 0$, *with probability at least* $1 - 2e^{-\xi}$, *the following holds:*

$$\begin{aligned} &\sum_{k \geq d+1} \lambda_k(K_1) - \frac{K}{K-1} \sum_{k \geq d+1} \lambda_k(K_{1,n}) \\ &\qquad \leq 6K \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4 \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{K\xi}{n} \left(5B_d + \frac{5M}{6} + 11M \right) . \end{aligned} \tag{3.40}$$

where $B_d = 2\sqrt{\mathbb{E}k^4(X,X')}/(\lambda_d - \lambda_{d+1})$.

Proof. Using (3.17) and Bernstein's type inequality (3.54) with $f(x) = \langle \Pi_{V_d^\perp}, C_x \rangle$ (see Theorem 3.9.3), the proof of inequality (3.39) uses the same arguments (e.g. inequality (3.20)) that the proof of the lower bound of (3.18). We now prove inequality (3.40).

Since we suppose $\lambda_d > \lambda_{d+1}$, $H_V = V_d$ for all $V \in \mathcal{V}_d$ (H_V is defined in the proof of Lemma 3.6.1). Moreover,

$$\sum_{k \geq d+1} \lambda_k(K_1) - \frac{K}{K-1} \sum_{k \geq d+1} \lambda_k(K_{1,n}) \leq R(\widehat{V}_d) - \frac{K}{K-1} R_n(\widehat{V}_d) \, .$$

Inequality (3.38) states that with probability at least $1 - e^{-\xi}$:

$$R(\widehat{V}_d) - \frac{K}{K-1} R_n(\widehat{V}_d) \leq R(V_d) - \frac{K}{K-1} R_n(V_d) \quad + \quad 6K \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4 \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{\xi(11M + 5B_d K)}{n} \, .$$

Moreover, Bernstein's type inequality (3.53) with $f(x) = \left\langle \Pi_{V_d^\perp}, C_x \right\rangle$ leads to

$$R(V_d) - \frac{K}{K-1} R_n(V_d) \leq \frac{5KM\xi}{6n} \, .$$

Gathering the three previous inequalities, we get inequality (3.40). □

3.4.2 Recentered Case

The goal of this section is to show that the rate of convergence obtained in Theorem 3.4.1 in the uncentered case is of the same order if we consider the empirical re-centering. In this case the Kernel-PCA algorithm solves the following optimization problem:

$$\widehat{\overline{V}}_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \frac{1}{n} \sum_{j=1}^{n} \| \overline{\varphi}_{X_j} - \Pi_V(\overline{\varphi}_{X_j}) \|^2 \, ,$$

where $\widehat{\overline{V}}_d$ is the vector space spanned by the first d eigenfunctions of $\overline{C}_{1,n}$. We also denote by $\overline{V_d}$ the space spanned by the first d eigenfunctions of $\overline{C}_1$:

$$\overline{V}_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \mathbb{E} \| \varphi_X - \mu - \Pi_V(\varphi_X - \mu) \|^2 \, .$$

We will adopt the following notation for the reconstruction error:

$$\overline{R}_n(V) = \frac{1}{n-1} \sum_{j=1}^{n} \| \overline{\varphi}_{X_j} - \Pi_V(\overline{\varphi}_{X_j}) \|^2 = \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle \, .$$

$$\overline{R}(V) = \mathbb{E} \| \varphi_X - \mu - \Pi_V(\varphi_X - \mu) \|^2 = P \left\langle \Pi_{V^\perp}, \overline{C}_X \right\rangle \, .$$

One has $\overline{R}_n(\widehat{\overline{V}}_d) = \frac{n}{n-1} \sum_{i>d} \lambda_i(\overline{K}_{1,n})$ and $\overline{R}(\overline{V}_d) = \sum_{i>d} \lambda_i(\overline{K}_1)$. Following the same line of reasoning as in Theorem 3.4.1 and using Theorem 3.3.5 to control the supremum yields the following result.

Theorem 3.4.4. *Under Assumption* **(A)***, for any* $\xi > 1$*, with probability greater than* $1 - 3e^{-x}$*,*

$$\overline{R}(\widehat{\overline{V}}_d) \leq \frac{n}{n-1} \sum_{i>d} \lambda_i(\overline{K}_{1,n}) + 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 18M\sqrt{\frac{\xi}{n}}\,.$$

Note that the leading complexity term is the same as in Theorem 3.4.1: hence recentering in kernel PCA essentially does not introduce additional complexity to the procedure.

3.5 Conclusion and Discussion

Comparison with Previous Work. [DP76] studied asymptotic convergence of PCA and proved almost sure convergence in operator norm of the empirical covariance operator to the population one. These results were further extended to PCA in a Hilbert space by [Bes91]. However, no finite sample bounds were presented. Moreover, the centering of the data was not considered.

Compared to the work of [KG00] and [Kol98], we are interested in non-asymptotic (i.e., finite sample sizes) results. Also, as we are only interested in the case where $k(x, y)$ is a positive definite function, we have the nice property of Theorem 3.2.3 which allows to consider the empirical operator and its limit as acting on the same space (since we can use covariance operators on the RKHS). This is crucial in our analysis and makes precise non-asymptotic computations possible unlike in the general case studied by [KG00, Kol98].

Comparing with [STWCK02, STWCK05], we overcome the difficulties coming from infinite dimensional feature spaces as well as those of dealing with kernel operators (of infinite rank). Moreover their approach for eigenvalues is based on the concentration around the mean of the empirical eigenvalues and on the relationship between the expectation of the empirical eigenvalues and the operator eigenvalues. Here we used a direct approach and extend their results to the recentered case and proved refined bounds and possible faster convergence rates for the uncentered case. In particular we show that there is a tight relation between how the (true or empirical) eigenvalues decay and the rate of convergence of the reconstruction error of the d-dimensional projection found by the kernel PCA procedure to the ideal one.

Open issues: the nagging problem of the choice of dimension in PCA. All along this chapter, the integer d (the number of eigenvalues summed, or the dimension of the space selected by PCA) was always considered fixed a priori.

It is tempting to interpret the bounds appearing in Theorems 3.4.1 and 3.4.2 as a classical statistical trade-off between approximation error (empirical reconstruction error, decreasing with the dimension d) and estimation error (complexity term, increasing with d). This point of view would suggest to select d as the dimension minimizing the bound. However, this view is an illusion since it is clear that the *true* reconstruction error $R(\widehat{V}_d)$ of the subspace selected empirically is a decreasing function of d (since $\widehat{V}_d \subset \widehat{V}_{d+1}$). This emphasizes two important points: first, that the (true) reconstruction error is by itself not a good criterion to select the dimension (of course, with this criterion the best choice would be not to project the data at all but to keep the whole space). Hence, an alternative and sensible criterion has to be found to define in a well-founded way what the optimal dimension would be.

A second consequence of this observation is that the bounds we found do not exhibit the correct behavior in terms of the dimension d (for a fixed sample size n), since they become *increasing* in d, for big enough d, while the true error is always *decreasing*. Because of the decreasing property of the true error, any quantity bounding the reconstruction error for dimension d is also a valid bound for any $d' > d$. Hence, if we denote $d(n)$ the dimension realizing the mininum of the bound of Theorem 3.4.1 (for example) for a fixed sample size n, then the bound obtained for $d(n)$ is also valid for any larger dimension and actually *more informative* than the bounds obtained directly for this larger dimension. This property was also noticed and used by [STWCK05]. To sum up, our bound on the estimation error is too pessimistic for larger dimensions and does not provide a correct qualitative explanation for what is really taking place. Obtaining a better understanding of the behavior of the estimation error for fixed n and varying d is a very interesting open problem, which could also eventually lead to a relevant dimension selection criterion (maybe by comparison of the relative importance of approximation error and estimation error for larger dimensions).

We conclude by mentioning additional open problems: it would be of interest to obtain relative convergence rates for the estimation of single eigenvalues, and to obtain nonasymptotic bounds for eigenspace estimation.

Acknowledgements

The authors are extremely grateful to Stéphane Boucheron for invaluable comments and ideas, as well as for motivating this work.

Appendix

3.6 Appendix A: Additional proofs.

3.6.1 Proofs for section 3.2

Proof of Theorem 3.2.2. For the existence of operator C and its basic properties, see, e.g., [Bax76]. We proceed to prove the last part of the Theorem. First, we have $\mathbb{E}\|Z \otimes Z^*\| = \mathbb{E}\|Z\|^2 < \infty$, so that $\mathbb{E}[Z \otimes Z^*]$ is well-defined. Now, for any $f, g \in \mathcal{H}$ the following holds by the definition of C:

$$\langle f, \mathbb{E}[Z \otimes Z^*] g\rangle = \mathbb{E}[\langle Z \otimes Z^*, f \otimes g^*\rangle] = \mathbb{E}[\langle Z, f\rangle \langle Z, g\rangle] = \langle f, Cg\rangle \ ;$$

this concludes the proof.

Proof of Theorem 3.2.3. It is a well-known fact that an integral kernel operator such as K_ϕ is Hilbert-Schmidt if and only if the kernel $k(x,y)$ (here equal to $\langle \Phi(x), \Phi(y)\rangle$) is an element of $L_2(\mathcal{X} \times \mathcal{X})$ (endowed with the product measure) . This is the case here since $k(x,y) \leq \|\Phi(x)\| \, \|\Phi(y)\|$ and $\mathbb{E}\|\Phi(X)\|^2 < \infty$ by assumption. We now characterize this operator more precisely.

Since $\mathbb{E}\|\Phi(X)\| < \infty$, $\Phi(X)$ has an expectation which we denote by $\mathbb{E}[\Phi(X)] \in \mathcal{H}$. Consider the linear operator $T : \mathcal{H} \to L_2(P)$ defined as $(Th)(x) = \langle h, \Phi(x)\rangle_{\mathcal{H}}$. By the Cauchy-Schwarz inequality, $\mathbb{E}\langle h, \Phi(X)\rangle^2 \leq \|h\|^2 \mathbb{E}\|\Phi(X)\|^2$. This shows that T is well-defined and continuous; therefore it has a continuous adjoint T^*. The variable $f(X)\Phi(X) \in \mathcal{H}$ has a well-defined expectation since f and $\|\Phi\|$ are in $L_2(P)$. But for all $g \in \mathcal{H}$, $\langle T^* f, g\rangle_{\mathcal{H}} = \langle f, Tg\rangle_{L_2(P)} = \mathbb{E}[\langle g, f(X)\Phi(X)\rangle_{\mathcal{H}}]$ which shows that $T^*(f) = \mathbb{E}[\Phi(X) f(X)]$.

We now show that $C = T^*T$ and $K_\Phi = TT^*$. By the definition of the expectation, for all $h, h' \in \mathcal{H}$, $\langle h, T^*T(h')\rangle = \langle h, \mathbb{E}[\Phi(X) \langle \Phi(X), h'\rangle]\rangle = \mathbb{E}[\langle h, \Phi(X)\rangle \langle h', \Phi(X)\rangle]$. Thus, by the uniqueness of the covariance operator, we get $C = T^*T$. Similarly $(TT^*f)(x) = \langle T^*f, \Phi(x)\rangle = \mathbb{E}\left[\langle f(X)\Phi(X), \Phi(x)\rangle\right] = \int f(y)\,\langle\Phi(y), \Phi(x)\rangle\, dP(y)$ so that $K_\Phi = TT^*$. This also implies that K_Φ is self-adjoint and positive.

We finally show that the nonzero eigenvalues of TT^* and T^*T coincide by a standard argument. Let $E_\mu(A) = \{x, Ax = \mu x\}$ be the eigenspace of the operator A associated with μ. Moreover, let $\lambda > 0$ be a positive eigenvalue of $K = TT^*$ and f an associated eigenvector. Then $(T^*T)T^*f = T^*(TT^*)f = \lambda T^*f$. This shows that $T^*(E_\lambda(TT^*)) \subset E_\lambda(T^*T)$ and conversely $T(E_\lambda(T^*T)) \subset E_\lambda(TT^*)$. Applying T^* to both terms of the last inclusion implies $E_\lambda(T^*T) \subset T^*(E_\lambda(TT^*))$ since $\lambda \neq 0$, and therefore $T^*T(E_\lambda(T^*T)) = E_\lambda(T^*T)$. Conversely, $E_\lambda(TT^*) \subset T(E_\lambda(T^*T))$ for$\lambda \neq 0$. Thus, $E_\lambda(T^*T) = T^*(E_\lambda(TT^*))$ and $E_\lambda(TT^*) = T(E_\lambda(T^*T))$ and finally $\dim(E_\lambda(T^*T)) = \dim(E_\lambda(TT^*))$. This shows that the multiplicity is the same. This concludes the proof. $\square$

3.6.2 Proof for section 3.3

Proof of Theorem 3.3.8. (Lower bound) We prove the lower bound for the largest eigenvalues. A similar proof gives the second statement.
Theorem 3.2.1 leads to

$$\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \geq \left\langle \overline{C}_{1,n}, \Pi_{\overline{V}_d}\right\rangle - \left\langle \overline{C}_1, \Pi_{\overline{V}_d}\right\rangle .$$

where the maximum $\max_{V\in\mathcal{V}_d} \left\langle \Pi_V, \overline{C}_1\right\rangle_{\mathrm{HS}(\mathcal{H})}$ is reached at $\overline{V}_d$. Using the decomposition (3.31), we get:

$$\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1)$$
$$\geq \left\langle C_{1,n} - C_1, \Pi_{\overline{V}_d}\right\rangle - \left\langle \Pi_{\overline{V}_d}, \frac{1}{n(n-1)} \sum_{i\neq j} \varphi_{X_i} \otimes \varphi_{X_j} - \mu \otimes \mu \right\rangle .$$

The first term is bounded by Hoeffding's inequality exactly as in the proof of Theorem 3.3.2. With probability greater than $1 - e^{-x}$,

$$\left\langle C_{1,n} - C_1, \Pi_{\overline{V}_d}\right\rangle = (P_n - P)\left\langle \Pi_{\overline{V}_d}, C_X\right\rangle \geq -M\sqrt{\frac{\xi}{2n}} .$$

For the second term, we apply Hoeffding's inequality for U-statistics (Theorem 3.9.2 with $r = 2$); with probability greater than $1 - e^{-\xi}$,

$$-\left\langle \Pi_{\overline{V}_d}, \frac{1}{n(n-1)} \sum_{i\neq j} \varphi_{X_i} \otimes \varphi_{X_j} - \mu \otimes \mu \right\rangle \geq -M\sqrt{\frac{\xi}{2\lceil \frac{n}{2}\rceil}} \geq -M\sqrt{\frac{\xi}{n}} .$$

We finally obtain

$$\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \geq -M\sqrt{\frac{\xi}{n}}\left(1 + \frac{1}{\sqrt{2}}\right) .$$

Finally using Lemma 3.2.5 with true and empirical distributions yields the conclusion.

3.6.3 Proofs for section 3.4

A key property necessary for the proof of Theorem 3.4.2 is established in the following Lemma:

Lemma 3.6.1. *Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues, and γ_d be defined as in equation (3.36). For any $V \in \mathcal{V}_d$, there exists $H_V \in \mathcal{V}_d$ such that*

$$R(H_V) = \min_{H \in \mathcal{V}_d} R(H) \,, \tag{3.41}$$

and

$$\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle^2\right] \leq 2\gamma_d^{-1} \sqrt{\mathbb{E}\left[k^4(X, X')\right]}\,\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle\right] .$$

Proof. Let us denote W_i the eigenspace associated to eigenvalue μ_i and $\overline{W}_j = \bigoplus_{i=1}^{j} W_i$. We first assume $\widetilde{d} > 1$ and denote k, ℓ the fixed integers such that $\lambda_{d-\ell} = \mu_{\widetilde{d}-1}$, $\lambda_{d-\ell+1} = \ldots = \lambda_d = \ldots = \lambda_{d+k} = \mu_{\widetilde{d}}$ and $\lambda_{d+k+1} = \mu_{\widetilde{d}+1}$.

Step 1: construction of H_V.

Let $(\phi_1, \ldots, \phi_{d-\ell})$ be an orthonormal basis of $\overline{W}_{\widetilde{d}-1}$. Let $V^{(1)}$ denote the orthogonal projection of $\overline{W}_{\widetilde{d}-1}$ on V; in other words, the space spanned by the projections of $(\phi_i)_{i \leq d-\ell}$ on V. The space $V^{(1)}$ is of dimension $d - \ell' \leq d - \ell$; let $(f_1, \ldots, f_{d-\ell'})$ denote an orthonormal basis of $V^{(1)}$. We complete this basis arbitrarily to an orthonormal basis $(f_i)_{i \leq d}$ of V.

Denote now $V^{(2)} = \operatorname{span}\{f_{d-\ell+1}, \ldots, f_d\}$. Note that by construction, $V^{(2)} \perp \overline{W}_{\widetilde{d}-1}$. Let $W_{\widetilde{d}}^{(2)}$ be the orthogonal projection of $V^{(2)}$ on $W_{\widetilde{d}}$. The space $W_{\widetilde{d}}^{(2)}$ is of dimension $\ell'' \leq \ell$; let $(\phi_{d-\ell+1}, \ldots, \phi_{d+\ell''-\ell})$ be an orthogonal basis of $W_{\widetilde{d}}^{(2)}$. We finally complete this basis arbitrarily to an orthonormal basis $(\phi_i)_{d-\ell+1 \leq i \leq d+k}$ of $W_{\widetilde{d}}$. Note that by construction, in particular $V^{(2)} \perp \operatorname{span}\{\phi_{d+1}, \ldots, \phi_{d+k}\}$.

We now define $H_V = \operatorname{span}\{\phi_i, 1 \leq i \leq d\}$. Obviously H_V is a minimizer of the reconstruction error over subspaces of dimension d. We have (using Lemma 3.2.4 (ii) at the first line)

$$\begin{aligned}
\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle^2\right] &= \left\langle \Pi_{H_V} - \Pi_V, C_2 \Pi_{H_V} - \Pi_V \right\rangle_{\mathrm{HS}(\mathcal{H}_k)} \\
&\leq \|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))} \|\Pi_{H_V} - \Pi_V\|^2_{\mathrm{HS}(\mathcal{H}_k)} \\
&= 2\|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))} (d - \langle \Pi_V, \Pi_{H_V} \rangle_{\mathrm{HS}(\mathcal{H}_k)}) \\
&= 2\, \|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))} \left(d - \sum_{i,j=1}^{d} \langle f_i, \phi_j \rangle^2 \right) ;
\end{aligned}$$

and on the other hand, using Lemma 3.2.4 (i):

$$\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle\right] = \langle \Pi_{H_V} - \Pi_V, C_1 \rangle = \sum_{i=1}^{d} \left(\lambda_i - \langle f_i, C_1 f_i \rangle \right) .$$

We will decompose the last sum into two terms, for indices i smaller or greater than $d - \ell$, and bound these separately.

Step 2a: indices $i \leq d-\ell$. In this case we decompose $f_i = \sum_{j\leq d-\ell} \langle f_i, \phi_j \rangle \phi_j + g_i$, with $g_i \in \overline{W}^{\perp}_{\widetilde{d}-1}$. We have

$$\langle g_i, C_1 g_i \rangle \leq \mu_{\widetilde{d}} \|g_i\|^2 = \mu_{\widetilde{d}} \left(1 - \sum_{j\leq d-\ell} \langle f_i, \phi_j \rangle^2 \right),$$

and

$$\begin{aligned}\sum_{i=1}^{d-\ell} (\lambda_i - \langle f_i, C_1 f_i \rangle) &\geq \sum_{i=1}^{d-\ell} \lambda_i \left(1 - \sum_{j=1}^{d-\ell} \langle f_i, \phi_j \rangle^2 \right) - \sum_{i=1}^{d-\ell} \mu_{\widetilde{d}} \left(1 - \sum_{j\leq d-\ell} \langle f_i, \phi_j \rangle^2 \right) \\ &\geq \left(\mu_{\widetilde{d}-1} - \mu_{\widetilde{d}} \right) \left(d - \ell - \sum_{i,j=1}^{d-\ell} \langle f_i, \phi_j \rangle^2 \right).\end{aligned}$$

Step 2b: indices $i > d-\ell$. In this case remember that $f_i \perp \phi_j$ for $1 \leq j \leq d-\ell$ and $d+1 \leq j \leq d+k$. We can therefore decompose $f_i = \sum_{j=d-\ell+1}^{d} \langle f_i, \phi_j \rangle \phi_j + g'_i$ with $g'_i \in \overline{W}^{\perp}_{\widetilde{d}}$. We have

$$\langle g'_i, C_1 g'_i \rangle \leq \mu_{\widetilde{d}+1} \left\| g'_i \right\|^2 = \mu_{\widetilde{d}+1} \left(1 - \sum_{j=d-\ell+1}^{d} \langle f_i, \phi_j \rangle^2 \right),$$

and

$$\begin{aligned}\sum_{i=d-\ell+1}^{d} (\lambda_i - \langle f_i, C_1 f_i \rangle) &= \mu_{\widetilde{d}} \left(\ell - \sum_{i,j=d-\ell+1}^{d} \langle f_i, \phi_j \rangle^2 \right) - \sum_{i=d-\ell+1}^{d} \langle g'_i, C_1 g'_i \rangle \\ &\geq \left(\mu_{\widetilde{d}} - \mu_{\widetilde{d}+1} \right) \left(\ell - \sum_{i,j=d-\ell+1}^{d} \langle f_i, \phi_j \rangle^2 \right).\end{aligned}$$

Finally collecting the results of steps 2a-b we obtain

$$\begin{aligned}\langle \Pi_{H_V} - \Pi_V, C_1 \rangle &\geq \min \left(\mu_{\widetilde{d}-1} - \mu_{\widetilde{d}}, \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1} \right) \left(d - \sum_{i,j=1}^{d-\ell} \langle f_i, \phi_j \rangle^2 - \sum_{i,j=d-\ell+1}^{d} \langle f_i, \phi_j \rangle^2 \right) \\ &\geq \min \left(\mu_{\widetilde{d}-1} - \mu_{\widetilde{d}}, \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1} \right) (2 \|C_2\|)^{-1} \mathbb{E} \left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle^2 \right].\end{aligned}$$

Finally, it holds that $\|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))} = \|K_2\|_{\mathrm{HS}(L_2(P))}$ by Lemma 3.2.4 (iv); since K_2 is an integral operator with kernel $k^2(x,y)$, we have $\|K_2\|^2_{\mathrm{HS}(L_2(P))} = \int k^4(x,y) dP(x) dP(y) = \mathbb{E}\left[k^4(X,X')\right]$. This concludes the proof of the Lemma when $\widetilde{d} > 1$. If $\widetilde{d} = 1$, the proof can be adapted with minor modifications, essentially removing step **(2a)**, so that in the final inequality only the second term of the minimum appears. □

3.7 Appendix B: Local Rademacher Complexities.

In this section we recall a fundamental Theorem that is the key to controlling deviations of empirical processes using local Rademacher averages defined either from the true or the empirical distribution. It is a simplified version of Theorems 3.3 and 4.1 of [BBM03]. In the terminology of the latter reference, a *sub-root* function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is nonnegative, nondecreasing, and such that $\psi(r)/\sqrt{r}$ is non increasing. Then it can be shown that the fixed point equation $\psi(r) = r$ has a unique positive solution (except for the trivial case $\psi \equiv 0$). Also we recall the following notation for Rademacher complexities:

$$R_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\,,$$

where (ε_i) are i.i.d. Rademacher.

Theorem 3.7.1 (Bartlett, Bousquet and Mendelson). *Let $\mathcal{F}$ be a class of functions with ranges in $[-1,1]$ and assume that there exists some constant $B>0$ such that for every $f\in\mathcal{F}$, $Pf^2 \leq BPf$. Let ψ be a sub-root function and r^* be the fixed point of ψ. If ψ satisfies*

$$\psi(r) \geq B\mathbb{E}_{X,\varepsilon}R_n\left\{f\in \mathrm{star}(\mathcal{F}) : Pf^2 \leq r\right\}\,,$$

then for any $K>1$ and $x>0$, with probability at least $1-e^{-x}$,

$$\forall f\in\mathcal{F},\quad Pf \leq \frac{K}{K-1}P_nf + \frac{6K}{B}r^* + \frac{x(11+5BK)}{n}\,; \tag{3.42}$$

also, with probability at least $1-e^{-x}$,

$$\forall f\in\mathcal{F},\quad P_nf \leq \frac{K+1}{K}Pf + \frac{6K}{B}r^* + \frac{x(11+5BK)}{n}\,. \tag{3.43}$$

Furthermore, if $\widehat{\psi}_n$ is a data-dependent sub-root function with fixed point $\widehat{r}^*$ such that

$$\widehat{\psi}_n(r) \geq 2(10\vee B)\mathbb{E}_\varepsilon R_n\left\{f\in \mathrm{star}(\mathcal{F})\,:\, P_nf^2 \leq 2r\right\} + \frac{(2(10\vee B)+11)x}{n}\,, \tag{3.44}$$

then with probability $1-2e^{-x}$, it holds that $\widehat{r}^* \geq r^*$; as a consequence, with probability $1-3e^{-x}$, inequality (3.42) *holds with r^* replaced by $\widehat{r}^*$; similarly for inequality* (3.43).

We complete this section with the following Lemma which can be used to obtain upper bounds on fixed points of functions of the form (3.44):

Lemma 3.7.2 (inspired by [Bou02a]). *Let ϕ be a sub-root function and let $\phi_1(r) = \alpha\phi(r)+\beta$ with $\alpha>1$ and $\beta>0$. Let r^* (resp. r_1^*) denote the fixed point of ϕ (resp. ϕ_1). We have:*

$$r_1^* \leq \inf_{K>1}\left(K\alpha^2 r^* + \frac{\sqrt{K}}{\sqrt{K}-1}\beta\right)\,.$$

Proof. [BBM03] has shown that the fixed point r^* of a sub-root function Ψ satisfies the following property:

$$r^* \leq r \text{ if and only if } \Psi(\mathrm{r}) \leq \mathrm{r}\,. \tag{3.45}$$

Let $a > 1$ and $b > 0$,

$$\alpha\phi(ar^* + b\beta) + \beta = \alpha\phi\left(a\left(r^* + \frac{b}{a}\beta\right)\right) + \beta \leq \alpha\sqrt{a}\phi\left(r^* + \frac{b}{a}\beta\right) + \beta\,,$$

where the inequality uses $a\left(r^* + \frac{b}{a}\beta\right) \geq r^* + \frac{b}{a}\beta$ and that $\phi(r)/\sqrt{r}$ is a decreasing function. Moreover, since $r^* + \frac{b}{a}\beta \geq r^*$, using property (3.45), we get

$$\phi\left(r^* + \frac{b}{a}\beta\right) \leq r^* + \frac{b}{a}\beta\,.$$

Finally, gathering the previous inequalities, the following holds

$$\phi_1(ar^* + b\beta) \leq \alpha\sqrt{a}r^* + \beta\left(1 + \alpha\frac{b}{\sqrt{a}}\right)\,.$$

Let $K > 1$. Choosing $a = K\alpha^2$ and $b = \frac{\sqrt{K}}{\sqrt{K}-1}$ yields $\phi_1(ar^* + b\beta) \leq ar^* + b\beta$. Since ϕ_1 is still sub-root, using again property (3.45), this entails Lemma 3.7.2. □

3.8 Appendix C: Localized Rademacher Averages on Ellipsoids.

In this section we group together results that deal with estimating localized Rademacher complexities of function classes given as ellipsoids of a reproducing kernel Hilbert space. We deduce as corollaries the results necessary for the proofs of Theorems 3.3.4 and 3.4.2.

Theorem 3.8.1. *Let $\mathcal{H}$ be a separable Hilbert space and $(Z_i)_{1\leq i\leq n} \in \mathcal{H}^n$. Let A be a compact self-adjoint positive linear operator of $\mathcal{H}$ and $(\Phi_i)_{i\geq 1}$ an orthonormal basis of $\mathcal{H}$ of eigenvectors of A. Denote $B_\alpha = \{\|v\| \leq \alpha\}$, $\mathcal{E}_r = \{\langle v, Av\rangle \leq r\}$ and let (ε_i) be an i.i.d. family of Rademacher random variables. Then for any integer $h \leq Rank(A)$, the following holds:*

$$\mathbb{E}_\varepsilon \sup_{v\in B_\alpha\cap\mathcal{E}_r} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle v, Z_i\rangle \leq \frac{\sqrt{r}}{n}\sqrt{\sum_{i=1}^h \frac{1}{\lambda_i(A)}\sum_{j=1}^n \langle Z_j, \Phi_i\rangle^2} + \frac{\alpha}{n}\sqrt{\sum_{i\geq h+1}\sum_{j=1}^n \langle Z_j, \Phi_i\rangle^2}\,. \quad (3.46)$$

Proof. For $v \in B_\alpha \cap \mathcal{E}_r$, we have

$$\begin{aligned}
\sum_{i=1}^n \varepsilon_i\langle v, Z_i\rangle &= \sum_{j=1}^h \langle v, \Phi_j\rangle\left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i Z_i\right\rangle + \sum_{j>h}\langle v, \Phi_j\rangle\left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i Z_i\right\rangle \\
&\leq \sqrt{r\sum_{i=1}^h \frac{1}{\lambda_i(A)}\left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i\right\rangle^2} + \alpha\sqrt{\sum_{i\geq h+1}\left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i\right\rangle^2}\,,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality for both terms, $\lambda_i > 0$ (since $h \leq Rank(A)$) and the equality $\langle v, Av\rangle = \sum_{i\geq 1}\lambda_i(A)\langle v, \Phi_i\rangle^2$. We now integrate over (ε_i); using Jensen's inequality the square roots are pulled outside of the expectation; finally, we have

$$\mathbb{E}_\varepsilon\left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i\right\rangle^2 = \sum_{j=1}^n \langle Z_j, \Phi_i\rangle^2\,.$$

since by independence the cross-terms vanish. This concludes the proof. □

We deduce the two following corollaries of Theorem 3.8.1:

Corollary 3.8.2. *Define $\mathcal{F}_d = \{x \mapsto \langle \Pi_V, C_x \rangle, V \in \mathcal{V}_d\}$. Then the following holds:*

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \mathrm{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right), \tag{3.47}$$

and

$$\mathbb{E}_{\varepsilon} R_n \left\{ f \in \mathrm{star}(\mathcal{F}_d), P_n f^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_{2,n})} \right). \tag{3.48}$$

Proof. The proof is the same for the two inequalities. We will apply Theorem 3.8.1 in the Hilbert space $\mathrm{HS}(\mathcal{H}_k)$. We can suppose that $h \leq Rank(C_2)$ since the obtained result extend straightforwardly to $h > Rank(C_2)$. We have for any $V \in \mathcal{V}_d$, $\|\Pi_V\|_{\mathrm{HS}(\mathcal{H}_k)} \leq \sqrt{d}$, and hence $\mathcal{F}_d \subset \{x \mapsto \langle \Gamma, C_x \rangle ; \Gamma \in B_{\sqrt{d}}(\mathrm{HS}(\mathcal{H}_k))\}$. Since the latter set is convex and contains the origin, it therefore also contains $\mathrm{star}(\mathcal{F}_d)$. Furthermore, by Lemma 3.2.4, $P \langle \Gamma, C_X \rangle^2 = \langle \Gamma, C_2 \Gamma \rangle$.

We can therefore apply Theorem 3.8.1 with $\alpha = \sqrt{d}$, $A = C_2$, $Z_i = C_{X_i}$, $v = \Pi_V$, leading to

$$\mathbb{E}_{\varepsilon} R_n \left\{ f \in \mathrm{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{\sqrt{r}}{n} \sqrt{\sum_{i=1}^{h} \frac{1}{\lambda_i(C_2)} \sum_{j=1}^{n} \langle C_{X_j}, \Phi_i \rangle^2} + \frac{\sqrt{d}}{n} \sqrt{\sum_{i \geq h+1} \sum_{j=1}^{n} \langle C_{X_j}, \Phi_i \rangle^2}.$$

Integrating with respect to Z leads to

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \mathrm{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right),$$

since $\mathbb{E}\left[\langle C_X, \Phi_i \rangle^2 \right] = \langle \Phi_i, C_2 \Phi_i \rangle = \lambda_i(C_2)$. We obtain (3.48) in the same way by taking $A = C_{2,n}$ instead of C_2. □

Corollary 3.8.3. *Define $\widetilde{\mathcal{F}}_d = \left\{ x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x \right\rangle, V \in \mathcal{V}_d \right\}$, where H_V is defined via Lemma 3.6.1. Then the following holds:*

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \mathrm{star}(\widetilde{\mathcal{F}}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left\{ \sqrt{rh} + 2\sqrt{d \sum_{j > h} \lambda_j(K_2)} \right\}, \tag{3.49}$$

and

$$\mathbb{E}_{\varepsilon} R_n \left\{ f \in \mathrm{star}(\widetilde{\mathcal{F}}_d), P_n f^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left\{ \sqrt{rh} + 2\sqrt{d \sum_{j > h} \lambda_j(K_{2,n})} \right\}. \tag{3.50}$$

Proof. We recall that H_V satisfies:

$$R(H_V) = \min_{H \in \mathcal{V}_d} R(H)\,.$$

Note that $\Pi_{V^\perp} - \Pi_{H_V^\perp} = \Pi_{H_V} - \Pi_V$. The proof is then almost the same as for Corollary 3.8.2, with the minor change $\alpha = 2\sqrt{d}$ since $\|\Pi_V - \Pi_{H_V}\|^2_{\mathrm{HS}(\mathcal{H}_k)} \leq 4d$. □

We finally give the following Lemma to estimate the fixed points of sub-root functions of the above form.

Lemma 3.8.4. *If $(\lambda_i)_{i>0}$ is a positive convergent series, denoting by ψ the function*

$$\psi(r) := \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left\{ \sqrt{hr} + \alpha \sqrt{\sum_{j \geq h+1} \lambda_j} \right\},$$

it holds that ψ is a sub-root function and the unique positive solution r^* of $\psi(r) = r/c$ where $c > 0$ satisfies

$$r^* \leq \inf_{h \geq 0} \left\{ \frac{c^2 h}{n} + \frac{2c\alpha}{\sqrt{n}} \sqrt{\sum_{j \geq h+1} \lambda_j} \right\}.$$

Proof. It is easy to see any infimum of sub-root functions is sub-root, hence $c\psi$ is sub-root. Existence and uniqueness of a solution is proved by [BBM03]. To obtain the announced bound, we solve $r^* \leq \frac{c}{\sqrt{n}} \left\{ \sqrt{hr^*} + \alpha \sqrt{\sum_{j \geq h+1} \lambda_j} \right\}$ for each $h \geq 0$ (by using the fact that $x \leq A\sqrt{x} + B$ implies $x \leq A^2 + 2B$), and take the infimum over h. □

3.9 Appendix D: Concentration Inequalities.

To make the chapter self-contained, some concentration inequalities used all along the chapter are now recalled.

Theorem 3.9.1 ([McD98]). *Let $X_1, ..., X_n$ be n independent random variables. Let us consider $Z = f(X_1, ..., X_n)$ where f is such that:*

$$\sup_{x_1, ..., x_n, x_i' \in X} |f(x_1, ..., x_n) - f(x_1, ., x_i', ., x_n)| \leq c_i\,,\ \forall\, 1 \leq i \leq n\,.$$

Then

$$P[Z - \mathbb{E}[Z] \geq \xi] \leq e^{-2\xi^2/(c_1^2 + ... + c_n^2)},$$

and

$$P[\mathbb{E}[Z] - Z \geq \xi] \leq e^{-2\xi^2/(c_1^2 + ... + c_n^2)}.$$

Theorem 3.9.2 ([Hoe63]). *Let $1 \leq r \leq n$ and $X_1, \ldots, X_n$ be n independent random variables. Denote*

$$U = \frac{1}{n(n-1)\ldots(n-r+1)} \sum_{i_1 \neq \ldots \neq i_r} g(X_{i_1}, \ldots, X_{i_r})\,.$$

If g has range in $[a,b]$ then

$$\mathbb{P}\left[U - \mathbb{E}\left[U\right] \geq t\right] \leq e^{-2\lceil n/r \rceil t^2/(b-a)^2}\,,$$

and

$$\mathbb{P}\left[\mathbb{E}\left[U\right] - U \geq t\right] \leq e^{-2\lceil n/r \rceil t^2/(b-a)^2}\,.$$

Theorem 3.9.3 ([Lug00]). *Let f be a bounded function. With probability at least $1-e^{-\xi}$,*

$$(P-P_n)(f) \leq \sqrt{\frac{2\xi Pf^2}{n}} + \frac{\|f\|_\infty \xi}{3n}\,, \tag{3.51}$$

and with probability at least $1-e^{-\xi}$,

$$(P_n-P)(f) \leq \sqrt{\frac{2\xi Pf^2}{n}} + \frac{\|f\|_\infty \xi}{3n}\,. \tag{3.52}$$

Moreover, supposing that $f(X) \geq 0$, we get that with probability at least $1-e^{-\xi}$, for all $K>1$,

$$P_n f - \frac{K-1}{K} Pf \geq -\frac{5K\|f\|_\infty \xi}{6n}\,, \tag{3.53}$$

and with probability at least $1-e^{-\xi}$, for all $K>1$,

$$Pf - \frac{K}{K+1} P_n f \geq -\frac{5K\|f\|_\infty \xi}{6n}\,. \tag{3.54}$$

Proof. Inequality (3.51) is proved in, e.g., [Lug00]. In order to prove inequality (3.52), it suffices to consider $-f$ in (3.51).

We now prove inequality (3.53). The proof of inequality (3.54) follows the same line: it suffices to use inequality (3.52) instead of (3.51). $f(X) \geq 0$ implies $Pf^2 \leq \|f\|_\infty Pf$. Moreover, using the elementary inequality $\sqrt{ab} \leq a/K + Kb/4$, we get

$$\sqrt{\frac{2\xi Pf^2}{n}} \leq \frac{1}{K} Pf + \frac{K\xi\|f\|_\infty}{2n}\,.$$

Gathering this inequality with (3.51) and solving the corresponding inequality, we get (3.53). □

Chapter 4

On the Convergence of Eigenspaces in Kernel Principal Component Analysis

Most of this chapter is already published ([BZ05]). The appendices provide technical backgrounds and a work in progress. It is a joint work with G. Blanchard.

Contents

Abstract

This chapter presents a non-asymptotic statistical analysis of Kernel-PCA with a focus different from the one proposed in previous work on this topic ([STWCK05], chapter 3). Here instead of considering the reconstruction error of KPCA we are interested in approximation error bounds for the eigenspaces themselves. We prove an upper bound depending on the spacing between eigenvalues but not on the dimensionality of the eigenspace. As a consequence this allows to infer stability results for these estimated spaces.

4.1 Introduction.

Principal Component Analysis (PCA for short in the sequel) is a widely used tool for data dimensionality reduction. It consists in finding the most relevant lower-dimension projection of some data in the sense that the projection should keep as much of the variance of the original data as possible. If the target dimensionality of the projected data is fixed in advance, say D – an assumption that we will make throughout the present chapter – the solution of this problem is obtained by considering the projection on the span S_D of the first D eigenvectors of the covariance matrix. Here by 'first D eigenvectors' we mean eigenvectors associated to the D largest eigenvalues counted with multiplicity; hereafter with some abuse the span of the first D eigenvectors will be called "D-eigenspace" for short when there is no risk of confusion.

The introduction of the 'Kernel trick' has allowed to extend this methodology to data mapped in a kernel feature space, then called KPCA [SSM98]. The interest of this extension is that, while still linear in feature space, it gives rise to *nonlinear* interpretation in original space – vectors in the kernel feature space can be interpreted as nonlinear functions on the original space.

For PCA as well as KPCA, the true covariance matrix (resp. covariance operator) is not known and has to be estimated from the available data, a procedure which in the case of Kernel spaces is linked to the so-called Nyström approximation [WS00]. The subspace given as an output is then obtained as D-eigenspace $\widehat{S}_D$ of the *empirical* covariance matrix or operator. An interesting question from a statistical or learning theoretical point of view is then, how reliable this estimate is.

This question has already been studied ([STWCK05], chapter 3) from the point of view of the *reconstruction error* of the estimated subspace. What this means is that (assuming the data is centered in Kernel space for simplicity) the average reconstruction error (square norm of the distance to the projection) of $\widehat{S}_D$ converges to the (optimal) reconstruction error of S_D and that bounds are known about the rate of convergence. However, this does not tell us much about the convergence of S_D to $\widehat{S}_D$ – since two very different subspaces can have a very similar reconstruction error, in particular when some eigenvalues are very close to each other (the gap between the eigenvalues will actually appear as a central point of the analysis to come).

In the present work, we set to study the behavior of these D-eigenspaces themselves: we provide finite sample bounds describing the closeness of the D-eigenspaces of the empirical covariance operator to the true one. There are several broad motivations for this analysis. First, the reconstruction error alone is a valid criterion only if one really plans to perform dimensionality reduction of the data and stop there. However, PCA is often used merely as a *preprocessing* step and the projected data is then submitted to further processing (which could be classification, regression or something else). In particular for KPCA, the projection subspace in the kernel space can be interpreted as a subspace of *functions* on the original space; one then expects these functions to be relevant for the data at hand and for some further task (see e.g. [BMVZ04]). In these cases, if we want to analyze the full procedure (from a learning theoretical sense), it is desirable to have a more precise information on the selected subspace than just its reconstruction error. In particular, from a learning complexity point of view, it is important to ensure that functions used for learning stay in a set of limited complexity, which is ensured if the selected subspace is stable (which is a consequence of its convergence).

The approach we use here is based on perturbation bounds and we essentially walk in the steps pioneered by Kolchinskii and Giné [KG00] (see also [DPR82]) using tools of operator perturbation theory [Kat66]. Similar methods have been used to prove consistency of spectral clustering [vLBB04b, vLBB04a]. An important difference here is that we want to study directly the convergence of the whole subspace spanned by the first D eigenvectors instead of the separate convergence of the individual eigenvectors; in particular we are interested in how D acts as a complexity parameter. The important point in our main result is that it does not: only the gap between the D-th and the $(D+1)$-th eigenvalue comes into account. This means that there in no increase in complexity (as far as this bound is concerned: of course we cannot exclude that better bounds can be obtained in the future) between estimating the D-th eigenvector alone or the span of the first D eigenvectors.

Our contribution in the present work is thus

- to adapt the operator perturbation result of [KG00] to D-eigenspaces.
- to get non-asymptotic bounds on the approximation error of Kernel-PCA eigenspaces thanks to the previous tool.

In section 4.2 we introduce shortly the notation, explain the main ingredients used and obtain a first bound based on controlling separately the first D eigenvectors, and depending on the dimension D. In section 4.3 we explain why the first bound is actually suboptimal and derive an improved bound as a consequence of an operator perturbation result that is more adapted to our needs and deals directly with the D-eigenspace as a whole. Section 4.4 concludes and discusses the obtained results. Mathematical proofs are found in the appendix.

4.2 First result.

Notation. The interest variable X takes its values in some measurable space $\mathcal{X}$, following the distribution P. We consider KPCA and are therefore primarily interested in the mapping of X into a reproducing kernel Hilbert space $\mathcal{H}$ with kernel function k through the feature mapping $\varphi(x) = k(x, \cdot)$. The objective of the kernel PCA procedure is to recover a D-dimensional subspace S_D of $\mathcal{H}$ such that the projection of $\varphi(X)$ on S_D has maximum averaged squared norm.

All operators considered in what follows are Hilbert-Schmidt and the norm considered for these operators will be the Hilbert-Schmidt norm unless precised otherwise. Furthermore we only consider symmetric nonnegative operators, so that they can be diagonalized and have a discrete spectrum.

Let C denote the covariance operator of variable $\varphi(X)$ (see Chapter 3 for an exact mathematical definition). To simplify notation we assume that nonzero eigenvalues $\lambda_1 > \lambda_2 > \dots$ of C are all simple (This is for convenience only. In the conclusion we discuss what changes have to be made if this is not the case). Let $\phi_1, \phi_2, \dots$ be the associated eigenvectors. It is well-known that the optimal D-dimensional reconstruction space is $S_D = \mathrm{span}\{\phi_1, \dots, \phi_D\}$. The KPCA procedure approximates this objective by considering the empirical covariance operator, denoted C_n, and the subspace $\widehat{S}_D$ spanned by its first D eigenvectors. We denote $P_{S_D}, P_{\widehat{S}_D}$ the orthogonal projectors on these spaces.

A first bound. Broadly speaking, the main steps required to obtain the type of result we are interested in are

1. A non-asympotic bound on the (Hilbert-Schmidt) norm of the difference between the empirical and the true covariance operators;

2. An operator perturbation result bounding the difference between spectral projectors of two operators by the norm of their difference.

The combination of these two steps leads to our goal. The first step consists in the following Lemma:

Lemma 4.2.1 (Corollary 5 of [STC03]). *Supposing that* $\sup_{x\in\mathcal{X}} k(x,x) \leq M$*, for all* $\xi > 0$*, with probability greater than* $1 - e^{-\xi}$*,*

$$\|C_n - C\| \leq \frac{2M}{\sqrt{n}}\left(1 + \sqrt{\frac{\xi}{2}}\right).$$

Proof of Lemma 4.2.1. The proof ensures that this lemma is available in infinite dimensional setting.
$\|C_n - C\| = \|\frac{1}{n}\sum_{i=1}^n C_{X_i} - \mathbb{E}\left[C_X\right]\|$ with $\|C_X\| = \|\varphi(X) \otimes \varphi(X)^*\| = k(X,X) \leq M$. If we denote $H(x_1,\cdots,x_n) = \|\frac{1}{n}\sum_{i=1}^n C_{x_i} - \mathbb{E}\left[C_x\right]\|$, then we have for any i_0:

$$|H(x_1,\cdots,x_n) - H(x_1,\cdots,x_{i_0-1},x'_{i_0},x_{i_0+1},\cdots,x_n)| \leq \frac{1}{n}\|C_{x'_i} - C_{x'_{i_0}}\| \leq \frac{2M}{n}.$$

Therefore, we can apply the bounded difference inequality (Theorem 3.9.1) to the variable $H(X_1,\cdots,X_n)$, so that, for all $\xi > 0$, with probability greater than $1 - e^{-\xi}$,

$$\|C_n - C\| \leq \mathbb{E}\left[\|C_n - C\|\right] + 2M\sqrt{\frac{\xi}{2n}}. \tag{4.1}$$

Moreover, by Jensen's inequality $\mathbb{E}\left[\|C_n - C\|\right] \leq \mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n C_{X_i} - \mathbb{E}\left[C_X\right]\|^2\right]^{\frac{1}{2}}$, and simple calculations leads to $\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n C_{X_i} - \mathbb{E}\left[C_X\right]\|^2\right] = \frac{1}{n}\mathbb{E}\left[\|C_X - \mathbb{E}\left[C_X\right]\|^2\right] \leq \frac{4M^2}{n}$. This concludes the proof of lemma 4.2.1. □

As for the second step, [KG00] provides the following perturbation bound (see also e.g. [vLBB04b]):

Theorem 4.2.2 (Simplified Version of [KG00], Theorem 5.2). *Let A be a symmetric positive Hilbert-Schmidt operator of the Hilbert space $\mathcal{H}$ with simple positive eigenvalues $\lambda_1 > \lambda_2 > \ldots$ For an integer r such that $\lambda_r > 0$, let $\widetilde{\delta}_r = \delta_r \wedge \delta_{r-1}$ where $\delta_r = \frac{1}{2}(\lambda_r - \lambda_{r+1})$. Let $B \in HS(\mathcal{H})$ be another symmetric operator such that $\|B\| < \widetilde{\delta}_r/2$ and $(A+B)$ is still a positive operator with simple nonzero eigenvalues.*

Let $P_r(A)$ (resp. $P_r(A+B)$) denote the orthogonal projector onto the subspace spanned by the r-th eigenvector of A (resp. $(A+B)$). Then, these projectors satisfy:

$$\|P_r(A) - P_r(A+B)\| \leq \frac{2\|B\|}{\widetilde{\delta}_r}.$$

Remark about the Approximation Error of the Eigenvectors: let us recall that a control over the Hilbert-Schmidt norm of the projections onto eigenspaces imply a control on the approximation errors of the eigenvectors themselves. Indeed, let ϕ_r, ψ_r denote the (normalized) r-th eigenvectors of the operators above with signs chosen so that $\langle \phi_r, \psi_r \rangle > 0$. Then

$$\|P_{\phi_r} - P_{\psi_r}\|^2 = 2(1 - \langle \phi_r, \psi_r \rangle^2) \geq 2(1 - \langle \phi_r, \psi_r \rangle) = \|\phi_r - \psi_r\|^2 .$$

Now, the orthogonal projector on the direct sum of the first D eigenspaces is the sum $\sum_{r=1}^{D} P_r$. Using the triangle inequality, and combining Lemma 4.2.1 and Theorem 4.2.2, we conclude that with probability at least $1 - e^{-\xi}$ the following holds:

$$\left\| P_{S_D} - P_{\widehat{S}_D} \right\| \leq \left(\sum_{r=1}^{D} \widetilde{\delta}_r^{-1} \right) \frac{4M}{\sqrt{n}} \left(1 + \sqrt{\frac{\xi}{2}} \right),$$

provided that $n \geq 16M^2 \left(1 + \sqrt{\frac{\xi}{2}} \right)^2 (\sup_{1 \leq r \leq D} \widetilde{\delta}_r^{-2})$. The disadvantage of this bound is that we are penalized on the one hand by the (inverse) gaps between the eigenvalues, and on the other by the dimension D (because we have to sum the inverse gaps from 1 to D). In the next section we improve the operator perturbation bound to get an improved result where only the gap δ_D enters into account.

4.3 Improved Result.

We first prove the following variant on the operator perturbation property which better corresponds to our needs by taking directly into account the projection on the first D eigenvectors:

Theorem 4.3.1. *Let A be a symmetric positive Hilbert-Schmidt operator of the Hilbert space $\mathcal{H}$ with simple nonzero eigenvalues $\lambda_1 > \lambda_2 > \ldots$ Let $D > 0$ be an integer such that $\lambda_D > 0$, $\delta_D = \frac{1}{2}(\lambda_D - \lambda_{D+1})$. Let $B \in HS(\mathcal{H})$ be another symmetric operator such that $\|B\| < \delta_D/2$ and $(A + B)$ is still a positive operator. Let $P^D(A)$ (resp. $P^D(A + B)$) denote the orthogonal projector onto the subspace spanned by the first D eigenvectors A (resp. $(A + B)$). Then these satisfy:*

$$\|P^D(A) - P^D(A + B)\| \leq \frac{\|B\|}{\delta_D} . \tag{4.2}$$

Proof of Theorem 4.3.1. The key property of Hilbert-Schmidt operators allowing to work directly in a infinite dimensional setting is that $HS(\mathcal{H})$ is a both right and left ideal of $\mathcal{L}_c(\mathcal{H}, \mathcal{H})$, the Banach space of all continuous linear operators of $\mathcal{H}$ endowed with the operator norm $\|.\|_{\text{op}}$. Indeed, $\forall T \in HS(\mathcal{H})$, $\forall S \in \mathcal{L}_c(\mathcal{H}, \mathcal{H})$, TS and ST belong to $HS(\mathcal{H})$ with

$$\|TS\| \leq \|T\| \, \|S\|_{\text{op}} \quad \text{and} \quad \|ST\| \leq \|T\| \, \|S\|_{\text{op}} . \tag{4.3}$$

The spectrum of an Hilbert-Schmidt operator T is denoted $\Lambda(T)$ and the sequence of eigenvalues in non-increasing order is denoted $\lambda(T) = (\lambda_1(T) \geq \lambda_2(T) \geq \ldots)$. In the following, $P^D(T)$ denotes the orthogonal projector onto the D-eigenspace of T.

The Hoffmann-Wielandt inequality in infinite dimensional setting ([BE94]) yields that:

$$\|\lambda(A) - \lambda(A+B)\|_{\ell_2} \leq \|B\| \leq \frac{\delta_D}{2}. \tag{4.4}$$

This implies in particular that

$$\forall i > 0, \qquad |\lambda_i(A) - \lambda_i(A+B)| \leq \frac{\delta_D}{2}. \tag{4.5}$$

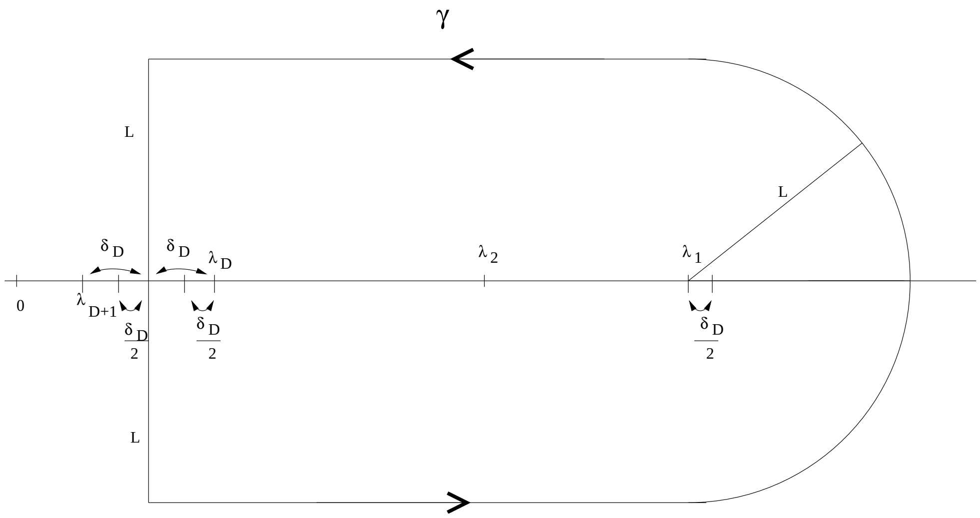
Results found in [Kat66] p.39 (more details can be found in appendix 4.6: see, e.g., Theorem 4.6.3 (2)) yield the formula

$$P^D(A) - P^D(A+B) = -\frac{1}{2i\pi}\int_\gamma (R_A(z) - R_{A+B}(z))dz \in \mathcal{L}_c(\mathcal{H},\mathcal{H}), \tag{4.6}$$

where $R_A(z) = (A - z\,Id)^{-1}$ is the resolvent of A, provided that γ is a simple closed curve in $\mathbb{C}$ enclosing exactly the first D eigenvalues of A and $(A+B)$. Moreover, the same reference (p.60) (see Theorem 4.6.3 (3)) states that for ξ in the complementary of $\Lambda(A)$,

$$\|R_A(\xi)\|_{\mathrm{op}} = dist(\xi, \Lambda(A))^{-1}. \tag{4.7}$$

The proof of the theorem now relies on the simple choice for the closed curve γ in (4.6), drawn in the picture below and consisting of three straight lines and a semi-circle of radius L. For all $L > \frac{\delta_D}{2}$, γ intersect neither the eigenspectrum of A (by equation (4.5)) nor the eigenspectrum of $A+B$. Moreover, the eigenvalues of A (resp. $A+B$) enclosed by γ are exactly $\lambda_1(A), \ldots, \lambda_D(A)$ (resp. $\lambda_1(A+B), \ldots, \lambda_D(A+B)$).



Moreover, for $z \in \gamma$, $T(z) = R_A(z) - R_{A+B}(z) = -R_{A+B}(z)BR_A(z)$ belongs to $HS(\mathcal{H})$ and depends continuously on z by ((4.3) and Theorem 4.6.3 (1)). Consequently,

$$\|P^D(A) - P^D(A+B)\| \leq \frac{1}{2\pi}\int_a^b \|(R_A - R_{A+B})(\gamma(t))\|\,|\gamma'(t)|dt.$$

Let $S_N = \sum_{n=0}^{N}(-1)^n (R_A(z)B)^n R_A(z)$. $R_{A+B}(z) = (Id + R_A(z)B)^{-1}R_A(z)$ and, for $z \in \gamma$ and $L > \delta_D$,

$$\|R_A(z)B\|_{\mathrm{op}} \le \|R_A(z)\|_{\mathrm{op}}\|B\| \le \frac{\delta_D}{2\,dist(z, \Lambda(A))} \le \frac{1}{2},$$

imply that $S_N \overset{\|\cdot\|_{\mathrm{op}}}{\longrightarrow} R_{A+B}(z)$ (uniformly for $z \in \gamma$). Using property (4.3), since $B \in HS(\mathcal{H})$, $S_N B R_A(z) \overset{\|\cdot\|}{\longrightarrow} R_{A+B}(z) B R_A(z) = R_{A+B}(z) - R_A(z)$. Finally,

$$R_A(z) - R_{A+B}(z) = \sum_{n\ge 1}(-1)^n (R_A(z)B)^n R_A(z)\,,$$

where the series converges in $HS(\mathcal{H})$, uniformly for $z \in \gamma$. Using again property (4.3) and (4.7) implies

$$\|(R_A - R_{A+B})(\gamma(t))\| \;\le\; \sum_{n\ge1}\|R_A(\gamma(t))\|_{\mathrm{op}}^{n+1}\|B\|^n \le \sum_{n\ge1}\frac{\|B\|^n}{dist^{n+1}(\gamma(t), \Lambda(A))}\,.$$

Finally, since for $L > \delta_D$, $\|B\| \le \frac{\delta_D}{2} \le \frac{dist(\gamma(t),\Lambda(A))}{2}$,

$$\|P^D(A) - P^D(A+B)\| \le \frac{\|B\|}{\pi}\int_a^b \frac{1}{dist^2(\gamma(t), \Lambda(A))}|\gamma'(t)|dt\,.$$

Splitting the last integral into four parts according to the definition of the contour γ, we obtain

$$\int_a^b \frac{1}{dist^2(\gamma(t),\Lambda(A))}|\gamma'(t)|dt \quad\le\quad \frac{2\arctan(\frac{L}{\delta_D})}{\delta_D} \;+\; \frac{\pi}{L} \;+\; 2\frac{\mu_1(A) - (\mu_D(A) - \delta_D)}{L^2}\,,$$

and letting L goes to infinity leads to the result. □

Theorem 4.3.1 then gives rise to our main result on KPCA:

Theorem 4.3.2. *Assume that* $\sup_{x\in\mathcal{X}} k(x,x) \le M$. *Let* $S_D, \widehat{S}_D$ *be the subspaces spanned by the first* D *eigenvectors of* C, *resp.* C_n *defined earlier. Denoting* $\lambda_1 > \lambda_2 > \ldots$ *the eigenvalues of* C, *if* $D > 0$ *is such that* $\lambda_D > 0$, *put* $\delta_D = \frac{1}{2}(\lambda_D - \lambda_{D+1})$ *and*

$$B_D = \frac{2M}{\delta_D}\left(1 + \sqrt{\frac{\xi}{2}}\right)\,.$$

Then provided that $n \ge B_D^2$, *the following bound holds with probability at least* $1 - e^{-\xi}$:

$$\left\|P_{S_D} - P_{\widehat{S}_D}\right\| \le \frac{B_D}{\sqrt{n}}\,. \tag{4.8}$$

This entails in particular

$$\widehat{S}_D \subset \left\{g + h, g \in S_D,\, h \in S_D^{\perp},\, \|h\|_{\mathcal{H}_k} \le B_D\; n^{-\frac{1}{2}}\|g\|_{\mathcal{H}_k}\right\}\,. \tag{4.9}$$

The important point here is that the approximation error now only depends on D through the (inverse) gap between the D-th and $(D+1)$-th eigenvalues. Note that using the results of section 4.2, we would have obtained exactly the same bound for estimating the D-th eigenvector only – or even a worse bound since $\widetilde{\delta}_D = \delta_D \wedge \delta_{D-1}$ appears in this case. Thus, at least from the point of view of this technique (which could still yield suboptimal bounds), there is no increase of complexity between estimating the D-th eigenvector alone and estimating the span of the first D eigenvectors.

Note that the inclusion (4.9) can be interpreted geometrically by saying that for any vector in $\widehat{S}_D$, the tangent of the angle between this vector and its projection on S_D is upper bounded by $B_D/\sqrt{n}$, which we can interpret as a stability property.

Comment about the Centered Case. In the actual (K)PCA procedure, the data is actually first empirically recentered, so that one has to consider the centered covariance operator $\overline{C}$ and its empirical counterpart $\overline{C}_n$. A result similar to Theorem 4.3.2 also holds in this case (up to some additional constant factors). Indeed, $\|\overline{C} - \overline{C}_n\| = \sup_{A, \|A\| \le 1} \langle \overline{C} - \overline{C}_n, A \rangle$ and the proof of Theorem 3.3.5 allows to get a result similar to Lemma 4.2.1. Combined again with Theorem 4.3.1, this allows to come to similar conclusions for the "true" centered KPCA.

Proof of Theorem 4.3.2. Lemma 4.2.1 and Theorem 4.3.1 yield inequality (4.8). Together with assumption $n \ge B_D^2$ it implies $\|P_{S_D} - P_{\widehat{S}_D}\| \le \frac{1}{2}$. Let $f \in \widehat{S}_D$: $f = P_{S_D}(f) + P_{S_D^\perp}(f)$,; Lemma 4.3.3 below with $F_D = S_D$ and $G_{D'} = \widehat{S}_D$ implies that

$$\|P_{S_D^\perp}(f)\|^2_{\mathcal{H}_k} \le \frac{4}{3}\|P_{S_D} - P_{\widehat{S}_D}\|^2 \|P_{S_D}(f)\|^2_{\mathcal{H}_k} .$$

Gathering the different inequalities, Theorem 4.3.2 is proved. □

Lemma 4.3.3. *Let F_D and $G_{D'}$ be two vector subspaces of $\mathcal{H}$ such that* $\dim(F_D) = D$ *and* $\dim(G_{D'}) = D'$. *Provided that $\|P_{F_D} - P_{G_{D'}}\| \le \frac{1}{2}$, the following bound holds:*

$$\forall\, f \in G_{D'} \,,\ \|P_{F_D^\perp}(f)\|^2_{\mathcal{H}} \le \frac{4}{3}\|P_{F_D} - P_{G_{D'}}\|^2 \|P_{F_D}(f)\|^2_{\mathcal{H}} .$$

Proof of Lemma 4.3.3. Let $g_1, \ldots, g_{D'}$ be an orthonormal basis of $G_{D'}$.

$\forall f \in G_{D'}$, $\|P_{F_D^\perp}(f)\|^2_{\mathcal{H}} = \left|\sum_{j,j'=1}^{D'} \langle f, g_j\rangle \overline{\langle f, g_{j'}\rangle} \langle P_{F_D^\perp}(g_j), P_{F_D^\perp}(g_{j'})\rangle\right|$. Thus, using twice the Cauchy-Schwarz inequality,

$$\|P_{F_D^\perp}(f)\|^2_{\mathcal{H}} \le \|f\|^2_{\mathcal{H}} \left(D' - \sum_{j=1}^{D'} \|P_{F_D}(g_j)\|^2_{\mathcal{H}} \right) .$$

Since $\|P_{F_D}\|^2 = D$, this yields

$$\|P_{F_D^\perp}(f)\|^2_{\mathcal{H}} \le \|f\|^2_{\mathcal{H}} \left(D' + D - 2\sum_{j=1}^{D'} \|P_{F_D}(g_j)\|^2_{\mathcal{H}} \right) .$$

Finally, $\|P_{F_D^\perp}(f)\|^2_{\mathcal{H}} \le \|f\|^2_{\mathcal{H}} \|P_{F_D} - P_{G_{D'}}\|^2$ and considering the Pythagoras decomposition $\|f\|^2_{\mathcal{H}} = \|P_{F_D}(f)\|^2_{\mathcal{H}} + \|P_{F_D^\perp}(f)\|^2_{\mathcal{H}}$, this concludes the proof. □

4.4 Conclusion and Discussion

In this chapter, finite sample size confidence bounds of the eigenspaces of Kernel-PCA (the D-eigenspaces of the empirical covariance operator) are provided using tools of operator perturbation theory. This provides a first step towards an in-depth complexity analysis of algorithms using KPCA as pre-processing, and towards taking into account the randomness of the obtained models (e.g. [BMVZ04]). We proved a bound in which the complexity factor for estimating the eigenspace S_D by its empirical counterpart depends only on the inverse gap between the D-th and $(D+1)$-th eigenvalues. In addition to the previously cited works, we take into account the centering of the data and obtain comparable rates.

In this work we assumed for simplicity of notation the eigenvalues to be simple. In the case the covariance operator C has nonzero eigenvalues with multiplicities $m_1, m_2, \ldots$ possibly larger than one, the analysis remains the same except for one point: we have to assume that the dimension D of the subspaces considered is of the form $m_1 + \cdots + m_r$ for a certain r. This could seem restrictive in comparison with the results obtained for estimating the sum of the first D eigenvalues themselves of chapter 3 (which is linked to the reconstruction error in KPCA) where no such restriction appears. However, it should be clear that we need this restriction when considering $D-$eigenspaces themselves since the target space has to be unequivocally defined, otherwise convergence cannot occur. Thus, it can happen in this special case that the reconstruction error converges while the projection space itself does not. Finally, a common point of the two analyses (over the spectrum and over the eigenspaces) lies in the fact that the bounds involve an inverse gap in the eigenvalues of the true covariance operator.

Finally, how tight are these bounds and do they at least carry some correct qualitative information about the behavior of the eigenspaces? Asymptotic results (central limit Theorems) in [Kol98, DPR82] always provide the correct goal to shoot for since they actually give the limit distributions of these quantities. They imply that there is still important ground to cover before bridging the gap between asymptotic and non-asymptotic. This of course opens directions for future work.

Appendix

4.5 Appendix A: a Mixed Strategy of Regularization .

This appendix provides a work in progress. It aims at giving a second step in order to analyze – from a learning theoretical sense – procedures using KPCA as pre-processing.

The study is specified to empirical risk minimizers of the following form:

$$\widehat{f}_{D,R} = \arg \min_{f \in \widehat{S}_{D,R}} \frac{1}{n} \sum_{i=1}^{n} \gamma(f, (X_i, Y_i)) \, ,$$

where γ is a lipschitz loss and

$$\widehat{S}_{D,R} = \widehat{S}_D \cap B_{\mathcal{H}}(0, R) \, .$$

$\gamma(f, (x, y))$ can be thought as the hinge loss $(1 - yf(x))_+$ or the quadratic loss $(y - f(x))^2$.

$\widehat{f}_{D,R}$ is obtained by a mixed strategy of regularization: chapter 5 shall explain that $\widehat{S}_D$ (resp. $B_{\mathcal{H}}(0, R)$) corresponds to a finite dimensional (resp. Tikhonov's) regularization. Ideally, it holds concurrently the advantages of these two types of regularization.

Let $\mathcal{F}$ be a class of functions. We use the following notation

$$R_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\,,$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. Rademacher. Roughly speaking, the measure of complexity of $\mathcal{F}$ used, e.g., in [BBM03] is the complexity radius r^*. It is a local notion of complexity relying on the control of the localized Rademacher average $\mathbb{E}R_n\{f\in\mathcal{F},\, Pf^2\le r\}$ by a sub-root function $\Psi(r)$. As defined in the latter reference, a sub-root function $\psi:\mathbb{R}_+\to\mathbb{R}_+$ is nonnegative, nondecreasing and such that $\psi(r)/\sqrt{r}$ is non increasing. r^* is defined as the unique (except for the trivial case $\Psi\equiv 0$) fixed point of Ψ. The latter reference proves the existence and uniqueness of the fixed point of a sub-root function.

The following result provides a first step towards the analysis of the estimation error of $\widehat{f}_{D,R}$ by controlling the complexity of $\widehat{S}_{D,R}$.

Theorem 4.5.1. *Let $\widehat{S}_{D,R}=\widehat{S}_D\cap B_{\mathcal{H}}(0,R)$. With the notations and the assumptions of Theorem 4.3.2, with probability at least $1-e^{-\xi}$,*

$$\widehat{S}_{D,R}\subset S_1(n,D,R)\,, \tag{4.10}$$

where $S_1(n,D,R)=\{g+h,\ g\in S_D \text{ and } h\in S_D^{\perp}\cap B_{\mathcal{H}}(0,\frac{B_D}{\sqrt{n}}R)\}$. Moreover,

$$\mathbb{E}R_n\{f\in S_1(n,D,R),\, Pf^2\le r\}\le\sqrt{\frac{rD}{n}}+\frac{1}{\sqrt{n}}\inf_{N\ge D}\left(\sqrt{r(N-D)}+\frac{B_DR}{\sqrt{n}}\sqrt{\sum_{j\ge N+1}\lambda_j}\right).$$

The upper-bound is a sub-root function. Its fixed point r^* satisfies:

$$r^*\le\frac{4}{n}\inf_{N\ge D}\left(N+2B_DR\sqrt{\sum_{j\ge N+1}\lambda_j}\right).$$

It is worth noticing that the complexity is of order n^{-1}. Moreover, if for all $i\ge D+1$ $\lambda_i=0$, this result entails $r^*\le 4\frac{D}{n}$.

For the sake of clarity, the result is stated when the eigenvalues of C are simple. Besides, in the general case, the same result holds assuming that D is of the form $m_1+\ldots+m_r$ for a certain r and changing the definition of B_D accordingly.

Proof. Since the assumptions of Theorem 4.3.2 are fulfilled, inclusion (4.9) straightforwardly implies inclusion (4.10).

We recall that $\phi_1,\phi_2,\ldots$ denotes an orthonormal basis of $\mathcal{H}$ of eigenfunctions of C associated with the eigenvalues $\lambda_1,\lambda_2,\ldots$. We have

$$\langle\phi_i,\phi_j\rangle_{L_2(P)}=\langle\phi_i,C\phi_j\rangle_{\mathcal{H}}=\lambda_j\langle\phi_i,\phi_j\rangle_{\mathcal{H}}=\delta_{i,j}\lambda_j\,, \tag{4.11}$$

where the first equality comes from the definition of chapter 3 of the covariance operator. Since $\|g\|^2_{L_2(P)}\le\mathbb{E}[k(X,X)]\,\|g\|^2_{\mathcal{H}}$ – reproducing property and Cauchy-Schwartz inequality in $\mathcal{H}$ – , the convergence in $\mathcal{H}$ is stronger than the convergence in $L_2(P)$. Thus, using equation (4.11), if $g\in S_D$ and $h\in S_D^{\perp}$, $\langle h,g\rangle_{L_2(P)}=0$. Finally, every $f\in\{S_1(n,R,D),\,Pf^2\le r\}$

satisfies $f = g + h$ with $Pg^2 \le r$ and $Pf^2 \le r$. Consequently, the supremum can be split in two parts:

$R_n\{f \in S_1(n, D, R), Pf^2 \le r\} \le R_n\{g \in S_D, Pg^2 \le r\} + R_n\{h \in G(n, D, R), Ph^2 \le r\}$,

where $G(n, D, R) = \{h \in S_D^{\perp}, \|h\|_{\mathcal{H}} \le \frac{B_D}{\sqrt{n}} R\}$.
Inequality (5.68) shall show

$$\mathbb{E}R_n\{g \in S_D, Pg^2 \le r\} \le \sqrt{\frac{rD}{n}}.$$

The second part is controlled as follows. We can suppose that for all $i \ge 1$, $\lambda_i > 0$ (otherwise, it suffices to define Ψ_i only for $\lambda_i > 0$ and the proof is readily the same). Let $\Psi_i = \frac{\phi_i}{\sqrt{\lambda_i}}$. For all $h \in S_D^{\perp}$, $h = \sum_{i \ge D+1} \alpha_i \Psi_i$, where $\alpha_i = \lambda_i \langle h, \Psi_i \rangle_{\mathcal{H}}$. Since $\langle \Psi, \Psi_j \rangle_{\mathcal{H}} = \delta_{i,j}/\lambda_j$, we have

$$\|h\|^2_{\mathcal{H}_k} = \sum_{i \ge D+1} \frac{\alpha_i^2}{\lambda_i}. \tag{4.12}$$

Moreover, the convergence in $\mathcal{H}$ is stronger than the convergence in $L_2(P)$ and equation (4.11) means $\langle \Psi_i, \Psi_j \rangle_{L_2(P)} = \delta_{i,j}$. Thus

$$\|h\|^2_{L_2(P)} = \sum_{i \ge D+1} \alpha_i^2. \tag{4.13}$$

Consequently, for $h \in G(n, D, R)$ satisfying $Ph^2 \le r$ and $N \ge D$, the following holds

$$\begin{aligned}
\sum_{i=1}^{n} \epsilon_i h(X_i) &= \sum_{j \ge D+1} \alpha_j \left(\sum_{i=1}^{n} \epsilon_i \Psi_j(X_i) \right) \\
&= \sum_{j \ge D+1}^{N} \alpha_j \left(\sum_{i=1}^{n} \epsilon_i \Psi_j(X_i) \right) + \sum_{j \ge N+1} \alpha_j \left(\sum_{i=1}^{n} \epsilon_i \Psi_j(X_i) \right) \\
&\le \sqrt{r} \sqrt{\sum_{j \ge D+1}^{N} \left(\sum_{i=1}^{n} \epsilon_i \Psi_j(X_i) \right)^2} + \frac{B_D R}{\sqrt{n}} \sqrt{\sum_{j \ge N+1} \lambda_j \left(\sum_{i=1}^{n} \epsilon_i \Psi_j(X_i) \right)^2},
\end{aligned}$$

where Cauchy-Schwarz inequality, inequalities (4.12) and (4.13) are used to obtain the inequality. Moreover, by Jensen's inequality,

$$\mathbb{E}_\epsilon R\{h \in G(n, D, R), Ph^2 \le r\} \le \frac{\sqrt{r}}{n} \sqrt{\sum_{j \ge D+1}^{N} \sum_{i=1}^{n} \Psi_j^2(X_i)} + \frac{B_D R}{n} \sqrt{\sum_{j \ge N+1} \lambda_j \frac{1}{n} \sum_{i=1}^{n} \Psi_j^2(X_i)}.$$

Using again Jensen's inequality and $\mathbb{E}\Psi_j^2(X) = 1$, we get:

$$\mathbb{E}R\{h \in G(n, D, R), Ph^2 \le r\} \le \frac{\sqrt{r(N-D)}}{\sqrt{n}} + \frac{B_D R}{n} \sqrt{\sum_{j \ge N+1} \lambda_j}.$$

This concludes the proof of the upper bound of the localized Rademacher average.

The bound on the fixed point r^* is a simple consequence of [Bou02a, Lemma 4.10] and of the reasoning of Lemma 3.8.4. □

4.6 Appendix B: Background of Functional Analysis.

In order to make the chapter self-contained, this appendix aims at gathering some background on analytic functions (see [GK02]) and integration of Banach space valued functions used in the perturbation of operator theory of [Kat66] and, consequently, all along the proof of Theorem 4.3.1.

4.6.1 Integration of Banach Space Valued Functions.

The plane curves considered in this chapter can be regarded as a set of complex points produced by a continuous and piecewise continuously differentiable transformation γ of a compact interval $[a, b]$ where $a < b$ i.e. for t between a and b, the point $\gamma(t)$ draws a curve $\gamma^* = \gamma([a, b])$ (the image of γ). The contour integral of a continuous function $f : \gamma^* \to \mathbb{C}$ along the curve γ is defined by the complex number

$$\int_\gamma f(z)dz := \int_a^b f(\gamma(t))\gamma'(t)dt\,.$$

In this chapter, we restricted ourselves to simple closed curves of finite length i.e. the curve does not intersect itself, satisfies $\gamma(a) = \gamma(b)$ and $\int_a^b |\gamma'(t)| < \infty$.

Let now B be a separable Banach space which is a vector space over $\mathbb{C}$ and B^* be its topological dual consisted of the continuous linear forms on B.

The integral along a curve γ is defined for a continuous Banach space valued function by duality through this theorem and definition. It is a special case of Theorem 3.27 of [Rud73] recalled in appendix 4.6.3.

Theorem 4.6.1. *Let $f : \gamma \to B$ be a continuous function in the separable Banach space B. Then*

- $\forall \Psi \in B^*$, *the scalar functions $\Psi(f)$ are integrable.*
- *There exists a unique element I of B such that $\forall \Psi \in B^*$, $\Psi(I) = \int_\gamma \Psi(f(\xi))d\xi$.*

We define $I = \int_\gamma f(\xi)d\xi$.

Let $\mathcal{L}_c(\mathcal{H}, \mathcal{H})$ be the Banach space of all continuous linear operators of $\mathcal{H}$ endowed with the operator norm $\|.\|_{\text{op}}$. The following result aims at showing that this definition of the integral preserves its natural properties.

Proposition 4.6.2 (Accordance of the definition). *(1) Let $T : \gamma \to \mathcal{L}_c(\mathcal{H}, \mathcal{H})$ be a continuous function and $h \in \mathcal{H}$: $\left(\int_\gamma B(\xi)d\xi\right)(h) = \int_\gamma B(\xi)(h)d\xi$.*

(2) Let $T : \gamma \to HS(\mathcal{H})$ be a continuous function for the $\|.\|-$topology. T is thus continuous for the $\|.\|_{\text{op}}$-topology and $\int_\gamma T(\xi)d\xi$ can be defined as an element of $\mathcal{L}_c(\mathcal{H}, \mathcal{H})$ or $\mathrm{HS}(\mathcal{H})$. Actually, this two definitions leads to the same quantity.

Proof. These assertions are straightforward consequences of the definition of the integral. Concerning (1), it suffices to notice that, if $\Psi \in \mathcal{H}^*$, then $\phi : \mathcal{L}_c(\mathcal{H}, \mathcal{H}) \to \mathbb{C}$ defined by $\phi(D) = \Psi(D(h))$ belongs to $\mathcal{L}_c(\mathcal{H}, \mathcal{H})^*$. As for (2), it is clear that the restriction of any element of $\mathcal{L}_c(\mathcal{H}, \mathcal{H})^*$ to $\mathrm{HS}(\mathcal{H})$ belongs to $\mathrm{HS}(\mathcal{H})^*$ since the convergence for the Hilbert-Schmidt norm is stronger than the convergence for the operator norm. $\square$

4.6.2 Use of the Resolvent .

The resolvent set of $T \in \mathcal{L}_c(\mathcal{H}, \mathcal{H})$ is the set of $z \in \mathbb{C}$ such that $T - z\,Id$ is invertible as an element of $\mathcal{L}_c(\mathcal{H}, \mathcal{H})$. The operator-valued function $R_T(z) = (T - z\,Id)^{-1}$ is defined on the resolvent set and is called the resolvent of T. The index of a point $z_0 \in \mathbb{C} \backslash \gamma^*$ with respect to a curve γ is defined as

$$\mathrm{Ind}_\gamma(\mathrm{z}_0) := \frac{1}{2\mathrm{i}\pi} \int_\gamma \frac{\mathrm{dz}}{\mathrm{z} - \mathrm{z}_0} .$$

Since γ is a closed curve, Ind_γ is an integer constant on each connected component of $\mathbb{C}\backslash\gamma^*$ and equal to zero on the unbounded connected component of $\mathbb{C}\backslash\gamma^*$. From an intuitive point of view, this contour integral counts the number of turns of the contour around the point z_0. For example, if $\gamma : t \to e^{int}$ $(0 \leq t \leq 2\pi)$ is the unit circle covered n times, simple calculations leads to $\mathrm{Ind}_\gamma(\mathrm{z}_0) = \mathrm{n}$ if $|z_0| < 1$ and 0 if $|z_0| > 1$.

Let Σ be the set of all simple closed curves γ of finite length satisfying $\mathrm{Ind}_\gamma(\mathrm{z}_0) = 1$ if z_0 is enclosed by γ. Two continuous closed curves γ_1 and γ_2 contained in an open set D of $\mathbb{C}$ are *homotopic* in D if one can be "continuously deformed" into the other; that is, if there exists a continuous map $F : [a, b] \times [0, 1] \to D$ such that $\forall\, t \in [a, b], \forall\, s \in [0, 1]$, $F(t, 0) = \gamma_1(t)$; $F(t, 1) = \gamma_2(t)$ and $F(a, s) = F(b, s)$. Such an F is called a homotopy between γ_1 and γ_2.

A result of Cauchy states that if f is an analytic function in the open connected set D of $\mathbb{C}$ and γ_1, γ_2 are homotopic in D, then

$$\int_{\gamma_1} f(z)dz = \int_{\gamma_2} f(z)dz .$$

Consequently, Σ contains all the curves continuously deformable into a circle in $\mathbb{C}\backslash\{z_0\}$ for all z_0 enclosed by γ.

Our work relies on the link between the projection on the eigenspaces of an operator and the resolvent of this operator. It is proved in [Kat66] by considering very few assumptions on the operator. However, we need this link only for a specific operator. In this case, it is a simple consequence of previous definitions and properties. The previous homotopy arguments imply that the curve considered in the proof of Theorem 4.3.1 satisfies the conditions of statement (2) of the following result.

Theorem 4.6.3. *(1) $R_T(z)$ depends continuously on z.*

(2) Let $T : \mathcal{H} \to \mathcal{H}$ be a self-adjoint compact operator and $\gamma \in \Sigma$ a curve included in the resolvent set of T. Then $-\frac{1}{2i\pi}\int_\gamma R_T(z)dz$ is the orthogonal projection onto the vector space spanned by the eigenvectors of T corresponding to eigenvalues enclosed by γ.

(3) The operator norm of the resolvent of a compact self-adjoint operator T is

$$\|R_T(z)\|_{op} = \frac{1}{dist(z, \Lambda(T))}, \tag{4.14}$$

where $dist(z, \Lambda(T)) = \inf\limits_{\lambda \in \Lambda(T)} |z - \lambda|$.

Proof. (1) comes from the standard property stating that the inverse function $A \to A^{-1}$ is continuous.

In order to prove (2), let $(\phi_i)_{i\geq 1}$ be an orthonormal basis of eigenfunctions of T corresponding to eigenvalues $(\lambda_i(T))_{i\geq 1}$ and $G = -\frac{1}{2i\pi}\int_\gamma R_T(z)dz \in \mathcal{L}_c(\mathcal{H},\mathcal{H})$ (defined thanks to Theorem 4.6.1). Due to proposition 4.6.2 (1),

$$G(\phi_i) = \left(-\frac{1}{2i\pi}\int_\gamma \frac{1}{\lambda_i - z}dz\right)\phi_i = \mathrm{Ind}_\gamma(\lambda_\mathrm{i})\phi_\mathrm{i}\,,$$

where $\mathrm{Ind}_\gamma(\lambda_\mathrm{i}) = 1$ if λ_i is enclosed by γ and 0 if it is outside (by definition of Σ). This concludes the proof of (2).

Since $R_T(z)$ is a normal operator, $\|R_T(z)\|_{\mathrm{op}} = \sup_{i\geq 1}|\lambda_i(R_T(z))|$. Moreover, $\lambda_i((T - zId)^{-1}) = \frac{1}{\lambda_i(T)-z}$. This concludes the proof of (3). □

4.6.3 Basic Results.

This appendix gathers results of the literature which are used in this chapter. The Hoffman-Wielandt inequality for finite dimensional spaces is stated in [Kol98] and is at the heart of the proof of its infinite dimension version.

Theorem 4.6.4 (Theorem 1 of [BE94]). *Let A and B be normal Hilbert-Schmidt operators and let $\{\alpha_1,\alpha_2,\cdots\}$ and $\{\beta_1,\beta_2,\cdots\}$ be enumerations of their eigenvalues (terms consist of all the eigenvalues each counted as often as its multiplicity). Then for each $\epsilon > 0$ there exists a permutation π of $\mathbb{N}$ such that*

$$\left(\sum_{i\geq 1}|\alpha_i - \beta_{\pi(i)}|^2\right)^{\frac{1}{2}} \leq \|A - B\| + \epsilon\,.$$

The following Theorem defines the integral of a topological space valued function under compactness and continuity assumptions. Its proof relies on a compactness argument.

Theorem 4.6.5 (Theorem 3.27 of [Rud73]). *Suppose*

- *X is a topological vector space on which X^* separates points, and*
- *μ is a finite real or complex probability measure on a compact Haussdorf space Q.*

If $f : Q \to X$ is continuous, and if the convex hull H of $f(Q)$ has compact closure $\overline{H}$ in X, then

- *$\forall\Psi \in X^*$, the scalar functions $\Psi(f)$ are integrable with respect to μ.*
- *There exists $y \in \overline{H}$ such that $\Psi(y) = \int_Q \Psi(f)d\mu$.*

We define $y = \int_Q f d\mu$.

This result implies Theorem 4.6.1 since a Banach space separates points and $[a,b]$ is a compact Haussdorf space. Moreover, in a Banach space, the closure of the convex hull of a compact space is compact (Theorem 3.24 of [Rud73]).

Part II

The Kernel Projection Machine

Chapter 5

Kernel Projection Machine: a New Tool for Pattern Recognition

This chapter is a joint work with G. Blanchard, R. Vert and P. Massart. Most of the main text is already published ([BMVZ04]). The results of the appendices aim at motivating the chapter 6.

Contents

Abstract

This chapter investigates the effect of the Kernel Principal Component Analysis (KPCA) within the classification framework, essentially the regularization properties of this dimensionality reduction method. KPCA has been previously used as a pre-processing step before applying an SVM but we point out that this method is somewhat redundant from a regularization point of view and we propose a new algorithm called *Kernel Projection Machine* to avoid this redundancy, based on an analogy with the statistical framework of regression for a Gaussian white noise model. Preliminary experimental results show that this algorithm reaches the same performances as an SVM.

5.1 Introduction .

Let $(x_i, y_i)_{i=1\dots n}$ be n given realizations of a random variable (X, Y) living in $\mathcal{X} \times \{-1; 1\}$. Let P denote the marginal distribution of X. The x_i's are often referred to as *inputs* (or *patterns*), and the y_i's as *labels*. Pattern recognition is concerned with finding a *classifier*, i.e. a function that assigns a label to any new input $x \in \mathcal{X}$ and that makes as few prediction errors as possible.

It is often the case with real world data that the dimension of the patterns is very large, and some of the components carry more noise than information. In such cases, reducing the dimension of the data before running a classification algorithm on it sounds reasonable. One of the most famous methods for this kind of pre-processing is PCA, and its kernelized version (KPCA), introduced in the pioneering work of Schölkopf, Smola and Müller [SSM98]. Now, whether the quality of a given classification algorithm can be significantly improved by using such pre-processed data still remains an open question. Some experiments have already been carried out to investigate the use of KPCA for classification purposes, and numerical results are reported in [SSM98]. The authors considered the USPS handwritten digit database and reported the test error rates achieved by the linear SVM trained on the data pre-processed with KPCA: the conclusion was that the larger the number of principal components, the better the performance. In other words, the KPCA step was useless or even counterproductive. This conclusion might be explained by a redundancy arising in their experiments: there is actually a double regularization, the first corresponding to the dimensionality reduction achieved by KPCA, and the other to the regularization achieved by the SVM. With that in mind it does not seem so surprising that KPCA does not help in that case: whatever the dimensionality reduction, the SVM anyway achieves a (possibly strong) regularization.

Still, de-noising the data using KPCA seems relevant. The aforementioned experiments suggest that KPCA should be used together with a classification algorithm that is not regularized (e.g. a simple empirical risk minimizer): in that case, it should be expected that the KPCA is by itself sufficient to achieve regularization, the choice of the dimension being guided by adequate model selection.

In this chapter, we propose a new algorithm, called the Kernel Projection Machine (KPM), that implements this idea: an optimal dimension is sought so as to minimize the test error of the resulting classifier. A nice property is that the training labels are used to select the optimal dimension – optimal means that the resulting D-dimensional representation of the data contains the right amount of information needed to classify the inputs. To sum up, the KPM can be seen as a dimensionality-reduction-based classification method that takes into

account the labels for the dimensionality reduction step.
This chapter is organized as follows: Section 5.2 gives some statistical background on regularized method vs. projection methods. Its goal is to explain the motivation and the "Gaussian intuition" that lies behind the KPM algorithm from a statistical point of view. Section 5.3 explicitly gives the details of the algorithm; experiments and results, which should be considered preliminary, are reported in Section 5.4.

5.2 Motivations for the Kernel Projection Machine.

5.2.1 The Gaussian Intuition: a Statistician's Perspective.

Regularization methods have been used for quite a long time in non parametric statistics since the pioneering works of Grace Wahba in the eighties (see [Wah90] for a review). Even if the classification context has its own specificity and offers new challenges (especially when the explanatory variables live in a high dimensional Euclidean space), it is good to remember what is the essence of regularization in the simplest non parametric statistical framework: the Gaussian white noise.
So let us assume that one observes a noisy signal $dY(x) = s(x)dx + \frac{1}{\sqrt{n}}dw(x)$, $Y(0) = 0$ on [0,1] where $dw(x)$ denotes standard white noise. To the reader not familiar with this model, it should be considered as nothing more but an idealization of the well-known fixed design regression problem $Y_i = s(i/n) + \varepsilon_i$ for $i = 1, \dots, n$, where $\varepsilon_i \sim N(0,1)$, where the goal is to recover the regression function s. (The white noise model is actually simpler to study from a mathematical point of view). The least square criterion is defined as

$$\gamma_n(f) = \|f\|^2 - 2\int_0^1 f(x)dY(x)$$

for every $f \in L_2([0,1])$.
Given a Mercer kernel k on $[0,1]\times[0,1]$, the regularization least square procedure proposes to minimize

$$\gamma_n(f) + \zeta_n \|f\|^2_{\mathcal{H}_k}\,, \tag{5.1}$$

where (ζ_n) is a conveniently chosen sequence and $\mathcal{H}_k$ denotes the RKHS induced by k. This procedure can indeed be viewed as a model selection procedure since minimizing $\gamma_n(f) + \zeta_n\|f\|_{\mathcal{H}_k}$ amounts to minimizing

$$\inf_{\|f\|\le R}\left[\gamma_n(f) + \zeta_n R^2\right]$$

over $R > 0$. In other words, regularization aims at selecting the "best" RKHS ball $\{f, \|f\| \le R\}$ to represent our data.
At this stage, it is interesting to realize that the balls in the RKHS space can be viewed as ellipsoids in the original Hilbert space $L_2([0,1])$. Indeed, let $(\phi_i)_{i=1}^{\infty}$ be some orthonormal basis of eigenfunctions for the compact and self adjoint operator

$$T_k : f \longrightarrow \int_0^1 k(x,y)f(x)dx$$

Then, setting $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ one has $\|f\|^2_{\mathcal{H}_k} = \sum_{j=1}^{\infty} \frac{\beta_j^2}{\lambda_j}$ where $(\lambda_j)_{j\geq 1}$ denotes the non increasing sequence of eigenvalues corresponding to $(\phi_j)_{j\geq 1}$. Hence

$$\{f, \|f\|_{\mathcal{H}_k} \leq R\} = \left\{ \sum_{j=1}^{\infty} \beta_j \phi_j \, ; \, \sum_{j=1}^{\infty} \frac{\beta_j^2}{\lambda_j} \leq R^2 \right\}.$$

Now, due to the approximation properties of the finite dimensional spaces $\{\phi_j, \, j \leq D\}$, $D \in \mathbb{N}^*$ with respect to the ellipsoids, one can think of penalized finite dimensional projection as an alternative method to regularization. More precisely, if $\widehat{s}_D$ denotes the projection estimator on $\langle \phi_j, \, j \leq D \rangle$, i.e. $\widehat{s}_D = \sum_{j=1}^{D} \left(\int \phi_j dY \right) \phi_j$ and one considers the penalized selection criterion

$$\widehat{D} = \underset{D}{\operatorname{argmin}} [\gamma_n(\widehat{s}_D) + \frac{2D}{n}], \tag{5.2}$$

then, it is proved in [BBM99] that the selected estimator $\widehat{s}_{\widehat{D}}$ obeys to the following oracle inequality

$$\mathbb{E}[\|s - \widehat{s}_{\widehat{D}}\|^2] \leq C \inf_{D\geq 1} \left[\mathbb{E}\|s - \widehat{s}_D\|^2 \right]$$

where C is some absolute constant.
The nice thing is that whenever s belongs to some ellipsoid

$$\mathcal{E}(c) = \left\{ \sum_{j=1}^{\infty} \beta_j \phi_j \, : \, \sum_{j=1}^{\infty} \frac{\beta_j^2}{c_j^2} \leq 1 \right\}$$

where $(c_j)_{j\geq 1}$ is a decreasing sequence tending to 0 as $j \to \infty$, then

$$\inf_{D\geq 1} \mathbb{E} \left[\|s - \widehat{s}_D\|^2 \right] = \inf_{D\geq 1} \left[\inf_{t \in S_D} \|s - t\|^2 + \frac{D}{n} \right] \leq \inf_{D\geq 1} \left[c_D^2 + \frac{D}{n} \right]$$

As shown in [DLM90] $\inf_{D\geq 1}[c_D^2 + \frac{D}{n}]$ is (up to some absolute constant) of the order of magnitude of the minimax risk over $\mathcal{E}(c)$.
As a consequence, the estimator $\widehat{s}_{\widehat{D}}$ is simultaneously minimax over the collection of *all* ellipsoids $\mathcal{E}(c)$, which in particular includes the collection $\{\mathcal{E}(\sqrt{\lambda}R), \, R > 0\}$.
To conclude and summarize, from a statistical performance point of view, what we can expect from a regularized estimator $\widehat{s}$ (i.e. a minimizer of (5.1)) is that a convenient device of ζ_n ensures that $\widehat{s}$ is simultaneously minimax over the collection of ellipsoids $\{\mathcal{E}(\sqrt{\lambda}R), \, R > 0\}$, (at least as far as asymptotic rates of convergence are concerned). The alternative estimator $\widehat{s}_{\widehat{D}}$ actually achieves this goal and even better since it is also adaptive over the collection of all ellipsoids and not only the family $\{\mathcal{E}(\sqrt{\lambda}R), \, R > 0\}$.

5.2.2 Extension to a General Classification Framework.

In this section we go back to classification framework as described in the introduction. First of all, it has been noted by several authors ([EPP00],[SS98]) that the SVM can be seen as a regularized estimation method, where the regularizer is the squared norm of the function in $\mathcal{H}_k$. Precisely, the SVM algorithm solves the following unconstrained optimization problem:

$$\min_{f \in \mathcal{H}_k^b} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|f\|^2_{\mathcal{H}_k}, \tag{5.3}$$

where $\mathcal{H}_k^b = \{f(x) + b,\ f \in \mathcal{H}_k, b \in \mathbb{R}\}$.
The above regularization can be viewed as a model selection process over RKHS balls, similarly to the previous section. Now, the line of ideas developed there suggests that it might actually be a better idea to consider a sequence of finite-dimensional estimators. Additionally, it has been shown in [BBM04] that the regularization term of the SVM is actually too strong. We therefore transpose the ideas of previous Gaussian case to the classification framework. Consider a Mercer kernel k defined on $\mathcal{X} \times \mathcal{X}$ and Let T_k denote the operator associated with kernel k in the following way

$$T_k \ : \ f(.) \in L_2(P) \longmapsto \int_{\mathcal{X}} k(x,.)f(x)dP(x) \in L_2(P)$$

Let $\phi_1, \phi_2, \ldots$ denote the eigenvectors of T_k, ordered by decreasing associated eigenvalues $(\lambda_i)_{1\geq 1}$. For each integer D, the subspace $\mathcal{F}_D$ defined by $\mathcal{F}_D = \operatorname{span}\{\mathbb{1}, \phi_1, \ldots, \phi_D\}$ (where $\mathbb{1}$ denotes the constant function equal to 1) corresponds to a subspace of $\mathcal{H}_k^b$ associated with kernel k, and $\mathcal{H}_k^b = \overline{\bigcup_{D=1}^{\infty} \mathcal{F}_D}$. Instead of selecting the "best" ball in the RKHS, as the SVM does, we consider the analogue of the projection estimator $\widehat{s}_D$:

$$\hat{f}_D = \arg\min_{f \in \mathcal{F}_D} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ \tag{5.4}$$

that is, more explicitly,

$$\hat{f}_D(.) = \sum_{j=1}^{D} \beta_j^* \phi_j(.) + b^*$$

with

$$(\beta^*, b^*) = \arg\min_{(\beta\in\mathbb{R}^D, b\in\mathbb{R})} \sum_{i=1}^{n} \left(1 - y_i \left(\sum_{j=1}^{D} \beta_j \phi_j(x_i) + b\right)\right)_+ \tag{5.5}$$

An appropriate D can then be chosen using an adequate model selection procedure such as penalization; we do not address this point in detail in the present work but it is of course the next step to be taken.

Unfortunately, since the underlying probability P is unknown, neither are the eigenfunctions $\phi_1, \ldots$, and it is therefore not possible to implement this procedure directly. We thus resort to considering empirical quantities as will be explained in more detail in section 5.3. Essentially, the unknown vectorial space spanned by the first eigenfunctions of T_k is replaced by the space spanned by the first eigenvectors of the normalized kernel Gram matrix $\frac{1}{n}(k(x_i, x_j))_{1\leq i,j\leq n}$. At this point we can see the relation appear with Kernel PCA. We next precise this relation and give an interpretation of the resulting algorithm in terms of dimensionality reduction.

5.2.3 Link with Kernel Principal Component Analysis.

Principal Component Analysis (PCA), and its non-linear variant, KPCA are widely used algorithms in data analysis. They extract from the input data space a basis $(v_i)_{i\geq 1}$ which is, in some sense, adapted to the data by looking for directions where the variance is maximized.

They are often used as a pre-processing on the data in order to reduce the dimensionality or to perform de-noising.
As will be made more explicit in the next section, the Kernel Projection Machine consists in replacing the ideal projection estimator defined by (5.4) by

$$\widehat{f}_D = \underset{f \in \widehat{S}_D}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(X_i))_+ \tag{5.6}$$

where $\widehat{S}_D$ is the space of dimension D chosen by the first D principal components chosen by KPCA in feature space. Hence, roughly speaking, in the KPM, the SVM penalization is replaced by dimensionality reduction.
Choosing D amounts to selecting the optimal D-dimensional representation of our data for the classification task, in other words to extracting the information that is needed for this task by model selection taking into account the relevance of the directions for the classification task.
To conclude, the KPM is a method of dimensionality reduction that takes into account the labels of the training data to choose the "best" dimension.

5.3 The Kernel Projection Machine Algorithm.

In this section, the empirical (and computable) version of the KPM algorithm is derived from the previous theoretical arguments.
In practice the true eigenfunctions of the kernel operator are not computable. But since only the values of functions $\phi_1, \ldots, \phi_D$ at points $x_1, \ldots, x_n$ are needed for minimizing the empirical risk over $\mathcal{F}_D$, the eigenvectors of the kernel matrix $K = (k(x_i, x_j))_{1 \le i,j \le n}$ will be enough for our purpose. Indeed, it is well known in numerical analysis (see [Bak77]) that the eigenvectors of the kernel matrix approximate the eigenfunctions of the kernel operator. This result has been pointed out in [Kol98] in a more probabilistic language. More precisely, if $V_1, \ldots, V_D$ denote the D first eigenvectors of K with associated eigenvalues $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \ldots \ge \widehat{\lambda}_D$, then for each V_i

$$V_i = \left(V_i^{(1)}, \ldots, V_i^{(n)}\right) \approx (\phi_i(x_1), \ldots, \phi_i(x_n)) \tag{5.7}$$

Hence, considering Equation (5.5), the empirical version of the algorithm described above will first consist of solving, for each dimension D, the following optimization problem:

$$(\beta^*, b^*) = \arg \min_{\beta \in \mathbb{R}^D, b \in \mathbb{R}} \sum_{i=1}^{n} \left(1 - y_i \left(\sum_{j=1}^{D} \beta_j V_j^{(i)} + b\right)\right)_+ \tag{5.8}$$

Then the solution should be

$$\hat{f}_D(.) = \sum_{j=1}^{D} \beta_j^* \phi_j(.) + b^* . \tag{5.9}$$

Once again the true functions ϕ_j's are unknown. At this stage, we can do an expansion of the solution in terms of the kernel similarly to the SVM algorithm, in the following way:

$$\hat{f}_D(.) = \sum_{i=1}^{n} \alpha_i^* k(x_i, .) + b^* \tag{5.10}$$

Narrowing expressions (5.9) and (5.10) at points $x_1, \ldots, x_n$ leads to the following equation:

$$\beta_1^* V_1 + \ldots + \beta_D^* V_D = K\alpha^* \tag{5.11}$$

which has a straightforward solution: $\alpha^* = \sum_{j=1}^{D} \frac{\beta_j^*}{\widehat{\lambda}_j} V_j$ (provided the first D eigenvalues are all strictly positive).
The KPM algorithm can now be summed up as follows:

1. given data $x_1, \ldots, x_n \in \mathcal{X}$ and a positive kernel k defined on $\mathcal{X} \times \mathcal{X}$, compute the kernel matrix K and its eigenvectors $V_1, \ldots, V_n$ together with its eigenvalues in decreasing order $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_n$.

2. for each dimension D such that $\widehat{\lambda}_D > 0$ solve the linear optimization problem

$$(\beta^*, b^*) = \arg\min_{\beta, b, \xi} \sum_{i=1}^{n} \xi_i \tag{5.12}$$

$$\text{under constraints } \forall i = 1 \ldots n, \xi_i \geq 0 \ , y_i \left(\sum_{j=1}^{D} \beta_j V_j^{(i)} + b \right) \geq 1 - \xi_i \ . \tag{5.13}$$

Next, compute $\alpha^* = \sum_{j=1}^{D} \frac{\beta_j^*}{\widehat{\lambda}_j} V_j$ and $\hat{f}_D(.) = \sum_{i=1}^{n} \alpha_i^* k(x_i, .) + b^*$

3. The last step is a model selection problem: choose a dimension $\hat{D}$ for which $\hat{f}_{\hat{D}}$ performs well. One can think of applying cross-validation, or to penalize the empirical loss by a penalty function depending on the dimension.

5.4 Experiments.

The KPM was implemented in Matlab using the free library GLPK for solving the linear optimization problem. Since the algorithm involves the eigendecomposition of the kernel matrix, only small datasets have been considered for the moment. The Matlab implementation of the KPM algorithm is available at http://www.lri.fr/∼vert.
In order to assess the performance of the KPM, we carried out experiments on benchmark datasets available on Gunnar Rätsch's web site [Rae99]. Several state-of-art algorithms have already been applied to these datasets, among which the SVM. All results are reported on the web site. To get a valid comparison with the SVM we used, on each classification task, the same kernel parameters as those used for SVM, so as to work with exactly the same geometry.

5.4.1 First Qualitative Examples.

The pictures below were obtained by training the KPM on the Banana problem, with a gaussian kernel: increasing D enables us to compute more and more complicated separating boundaries. Three values for D were tested: 3, 15 and 50 respectively.

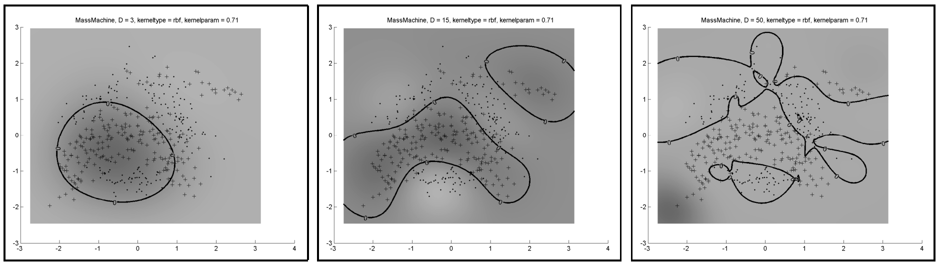


Figure 5.1: D controls the smoothness of the decision boundary: from the left to the right, $D = 3$, 15 (optimal), and 50. The dataset is 'banana'.

The left part (resp. the right part) of graphic below shows both training (dotted line) and test errors as functions of parameter D (resp. $C = (n\lambda)^{-1}$) on the Flare-Solar dataset. It highlights that D acts as a complexity term in the KPM as for λ in the SVM. Concerning the KPM, the correlated shapes of the two curves suggest alternatives to re-sampling methods for parameter selection, such as penalization or slope heuristic.

Interestingly, the graphic on the left in the figure below shows that our procedure is very different from the one of [SSM98]: when D is very large, our risk increases (leading to the existence of a minimum) while the risk of [SSM98] always decreases with D.

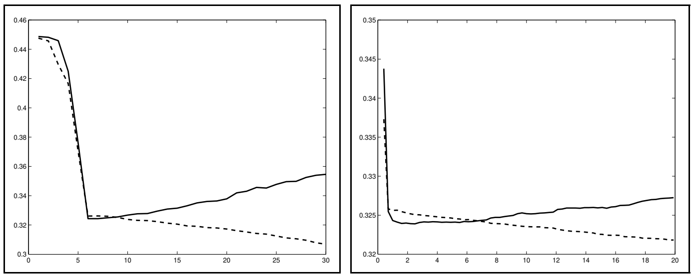


Figure 5.2: **Left**: KPM risk (solid) and empirical risk (dashed) versus dimension D. **Right**: SVM risk and empirical risk versus the parameter C ($C = (n\lambda)^{-1}$). Both on dataset 'flare-solar'.

5.4.2 Quantitative Experiments.

There is a subtle, but important point arising here. In the SVM performance reported by G. Rätsch, the regularization parameter C was first determined by cross-validation on the first 5 realizations of each dataset; the median of these values was then taken as a fixed value for the other realizations. This was done apparently in order to save computation time. However, this might lead to an over-optimistic estimation of the performances since in some sense some extraneous information is then available to the algorithm and the variation due to the choice of λ is reduced to almost zero. We first tried to mimic this methodology by applying it, in our case, to the choice of D itself (the median of 5 D values obtained by cross-validation on the first realizations was then used on the other realizations).

One might then argue that in this way we are selecting a *parameter* by this method instead of a *meta-parameter* for the SVM, so that the comparison is unfair. However, this distinction

being loose, this a rather moot point. To avoid this kind of debate and to obtain fair results, we decided to re-run the SVM tests by selecting systematically the regularization parameter by a 5-fold cross-validation on each training set, and for our method, apply the same procedure to select D. Note that there is still extraneous information in the choice of the kernel parameters, but at least it is the same for both algorithms.
Results relative to the first methodology are reported in table 5.1. Results relative to the second are reported in table 5.2. It is worth noting that the parameter C of the SVM was sought on a geometric grid of only 101 values, ranging from $C_G/100$ to $100\,C_G$ and containing C_G. C_G denotes the optimal value given in [Rae99].

The globally worst performances exhibited in the second table show that the first procedure may indeed be too optimistic.

In the following experiments, the dimension of the KPM algorithm is selected by **hold-out**: this is an elementary step of cross-validation. We can suppose that n is even. Each training sample is split in two parts $(X_1, Y_1), \cdots, (X_{n/2}, Y_{n/2})$ and $(X_{n/2+1}, Y_{n/2+1}, \cdots, (X_n, Y_n)$. The classifiers $\{\widehat{f}_D\}_{1 \leq D \leq n}$ are computed using $(X_1, Y_1), \cdots, (X_{n/2}, Y_{n/2})$ and the dimension is simply selected by minimizing $\frac{2}{n} \sum_{i=n/2+1}^{n} \mathbb{1}_{\widehat{f}_D(X_i) \neq Y_i}$ over D. The test errors are presented in table 5.3. Hold-out is known to be unstable. This translates into a larger variance but, interestingly, performances are comparable to cross-validation (Table 5.2).

Table 5.1: Test errors of the KPM on several benchmark datasets, compared with SVM, using G.Rätsch's parameter selection procedure (see text). As an indication the best of the six results presented in [Rae99] are also reported.

	KPM	(selected D)	**SVM**	**Best of 6**
Banana	10.73 ± 0.42	15	11.53 ± 0.66	10.73 ± 0.43
Breast Cancer	26.51 ± 4.75	24	26.04 ± 4.74	24.77 ± 4.63
Diabetis	23.37 ± 1.92	11	23.53 ± 1.73	23.21 ± 1.63
Flare Solar	32.43 ± 1.85	6	32.43 ± 1.82	32.43 ± 1.82
German	23.59 ± 2.15	14	23.61 ± 2.07	23.61 ± 2.07
Heart	16.89 ± 3.53	10	15.95 ± 3.26	15.95 ± 3.26

Table 5.2: Test errors

	SVM	**KPM**
Banana ($\sigma = 0.7071$)	10.69 ± 0.67	10.91 ± 0.57
Breast Cancer($\sigma = 5$)	26.68 ± 5.23	28.73 ± 4.42
Diabetis ($\sigma = 3.1623$)	23.79 ± 2.01	23.77 ± 1.69
Flare Solar ($\sigma = 3.8730$)	32.62 ± 1.86	32.52 ± 1.78
German ($\sigma = 5.2440$)	23.79 ± 2.12	24.09 ± 2.38
Heart ($\sigma = 7.7460$)	16.23 ± 3.18	17.35 ± 3.54

All the presented experimental results are to be considered as preliminary, and in no way should they be used to establish a significant difference between the performances of the

Table 5.3: hold-out

	KPM
Banana ($\sigma = 0.7071$)	11.09±0.94
Breast Cancer($\sigma = 5$)	27.45±4.79
Diabetis ($\sigma = 3.1623$)	24.53±2.14
Flare Solar ($\sigma = 3.8730$)	33.31±2.73
German ($\sigma = 5.2440$)	24.30 ±2.23
Heart ($\sigma = 7.7460$)	17.69±3.69

KPM and the SVM.

5.5 Conclusion and Discussion .

To summarize, one can see the KPM as an alternative to the regularization of the SVM: regularization using the RKHS norm can be replaced by finite dimensional projection. Moreover, this algorithm performs KPCA towards classification and thus offers a criterion to decide what is the right order of expansion for the KPCA.
Dimensionality reduction can thus be used for classification but it is important to keep in mind that it behaves like a regularizer. Hence, it is clearly useless to plug it into a classification algorithm that is already regularized: the effect of the dimensionality reduction may be canceled as noted by [SSM98].
Our experiments explicitly show the regularizing effect of KPCA: no other smoothness control has been added in our algorithm, but nonetheless it gives performances comparable to the one of SVM provided that the dimension D is picked correctly. We only considered here selection of D by cross-validation; other methods such as penalization will be studied in future works. Moreover, with this algorithm, we obtain a D-dimensional representation of our data which is optimal for the classification task. Thus KPM can be seen as a de-noising method which takes into account the labels.
This version of the KPM only considers one kernel and thus one vector space by dimension. A more advanced version of this algorithm considers several kernels and thus chooses among a bigger family of spaces. This family then contains more than one space by dimension and will allow to directly compare the performance of different kernels on a given task, thus improving the efficiency of the dimensional reduction while taking into account the labels.

Appendix

This appendix aims at studying an alternative to re-sampling methods for dimension selection: the penalized empirical loss minimization. Throughout this appendix, a deterministic collection of models $\{S_D\}_{D\geq 1}$ is considered, where S_D is a vector subspace of $L_2(P)$ of dimension at most D and $S_D \subset S_{D+1}$. By analogy with the penalized selection criterion (5.2), the dimension is chosen by the following criterion:

$$\widehat{D} = \arg\min_{D\geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widehat{f}_D(X_i))_+ + \mathrm{pen}(D) \right) ,$$

where, by analogy with (5.6),

$$\widehat{f}_D = \arg\min_{f\in S_D} \frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ , \tag{5.14}$$

and the function *pen* has to be calibrated according to this classification framework.

Let $p(x) = \mathbb{P}(Y=1|X=x)$ and $f^*(x) = 2\mathbb{1}_{\{x, p(x)\geq \frac{1}{2}\}} - 1$ be the Bayes classifier. It is the best classifier since it satisfies $f^* = \arg\min_{f \text{classifier}} \mathbb{P}(Y \neq f(X))$.

5.6 Appendix A: Assumptions and First Theoretical Results for the Penalized Criterion.

From now on, $\|f\|_2^2 = \mathbb{E}\left[f^2(X)\right]$, $\|f\|_1 = \mathbb{E}\left[|f(X)|\right]$. The excess risk is denoted L:

$$L(f,f^*) = \mathbb{E}\left[(1-Yf(X))_+\right] - \mathbb{E}\left[(1-Yf^*(X))_+\right].$$

5.6.1 Assumptions.

The first assumption concerns the models.

H(τ_D). There exists τ_D such that

$$\forall f\in S_D,\ \|f\|_\infty \leq \tau_D \|f\|_2 .$$

If each function of the model S_D is bounded, then τ_D exists since S_D is a finite dimensional vector space. This assumption aims at highlighting the order of magnitude of τ_D: it will be crucial in the sequel. Moreover, it is connected to a property of the orthonormal bases of S_D. Let $(g_i)_{1\leq i\leq D}$ denote an $L_2(P)$-orthonormal basis of S_D. As observed by [BBM99], Cauchy-Schwarz inequality implies that $\|\sum_{i=1}^{D} g_i^2\|_\infty^{\frac{1}{2}} = \sup_{f\in S_D\backslash 0}\frac{\|f\|_\infty}{\|f\|_2}$. Thus, $\|\sum_{i=1}^{D} g_i^2\|_\infty \leq \tau_D^2$. However, $\|\sum_{i=1}^{D} g_i^2\|_\infty \geq \|\sum_{i=1}^{D} g_i^2\|_1 = D$. Consequently, $\tau_D \geq \sqrt{D}$; and so the optimal order of magnitude of τ_D is $\sqrt{D}$. In the following, we consider the smallest possible τ_D:

$$\tau_D = \sup_{f\in S_D\backslash 0}\frac{\|f\|_\infty}{\|f\|_2} = \|\sum_{i=1}^{D} g_i^2\|_\infty^{\frac{1}{2}} .$$

where $(g_i)_{1\leq i\leq D}$ denotes any $L_2(P)$-orthonormal basis of S_D.

The second assumption is inspired by the *Tsybakov margin condition* ([MT95]).

H(mar). There exists $p_0 > 0$ such that $p(x) = P[Y=1|X=x]$ satisfies $|p(x) - \frac{1}{2}| \geq p_0$, $p(x) \geq p_0$ and $p(x) \leq 1-p_0$.

5.6.2 Results.

Let $S_{D,n} = S_D \cap \{f, \|f\|_{L_2(P)} \leq n^2\}$. The next result provides a risk bound on a single model. The main advantage of this theorem is the simplicity of its proof.

Theorem 5.6.1. *Let* $\widehat{g}_D = \underset{g \in S_{D,n}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (1 - Y_i g(X_i))_+$ *and* $K > 1$*. Supposing that the assumptions* $\mathbf{H}(\tau_D)$ *and* $\mathbf{H}(\mathbf{mar})$ *are satisfied, that*

$$\tau_D \sqrt{D} < \frac{p_0 \sqrt{n}}{337(1+K)\log n}\,, \tag{5.15}$$

and that

$$0 < x \leq \log^2 n\,, \tag{5.16}$$

then, with probability greater than $1 - e^{-x}$*,*

$$L(\widehat{g}_D, f^*) \leq \frac{K+2}{K-1} \inf_{g \in S_{D,n}} L(g, f^*) + C_1 \frac{(K+1)^2}{K-1} \frac{D\tau_D^2 \log^2 n}{np_0^2} + C_2 \frac{x}{n} \frac{\tau_D^2 (K+1)^2}{p_0^2 (K-1)}\,.$$

where C_1 *and* C_2 *are universal constants.*

The main result of model selection is now stated. We suppose that $n \geq 8$.

Theorem 5.6.2 (Deviation Result). *Let* $S_0 = \{0\}$, $\widehat{g}_D = \underset{g \in S_{D,n}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (1 - Y_i g(X_i))_+$ *and*

$$\widehat{D} = \arg\min_{D \geq 0} \left(\frac{1}{n} \sum_{i=1}^n (1 - Y_i \widehat{g}_D(X_i))_+ + \mathrm{pen}(D) \right)\,.$$

Assumptions $\mathbf{H}(\tau_D)$ *and* $\mathbf{H}(\mathbf{mar})$ *are supposed to be satisfied. Let* $K > 1$ *and* $x > 0$ *be such that*

$$x \leq \frac{9}{10} \log^2 n\,. \tag{5.17}$$

Then there exist constants C_K *and* $C_{K,1}$ *such that the following holds: if* $\forall D \geq 1$*,*

$$\mathrm{pen(D)} \geq \mathrm{C_{K,1}} \left(\frac{\mathrm{D}\tau_{\mathrm{D}}^2 \log^2 \mathrm{n}}{\mathrm{p}_0^2 \mathrm{n}} \right) + \mathrm{C_K} \left(\frac{\tau_{\mathrm{D}}^2 (\log(\mathrm{D}) + \mathrm{x})}{\mathrm{p}_0^2 \mathrm{n}} \right) \text{ and } \mathrm{pen}(0) = 0\,, \tag{5.18}$$

then with probability greater than $1 - 2e^{-x}$*,*

$$L(\widehat{g}_{\widehat{D}}, f^*) \leq K \inf_{D \geq 0} \big(\inf_{g \in S_{D,n}} L(g, f^*) + 2\mathrm{pen}(D) \big)\,.$$

Up to a larger penalty, the condition (5.17) does not prevent from obtaining an average performance bound. Such a result (Theorem 5.9.1) is stated in section 5.9.

The gaussian criterion (5.2) recommends a choice of a penalty function *pen* of order $\frac{D}{n}$. The previous result is not entirely satisfactory since it fails to justify it in our case. Even if τ_D has the optimal order of magnitude $\sqrt{D}$, the penalty is of order D^2/n which is not linear with respect to the dimension. Chapter 6 shall justify the linear penalty using a clip procedure.

5.6.3 Experiments.

First experiments related to this approach have been carried out. Classifiers $\widehat{f}_D$ are computed using the first two steps of the KPM given in section 5.3. Next, the dimension is chosen by the penalized empirical loss minimization:

$$\widehat{D} = \arg\min_{D\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\widehat{f}_D(X_i))_+ + \lambda D^2\right) .$$

where the squared dimension is inspired by the penalty of Theorem 5.6.2. The parameter λ is chosen by 5-fold cross-validation on each of the training sample. This requires the choice of a grid of λ. Interestingly, the computation time is almost independent of the size of this grid: the longest task is to compute the various classifiers. Then, it suffices to minimize a vector of small size to get the selected dimension corresponding to each λ. The test errors, reported in table 5.4, have been obtained with a grid of 5001 points regularly spaced out from 0 to $0,5$.

Table 5.4: pen(D)$\propto$ D^2

	KPM
Banana ($\sigma = 0.7071$)	10.90±0.48
Breast Cancer($\sigma = 5$)	27.45±4.86
Diabetis ($\sigma = 3.1623$)	23.67±1.71
Flare Solar ($\sigma = 3.8730$)	32.50±1.77
German ($\sigma = 5.2440$)	23.97±2.38
Heart ($\sigma = 7.7460$)	17.25±3.56

The performances of this procedure and those of cross-validation directly on the dimension (table 5.2) are similar.

5.7 Appendix B: Semi-supervised Learning.

In the semi-supervised setting, some part (e.g. half to be simple) of the data is unlabeled: a decision rule is to be learned from $\{(X_i, Y_i),\ i = 1\cdots n/2\}$ and $\{X_i,\ i = n/2+1\cdots n\}$.

The finite dimensional procedure allows to take into account the information given by the unlabeled data. To begin with, a KPCA is performed on $\{X_i,\ i = n/2+1\cdots n\}$. Let S_D be the space of dimension D spanned by the first D principal components chosen by this KPCA. By analogy with (5.14), the classifiers are defined as

$$\widehat{f}_D = \operatorname*{argmin}_{f\in S_D}\frac{2}{n}\sum_{i=1}^{n/2}(1 - Y_i f(X_i))_+ .$$

The dimension is selected by

$$\widehat{D} = \arg\min_{D\geq 1}\left(\frac{2}{n}\sum_{i=1}^{n/2}(1 - Y_i\widehat{f}_D(X_i))_+ + \mathrm{pen}(D)\right) .$$

If we argue conditionally to the unlabeled data, then the models S_D are deterministic. Consequently, Theorem 5.6.2 states an oracle inequality conditionally on $\{X_i,\, i = n/2+1\cdots n\}$ for $\widehat{f}_{\widehat{D}}$ provided that $\mathrm{pen}(D)$ is proportional to $D\tau_D^2$. However, in practice, τ_D is unknown since it depends on the underlying law of the data. Only a basis $f_1, \cdots, f_D$ of S_D (coming from KPCA) is available. The following aims at showing that results of chapter 3 allows to get data-dependent bounds on τ_D.

Let M^D (resp. $\widehat{M}^D$) be the $L_2(P)$-matrix (resp. $L_2(P_{n/2})$-matrix) associated with the basis $(f_i)_{1\le i\le D}$. It is the $D\times D$ matrix with coefficients $(M^D)_{i,j} = \mathbb{E}[f_i(X)f_j(X)]$ (resp. $(\widehat{M}^D)_{i,j} = \frac{2}{n}\sum_{\ell=1}^{n/2} f_i(X_\ell)f_j(X_\ell)$). Endowed with the dot product

$$\forall\alpha,\,\beta\in\mathbb{R}^D,\ \ \langle\sum_{j=1}^D \alpha_j f_j, \sum_{j=1}^D \beta_j f_j\rangle := \sum_{j=1}^D \alpha_j\beta_j\,,$$

S_D is the RKHS associated with the kernel

$$k_D(x,x') = \sum_{i=1}^D f_i(x)f_i(x')\,.$$

Indeed, S_D is a Hilbert space such that, for all $f\in S_D$ and for all $x\in\mathcal{X}$, $\langle f, k_D(x,.)\rangle = f(x)$. Moreover, as defined in chapter 3, let C_1 (resp. $C_{1,\frac{n}{2}}$) be the covariance operator (resp. empirical covariance) associated with k_D. The matrix of C_1 (resp. $C_{1,\frac{n}{2}}$) in the basis $(f_1,\cdots,f_D)$ is M^D (resp. $\widehat{M}^D$). Consequently, chapter 3 yields concentration inequalities for the eigenvalues of $\widehat{M}^D$ around those of M^D. This remark will be useful to prove the following upper-bounds on τ_D.

Theorem 5.7.1. *Let $f_1,\ldots,f_D$ be a basis of S_D satisfying $\|\sum_{i=1}^D f_i^2\|_\infty \le B_D^2$. In the following statements, we suppose that n is sufficiently large. We have:*
$\forall\xi>0$, with probability greater than $1-e^{-\xi}$,

$$\tau_D^2 \le B_D^2\left(\lambda_D(\widehat{M}^D) - \frac{2\sqrt{2}B_D^2}{\sqrt{n}} - 2B_D^2\sqrt{\frac{\xi}{n}}\right)^{-1}, \tag{5.19}$$

and with probability greater than $1-2e^{-\xi}$,

$$\tau_D^2 \le B_D^2\left(\lambda_D(\widehat{M}^D) - \sqrt{\frac{2}{n}}\sqrt{\frac{2}{n}\sum_{\ell=1}^{n/2}\left(\sum_{i=1}^D f_i^2(X_\ell)\right)^2} - 3B_D^2\sqrt{\frac{\xi}{n}}\right)^{-1}. \tag{5.20}$$

This result is proved in the following section. The second inequality provides a data-dependent bound on τ_D.

5.8 Appendix C: Proofs.

All along this section, the following notations are used. The hinge loss is denoted by $\gamma(f,(x,y)) = (1-yf(x))_+$. Depending on the setup, Pf denotes either $\mathbb{E}[f(X)]$ or $\mathbb{E}[f(X,Y)]$ and P_nf is either $\frac{1}{n}\sum_{i=1}^n f(X_i)$ or $\frac{1}{n}\sum_{i=1}^n f(X_i,Y_i)$. Moreover, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a linear operator.

5.8.1 Proof of Theorem 5.7.1 .

Since $f_1, \cdots, f_D$ are linearly independent, M^D is a non-singular matrix. Let $F_1, \cdots, F_D$ be the following linear combination of the functions $f_1, \cdots, f_D$:

$$F = (F_1, \cdots, F_D)' = (M^D)^{-\frac{1}{2}} f\,,$$

where $f(x) = (f_1(x), \cdots, f_D(x))'$. $(F_1, \cdots, F_D)$ is an $L_2(P)$-orthonormal basis of S_D since

$$\begin{aligned} \langle F_i, F_j \rangle_{L_2(P)} = (\mathbb{E}\left[F(X)F(X)'\right])_{i,j} &= (M^D)^{-1/2}\mathbb{E}\left[f(X)f(X)'\right](M^D)^{-1/2} \\ &= (M^D)^{-1/2} M^D (M^D)^{-1/2} = I_D\,, \end{aligned}$$

where I_D is the identity matrix of $\mathbb{R}^D$. Moreover, by definition of F,

$$\sum_{i=1}^D F_i^2(x) = \langle (M^D)^{-\frac{1}{2}} f(x), (M^D)^{-\frac{1}{2}} f(x)\rangle = \langle (M^D)^{-1} f(x), f(x))\rangle\,,$$

and the Cauchy-Schwarz inequality yields

$$\|\sum_{i=1}^D F_i^2\|_\infty \leq \|(M^D)^{-1}\|_{\mathrm{op}} \|\sum_{i=1}^D f_i^2\|_\infty\,.$$

Since $(M^D)^{-1}$ is a positive definite symmetric matrix, $\|(M^D)^{-1}\|_{\mathrm{op}} = \lambda_1((M^D)^{-1})$ (see [Rud73] for example). Finally,

$$\tau_D^2 \leq \frac{B_D^2}{\lambda_D(M^D)}\,. \tag{5.21}$$

Rayleigh's formula (Inequality (3.8) of Theorem 3.2.1) yields

$$\lambda_D(M^D) \geq \lambda_D(\widehat{M}^D) - \|M^D - \widehat{M}^D\|_{\mathrm{op}}\,. \tag{5.22}$$

Lemma 4.2.1 is now specified to the finite dimensional case where the kernel is $k_D(x,x') = \sum_{i=1}^D f_i(x)f_i(x')$ and the associated RKHS is S_D. The bounded difference concentration inequality (Theorem 3.9.1) is applied to the random variable $\|M^D - \widehat{M}^D\|_{\mathrm{op}}$. As for inequality (4.1) (with $M = B_D^2$), we get that for all $\xi > 0$, with probability greater than $1 - e^{-\xi}$,

$$\|M^D - \widehat{M}^D\|_{\mathrm{op}} \leq \mathbb{E}\left[\|M^D - \widehat{M}^D\|_{\mathrm{op}}\right] + 2B_D^2\sqrt{\frac{\xi}{n}}\,. \tag{5.23}$$

Finally, Lemma 4.2.1 entails

$$\|\widehat{M}^D - M^D\|_{\mathrm{op}} \leq \frac{2\sqrt{2}B_D^2}{\sqrt{n}}\left(1 + \sqrt{\frac{\xi}{2}}\right)\,.$$

This entails that $\|\widehat{M}^D - M^D\|_{\mathrm{op}}$ tends to 0 when n goes to infinity. Thus, the lower bound of inequality (5.22) is strictly positive for large n (since $\lambda_D(\widehat{M}^D) > 0$). Gathering the previous inequality with (5.22) and (5.21) leads to in equality (5.19).

In order to prove the data-dependent bound (5.20), $\mathbb{E}\left[\|M^D - \widehat{M}^D\|_{\mathrm{op}}\right]$ is empirically controlled by specifying some reasonings of chapter 3 to the finite dimensional case where the kernel is $k_D(x,x') = \sum_{i=1}^D f_i(x)f_i(x')$ and the associated RKHS is S_D. We have

$$\begin{aligned}
\mathbb{E}\left[\|M^D - \widehat{M}^D\|_{\mathrm{op}}\right] &= \mathbb{E}\left[\sup_{\|v\|=1} |\langle v, (M^D - \widehat{M}^D)v\rangle|\right] \\
&= \mathbb{E}\left[\sup_{\|v\|=1}\left|P\left(\sum_{i,j=1}^D v_i v_j f_i f_j\right) - P_{n/2}\left(\sum_{i,j=1}^D v_i v_j f_i f_j\right)\right|\right] \\
&\leq \frac{4}{n}\mathbb{E}\left[\sup_{\|v\|=1}\left|\sum_{\ell=1}^{n/2}\epsilon_\ell\left(\sum_{i,j=1}^D v_i v_j f_i(X_\ell) f_j(X_\ell)\right)\right|\right], \qquad (5.24)
\end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_{n/2}$ are independent identically distributed Rademacher variables. ($\mathbb{P}[\varepsilon_1 = 1] = \mathbb{P}[\varepsilon_1 = -1] = 1/2$). The first equality holds since $M^D - \widehat{M}^D$ is a symmetric operator, the second one follows from the definitions of M^D and $\widehat{M}^D$. The last upper-bound is the standard symmetrization inequality. Similarly to inequality (3.23), we use the bounded difference inequality a second time to get that with probability greater than $1 - e^{-\xi}$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|v\|=1}\frac{2}{n}\left|\sum_{\ell=1}^{n/2}\epsilon_\ell\left(\sum_{i,j=1}^D v_i v_j f_i(X_\ell f_j(X_\ell\right)\right|\right] &\leq \mathbb{E}_\varepsilon\left[\sup_{\|v\|=1}\frac{2}{n}\left|\sum_{\ell=1}^{n/2}\epsilon_\ell\left(\sum_{i,j=1}^D v_i v_j f_i(X_\ell) f_j(X_\ell)\right)\right|\right] \\
&+ B_D^2\sqrt{\frac{\xi}{n}}, \qquad (5.25)
\end{aligned}$$

where $\mathbb{E}_\varepsilon$ means that the expectation is considered only with respect to the variables ε: $X_1, \ldots, X_{n/2}$ are "fixed". The control of the expectation appearing in the upper bound of the previous inequality is sketched since it follows the same line that the proof of the first inequality of Lemma 3.3.3.

$$\begin{aligned}
\sup_{\|v\|=1}\left|\sum_{\ell=1}^{n/2}\epsilon_\ell\left(\sum_{i,j=1}^D v_i v_j f_i(X_\ell) f_j(X_\ell)\right)\right| &= \sup_{\|v\|=1}\left|\sum_{i,j=1}^D v_i v_j\left(\sum_{\ell=1}^{n/2}\epsilon_\ell f_i(X_\ell) f_j(X_\ell)\right)\right| \\
&\leq \sqrt{\sum_{i,j=1}^D\left(\sum_{\ell=1}^{n/2}\epsilon_\ell f_i(X_\ell) f_j(X_\ell)\right)^2},
\end{aligned}$$

where Cauchy-Schwarz inequality is used. Finally, by Jensen's inequality,

$$\mathbb{E}_\varepsilon\left[\sup_{\|v\|=1}\left|\sum_{\ell=1}^{n/2}\epsilon_\ell\left(\sum_{i,j=1}^D v_i v_j f_i(X_\ell) f_j(X_\ell)\right)\right|\right] \leq \sqrt{\sum_{\ell=1}^{n/2}\left(\sum_{j=1}^D f_j^2(X_\ell)\right)^2}.$$

Gathering the last inequality with (5.25), (5.24), (5.23), (5.22) and (5.21), this concludes the proof of inequality (5.20). □

5.8.2 Proofs of Theorems 5.6.1 and 5.6.2 .

The proofs of these results use some common material provided in the sequel.

Proof of Theorem 5.6.1 . The definition of $\widehat{g}_D$ leads to:

$$L(\widehat{g}_D, f^*) \leq L(g_D^*, f^*) + (P - P_n)(\gamma(\widehat{g}_D) - \gamma(g_D^*)) \,, \tag{5.26}$$

where $g_D^* = \arg \min_{f \in S_{D,n}} P\gamma(f)$.

Since $\widehat{g}_D$ is a random element of $S_{D,n}$, we have to control

$$\{(P - P_n)(\gamma(g) - \gamma(g_D^*)),\, g \in S_{D,n}\} \,.$$

Theorem 5.8.2 below is applied to the class of functions $\mathcal{F} = \{\gamma(g) - \gamma(g_D^*),\, g \in S_{D,n}\}$. The weight is $\omega(f) = \|g - g_D^*\|_2^2$. Since γ is 1-lipschitz, $\|f\|_2^2 \leq \omega(f)$ and $\mathbf{H}(\tau_D)$ implies that $\|f\|_\infty \leq \tau_D \sqrt{\omega(f)}$. By definition of $S_{D,n}$, $\omega(f) \leq 4n^4 = R$. Besides, Theorem 5.8.5 below leads to $r^* = 16\frac{D}{n}$. Thus , with probability greater than $1 - e^{-x}$,

$$\forall g \in S_{D,n}, (P - P_n)(\gamma(g) - \gamma(g_D^*)) \leq K_D \|g - g_D^*\|_2 \vee K_D^2 \,,$$

where

$$K_D = 8\sqrt{\frac{D}{n}} \left(1 + e + \frac{e}{2}\left(\log \frac{n^5}{4D}\right)_+\right) + \sqrt{\frac{2x}{n}} + \frac{8x\tau_D}{3n} \,.$$

Combining this inequality with Inequality (5.26), we obtain:

$$L(\widehat{g}_D, f^*) \leq L(g_D^*, f^*) + K_D \|\widehat{g}_D - g_D^*\|_2 \vee K_D^2 \,. \tag{5.27}$$

$0 \in S_{D,n}$ implies that $L(g_D^*, f^*) \leq \sqrt{L(g_D^*, f^*)}$. Combining this inequality with Theorem 5.8.3 below yields

$$\|f^* - g_D^*\|_2 \leq 2\frac{L(g_D^*, f^*)^{\frac{1}{2}}}{p_0} \,. \tag{5.28}$$

The triangle inequality now implies:

$$\|\widehat{g}_D - g_D^*\|_2 \leq \|\widehat{g}_D - f^*\|_2 + \|f^* - g_D^*\|_2 \,.$$

Combining Inequality (5.28), Theorem 5.8.3 below and the last inequality yields

$$\|\widehat{g}_D - g_D^*\|_2 \leq (1 + \tau_D) \left[\frac{L(\widehat{g}_D, f^*)}{p_0} + \left(\frac{L(\widehat{g}_D, f^*)}{p_0}\right)^{\frac{1}{2}} + 2\frac{L(g_D^*, f^*)^{\frac{1}{2}}}{p_0}\right] \,.$$

Moreover, $x \leq a + b$ implies that $\forall K > 0$, $x \leq [(1 + \frac{1}{K})a] \vee [(1 + K)b]$. Using this elementary inequality, Inequality (5.27) and the previous one entails that:

either

$$L(\widehat{g}_D, f^*) \leq (1 + \frac{1}{K}) \left(L(g_D^*, f^*) + 2K_D\tau_D \left(\frac{L(\widehat{g}_D, f^*)}{p_0}\right)^{\frac{1}{2}} + \frac{4\tau_D K_D L(g_D^*, f^*)^{\frac{1}{2}}}{p_0} + K_D^2\right) \tag{5.29}$$

or

$$L(\widehat{g}_D, f^*) \leq (1+K)\frac{K_D(1+\tau_D)}{p_0} L(\widehat{g}_D, f^*)\,. \tag{5.30}$$

Conditions (5.15) and (5.16) imply that Inequality (5.30) is not true. Precisely, since $a + b \leq 2(a \vee b)$,

$$K_D\tau_D \leq \left[16\tau_D\sqrt{\frac{D}{n}}\left(1+e+\frac{e}{2}\left(\log\frac{n^5}{4D}\right)_+\right)\right] \vee \left[2\tau_D\sqrt{\frac{2x}{n}}+\frac{16x\tau_D^2}{3n}\right], \tag{5.31}$$

and $4D \geq 1$ implies

$$16\sqrt{\frac{D}{n}}\left(1+e+\frac{e}{2}\left(\log\frac{n^5}{4D}\right)_+\right) \leq 169\sqrt{\frac{D}{n}}\log n\,.$$

Thus, by Condition (5.15),

$$16\tau_D\sqrt{\frac{D}{n}}\left(1+e+\frac{e}{2}\left(\log\frac{n^5}{4D}\right)_+\right) < \frac{p_0}{2(1+K)}\,. \tag{5.32}$$

Moreover, Condition (5.16) imposes:

$$2\tau_D\sqrt{\frac{2x}{n}}+\frac{16x\tau_D^2}{3n} \leq 2\tau_D\sqrt{2}\frac{\log n}{\sqrt{n}}+\frac{16\tau_D^2\log^2 n}{3n}\,,$$

and Condition (5.15) leads to

$$2\tau_D\sqrt{\frac{2x}{n}}+\frac{16x\tau_D^2}{3n} \leq \frac{\sqrt{2}p_0}{169(1+K)}+\frac{16}{3}\frac{p_0^2}{337^2(1+K)^2}\,,$$

Besides, $p_0 \leq 1$ entails

$$2\tau_D\sqrt{\frac{2x}{n}}+\frac{16x\tau_D^2}{3n} < \frac{p_0}{2(1+K)}\,. \tag{5.33}$$

Finally, combining Inequalities (5.31), (5.32) and (5.33), we get $K_D\tau_D < \frac{p_0}{2(1+K)}$. This shows that (5.30) is not possible.

Consequently, Inequality (5.29) occurs. Using the elementary inequality $2ab \leq a^2 + b^2$, we have

$$2K_D\tau_D\left(\frac{L(\widehat{g}_D, f^*)}{p_0}\right)^{\frac{1}{2}} \leq \frac{L(\widehat{g}_D, f^*)}{(1+\frac{1}{K})K}+\frac{4K_D^2\tau_D^2}{p_0}\frac{(1+\frac{1}{K})K}{4}\,,$$

and

$$2\frac{2K_D\tau_D}{p_0}L(g_D^*, f^*)^{\frac{1}{2}} \leq \frac{L(g_D^*, f^*)}{(1+\frac{1}{K})K}+\frac{4K_D^2\tau_D^2}{p_0^2}(1+\frac{1}{K})K\,.$$

Plugging these inequalities into Inequality (5.29) and solving the corresponding equation, we get

$$L(\widehat{g}_D, f^*) \leq \frac{K+2}{K-1}L(g_D^*, f^*)+\frac{(K+1)^2}{K-1}K_D^2(1+5\frac{\tau_D^2}{p_0^2})\,. \tag{5.34}$$

Moreover,

$$K_D^2(1+5\frac{\tau_D^2}{p_0^2}) \le 10K_D^2\frac{\tau_D^2}{p_0^2}.$$

Condition (5.16) leads to

$$K_D \le 85\sqrt{\frac{D}{n}}\log n + \sqrt{\frac{2x}{n}} + \frac{8\sqrt{x}\tau_D\log n}{3n},$$

and condition (5.15) entails

$$K_D \le 85\sqrt{\frac{D}{n}}\log n + \sqrt{\frac{2x}{n}} + \frac{p_0\sqrt{x}}{\sqrt{10}\sqrt{n}}.$$

Thus

$$K_D^2 \le 3\times 85^2\frac{D}{n}\log^2 n + \frac{6x}{n} + \frac{3p_0^2x}{10n},$$

and finally

$$10K_D^2\frac{\tau_D^2}{p_0^2} \le 30\times 85^2\frac{D\tau_D^2}{np_0^2}\log^2 n + \frac{3x\tau_D^2}{n}\left(\frac{20}{p_0^2}+1\right). \tag{5.35}$$

Combining inequalities (5.34) and (5.35), this concludes the proof of Theorem 5.6.1 with $C_1 = 216750$ and $C_2 = 120$. □

Proof of Theorem 5.6.2. The definitions of $\widehat{D}$ and $\widehat{g}_D$ imply that $\forall\, D \ge 0$ and for $\mathrm{g_D} \in \mathrm{S_{D,n}}$,

$$L(\widetilde{g}, f^*) \le L(g_D, f^*) + \mathrm{pen}(D) - \mathrm{pen}(\widehat{D}) + (P-P_n)(\gamma(\widetilde{g}) - \gamma(g_D)). \tag{5.36}$$

The following decomposition holds:

$$(P-P_n)(\gamma(\widetilde{g}) - \gamma(g_D)) = (P-P_n)(\gamma(u_{\widehat{D}}) - \gamma(g_D)) + (P-P_n)(\gamma(\widetilde{g}) - \gamma(u_{\widehat{D}})), \tag{5.37}$$

where u_D satisfies

$$\|u_D - f^*\|_2 = \inf_{g\in S_{D,n}} \|g - f^*\|_2,$$

(if the infimum is not reached we can use a dominate convergence argument) and $g_D = \operatorname*{argmin}_{g\in S_{D,n}} P\gamma(g)$. The terms $(1) = (P-P_n)(\gamma(u_{\widehat{D}}) - \gamma(g_D))$ and $(2) = (P-P_n)(\gamma(\widetilde{g}) - \gamma(u_{\widehat{D}}))$ are controlled separately. Let us consider the following weights on each model S_D:

$$x_0 = 1\,,\ x_D = \log(3D^2)\ \mathrm{for\,D} \ge 1\,.$$

They satisfy $\sum_{D\ge 0} e^{-x_D} \le 1$.

Let $D_1 \ge 0$. We have

$$\|\gamma(u_{D_1}) - \gamma(g_D)\|_\infty \le \|u_{D_1} - g_D\|_\infty \le (\tau_{D_1}\vee\tau_D)\|u_{D_1} - g_D\|_2,$$

where the first inequality occurs since γ is 1-lipschitz. The second one uses that the collection of models $(S_D)_{D\geq 0}$ is an increasing sequence of subsets satisfying assumption $\mathbf{H}(\tau_D)$. Bernstein's inequality yields that with probability greater than $1-e^{-x_{D_1}-x_D-x}$,

$$(P-P_n)(\gamma(u_{D_1})-\gamma(g_D)) \leq \sqrt{\frac{2(x_{D_1}+x_D+x)\mathrm{Var}(\gamma(u_{D_1})-\gamma(g_D))}{n}} + \frac{(\tau_{D_1}\vee\tau_D)\|u_{D_1}-g_D\|_2(x_{D_1}+x_D+x)}{6n}. \tag{5.38}$$

The definition of u_{D_1} leads to:

$$\|u_{D_1}-g_D\|_2 \leq \|u_{D_1}-f^*\|_2+\|f^*-g_D\|_2 \leq \|g_1-f^*\|_2+\|f^*-g_D\|_2\,, \tag{5.39}$$

for all $g_1 \in S_{D_1,n}$. Combining inequality (5.39) with inequality (5.38), we find that, with probability greater than $1-e^{-x_{D_1}-x_D-x}$,

$$\begin{aligned}(P-P_n)(\gamma(u_{D_1})-\gamma(g_D)) \leq\ & \sqrt{2\frac{(x_{D_1}+x_D+x)}{n}}(\|g_1-f^*\|_2+\|f^*-g_D\|_2)\\ &+(\tau_{D_1}\vee\tau_D)\frac{x_{D_1}+x_D+x}{6n}\,(\|g_1-f^*\|_2+\|f^*-g_D\|_2)\,.\end{aligned}$$

Using twice union-of-event bound and choosing the random element $\widehat{D}=D_1$ and $g_1=\widetilde{g}$, we find that with probability greater than $1-e^{-x}$, $\forall D\geq 0$

$$(1) \leq B_{D,\widehat{D}}(\|\widetilde{g}-f^*\|_2+\|g_D-f^*\|_2)\,, \tag{5.40}$$

where

$$B_{D,\widehat{D}} = \sqrt{\frac{2(x_D+x_{\widehat{D}}+x)}{n}} + (\tau_D\vee\tau_{\widehat{D}})\frac{x_D+x_{\widehat{D}}+x}{6n}\,. \tag{5.41}$$

Let $D'\geq 0$. Theorem 5.8.2 below is applied to the class of functions $\mathcal{F}_{D'}=\{(\gamma(g)-\gamma(u_{D'})),\ g\in S_{D',n}\}$. The weight $\omega(f)=\|g-u_{D'}\|_2^2$ satisfies $\|f\|_2^2\leq\omega(f)\leq 4n^2=R$. Moreover, assumption $\mathbf{H}(\tau_{D'})$ implies that $\mu=\tau_{D'}$ since $\|f\|_\infty\leq\|g-u_{D'}\|_\infty\leq\tau_{D'}\sqrt{\omega(f)}$. Finally, Theorem 5.8.5 below leads to $r^*=16\frac{D'}{n}$. It follows that with probability greater than $1-e^{-x-x_{D'}}$, $\forall g\in S_{D',n}$

$$(P-P_n)(\gamma(g)-\gamma(u_{D'})) \leq K_{D'}\|g-u_{D'}\|_2\vee K_{D'}^2\,, \tag{5.42}$$

where

$$K_{D'} \leq 8\sqrt{\frac{D'}{n}}\left(1+e+\frac{e}{2}\left(\log\frac{n^5}{4D'}\right)_+\right)+\sqrt{\frac{2(x+x_{D'})}{n}}+\frac{8(x+x_{D'})\tau_{D'}}{3n}\,.$$

Using the union-of-event bound and choosing $D'=\widehat{D}$ and $g=\widetilde{g}$, the following holds: with probability greater than $1-e^{-x}$,

$$(2) \leq K_{\widehat{D}}\|\widetilde{g}-u_{\widehat{D}}\|_2\vee K_{\widehat{D}}^2\,, \tag{5.43}$$

with

$$K_{\widehat{D}} \leq 85\sqrt{\frac{\widehat{D}}{n}}\log n+\sqrt{\frac{2(x+x_{\widehat{D}})}{n}}+\frac{8(x+x_{\widehat{D}})\tau_{\widehat{D}}}{3n}\,. \tag{5.44}$$

The definition of $u_{\widehat{D}}$ leads to:

$$\|\widetilde{g} - u_{\widehat{D}}\|_2 \leq \|\widetilde{g} - f^*\|_2 + \|f^* - u_{\widehat{D}}\|_2 \leq 2\|\widetilde{g} - f^*\|_2 \,. \tag{5.45}$$

Combining Theorem 5.8.3 below and inequalities (5.40), (5.43) and (5.45) yields

$$(1) \leq B_{D,\widehat{D}}\left((1+\tau_{\widehat{D}})\left(\frac{L(\widetilde{g}, f^*)}{p_0} + \frac{L(\widetilde{g}, f^*)^{\frac{1}{2}}}{\sqrt{p_0}}\right) + (1+\tau_D)\left(\frac{L(g_D, f^*)}{p_0} + \frac{L(g_D, f^*)^{\frac{1}{2}}}{\sqrt{p_0}}\right)\right), \tag{5.46}$$

and

$$(2) \leq 2K_{\widehat{D}}(1+\tau_{\widehat{D}})\left(\frac{L(\widetilde{g}, f^*)}{p_0} + \frac{L(\widetilde{g}, f^*)^{\frac{1}{2}}}{\sqrt{p_0}}\right) + K_{\widehat{D}}^2 \,. \tag{5.47}$$

$0 \in S_{D,n}$ implies that $L(g_D, f^*) \leq \sqrt{L(g_D, f^*)}$. Combining this with inequalities (5.46), (5.47) (5.37) and (5.36) implies that, with probability at least $1 - 2e^{-x}$,

$$\begin{aligned} L(\widetilde{g}, f^*) \quad \leq \quad & L(g_D, f^*) + \mathrm{pen}(D) - \mathrm{pen}(\widehat{D}) + (2K_{\widehat{D}} + B_{D,\widehat{D}})(1+\tau_{\widehat{D}})\left[\frac{L(\widetilde{g}, f^*)}{p_0} + \frac{L(\widetilde{g}, f^*)^{\frac{1}{2}}}{\sqrt{p_0}}\right] \\ & + 2B_{D,\widehat{D}}\frac{1+\tau_D}{p_0}[L(g_D, f^*)^{\frac{1}{2}}] + K_{\widehat{D}}^2 \,. \end{aligned}$$

If $x \leq a+b$ then $\forall \theta > 0$, $x \leq [(1+\theta)a] \vee [(1+\frac{1}{\theta})b]$. Using this elementary inequality with $\theta_K = \frac{K-1}{(K+2)^2}$, we have that, with probability at least $1 - 2e^{-x}$,

either

$$\begin{aligned} L(\widetilde{g}, f^*) \leq (1+\theta_K)\{L(g_D, f^*) + \quad \mathrm{pen}(D) \quad & -\mathrm{pen}(\widehat{D}) + \frac{2\tau_{\widehat{D}}(2K_{\widehat{D}} + B_{D,\widehat{D}})}{\sqrt{p_0}} L(\widetilde{g}, f^*)^{\frac{1}{2}} \\ + \quad & \frac{4B_{D,\widehat{D}}\tau_D}{p_0} L(g_D, f^*)^{\frac{1}{2}} + K_{\widehat{D}}^2\} \,, \end{aligned} \tag{5.48}$$

or

$$L(\widetilde{g}, f^*) \leq (1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}(2K_{\widehat{D}} + B_{D,\widehat{D}})}{p_0} L(\widetilde{g}, f^*) \,. \tag{5.49}$$

In order to exclude the case where (5.49) is true, the considered dimensions are momentarily restricted. To be precise, we consider only the dimensions satisfying

$$\tau_D\sqrt{D} \leq \frac{\Gamma_K\sqrt{n}}{\log n} \,, \tag{5.50}$$

where $\Gamma_K = \frac{(K-1)p_0}{5441(K+2)^2}$. Moreover, we suppose that the selected dimension satisfies this condition. This assumption is denoted by $\mathrm{H}(\widehat{D})$.

$$H(\widehat{D}): \ \tau_{\widehat{D}}\sqrt{\widehat{D}} \leq \frac{\Gamma_K\sqrt{n}}{\log n} \,. \tag{5.51}$$

The proof is now divided into three steps:

- **Step 1** shows that the inequality (5.49) is not true under the previous assumptions.
- **Step 2** provides an oracle inequality for dimensions satisfying condition (5.50).
- **Step 3** provides an oracle inequality without restrictions on the dimension.

Step 1. We show that

$$(1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}(2K_{\widehat{D}}+B_{D,\widehat{D}})}{p_0} < 1\,. \tag{5.52}$$

Condition (5.17), $\tau_D \geq \sqrt{D}$ and $n \geq 8$ implies that

$$x + x_D \leq 2\log^2 n \,\mathrm{if}\, \mathrm{\tau_D\sqrt{D} \leq \Gamma_K\frac{\sqrt{n}}{\log n}}\,. \tag{5.53}$$

On the one hand, Inequality (5.44) and condition (5.53) yields

$$K_{\widehat{D}} \leq 85\sqrt{\frac{\widehat{D}}{n}}\log n + \sqrt{\frac{2(x+x_{\widehat{D}})}{n}} + \frac{4\log n\sqrt{x+x_{\widehat{D}}}\tau_{\widehat{D}}}{n}\,,$$

and Assumption $\mathrm{H}(\widehat{D})$ leads to:

$$K_{\widehat{D}} \leq 85\sqrt{\frac{\widehat{D}}{n}}\log n + \sqrt{\frac{2(x+x_{\widehat{D}})}{n}} + \frac{5\Gamma_K\sqrt{x+x_{\widehat{D}}}}{\sqrt{n}} \leq 85\sqrt{\frac{\widehat{D}}{n}} + 5\sqrt{\frac{x+x_{\widehat{D}}}{n}}\,.$$

Thus

$$K_{\widehat{D}} \leq 85 f(D) + 5g(D) + 85 f(\widehat{D}) + 5g(\widehat{D})\,, \tag{5.54}$$

where $f(D) = \sqrt{\frac{D}{n}}\log n$ and $g(D) = \sqrt{\frac{(x+x_D)}{n}}$. On the other hand, inequality (5.41) entails

$$B_{D,\widehat{D}} \leq 2\left(\sqrt{\frac{x_D+x}{n}} \vee \sqrt{\frac{x_{\widehat{D}}+x}{n}} + (\tau_D \vee \tau_{\widehat{D}})(\frac{x_D+x}{6n} \vee \frac{x_{\widehat{D}}+x}{6n})\right)$$

The definition of τ_D permits us to suppose that the sequence $(\tau_D)_{D\geq 1}$ is increasing. Lemma 5.8.1 thus implies that

$$B_{D,\widehat{D}} \leq 2\left(\sqrt{\frac{x_D+x}{n}} \vee \sqrt{\frac{x_{\widehat{D}}+x}{n}} + \tau_D\frac{x_D+x}{6n} + \tau_{\widehat{D}}\frac{x_{\widehat{D}}+x}{6n}\right)\,.$$

Moreover, Assumption $\mathrm{H}(\widehat{D})$ and Inequality (5.53) yield $\tau_D\frac{x_D+x}{n} \leq 5\sqrt{\frac{x_D+x}{n}}$. Thus

$$B_{D,\widehat{D}} \leq 4g(D) + 4g(\widehat{D})\,. \tag{5.55}$$

Inequalities (5.54) and (5.55) yield

$$(1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}(2K_{\widehat{D}}+B_{D,\widehat{D}})}{p_0} \leq (1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}}{p_0}\left(170f(D)+14g(D)+170f(\widehat{D})+14g(\widehat{D})\right)\,,$$

and so

$$(1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}(2K_{\widehat{D}}+B_{D,\widehat{D}})}{p_0} \le 4(1+\theta_K^{-1})\frac{\tau_D \vee \tau_{\widehat{D}}}{p_0}\left(170\left(f(D)\vee f(\widehat{D})\right)+14\left(g(D)\vee g(\widehat{D})\right)\right).$$

Since $\sqrt{D}$, τ_D and x_D increase with D, Lemma 5.8.1 implies that

$$(1+\theta_K^{-1})\frac{2\tau_{\widehat{D}}(2K_{\widehat{D}}+B_{D,\widehat{D}})}{p_0} \le (1+\theta_K^{-1})\frac{1360}{p_0}(\tau_D f(D)\vee \tau_{\widehat{D}} f(\widehat{D})\vee \tau_D g(D)\vee \tau_{\widehat{D}} g(\widehat{D})) . \tag{5.56}$$

Condition (5.50) implies that

$$(1+\theta_K^{-1})\frac{1360}{p_0}\tau_D f(D) < 1\,, \tag{5.57}$$

and Assumption H($\widehat{D}$) yields

$$(1+\theta_K^{-1})\frac{1360}{p_0}\tau_{\widehat{D}} f(\widehat{D}) < 1\,. \tag{5.58}$$

Conditions (5.53) and (5.50) entail

$$(1+\theta_K^{-1})\frac{1360}{p_0}\tau_D g(D) \le (1+\theta_K^{-1})\frac{1360}{p_0}\tau_D\frac{\log n}{\sqrt{n}} < 1\,, \tag{5.59}$$

and Assumption H($\widehat{D}$) implies that

$$(1+\theta_K^{-1})\frac{1360}{p_0}\tau_{\widehat{D}} g(\widehat{D}) < 1\,, \tag{5.60}$$

Combining inequalities (5.57), (5.58), (5.59), (5.60) and (5.56), we get inequality (5.52) which ensures that (5.49) does not occur. Consequently, inequality (5.48) is true.

Step 2. Using $2ab \le a^2+b^2$, we get:

$$(2K_{\widehat{D}}+B_{D,\widehat{D}})\frac{2\tau_{\widehat{D}}}{\sqrt{p_0}}L(\widetilde{g},f^*)^{\frac{1}{2}} \le \theta_K L(\widetilde{g},f^*) + \frac{\tau_{\widehat{D}}^2(2K_{\widehat{D}}+B_{D,\widehat{D}})^2}{p_0\theta_K}\,,$$

and

$$4B_{D,\widehat{D}}\frac{\tau_D}{p_0}L(g_D,f^*)^{\frac{1}{2}} \le \theta_K L(g_D,f^*) + \frac{4B_{D,\widehat{D}}^2\tau_D^2}{\theta_K p_0^2}\,.$$

Combining these inequalities with Inequality (5.48), we get: with probability at least $1-2e^{-x}$,

$$\begin{aligned} L(\widetilde{g},f^*) \quad &\le \quad (1+\theta_K)\{(1+\theta_K)L(g_D,f^*)+\mathrm{pen}(D)-\mathrm{pen}(\widehat{D}) \\ &\quad +\theta_K L(\widetilde{g},f^*) + \frac{2\tau_{\widehat{D}}^2(2K_{\widehat{D}}+B_{D,\widehat{D}})^2}{p_0\theta_K} + \frac{4B_{D,\widehat{D}}^2\tau_D^2}{\theta_K p_0^2}\}\,. \end{aligned}$$

Since $0<\theta_K<(\sqrt{5}-1)/2$, solving the previous inequality entails

$$L(\widetilde{g},f^*) \le \frac{(1+\theta_K)^2}{1-(1+\theta_K)\theta_K}L(g_D,f^*) + \frac{(1+\theta_K)}{1-(1+\theta_K)\theta_K}\left(\mathrm{pen}(D)-\mathrm{pen}(\widehat{D})+\frac{1}{\theta_K}(C)\right) \tag{5.61}$$

where

$$(C) = \frac{2(2K_{\widehat{D}} + B_{D,\widehat{D}})^2 \tau_{\widehat{D}}^2}{p_0} + 4B_{D,\widehat{D}}^2 (\frac{\tau_D}{p_0})^2 .$$

This shows that, with probability at least $1 - 2e^{-x}$,

$$L(\widetilde{g}, f^*) \leq KL(g_D, f^*) + \frac{(1+\theta_K)}{1-(1+\theta_K)\theta_K} \left(\mathrm{pen}(D) - \mathrm{pen}(\widehat{D}) + \frac{1}{\theta_K}(C) \right) . \tag{5.62}$$

Inequalities (5.54) and (5.55) imply

$$\frac{(2K_{\widehat{D}} + B_{D,\widehat{D}})^2}{2} \leq 57800 \frac{D \log^2 n}{n} + 392 \frac{x + x_D}{n} + 57800 \frac{\widehat{D} \log^2 n}{n} + 392 \frac{x + x_{\widehat{D}}}{n} ,$$

thus

$$\frac{2(2K_{\widehat{D}} + B_{D,\widehat{D}})^2 \tau_{\widehat{D}}^2}{p_0} \leq \frac{8(\tau_D^2 \vee \tau_{\widehat{D}}^2)}{np_0} \left(57800 (D \log^2 n \vee \widehat{D} \log^2 n) + 392 (x + x_D) \vee (x + x_{\widehat{D}}) \right) .$$

Since τ_D and x_D increase with D, we have

$$\frac{2(2K_{\widehat{D}} + B_{D,\widehat{D}})^2 \tau_{\widehat{D}}^2}{p_0} \leq \frac{8}{np_0} \left(c_1 (\tau_D^2 D \log^2 n + \tau_{\widehat{D}}^2 \widehat{D} \log^2 n) + c_2 (\tau_D^2 (x + x_D) + \tau_{\widehat{D}}^2 (x + x_{\widehat{D}})) \right) ,$$

where $c_1 = 57800$ and $c_2 = 392$. The same reasoning leads to:

$$4B_{D,\widehat{D}}^2 (\frac{\tau_D}{p_0})^2 \leq \frac{256}{p_0^2} \left(\frac{\tau_D^2 (x + x_D)}{n} + \frac{\tau_{\widehat{D}}^2 (x + x_{\widehat{D}})}{n} \right) .$$

Finally,

$$(C) \leq \frac{\kappa_1}{p_0} \left(\frac{\tau_D^2 D \log^2 n}{n} + \frac{\tau_{\widehat{D}}^2 \widehat{D} \log^2 n}{n} \right) + \frac{\kappa_2}{p_0^2} \left(\frac{\tau_D^2 (x + x_D)}{n} + \frac{\tau_{\widehat{D}}^2 (x + x_{\widehat{D}})}{n} \right) , \tag{5.63}$$

where $\kappa_1 = 462400$ and $\kappa_2 = 3392$. Considering inequality (5.61), if

$$\mathrm{pen}(D) \geq \frac{462400}{\theta_K p_0} \frac{\tau_D^2 D \log^2 n}{n} + \frac{3392}{\theta_K p_0^2} \frac{\tau_D^2 (x + x_D)}{n} ,$$

then with probability at least $1 - 2e^{-x}$,

$$L(\widetilde{g}, f^*) \leq KL(g_D, f^*) + \frac{2(1+\theta_K)}{1-(1+\theta_K)\theta_K} \mathrm{pen}(D) .$$

Moreover,

$$\frac{(1+\theta_K)}{1-(1+\theta_K)\theta_K} \leq K ,$$

thus with probability at least $1 - 2e^{-x}$,

$$L(\widetilde{g}, f^*) \leq KL(g_D, f^*) + 2K \mathrm{pen}(D) .$$

We have shown that with probability greater than $1 - 2e^{-x}$, $\forall D$ such that $\Gamma_K \frac{\sqrt{n}}{\log n} \geq \sqrt{D} \tau_D$,

$$L(\widetilde{g}, f^*) \leq K\left(L(g_D, f^*) + 2\text{pen}(D)\right) .$$

Consequently, with probability greater than $1 - 2e^{-x}$,

$$L(\widetilde{g}, f^*) \leq K \inf_{\Gamma_K \frac{\sqrt{n}}{\log n} \geq D \geq 0} \left(L(g_D, f^*) + 2\text{pen}(D)\right) .$$

Step 3. From now on, $\text{pen}(D)$ satisfies inequality (5.18) with $C_{K,1} \geq \frac{5441^2 (K+2)^4}{(K-1)^2}$. Consequently, if $\sqrt{D}\tau_D > \frac{\Gamma_K \sqrt{n}}{\log n}$, then

$$\text{pen}(D) \geq C_{K,1}\Gamma_K^2 \geq 1 . \tag{5.64}$$

Besides,

$$P_n(\widehat{g}_0) + \text{pen}(0) = 1 .$$

Consequently, since $\widehat{D}$ is defined in Theorem 5.6.2 by

$$\widehat{D} = \arg\min_{D \geq 0} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(\widehat{g}_D, (X_i, Y_i)) + \text{pen}(D) \right) ,$$

it satisfies Assumption $\text{H}(\widehat{D})$ (inequality (5.51)). Thus, step 2 entails that with probability greater than $1 - 2e^{-x}$,

$$L(\widetilde{g}, f^*) \leq K \inf_{\Gamma_K \frac{\sqrt{n}}{\log n} \geq D \geq 0} \left(L(g_D, f^*) + 2\text{pen}(D)\right) .$$

However, if $\sqrt{D}\tau_D > \frac{\Gamma_K \sqrt{n}}{\log n}$, inequality (5.64) yields

$$L(g_D, f^*) + 2\text{pen}(\text{D})) \geq -\mathbb{E}(\gamma(\text{f}^*, \text{X}, \text{Y})) + 2\text{pen}(\text{D})) \geq 2 - \mathbb{E}[\gamma(\text{f}^*, \text{X}, \text{Y})] .$$

Moreover,

$$L(g_0, f^*) + 2\text{pen}(0) = 1 - \mathbb{E}[\gamma(\text{f}^*, \text{X}, \text{Y})] .$$

Consequently,

$$\inf_{\frac{\Gamma_K \sqrt{n}}{\log n} \geq \sqrt{D}\tau_D} \left(L(g_D, f^*) + 2\text{pen}(\text{D})\right) = \inf_{\text{D} \geq 0} \left(\text{L}(\text{g}_\text{D}, \text{f}^*) + 2\text{pen}(\text{D})\right) .$$

Theorem 5.6.2 is thus proved with $C_{K,1} = \frac{5441^2 (K+2)^4}{(K-1)^2}$ and $C_K = \frac{8440 (K+2)^2}{(K-1)}$. □

Lemma 5.8.1. *Let f and g be two positive increasing functions or two positive decreasing functions. Then*

$$(f(x) \vee f(y)) \times (g(x) \vee g(y)) \leq f(x)g(x) + f(y)g(y) .$$

proof Let $x^* \in \{x, y\}$ such that $f(x) \vee f(y) = f(x^*)$ and $g(x) \vee g(y) = g(x^*)$. We have $(f(x) \vee f(y)) \times (g(x) \vee g(y)) = f(x^*)g(x^*) \leq f(x)g(x) \vee f(y)g(y)$. □

5.8.3 Concentration Inequality .

Theorem 5.8.2 (Concentration Result). *Let $\mathcal{F}$ be a set of functions such that there exists a weight $\omega : \mathcal{F} \longrightarrow \mathbb{R}_+$ satisfying*

$$R \geq \omega(f) \geq \|f\|_2^2, \; \|f\|_\infty \leq \mu\sqrt{\omega(f)} ,$$

and $\forall r \in [r^*, R]$,

$$\mathbb{E} \sup_{f\in\mathcal{F},\omega(f)\leq r} |(P-P_n)(f)| \leq \phi(r) ,$$

where ϕ is a sub-root function (defined in [BBM03] and in Appendix B of chapter 3).
Then with probability greater than $1 - e^{-x}$,

$$\forall f \in \mathcal{F} \; , \; (P-P_n)(f) \leq A\sqrt{\omega(f)} \vee A^2 ,$$

where

$$A = 2\sqrt{r^*}\left(1 + e + \frac{e}{2}\left(\log\frac{R}{r^*}\right)_+\right) + \sqrt{\frac{2x}{n}} + \frac{8x\mu}{3n} ,$$

and r^* is the fixed point of ϕ.

proof Let V_r be the following normalized supremum of empirical process

$$V_r = \sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{\sqrt{r\omega(f)} \vee r} .$$

In order to use Bousquet's version of Talagrand's inequality ([Bou02b]), we need an upper bound on V_r as well on its variance.

$$\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} \frac{Pf - f(x)}{\sqrt{r\omega(f)} \vee r} \leq \frac{2\mu\sqrt{\omega(f)}}{\sqrt{r\omega(f)} \vee r} \leq \frac{2\mu}{\sqrt{r}} ,$$

and

$$\mathrm{Var}\left(\frac{\mathrm{f}}{\sqrt{\mathrm{r}\omega(\mathrm{f})} \vee \mathrm{r}}\right) \leq \frac{\omega(\mathrm{f})}{\mathrm{r}\omega(\mathrm{f}) \vee \mathrm{r}^2} \leq \frac{1}{\mathrm{r}} .$$

Thus $\forall x > 0$, with probability greater than $1 - e^{-x}$,

$$V_r \leq \mathbb{E}V_r + \sqrt{\frac{2x}{rn} + \frac{8x\mathbb{E}V_r}{n}\frac{\mu}{\sqrt{r}}} + \frac{2x\mu}{3n\sqrt{r}} .$$

Finally, using $2ab \leq a^2 + b^2$: with probability greater than $1 - e^{-x}$,

$$V_r \leq 2\mathbb{E}V_r + \sqrt{\frac{2x}{rn}} + \frac{8x\mu}{3\sqrt{r}n} .$$

Since ϕ is a sub-root function,

$$\forall r \geq r^*, \frac{\phi(r)}{r} \leq \sqrt{\frac{r^*}{r}} ,$$

thus, Theorem 5.8.6 (case 3) entails that $\forall r \geq r^*$,

$$V_r \leq \frac{A}{\sqrt{r}},$$

where

$$A = 2\sqrt{r^*}\left(1 + e + \frac{e}{2}\left(\log\frac{R}{r^*}\right)_+\right) + \sqrt{\frac{2x}{n}} + \frac{8x\mu}{3n}.$$

The choice $r = A^2$ concludes the proof. □

5.8.4 Communication between the Variance and the Risk.

Theorem 5.8.3. *If assumptions* $\mathbf{H}(\tau_D)$ *and* $\mathbf{H}(\mathbf{mar})$ *are satisfied, then the following holds:*

$$\forall\ g \in S_D\ , \|g - f^*\|_2 \leq (1 + \tau_D)\left(\frac{L(g, f^*)}{p_0} + \sqrt{\frac{L(g, f^*)}{p_0}}\right).$$

Proof. Lemma 5.8.4 below with $C = S_D$, $\tau = \tau_D$ and $f = f^*$ implies that: $\forall g \in S_D$,

$$\|g - f^*\|_2 \leq \tau_D \|g - f^*\|_1 + \|f^* - P_C(f^*)\|_\infty^{\frac{1}{2}}(1 + \tau)^{\frac{1}{2}}\sqrt{\|g - f^*\|_1}.$$

Moreover, $\|f^* - P_C(f^*)\|_\infty \leq 1 + \|P_C(f^*)\|_\infty \leq 1 + \tau_D$ and the proof of Lemma 4 of [BBM04] leads to

$$\|g - f^*\|_1 \leq \frac{L(g, f^*)}{p_0}.$$

This concludes the proof. □

Lemma 5.8.4 (Communication between L_2 and L_1 norms). *Let* $f \in L_\infty(\mathcal{X})$ *and let* C *be a closed convex subspace of* $L_2(P)$*. If there exists* τ *(depending on the convex set) satisfying* $\forall c, c' \in C$,

$$\|c - c'\|_\infty \leq \tau \|c - c'\|_2,$$

then the following inequality holds:

$$\forall\ c \in C\ , \|c - f\|_2 \leq \tau\|c - f\|_1 + \|f - P_C(f)\|_\infty^{\frac{1}{2}}(1 + \tau)^{\frac{1}{2}}\sqrt{\|c - f\|_1}.$$

Proof. Let P_C be the projection onto the closed convex set C. The projection $P_C(g)$ is the element of C satisfying: $\forall\, c \in C$,

$$\langle g - P_C(g), c - P_C(g)\rangle \leq 0. \tag{5.65}$$

$P_C(h)$ satisfies: $\forall\, c \in C$,

$$\langle h - P_C(h), c - P_C(h)\rangle \leq 0. \tag{5.66}$$

By summing equation (5.65) with $c = P_C(h)$ and equation (5.66) with $c = P_C(g)$, we obtain:

$$\begin{aligned}
\|P_C(g) - P_C(h)\|_2^2 &\leq \langle h - g, P_C(h) - P_C(g)\rangle \\
&= \int_{x\in\mathcal{X}} (h - g)(x)(P_C(h) - P_C(g))(x)dP(x) \\
&\leq \tau\|g - h\|_1 \|P_C(g) - P_C(h)\|_2.
\end{aligned}$$

Finally,

$$\|P_C(g) - P_C(h)\|_2 \leq \tau \|g - h\|_1 \,. \tag{5.67}$$

Moreover, $\forall c \in C$,

$$\begin{aligned}
\|c - f\|_2 &\leq \|c - P_C(f)\|_2 + \|f - P_C(f)\|_2 \\
&\leq \|c - P_C(f)\|_2 + \|f - P_C(f)\|_\infty^{\frac{1}{2}} \, \|f - P_C(f)\|_1^{\frac{1}{2}} \\
&\leq \|c - P_C(f)\|_2 + \|f - P_C(f)\|_\infty^{\frac{1}{2}} \, \sqrt{\|f - c\|_1 + \|c - P_C(f)\|_1} \\
&\leq \|c - P_C(f)\|_2 + \|f - P_C(f)\|_\infty^{\frac{1}{2}} \, \sqrt{\|f - c\|_1 + \|c - P_C(f)\|_2} \,.
\end{aligned}$$

Lemma 5.8.3 follows by combining Equation (5.67) and the last inequality. □

5.8.5 Localized Rademacher Average .

The next result provides a complexity control of a linear space of finite dimension.

Theorem 5.8.5. *$\forall u \in S_D$,*

$$\mathbb{E} \sup_{g \in S_D, \, \|g-u\|_2^2 \leq r} |(P - P_n)(\gamma(g) - \gamma(u))| \leq 4\sqrt{\frac{rD}{n}} = \phi(r) \,.$$

The fixed point of ϕ is $r^* = 16\frac{D}{n}$.

Proof Let $S_D(r) = \{g \in S_D, \, \|g - u\|_2^2 \leq r\}$. By symmetrization,

$$\mathbb{E}\left[\sup_{g \in S_D(r)} |(P - P_n)(\gamma(g)) - \gamma(u))|\right] \leq \frac{2}{n}\mathbb{E}\left[\sup_{g \in S_D(r)} |\sum_{i=1}^n \epsilon_i(\gamma(g, (X_i, Y_i)) - \gamma(u, (X_i, Y_i)))|\right] .$$

Let $R_r(S_D) = \sup_{g \in S_D(r)} \sum_{i=1}^n \epsilon_i(\gamma(g, (X_i, Y_i)) - \gamma(u, (X_i, Y_i)))$. Using the symmetry of the Rademacher random variables and $|x| = x_+ + x_-$, we have

$$\mathbb{E}\left[\sup_{g \in S_D(r)} |\sum_{i=1}^n \epsilon_i(\gamma(g, (X_i, Y_i)) - \gamma(u, (X_i, Y_i)))|\right] \leq 2\mathbb{E}\left[(R_r(S_D))_+\right] .$$

Since $R_r(S_D) \geq 0$ (consider $g = u$), $\mathbb{E}\left[(R_r(S_D))_+\right] = \mathbb{E}\left[R_r(S_D)\right]$.

Finally,

$$\mathbb{E}\left[\sup_{g \in S_D(r)} |(P - P_n)(\gamma(g)) - \gamma(u))|\right] \leq \frac{2}{n}\mathbb{E}\left[\sup_{g \in S_D(r)} \sum_{i=1}^n \epsilon_i(\gamma(g, (X_i, Y_i)))\right] .$$

Since γ is 1-lipschitz, the contraction principle ([LT91]) leads to

$$\begin{aligned}
\mathbb{E}\left[\sup_{g \in S_D(r)} |(P - P_n)(\gamma(g)) - \gamma(u))|\right] &\leq \frac{4}{n}\mathbb{E}\left[\sup_{g \in S_D(r)} \sum_{i=1}^n \epsilon_i g(X_i)\right] \\
&= \frac{4}{n}\mathbb{E}\left[\sup_{g \in S_D(r)} \sum_{i=1}^n \epsilon_i (g(X_i) - u(X_i))\right] \\
&= \mathbb{E}\left[\sup_{f \in S_D, \, \|f\|^2 \leq r} \sum_{i=1}^n \epsilon_i f(X_i)\right] .
\end{aligned}$$

Let $(\phi_i)_{1\leq i\leq D}$ be an orthonormal basis of S_D.

$$\mathbb{E}\left[\sup_{f\in S_D, \|f\|^2\leq r}\sum_{i=1}^n \epsilon_i f(X_i)\right] = \mathbb{E}\left[\sup_{\alpha\in\mathbb{R}^D, \sum_{i=1}^D \alpha_i^2\leq r}\sum_{j=1}^D \alpha_j \sum_{i=1}^n \epsilon_i \phi_j(X_i)\right].$$

Using the Cauchy-Schwarz inequality, we obtain

$$\sup_{\alpha\in\mathbb{R}^D, \sum_{i=1}^D \alpha_i^2\leq r}\sum_{j=1}^D \alpha_j \sum_{i=1}^n \epsilon_i \phi_j(X_i) \leq \sqrt{r}\left(\sum_{j=1}^D\left(\sum_{i=1}^n \epsilon_i\phi_j(X_i)\right)^2\right)^{\frac{1}{2}}.$$

Using Jensen's inequality twice now yields:

$$\mathbb{E}\left[\left(\sum_{j=1}^D\left(\sum_{i=1}^n \epsilon_i\phi_j(X_i)\right)^2\right)^{\frac{1}{2}}\right] \leq \mathbb{E}_X\left(\sum_{j=1}^D\sum_{i=1}^n \phi_j(X_i)^2\right)^{\frac{1}{2}},$$

and

$$\mathbb{E}_X\left(\sum_{j=1}^D\sum_{i=1}^n \phi_j(X_i)^2\right)^{\frac{1}{2}} \leq \sqrt{nD}.$$

Finally,

$$\mathbb{E}\left[\sup_{f\in S_D, \|f\|_2^2\leq r}\sum_{i=1}^n \epsilon_i f(X_i)\right] \leq \sqrt{rnD}. \tag{5.68}$$

This concludes the proof. □

5.8.6 Peeling Device.

The purpose of claims 1 and 2 of the following result is to explain the choice of the considered weight function (claim 3).

Theorem 5.8.6 (Peeling Device). *Let ω denote a positive function (a weight) on the class of functions $\mathcal{F}$. We assume that there exists a sub-root function ϕ and a positive number r^* such that:*

$$\forall r\geq r^*,\ \mathbb{E}\sup_{f\in\mathcal{F}, \omega(f)\leq r}|(P-P_n)(f))| \leq \phi(r). \tag{5.69}$$

Then the following holds.

1. *For all α such that $1\geq\alpha>\frac{1}{2}$,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{(P-P_n)(f))}{r^{1-\alpha}\omega(f)^\alpha\vee r} \leq \frac{\phi(r)}{r}\left(1+\frac{(2\alpha)^{\frac{1}{2\alpha-1}}}{1-(2\alpha)^{-1}}\right).$$

Theorem 5.8.3 implies that the case of interest is $\alpha=\frac{1}{2}$. However,

$$\lim_{\alpha\to\frac{1}{2}}\left(1+\frac{(2\alpha)^{\frac{1}{2\alpha-1}}}{1-(2\alpha)^{-1}}\right) = +\infty.$$

2. *If $\phi(r) = cr^\beta$ with $\beta < \frac{1}{2}$, for all $x > 1$,*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{\sqrt{r\omega(f)} \vee r}\right) \le \frac{\phi(r)}{r}\left(1 + \frac{x^\beta}{1 - x^{\beta-\frac{1}{2}}}\right) .$$

Theorem 5.8.5 incites us to consider the case $\beta = \frac{1}{2}$.

However, $\lim_{\beta\to\frac{1}{2}}\left(1 + \frac{1}{1-x^{\beta-\frac{1}{2}}}\right) = +\infty$. This is the reason why we consider a boundedness condition on the weight.

3. *If $\forall f \in \mathcal{F}$, $\omega(f) \in [0, R]$ and Equation (5.69) is available $\forall\, r \in [r^*, R]$ then*

$$\forall r \ge r^*,\ \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{\sqrt{r\omega(f)} \vee r}\right) \le \frac{\phi(r)}{r}\left(1 + e + \frac{e}{2}\left(\log\frac{\mathrm{R}}{\mathrm{r}}\right)_+\right) .$$

Proof. We prove the three points together: in cases 1 and 2, $R = \infty$.
Case 3 when $r \ge R$ is clear since

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{\sqrt{r\omega(f)} \vee r}\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F},\omega(f)\le r} \frac{(P-P_n)(f)}{r}\right] \le \frac{\phi(R)}{r} \le \frac{\phi(r)}{r} .$$

We now assume that $r < R$. Let $x > 1$ be a real number to be specified later.

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{r^{1-\alpha}\omega(f)^\alpha \vee r}\right] &\le \mathbb{E}\left[\sup_{f\in\mathcal{F},\omega(f)\le r} \frac{|(P-P_n)(f)|}{r}\right] \\
&+ \sum_{k=0}^{N_\alpha} \mathbb{E}\left[\sup_{f\in\mathcal{F},rx^k\le\omega(f)\le rx^{k+1}} \frac{|(P-P_n)(f)|}{r^{1-\alpha}\omega(f)^\alpha}\right] ,
\end{aligned}$$

where $N_\alpha = \lfloor \log_x \frac{R}{r} \rfloor$ in case 3 and $N_\alpha = \infty$ in cases 1 and 2.
Thus,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{r^{1-\alpha}\omega(f)^\alpha \vee r}\right] &\le \frac{\phi(r)}{r} + \sum_{k=0}^{N_\alpha} \mathbb{E}\left[\sup_{f\in\mathcal{F},\omega(f)\le rx^{k+1}} \frac{|(P-P_n)(f)|}{rx^{\alpha k}}\right] \\
&\le \frac{1}{r}\left(\phi(r) + \sum_{k=0}^{N_\alpha} \frac{\phi(rx^{k+1})}{x^{\alpha k}}\right) . \qquad (5.70)
\end{aligned}$$

If $\phi(r) = cr^\beta$ with $\beta < \frac{1}{2}$ (case 2), then

$$\frac{1}{r}\left(\phi(r) + \sum_{k=0}^{N_\alpha} \frac{\phi(rx^{k+1})}{x^{\frac{k}{2}}}\right) = \frac{\phi(r)}{r}\left(1 + x^\beta \sum_{k\ge0} x^{k(\beta-\frac{1}{2})}\right) ,$$

and we obtain case 2 using Inequality (5.70).
In cases 1 and 3, since ϕ is a sub-root function, $\frac{\phi(rx^{k+1})}{\sqrt{rx^{k+1}}} \le \frac{\phi(r)}{\sqrt{r}}$.

It follows that $\phi(rx^{k+1}) \leq \phi(r)x^{\frac{k+1}{2}}$.
Consequently, Inequality (5.70) entails

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{r^{1-\alpha}\omega(f)^\alpha \vee r}\right] \leq \frac{\phi(r)}{r}\left(1+\sqrt{x}\sum_{k=0}^{N_\alpha} x^{k(\frac{1}{2}-\alpha)}\right).$$

In the case 3, this yields

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{\sqrt{r\omega(f)}\vee r}\right] \leq \frac{\phi(r)}{r}\left(1+\sqrt{x}(1+\log_x\frac{R}{r})\right).$$

We choose $x=e^2$ and this concludes the proof of case 3.
In the case 2, we obtain:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{(P-P_n)(f)}{r^{1-\alpha}\omega(f)^\alpha \vee r}\right] \leq \frac{\phi(r)}{r}\left(1+\frac{\sqrt{x}}{1-x^{\frac{1}{2}-\alpha}}\right).$$

The optimal choice is $x=(2\alpha)^{\frac{1}{\alpha-\frac{1}{2}}}$ and we finally get case 1. □

5.9 Appendix D: Average Bound .

Increasing the penalty with a logarithmic term, we obtain an average performance bound. In the following result, we consider $n \geq 55$.

Theorem 5.9.1. *Let $S_0=\{0\}$ and $K>1$. Let $\widehat{g}_D = \operatorname*{argmin}_{g\in S_{D,n}} \frac{1}{n}\sum_{i=1}^n (1-Y_i g(X_i))_+$ and*

$$\widehat{D} = \arg\min_{D\geq 0}\left(\frac{1}{n}\sum_{i=1}^n (1-Y_i\widehat{g}_D(X_i))_+ + \mathrm{pen}(D)\right).$$

We assume that Hypotheses $\mathbf{H}(\tau_D)$ and $\mathbf{H}(\mathbf{mar})$ are satisfied.
Then there exist constants C_K and $C_{K,1}$ such that the following holds: if $\forall D\geq 1$,

$$\mathrm{pen(D)} \geq \frac{\mathrm{C_{K,1} D \tau_D^2 \log^2 n}}{\mathrm{np_0^2}} + \frac{\mathrm{C_K \tau_D^2 \log(D\tau_D)}}{\mathrm{np_0^2}} \,\mathrm{and\ pen(0) = 0}\,,$$

then

$$\mathbb{E}L(\widetilde{g}, f^*) \leq K \inf_{D\geq 0}\left(\inf_{g\in S_{D,n}} L(g,f^*) + 2\mathrm{pen}(D)\right).$$

Proof of Theorem 5.9.1 . We follow the line of the proof of Theorem 5.6.2. The product $\tau_D^2 x$ is linearized by using the following Young's inequality:

$$ab \leq e^{\frac{9a}{10}} + \frac{10b}{9}\log(\frac{5b}{9}),$$

with $a=x$ and $b=\tau_D^2$. Combined with inequality (5.63), it yields

$$\begin{aligned}(C) \quad &\leq \frac{\kappa_1}{p_0}\left(\frac{\tau_D^2 D\log^2 n}{n} + \frac{\tau_{\widehat{D}}^2 \widehat{D}\log^2 n}{n}\right)\\ &+ \frac{\kappa_2}{np_0^2}\left(\tau_D^2 x_D + 2e^{\frac{9x}{10}} + \frac{10\tau_D^2}{9}\log(\frac{5\tau_D^2}{9}) + \tau_{\widehat{D}}^2 x_{\widehat{D}} + \frac{10\tau_{\widehat{D}}^2}{9}\log(\frac{5\tau_{\widehat{D}}^2}{9})\right).\end{aligned}$$

Consequently, if

$$\mathrm{pen}(D) \geq \frac{\kappa_1 (K+2)^2}{K-1} \frac{\tau_D^2 D \log^2 n}{p_0 n} + \frac{\kappa_2 (K+2)^2}{(K-1) n p_0^2} \left(\tau_D^2 x_D + \frac{10 \tau_D^2}{9} \log(\frac{5\tau_D^2}{9}) \right) ,$$

inequality (5.62) entails that with probability greater than $1-2e^{-x}$, $\forall D$, such that $\Gamma_K \frac{\sqrt{n}}{\log n} \geq D \geq 0$

$$L(\tilde{g}, f^*) \leq K \left(L(g_D, f^*) + 2\mathrm{pen}(D) + \frac{2\kappa_2 e^{\frac{9x}{10}}}{p_0^2 n} \right) . \tag{5.71}$$

In order to integrate this bound, we have to be careful that it is available only for $x \leq \frac{9}{10} \log^2(n)$. Let

$$G = n \left(L(\tilde{g}, f^*) - K \left(L(g_D, f^*) + 2\mathrm{pen}(D) \right) \right) .$$

An integration by parts implies that

$$\mathbb{E}G \leq \mathbb{E}G_+ = \int_0^\infty P[G > t] dt = \frac{9\kappa_2 K}{5 p_0^2} \int_{-\infty}^{+\infty} P[G > \frac{2\kappa_2 K e^{\frac{9x}{10}}}{p_0^2}] e^{\frac{9x}{10}} dx .$$

The integral is split up into three parts
$(I) = \int_{-\infty}^0 P[G > \frac{2\kappa_2 K e^{\frac{9x}{10}}}{p_0^2}] e^{\frac{9x}{10}} dx$; $(II) = \int_0^{\frac{9}{10}\log^2(n)} P[G > \frac{2\kappa_2 K e^{\frac{9x}{10}}}{p_0^2}] e^{\frac{9x}{10}} dx$ and $(III) = \int_{\frac{9}{10}\log^2(n)}^{+\infty} P[G > \frac{2\kappa_2 K e^{\frac{9x}{10}}}{p_0^2}] e^{\frac{9x}{10}} dx$. It is straightforward that $(I) \leq \int_{-\infty}^0 e^{\frac{9x}{10}} dx = \frac{10}{9}$. Moreover, inequality (5.71) yields $(II) \leq 2 \int_0^{\frac{9}{10}\log^2(n)} e^{\frac{-x}{10}} dx \leq 20$. Finally,

$$G \leq nL(\tilde{g}, f^*) \leq n\mathbb{E}|\tilde{g} - f^*| \leq n \|\tilde{g} - f^*\|_2 \leq n(1+n^2) \leq 2n^3 .$$

and, since $n \geq 55$, $x \geq \frac{9}{10}\log^2(n)$ implies that $x \geq \frac{10}{3}\log(n)$ and consequently that

$$\frac{2K\kappa_2 e^{\frac{9x}{10}}}{p_0^2} > 2n^3 .$$

This implies that $(III) = 0$. Summing the upper bounds on (I), (II) and (III), we obtain

$$\mathbb{E}G \leq \frac{\kappa_1 K}{p_0^2} ,$$

where $\kappa_1 = 135688$. Thus

$$\mathbb{E}L(\tilde{g}, f^*) \leq K \left(L(g_D, f^*) + 2\mathrm{pen}(D) \right) + \frac{\kappa_1 K}{n p_0^2} ,$$

and

$$\mathbb{E}L(\tilde{g}, f^*) \leq K \left(\inf_{\Gamma_K \frac{\sqrt{n}}{\log n} \geq D \geq 0} \left(L(g_D, f^*) + 2\mathrm{pen}(D) \right) + \frac{\kappa_1}{n p_0^2} \right) .$$

Following the reasoning of step 3 of the proof of Theorem 5.6.2, in order to consider the whole family of models, it suffices to consider

$$\mathrm{pen}(D) \geq \frac{(K+2)^4}{(K-1)^2} \frac{5441^2 \tau_D^2 D \log^2 n}{p_0^2 n} + \frac{(K+2)^2 \kappa_2}{(K-1) n p_0^2} \left(\tau_D^2 x_D + \frac{10\tau_D^2}{9} \log(\frac{5\tau_D^2}{9}) \right) .$$

Finally, $C_{K,1} = \frac{5441^2 (K+2)^4}{(K-1)^2}$ and $C_K = \frac{7000(K+2)^2}{K-1}$ are suitable for Theorem 5.9.1. $\square$

Chapter 6

Finite Dimensional Projection for Classification and Statistical Learning

This chapter is intended for submission. This is a joint work with G. Blanchard.

Contents

6.1 Introduction.

6.1.1 The Classification Framework.

In this chapter, we study the supervised binary classification problem. Let (X, Y) denote a random variable with values in $\mathcal{X} \times \{-1, +1\}$ distributed according to P. The marginal distribution of X is denoted by Q. Y is the *label* associated to the *input variable* X. We observe a set of n independent and identically distributed (i.i.d.) pairs $(X_i, Y_i)_{i=1}^n$ sampled according to P. These observations form the *training set*. We will suppose $n \geq 3$.

A *classifier* is a map f from $\mathcal{X}$ to $\{-1, +1\}$ that assigns to every element $x \in \mathcal{X}$ a prediction of its label. The quality of such a classifier is naturally measured by its *generalization error* $\mathbb{P}[f(X) \neq Y]$. Consequently, from a statistical point of view, classification aims at estimating the classifier of minimal generalization error f^* called *Bayes classifier* using only the training set. [DGL96] reported that $f^*(x) = 2\mathbb{1}_{\{x \in \mathcal{X}, p(x) > 1/2\}} - 1$ Q-a.s. on the set $\{p(x) \neq 1/2\}$ where $p(x) = \mathbb{P}[Y = 1|X = x]$.

A natural procedure (Empirical Risk Minimization of [Vap95]) is to find a classifier t minimizing the empirical classification error $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i) \neq Y_i}$: that is the reason why the *hard loss* $\ell(f,(x,y)) = \mathbb{1}_{f(x)\neq y}$ is the natural loss associated to a classification problem. For algorithmic reasons, numerous actual classification algorithms (e.g. SVM, boosting) replace this loss by a *convex* surrogate one γ over some real-valued function spaces. They minimize $\frac{1}{n}\sum_{i=1}^n \gamma(f,(X_i,Y_i))$ over a set of functions $\mathcal{F}$ and the sign of the obtained function leads to a classifier.

Overfitting is avoided by regularization that consists in constraining the obtained real values function to be smooth. More precisely, the renowned SVM uses Arsenin-Tikhonov's regularization (Tikhonov's regularization for short in the sequel) already investigated in a regression framework ([Wah90]): we aim at providing an alternative to this regularization procedure by considering finite dimensional projection. It has already been studied in numerous domains (see [BBM99] for a review and [Bar02] for the regression on a random design): in this chapter, we investigate it in the classification framework.

To begin with, the model selection aspect is studied from a statistical point of view using recent theoretical tools of M-estimation ([Mas00b],[BBM03]). Next, Kernel Principal Component Analysis is used to obtain the Kernel Projection Machine algorithm (KPM): the model selection used in this new algorithm is guided by the previous theoretical result. Moreover, its performances are comparable to those of SVM on the data set we have considered.

With some abuse of notations, depending on the setup, we may use the notation $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ or $\frac{1}{n}\sum_{i=1}^n f(X_i,Y_i)$ and similarly, $Pf = \mathbb{E}[f(X)]$ or $\mathbb{E}[f(X,Y)]$. For the remainder of this chapter, C denotes a universal constant.

6.1.2 The SVM Algorithm.

In this subsection, we recall briefly some points crucial to the Support Vector Machines (SVM). They have been introduced by [BGV92] and the SVM loss is the *hinge loss* $\gamma_h(g,(x,y)) = (1 - yg(x))_+$. The SVM loss is consistent in the sense that the Bayes classifier satisfies (see [Lin99])

$$f^* = \arg \min_{g \text{ measurable}} P\gamma_h(g)\,,$$

and

$$L_h(g,f^*) \geq P(Yg(X) \leq 0) - P(Yf^*(X) \leq 0)\,, \tag{6.1}$$

where $L_h(g,f^*) = P\gamma_h(g) - P\gamma_h(f^*)$ denotes the excess risk (also called *risk*) of g for the hinge loss function. This means that rate of convergence for the SVM-excess risk implies the same rate for the prediction error excess risk.

A *kernel function* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a symmetric and positive semi-definite function, in the following sense

$$\forall\, n, \forall\,(x_1,\ldots,x_n)\in\mathcal{X}^n, \forall\,(a_1,\ldots,a_n)\in\mathbb{R}^n\,,\ \sum_{i,j=1}^n a_i a_j k(x_i,x_j) \geq 0\,.$$

It can be proved (see e.g. [Jan97]) that such a function defines a unique Hilbert space $\mathcal{H}$ of functions such that $\forall x\in\mathcal{X}$, $k(x,.)\in\mathcal{H}$ and

$$\forall g\in\mathcal{H}, \langle g, k(x,\cdot)\rangle_{\mathcal{H}} = g(x)\,. \tag{6.2}$$

$\mathcal{H}$ is called the Reproducing Kernel Hilbert Space (RKHS for short in the sequel) associated to k.

In order to study its statistical performances, the (soft-margin) SVM algorithm can be summed up by formulating it as the minimization of a regularized functional of Tikhonov's style ([EPP00],[SS98]) :

$$\hat{g} = \arg\min_{g \in \mathcal{H}^b} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ + C\|g\|_{\mathcal{H}}^2 , \tag{6.3}$$

where $\mathcal{H}^b = \{g(x) + c,\ g \in \mathcal{H}, c \in \mathbb{R}\}$ and $\mathcal{H}$ denotes the RKHS associated to the kernel k. The SVM classifier is $\hat{f} = \text{sign}(\hat{g})$ where $\text{sign}(a) = 2\mathbb{1}_{a \geq 0} - 1$.

Now, it is straightforward that the optimization problem (6.3) can be written in the following way: $\hat{g} = \hat{g}_{\hat{R}}$ where

$$\hat{g}_R = \arg\min_{g \in \mathcal{E}(R)} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ ,$$

and

$$\widehat{R} = \arg\min_{R \geq 0} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \hat{g}_R(X_i))_+ + CR^2 \right) . \tag{6.4}$$

$\mathcal{E}(R)$ are "semi-norm balls" of radius R in $\mathcal{H}^b$. This gives rise to the interpretation of the above regularization as *model selection*, where the models are "semi-norm balls" in $\mathcal{H}^b$. The criterion (6.4) is used to select the ellipsoid: it corresponds to a *penalized minimization of the empirical loss*. The choice of the squared penalty CR^2 is essentially motivated by algorithmic reasons. [SS03] studied this squared penalization but it is natural to ask about the statistical relevance of this choice: [BBM04] have shown that from a statistical point of view, under margin assumption, this squared penalization is too heavy. It would be better to replace it by a linear penalty $C'R$.

At this stage, it is interesting to realize that the balls in the RKHS space can be viewed as ellipsoids in the original space $L_2(Q)$. We suppose that k is a Mercer's kernel, i.e., $\mathcal{X}$ is a compact metric space and k a continuous kernel. Let T_k be the following compact and self-adjoint kernel operator on $L_2(Q)$:

$$T_k : g \to \int_{\mathcal{X}} k(x, .) g(x) dQ(x) .$$

Let $(\Psi_i)_{i \geq 1}$ be an orthonormal basis of eigenfunctions corresponding to the non-increasing sequence of eigenvalues $(\lambda_i)_{i \geq 1}$. Mercer's Theorem allows to get a representation of $\mathcal{H}$ in terms of spectral quantities associated to T_k. More precisely,

$$\mathcal{H} = \left\{ f \in L_2(Q) \ : \ g = \sum_{i \geq 1} a_i \Psi_i \,\text{such that}\, \|g\|_{\mathcal{H}}^2 = \sum_{i \geq 1} \frac{a_i^2}{\lambda_i} < \infty \right\} . \tag{6.5}$$

The uniqueness of the RKHS associated to a kernel implies that it does not depend on the underlying measure Q. We have

$$\{g \in \mathcal{H}, \|g\|_{\mathcal{H}} \leq R\} = \left\{ \sum_{i \geq 1} a_i \Psi_i \, ; \, \sum_{i \geq 1} \frac{a_i^2}{\lambda_i} \leq R^2 \right\} .$$

Consequently, a ball of the RKHS is an ellipsoid of $L_2(P)$ with principal axis the eigenfunctions of T_k.

The finite dimensional projection is now presented in the framework of classification. Roughly speaking, keeping the minimization of the hinge loss, the ellipsoids involved in the collection of models of SVM are replaced by finite dimensional spaces.

6.1.3 The Finite Dimensional Approach.

Due to the approximation properties of finite dimensional spaces $S_D = \langle \Psi_1, \cdots, \Psi_D \rangle$, $D \in \mathbb{N}^*$ with respect to the ellipsoids, one can think of penalized finite dimensional projection as an alternative to Tikhonov's regularization. The following classifier is associated to each dimension by minimizing the SVM empirical risk over S_D:

$$\hat{f}_D = \arg\min_{f \in S_D} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ . \tag{6.6}$$

It proves meaningful to think about the eigenfunctions as a smooth basis: the fewer the number of coefficients, the smoother the function.

Non-compactness of vector spaces added to the unboundedness of the hinge loss yields questions about the relevance of $\hat{f}_D$. On the one hand, from a practical point of view, this linear optimization problem could lead to some instability that does not occurs in the SVM case since the ellipsoids are compact spaces. On the other hand, analysis of the statistical properties of $\hat{f}_D$ raises technical difficulties relying on boundedness problems. They occur in the use of concentration inequalities and of complexity control: the classical *global* notions of complexity (i.e. taking into account the whole space S_D) are not relevant due the non-compactness of S_D.

However, it has been shown recently ([Mas00b],[BBM03]) that, in order to obtain fast rate of convergence, one has to analyze the behavior of the empirical minimizer on balls which are compact since S_D is a finite dimensional space. Consequently, the analysis of the statistical behavior of $\hat{f}_D$ would fully exploit the *local* notions of complexity by considering the balls of S_D. This idea has been investigated in appendices of chapter 5. However, these results were not entirely satisfactory. By shrinking the estimators $\hat{f}_D$, the following procedure allows to get more easily interpretable bounds.

In order to make up for the unboundedness of the hinge loss, the function "clip" is used :

$$\mathrm{clip}(g(x)) = \begin{cases} 1 & \text{if } g(x) \geq 1 \\ g(x) & \text{if } -1 < g(x) < 1 \\ -1 & \text{if } g(x) \leq -1 , \end{cases}$$

It has been already considered in [GKKW02] in a regression framework. The estimator $\widetilde{f}_D = \mathrm{clip}(\hat{f}_D)$ is associated to each dimension. This yields no loss of generality since $\hat{f}_D$ and $\widetilde{f}_D$ correspond to the same classifier.

The model selection step is performed by penalized minimization of the clipped empirical loss. The final classifier is $\mathrm{sign}(\widetilde{f}_{\widehat{D}})$ where

$$\widehat{D} = \arg\min_{D \geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{f}_D(X_i))_+ + \lambda D \right) , \tag{6.7}$$

and the constant λ has to be suitably chosen. As explained in the sequel, the linear penalty λD is chosen for its statistical properties.

In section 6.2.2, S_D is described in terms of the eigenfunctions of the covariance operator C_1 associated to $k(X,.)$. This leads to an interpretation of the selected dimension as an optimal dimension in a dimensionality reduction method (KPCA explained therein) towards classification. Moreover, this point of view will be useful for the effective algorithm (KPM) since the operator C_1 has a natural empirical counterpart.

To conclude, this work highlights by two ways that the finite dimensional projection is a credible alternative to the Tikhonov's regularization used in SVM.

1. It provides a statistical analysis of finite dimensional regularization in the framework of clipped empirical risk minimization.
2. This regularization principle is used to design a new algorithm of classification called Kernel Projection Machine (KPM) which is partially justified by the previous theoretical study. The performances of KPM and SVM are similar on the considered data sets.

The chapter is organized as follows. Section 6.2 presents a statistical study of finite dimensional regularization in the classification framework by stating Theorem 6.2.1. Section 6.3 describes precisely the KPM algorithm and section 6.4 provides some numerical experiments. Proofs are relegated in the appendix.

6.2 Main Result.

We are now ready to formulate our main theorem: it aims at studying the model selection procedure involved in the finite dimensional regularization.

Theorem 6.2.1. *Let $(S_D)_{D\geq 1}$ be a family of linear subspaces of $L_2(P)$ where S_D is of dimension at most D and $\hat{f}_D$ denotes the minimizer of the empirical risk:*

$$\hat{f}_D = \arg\min_{f\in S_D} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ \,. \tag{6.8}$$

Let p be defined as $p(x) = P[Y = 1|X = x]$. We suppose that the following margin condition holds:

$$\exists\, h_0 > 0,\, \forall x \in \mathcal{X},\, \left|p(x) - \frac{1}{2}\right| \geq h_0 \,.$$

Let $K > 1$ and $\widetilde{f}_D = \mathrm{clip}(\hat{f}_D)$. The dimension is chosen by

$$\widehat{D} = \arg\min_{D\geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{f}_D(X_i))_+ + \mathrm{pen}_n(D) \right) \,. \tag{6.9}$$

Then, there exist universal constants C_1 and C_2 such that the following holds. If

$$\forall\, D \geq 1,\, \mathrm{pen}_n(D) \geq \frac{C_1 K}{h_0} \frac{D \log n}{n} \,,$$

then

$$\mathbb{E}\left[L_h(\widetilde{f}_{\widehat{D}}, f^*) \right] \leq \frac{K}{K-1} \left(\inf_{D\geq 1} \left(\inf_{f\in S_D} L_h(f, f^*) + 2\mathbb{E}[\mathrm{pen}_n(D)] \right) \right) + \frac{C_2 K}{h_0 n} \,. \tag{6.10}$$

Section 6.6.3 shows that this result is a consequence of a more general one (Theorem 6.6.2 stated and proved in appendix A). Some comments are now provided.

- This result means that the penalty λD linear with the dimension used in the criterion (6.7) is statistically convenient. However, even if inequality (6.10) is optimized over K, the constants obtained by this theoretical study are too coarse to be used in practice.
- Its main drawback is the dependence of the penalty upon the unknown margin parameter h_0. Results of [NM03] (inspired by [MT95, Tsy04]) states that this parameter is unavoidable. However, this leads to high difficulties to choose the constant λ in front of the dimension D in criterion (6.7). The study of technics calibrating the penalization constant is outside the scope of the present chapter though it is an interesting future direction of work.
- Because of inequality (6.1), inequality (6.10) provides an upper bound of the prediction error excess risk. However, the approximation error $\inf_{f\in S_D} L_h(f, f^*)$ has to be controlled to get a rate of convergence. This is also an interesting future direction.
- In order to get a linear optimization problem, the empirical risk is minimized over S_D and not $\text{clip}(S_D) = \{\text{clip}(f),\ f \in S_D\}$. Consequently, the oracle inequality satisfied by the linearly penalized estimator $\widetilde{f}_{\widehat{D}}$ calls for the approximation error of the linear space S_D.
- The result provides a bound on the expected excess risk. In order to get a deviation bound, an hypothesis on the infinity norm of the functions of S_D is necessary. For the sake of clarity, we only state the average bound.
- Considering $S_D = \langle \Psi_1, \dots, \Psi_D \rangle$ in the finite dimensional procedure, the kernel is used only through the eigenfunctions of the kernel operator T_k whereas SVM uses eigenvalues as well.

6.2.1 Comparison with the Risk Bound of SVM.

This section is devoted to exhibit a simple situation where the finite dimensional regularization gets a better rate of convergence than the SVM one's. Let us consider a one-dimensional situation where X is a $[0,1]$-valued random function. Let k be the following kernel:

$$k(x, y) = 1 + 2 \sum_{j\geq 1} \lambda_j \left(\cos(2\pi jx)\cos(2\pi jy) + \sin(2\pi jx)\sin(2\pi jy)\right),$$

where, for $j \geq 2$, $\lambda_j = (2\pi j)^{-2\gamma}$ with γ a positive integer. In this case, $\{\Psi_i\}_{i\geq 1}$ is the trigonometric basis of $L_2([0,1])$ endowed with the Lebesgue's measure.

It is known (see e.g. chapter 2 of [Wah90]) that, if γ is an integer, the RKHS associated with this kernel is the Sobolev space of order γ with periodic boundary conditions $H^{(\gamma)}_{\text{per}}$. Precisely, $H^{(\gamma)}_{\text{per}}$ is the set of functions of $L_2([0,1])$ with $\gamma - 1$ continuous derivatives satisfying $f(0) = f(1), \cdots, f^{(\gamma-1)}(0) = f^{(\gamma-1)}(0)$, and with $f^{(\gamma)} \in L_2([0,1])$. It is endowed with the Sobolev norm $\|f\|^2_{\text{Sob}} = \int_0^1 |f^{(\gamma)}(t)|^2 dt + \left(\int_0^1 f(t)dt\right)^2$. By Parseval Theorem $\|f\|^2_{\text{Sob}} = \sum_{j\geq 1} (f_j^{(\gamma),c})^2 + (f_j^{(\gamma),s})^2 + \left(\int_0^1 f(t)dt\right)^2$ where $f_j^c = \sqrt{2}\int_0^1 f(t)\cos(2\pi jt)dt$ and $f_j^s = \sqrt{2}\int_0^1 f(t)\sin(2\pi jt)dt$ are the Fourier coefficients of f. Gathering integrations by parts and the periodic boundary conditions leads to $\|f\|^2_{\text{Sob}} = \sum_{j\geq 1}(2\pi j)^{2\gamma}\left((f_j^c)^2 + (f_j^s)^2\right) + \left(\int_0^1 f(t)dt\right)^2 = \|f\|^2_{\mathcal{H}_k}$ where the last equality follows from the definition of the RKHS norm in equation (6.5).

In order to control the approximation error, we suppose that:

$$f^* \in \mathcal{F}_\alpha = \{f \in L_2([0,1]),\, ci^{-\alpha} \leq |f_j^c| \leq Ci^{-\alpha} \, and \, ci^{-\alpha} \leq |f_j^s| \leq Ci^{-\alpha}\}\,,$$

where C and c are universal constants. $\|f^*\|_{L_2([0,1])} = 1$ implies $\alpha > 1/2$. Moreover, non-continuity of f^* avoids normal convergence of its Fourier series: that's why we suppose $\alpha \leq 1$. Finally, spaces $\mathcal{F}_\alpha$ are considered for $\alpha \in]1/2, 1]$. The order of magnitude of the risk bound obtained in [BBM04] for the SVM classifier $\hat{f}$ is

$$\mathbb{E}[L_h(\hat{f}, f^*)] \leq C \inf_{g\in\mathcal{H}_k} \left(L_h(g, f^*) + \Lambda_n \|g\|^2_{\mathcal{H}}\right)\,,$$

where $\Lambda_n \sim n^{-\frac{4\gamma}{4\gamma+2}}$ if $\lambda_j \sim j^{-2\gamma}$ for $\gamma \geq 1$. Using that the hinge loss is 1-Lipschitz, this bound ensures

$$\mathbb{E}[L_h(\hat{f}, f^*)] \leq C \inf_{g\in\mathcal{H}_k} \left(d(g, f^*) + \Lambda_n \|g\|^2_{\mathcal{H}_k}\right)\,,$$

and Theorem 6.2.1 yields

$$\mathbb{E}[L_h(\widetilde{f}, f^*)] \leq C \inf_{D\geq 1} \left(\inf_{g\in S_D} d(g, f^*) + \frac{D}{n}\right)\,,$$

where $\widetilde{f}$ is a shortcut for $\widetilde{f}_{\widehat{D}}$ and $d(g, f^*) = \|g - f^*\|_2$.

Straightforwardly, Theorem 6.2.1 implies

$$\sup_{f^*\in\mathcal{F}_\alpha} \mathbb{E}[L_h(\widetilde{f}, f^*)] \leq Cn^{-\frac{2\alpha-1}{2\alpha+1}}\,.$$

Moreover, calculations sketched in Appendix 6.9 yield

$$\inf_{f^*\in\mathcal{F}_\alpha} \inf_{g\in H^{(\gamma)}_{\text{per}}} \left(d(g, f^*) + \Lambda_n \|g\|^2_{\mathcal{H}}\right) \geq Cn^{-\frac{2(2\alpha-1)\gamma}{(2\gamma+1)(4\gamma-2\alpha+1)}}\,. \tag{6.11}$$

We easily check that $\frac{2\alpha-1}{2\alpha+1} > \frac{2(2\alpha-1)\gamma}{(2\gamma+1)(4\gamma-2\alpha+1)}$. Consequently, in this case, the finite dimensional bound entails better rate of convergence uniformly over $\mathcal{F}_\alpha$.

Arguably, we could use the minimizer of the empirical risk over any finite dimensional models since we do not make any assumption on the vector spaces. However, it would be natural to choose $S_D = \langle \Psi_1, \ldots, \Psi_D \rangle$. Since the eigenfunctions $(\Psi_i)_{i\geq 1}$ of the kernel operator T_k depend on the underlying probability of the data, in order to use the finite dimensional regularization in practice, it is necessary to design finite dimensional spaces approximating S_D from the data. [Kol98] notices that the empirical counterpart of T_k is the kernel matrix $K_{1,n} = \frac{1}{n}(k(X_i, X_j))_{1\leq i,j\leq n}$. The KPM algorithm is formulated by this way in chapter 5. It is worth noticing that the kernel operator T_k acts on $L_2(Q)$ whereas its empirical counterpart $K_{1,n}$ acts on $\mathbb{R}^n$: in this case, functions are approximated by vectors.

The KPM algorithm is now presented in the following section by formulating differently the empirical approximation of S_D. A dimensionality reduction method (KPCA) is exploited to construct data dependent models. On the one hand, they approximate S_D and on the other hand, they fit to the structure of the input data by maximizing the variance.

6.2.2 Link with KPCA using Kernel Models.

Let C_1 denote the covariance operator of $k(X,.)$. It is defined by its quadratic form on $\mathcal{H}$:

$$\forall f,g \in \mathcal{H}, \langle f, C_1 g\rangle_{\mathcal{H}} = \mathbb{E}\left[f(X)g(X)\right]. \tag{6.12}$$

In order to define it, integrability conditions are imposed on the kernel: $\mathbb{E}\left[k(X,X)\right] < \infty$ is enough (see e.g. chapter 3 for more details). Moreover, the proof of Theorem 3.2.3 implies that the integral operator T_k and the covariance operator C_1 associated to the kernel k have the same non-zero eigenvalues and, above all, the same eigenspaces: $\forall \lambda \neq 0$, $E_\lambda(T_k) = E_\lambda(C_1)$ where $E_\lambda(A) = \{x, Ax = \lambda x\}$ for a linear operator A. Consequently,

$$S_D = \langle \phi_1, \cdots, \phi_D \rangle , \tag{6.13}$$

where $(\phi_i)_{i\geq 1}$ denotes an orthonormal basis of $\mathcal{H}$ of eigenfunctions of C_1.

The covariance associated to a kernel allows to get models taking into account possibly non-linear relevant structure of the input data. Moreover, this characterization of S_D links it with the Kernel Principal Component Analysis (KPCA for short in the sequel): KPCA is nothing more but the diagonalization of C_1 (it is a particular case of Principal Component Analysis of a Hilbert values random variable, first defined in [DPR82]). Consequently, S_D gets the optimality properties of KPCA: in particular, this space of dimension D is catching most of the variance of the data.

A natural question raised in [STWCK05] and in chapter 3 as an open problem is: what is the best dimension representing the data? Indeed, there exists no clear criterion to choose an optimal dimension considering only the input data. Roughly speaking, in order to keep most of the information of the data, the best choice would be not to project the data at all but to keep the whole space. Consequently, the selection of a dimension D by criterion (6.9) in the finite dimensional projection method amounts to selecting the optimal D-dimensional representation of our data for the classification task. In other words, to extracting the information that is needed for this task by model selection taking into account the relevance of the directions for the classification task.

6.2.3 Advantages of The Finite Dimensional Regularization.

The finite dimensional approach allows to easily compare the performances of different kernels without involving the perturbations due to model selection. Indeed, a family of computable classifiers is associated with each kernel: it consists in the minimizers of the empirical risk over each finite dimensional space. In order to compare the kernels, it suffices to compare the corresponding families.

The main advantage is that this procedure enables to easily use *simultaneously* different kernels by considering finite dimensional spaces spanned by eigenfunctions of covariance operator associated to different kernels.

As explained in the following section, the KPM algorithm is implemented by considering empirical models in the penalized minimization of the clipped empirical loss.

6.3 Kernel Projection Machine.

In order to get models adapted to the underlying law of the input data, let us consider an approximation of the spaces S_D. Due to its definition (6.12), it is natural to approach the theoretical covariance operator C_1 by its empirical counterpart: the empirical covariance operator $C_{1,n}$ defined by

$$\forall f, g \in \mathcal{H}, \langle f, C_{1,n} g\rangle = \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i)\,.$$

The following approximation is considered:

$$S_D = \langle \phi_1, \cdots, \phi_D\rangle \sim \langle \hat{\phi}_1, \cdots, \hat{\phi}_D\rangle\,, \tag{6.14}$$

where $(\hat{\phi}_i)_{i\geq 1}$ denotes an orthonormal basis of $\mathcal{H}$ of eigenfunctions associated to the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$ of the empirical covariance operator. It is called Nyström approximation ([Bak77]). Results of [Kol98, DPR82] and chapter 4 ensures that it is reasonable to approach S_D by such a way. Moreover, it is numerically illustrated on figure 6.1.

The following aims at giving a detailed account of the obtained algorithm, called Kernel Projection Machine.

The first step consists in computing the empirical minimizers over $\langle \mathbb{1}, \hat{\phi}_1, \cdots, \hat{\phi}_D\rangle$:

$$\hat{f}_D = \arg \min_{f\in\langle \mathbb{1}, \hat{\phi}_1,\cdots,\hat{\phi}_D\rangle} \sum_{i=1}^{n}(1 - Y_i f(X_i))_+\,. \tag{6.15}$$

Note that the constant function $\mathbb{1}$ equal to 1 is systematically added to the models in order to take into account translation on the data: this function corresponds to the threshold c in the SVM algorithm. This optimization problem is a linear programming one since it consists in the minimization of a convex function with linear constraints.

Besides, the data-dependent functions $\hat{\phi}_i$ are eigenfunctions of the linear map $C_{1,n}$ which acts in high dimension: typically, $\mathcal{H}$ is of infinite dimension. Thus, from an algorithmic point of view, it could be tricky to obtain these functions. However, KPCA is precisely the diagonalization of the operator $C_{1,n}$. The very convenient fact, precisely explained in [SSM96] for example, used to perform it efficiency is to note that it suffices to diagonalize the $n \times n$ kernel Gram matrix $K_{1,n} = \frac{1}{n}(k(X_i, X_j))_{1\leq i,j\leq n}$ to obtain the eigenfunctions of $C_{1,n}$. Indeed, for $j \geq 1$ such that $\hat{\lambda}_j > 0$,

$$\hat{\phi}_j = \frac{1}{\sqrt{\hat{\lambda}_j n}}\sum_{i=1}^{n} V_j^{(i)} k(X_i, .)\,, \tag{6.16}$$

where $(V_j)_{1\leq j\leq n}$ is an orthonormal basis of eigenvectors of $K_{1,n}$ associated to the eigenvalues $(\hat{\lambda}_j)_{1\leq j\leq n}$ sorted in decreasing order. The normalization ensures $\|\hat{\phi}_j\|_{\mathcal{H}} = 1$. Considering Equation (6.15), we have $\hat{f}_D = \sum_{j=1}^{D} \frac{\gamma_j^*}{\sqrt{\hat{\lambda}_j}}\hat{\phi}_j + b^*$ where

$$(\gamma^*, b^*) = \arg \min_{\gamma\in\mathbb{R}^D, b\in\mathbb{R}} \sum_{i=1}^{n}\left(1 - Y_i\left(\sum_{j=1}^{D}\frac{\gamma_j}{\sqrt{\hat{\lambda}_j}}\hat{\phi}_j(X_i) + b\right)\right)_+\,.$$

However, Equation (6.16) leads to $\hat{\phi}_j(X_i) = \sqrt{n\hat{\lambda}_j}V_j^{(i)}$ and to an expansion of $\hat{f}_D$ with respect to the simplest functions $k(X_i, .)$.

To conclude, the KPM algorithm can be summed up as follows:

1. given data $X_1, \dots, X_n \in \mathcal{X}$ and a positive kernel k defined on $\mathcal{X}\times\mathcal{X}$, compute the kernel matrix $K_{1,n}$ and its eigenvectors $V_1, \dots, V_n$ together with its eigenvalues in decreasing order $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n$.

2. for each dimension D such that $\hat{\lambda}_D > 0$ solve the linear optimization problem

$$(\gamma^*, b^*) = \arg\min_{\gamma,b,\xi} \sum_{i=1}^{n} \xi_i \tag{6.17}$$

$$\text{under constraints } \forall i = 1 \dots n, \xi_i \geq 0 \text{ , } Y_i \left(\sum_{j=1}^{D} \sqrt{n}\gamma_j V_j^{(i)} + b \right) \geq 1 - \xi_i \text{ .} \tag{6.18}$$

Next, compute $\alpha^* = \displaystyle\sum_{j=1}^{D} \frac{\gamma_j^*}{\sqrt{n}\hat{\lambda}_j} V_j$ and $\hat{f}_D = \sum_{i=1}^{n} \alpha_i^* k(x_i, .) + b^*$.

3. The last step is a model selection problem consisting in choosing the dimension D by an adequate procedure. Theorem 6.2.1 suggests the following criterion:

$$\widetilde{\hat{D}} = \arg\min_{D \geq 1} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \widetilde{\hat{f}}_D(X_i))_+ + \lambda D \right) \text{ .} \tag{6.19}$$

where $\widetilde{\hat{f}}_D = \text{clip}(\hat{f}_D)$. λ has to be chosen by an adequate procedure. The classifier obtained by the KPM is $\text{sign}(\hat{f}_{\widetilde{\hat{D}}})$.

6.4 Numerical Results.

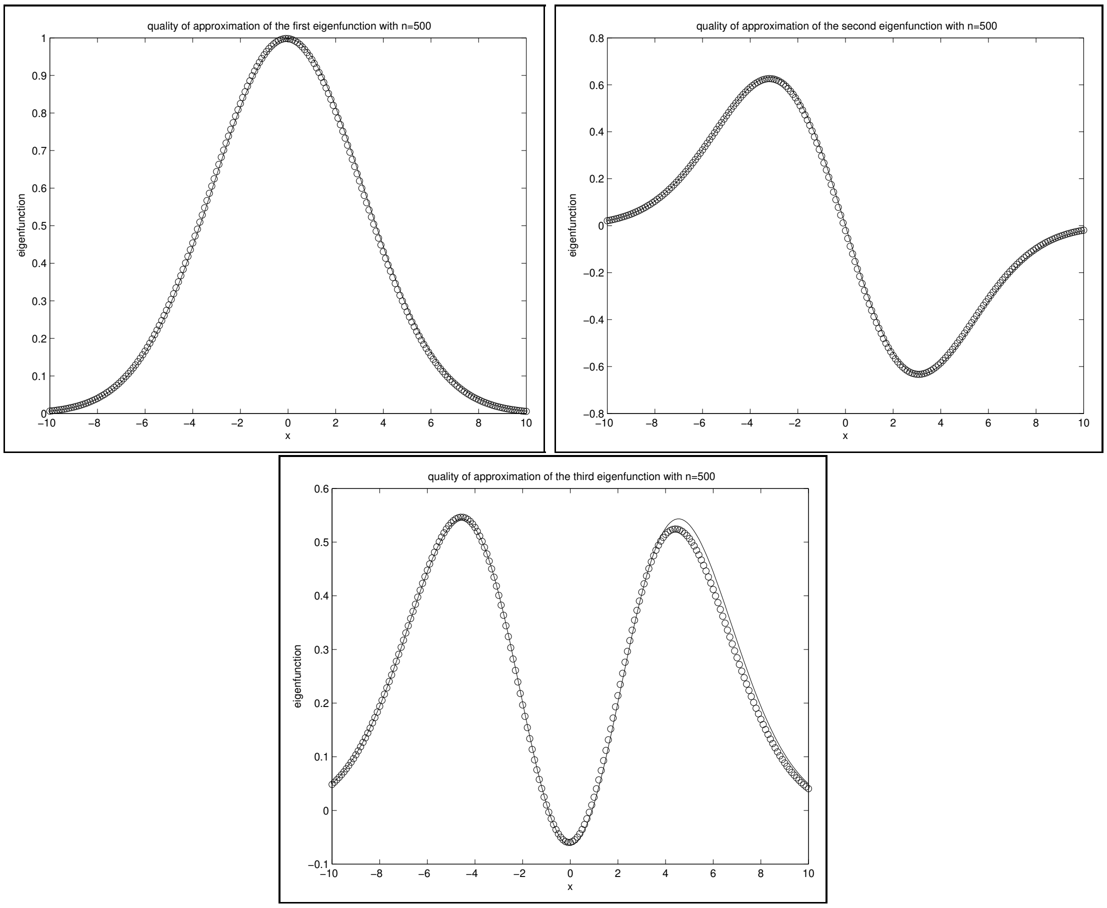
Two types of experiments are presented. The first one consists in investigating graphically the Nyström approximation. As for the second, it compares the performances of KPM and SVM. To begin with, the family of classifiers of the two algorithms are likened without selecting the regularization parameter (denoted by C in equation (6.3) for the SVM and λ for the KPM). Indeed, the involved method could advantage one or the other algorithm. Next, the regularization parameters are chosen by different ways.

6.4.1 Numerical Experiments for the Nyström Approximation.

The KPM algorithm relies on the Nyström approximation formulated in equation (6.14). In order to illustrate it, a gaussian kernel is considered along with a gaussian underlying law. In this case, Theorem 6.10.1 recalled in appendix 6.10 gives explicitly the eigenfunctions; they are of multiplicity 1. In order to compare two functions of norm 1 of $\mathcal{H}$, $\sqrt{\lambda_i}\Psi_i$ and $\hat{\phi}_i$ are drawn. Indeed, the definition (6.12) of C_1 implies $\mathbb{E}\left[\Psi_i^2(X)\right] = \langle \Psi_i, C_1 \Psi_i \rangle_{\mathcal{H}}$ and $\Psi_i \in E_{\lambda_i}(T_k) = E_{\lambda_i}(C_1)$ yields $\|\Psi_i\|_2^2 = \lambda_i \|\Psi_i\|_{\mathcal{H}}^2$. The empirical eigenfunctions are computed by using formula (6.16) and a random draw of 500 points. Figure 6.1 is obtained with $a = \frac{1}{4}$ and $b = \frac{1}{18}$ where b determines the width of the gaussian kernel $k(x, y) = e^{-b(x-y)^2}$ and a the gaussian law $dP(x) = \frac{1}{\sqrt{2\pi}} e^{-2ax^2} dx$ of X.

The straight line (resp. the circle) represents the theoretical (resp. empirical) eigenfunction. The empirical eigenfunctions corresponding to the two largest eigenvalues fits exactly the true one whereas the empirical eigenfunction corresponding to the third eigenvalue does not. Consequently, these graphics highlight a consequence of some results of [Kol98]: the accuracy of Nyström approximation decreases with the eigenvalues suggesting that the approximation of the theoretical eigenfunctions by the empirical ones is convenient at least for large eigenvalues i.e. for the first eigenfunctions.

Figure 6.1: From the left to the right: approximation of the first, the second and the third eigenfunction of the covariance operator in the gaussian case of Theorem 6.10.1.



6.4.2 Numerical Results for the KPM Algorithm.

The KPM was implemented in Matlab using the free library GLPK for solving the linear optimization problem. Since the algorithm involves the eigendecomposition of the kernel matrix, only small datasets have been considered for the moment. The Matlab implementation of the KPM algorithm is available at http://www.lri.fr/~vert.

It has been tested on benchmark datasets available on Gunnar Rätsch's web site coming from the UCI repository ([Rae99]). Several state-of-the-art algorithms have already been applied to those datasets among which the SVM with parameters C_G and σ chosen by cross-validation and depending on the datasets. In all experiments presented here, the gaussian kernel $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ is used with the σ proposed by [Rae99]. All datasets consist of 100 samples, each sample being split into a training sample and a test sample.

Without selecting the regularization parameter, the family of classifiers obtained by the KPM algorithm is $(\widetilde{\hat{f}}_D)_{D\geq 1}$ and the one obtained by the SVM algorithm is $(\hat{f}_C)_{C\in\mathcal{C}}$. The regularization constants contained in $\mathcal{C}$ are tailored to the data set by forming a geometric sequence of 101 points running over $C_G/100$ to $100\,C_G$ and containing C_G.

The first results aim at comparing the two families by shunting off the model selection procedure.

Comparison of the families of SVM's classifiers and KPM's classifiers. In table 6.1, the goal is to compare the approximation powers of the two families of classifiers. For each sample, the smallest test error of KPM (w.r.t. parameter D) is compared with the smallest test error of SVM (w.r.t. parameter C). Each time the winner is given one point.

Table 6.1: Approximation Powers

	SVM	**KPM**
Banana	67	31
Breast Cancer	50	44
Diabetis	42	55
Flare Solar	63	19
German	49	43
Heart	64	27

Selection of the parameter by cross-validation. Table 6.2 presents results of SVM (resp. KPM) where the regularization parameters C (resp. λ) is chosen by 5-fold cross-validation on each of the samples. Concerning the KPM, the test errors have been obtained with a grid of λ made up of 5001 points regularly spaced out from 0 to 0.5. The results are presented in the form { mean of the 100 test errors } $\pm$ { variance of the 100 test errors }.

Table 6.2: Test errors

	σ	**SVM**	**KPM**
Banana	0.7071	10.69 ± 0.67	10.91 ± 0.57
Breast Cancer	5	26.68 ± 5.23	28.73 ± 4.42
Diabetis	3.1623	23.79 ± 2.01	23.77 ± 1.69
Flare Solar	3.8730	32.62 ± 1.86	32.52 ± 1.78
German	5.2440	23.79 ± 2.12	24.09 ± 2.38
Heart	7.7460	16.23 ± 3.18	17.35 ± 3.54

These two tables highlight that the performances of KPM are comparable with those of SVM: considering table 6.1, SVM seems to have a certain advantage on KPM but table 6.2 shows that the average differences are quite small. Moreover, it is worth noticing that the same parameter σ was used for KPM and SVM whereas it is tailored only to SVM.

The cross-validation requires the choice of a grid of λ. Contrary to the SVM, the computation time of the KPM used with cross-validation is almost independent of the size of this grid. Indeed, the longest task is to compute the various classifiers. Then, it suffices to minimize a vector of small size to get the selected dimension corresponding to each λ. Consequently, in this case, the finite dimensional procedure allows to save computation time.

Results obtained in tables 5.4 and 6.2 concerning the KPM are similar. This remark highlights that the cross-validation makes up for some possible sub-optimality of the order of magnitude of the chosen penalty. Consequently, another way of choosing λ from the data in the KPM (see equation (6.7)) is now provided: the slope heuristic. Contrary to cross-validation, it relies on the choice of a linear penalty. No theoretical result justifies this method but it has been already studied in different frameworks. In particular, it has proved its practical efficiency for the detection of multiple change points ([Leb02]). It is inspired by the studies of trees in [BFOS84].

6.4.3 Slope Heuristic for the Kernel Projection Machine.

To begin with, the slope heuristic relies on natural ideas provided in the next sections where the vector spaces are supposed to be deterministic. Next, some numerical results are presented.

Conjectures and Comments.

Let D_{opt} be the dimension corresponding to the best classification error of the family $\{\widetilde{f}_D\}_{D\geq 1}$:

$$D_{\text{opt}} = \arg\min_{D\geq 1} \mathbb{P}[Y \neq \text{sign}(\widetilde{f}_D(X))] .$$

Any dimension selection procedure aims at estimating D_{opt}. Let

$$D_1 = \arg\min_{D\geq 1} P\gamma(\widetilde{f}_D) . \tag{6.20}$$

Moreover, $f_D = \arg\min_{f\in S_D} P\gamma(f)$ and $g_D = \text{clip}(\text{f}_\text{D})$. Let's recalled the definition of $\widehat{D}$

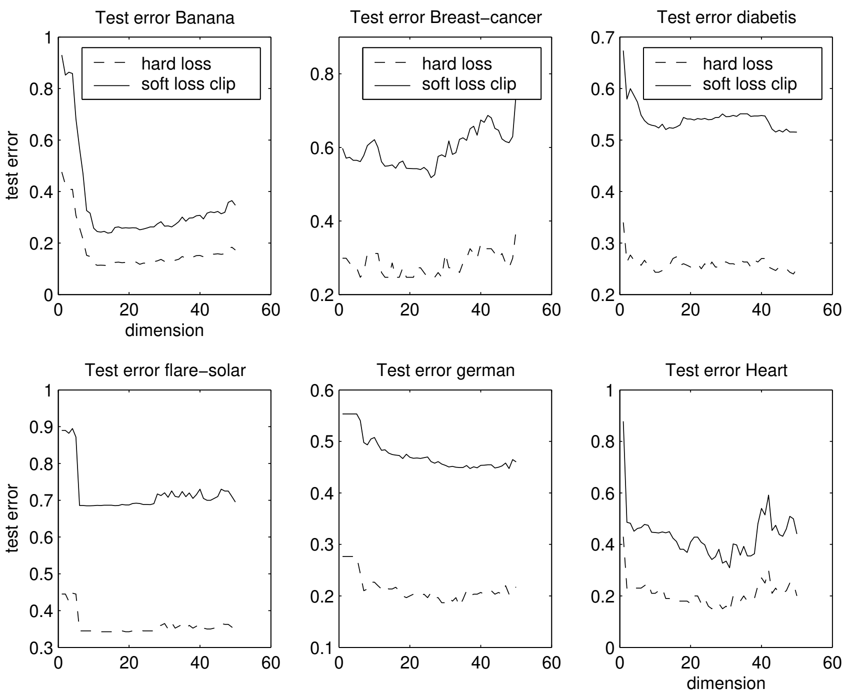
$$\widehat{D} = \arg\min_{D\geq 1} \left(P_n\gamma(\widetilde{f}_D) + \text{pen}(D)\right) , \tag{6.21}$$

where $\hat{f}_D = \arg\min_{f\in S_D} P_n\gamma(f)$ and $\widetilde{f}_D = \text{clip}(\hat{\text{f}}_\text{D})$. The slope heuristic relies on three conjectures.

- Conjecture 1: $D_{opt} \sim D_1$.
- Conjecture 2: $P_n(\gamma(g_D) - \gamma(\widetilde{f}_D)) \sim P(\gamma(\widetilde{f}_D) - \gamma(g_D))$.
- Conjecture 3: $P_n(\gamma(g_D) - \gamma(\widetilde{f}_D))$ behaves like an affine function (with respect to D) for reasonable dimensions.

Some comments are now provided to explain why these conjectures sound reasonable. Conjecture 1 is due to the convexification of the risk. Figure 6.2 represents the test error associated with the hard loss and with the soft-clipped loss. They behave similarly and have approximatively the same minimizer. In view of this figure, the considered data satisfies Conjecture 1. Conjecture 2 is natural since the first term corresponds to the second by swapping the true probability measure P with the empirical one P_n (even in the definition of the minimizers $\hat{f}_D$ and f_D). The study of this statement requires powerful tools of concentration. Conjecture 3 is the consequence of two facts. On the one hand, by definition of g_D, for large dimensions, $P_n(\gamma(g_D))$ becomes almost independent of the dimension. On the other hand, the straight line of the left part of figure 6.3 represents the empirical risk $P_n(\gamma(\widetilde{f}_D))$ with respect to D. We can observe that for reasonable dimensions ($6 \leq D \leq 35$), the decrease is almost linear.

Figure 6.2: Influence of the Loss Function on the Test Error



The Slope Heuristic.

The choice of the penalty function *pen* recommended by the slope heuristic is now derived from the previous conjectures. The approach relies on the classical trade-off approximation error-estimation error. The penalty has to be chosen in such a way that it leads to a selected dimension $\widehat{D}$ as close as possible to D_{opt}. However, for computational reasons, the penalty function acts on the empirical risk computed for the soft loss function (see (6.21)). Consequently, we can only estimate D_1 from the data. Conjecture 1 claims that it achieves the right goal. By the definition (6.20) of D_1, we get

$$D_1 = \arg\min_{D \geq 1} \left\{ P(\gamma(g_D) - \gamma(f^*)) + P(\gamma(\widetilde{f}_D) - \gamma(g_D)) \right\} .$$

Moreover, the definition (6.21) of $\widehat{D}$ yields

$$\widehat{D} = \arg\min_{D\geq 1} \left\{ P_n(\gamma(g_D) - \gamma(f^*)) - P_n(\gamma(g_D) - \gamma(\widetilde{f}_D)) + \mathrm{pen}(D) \right\} .$$

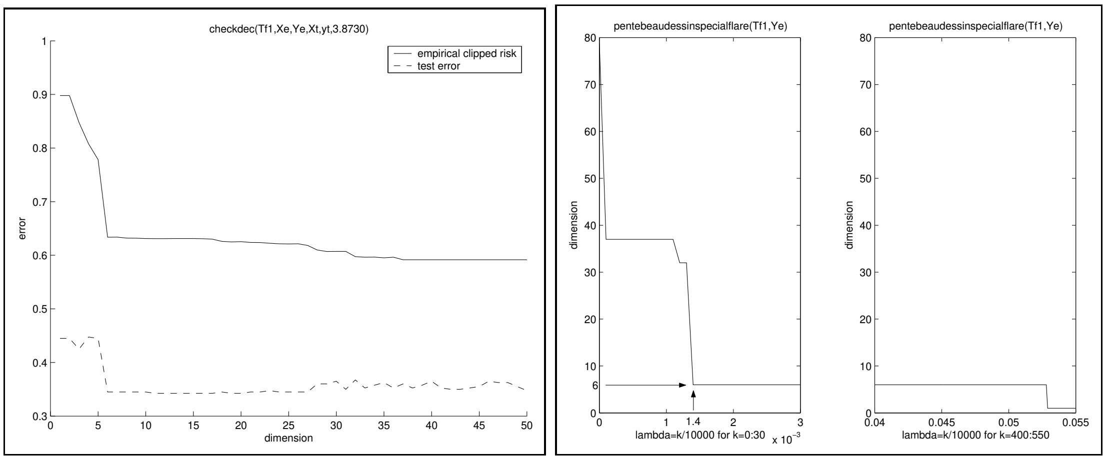
Since g_D is a deterministic function, $P(\gamma(g_D) - \gamma(f^*))$ is conveniently estimated by $P_n(\gamma(g_D) - \gamma(f^*))$. Consequently, in order to obtain the optimal dimension D_1, the ideal penalty would be:

$$\mathrm{pen}(D) = P_n(\gamma(g_D) - \gamma(\widetilde{f}_D)) + P(\gamma(\widetilde{f}_D) - \gamma(g_D)) .$$

Conjecture 2 claims that the two terms have the same order. Moreover, Theorem 6.2.1 incites us to choose a linear penalty $\mathrm{pen}(D) = \lambda D$. In view of conjecture 3, once we compute the slope $\hat{m}$ of the empirical risk for reasonable dimensions, the slope heuristic recommends $\lambda = -2\hat{m}$.

The computation of $\hat{m}$ relies on the behavior of the empirical risk. The three rates of decrease of the empirical risk observed on the straight line of the left part of figure 6.3 are typical. For small dimensions, $(1 \leq D \leq 6)$ the empirical risk decreases quickly. Then, for reasonable ones $(6 \leq D \leq 35)$, the rate becomes almost linear. Finally, for large dimensions $(D \geq 35)$, the empirical risk is almost constant. In order to detect them, the right part of figure 6.3 draws the selected dimension $\widehat{D}$ with respect to λ. By definition (6.7), the three rates of decrease of the empirical risk correspond to the three jumps of dimension observed in the right part of figure 6.3. For $\lambda = 0$, $\widehat{D}$ is very large (around 80 in figure 6.3). Then, a small value of λ compensates for the third rate (almost constant) of decrease of the empirical risk: the corresponding selected dimension drops to 35. The selected dimension is then almost constant until λ is large enough to compensate for the second (almost linear) rate of decrease. We observe that λ is large enough by noting that the selected dimension drops again to a smaller dimension ($D = 6$ in figure 6.3). The corresponding value (1.4 in figure 6.3) is the slope $\hat{m}$.

Figure 6.3: **Left:** The straight line is the empirical risk for the clipped soft loss function $P_n\gamma(\widetilde{f}_D)$ and the dashed line is the test error. Both on 'flare-solar' **Right:** Selected dimension $\widehat{D}$ versus λ using (6.21).



Numerical Results for the Slope Heuristic.

In order to assess the performances of the slope heuristic, experiments were carried out on the benchmark datasets [Rae99]. It is worth noticing that the corresponding algorithm does not restrict λ to belong to a grid a priori defined. In order to compute the slope, the following procedure was used. To begin with, one easily shows that the possible dimensions $\widehat{D}$ selected by the criterion (6.7) have the following explicit expression:

$$\widehat{D}_0 = 1 \, and \, \widehat{D}_{i+1} = \arg\max_{D>\widehat{D}_i} \frac{J(\widehat{D}_i) - J(D)}{D - \widehat{D}_i},$$

where $J(D) = P_n\gamma(\widetilde{f}_D)$. From now on, the considered dimensions are bounded from above by a maximal dimension $Dmax$. $Dmax$ is chosen a priori to avoid the third rate of decrease (almost constant) of the empirical risk. As suggested by figure 6.3, the differences $\widehat{D}_{i+1} - \widehat{D}_i$ are computed for $i \geq 0$ and the maximum is reached at $i = \widehat{N}$. Consequently, the slope $\hat{m}$ is $\max_{D>\widehat{D}_{\widehat{N}}} \frac{J(\widehat{D}_{\widehat{N}}) - J(D)}{\widehat{D}_{\widehat{N}} - D}$.

The following table presents the obtained results when the dimension is chosen thanks to criterion (6.7) with $\lambda = -2\hat{m}$ and $\hat{m} = \max_{D>\widehat{D}_{\widehat{N}}} \frac{J(\widehat{D}_{\widehat{N}}) - J(D)}{\widehat{D}_{\widehat{N}} - D}$. λ is computed on each of the 100 training samples. The test errors are similar to the one obtained using a cross-validation

Table 6.3: Test errors of KPM with the slope heuristic

	σ	$Dmax$	**Test errors of KPM**
Banana	0.7071	50	10.89±0.53
Breast Cancer	5	50	28.40±4.43
Diabetis	3.1623	50	23.70±1.68
Flare Solar	3.8730	32	32.60±1.82
German	5.2440	50	23.88±2.27
Heart	7.7460	32	17.62±3.59

(see table 6.2). The main drawback of this procedure lies in the dependence on the choice of $Dmax$.

To conclude, the slope heuristic provides a data dependent procedure to perform the model selection. It is adaptive to the margin and, contrary to the re-sampling methods, e.g., cross-validation, it considers *simultaneously* all the training data. It is worth noticing that the computation time for the slope heuristic is shorter than for cross-validation since the classifiers are computed only one time (from the whole training data) whereas 5-fold cross-validation requires calulation of 5 families of classifiers.

6.5 Conclusion and Discussion.

Highlight of our Work. We described the finite dimensional approach in the classification framework and designed an effective algorithm: the KPM. The model selection is tackled using the penalized minimization of the clipped empirical loss. The linear penalty used in the KPM algorithm is partially theoretically justified by Theorem 6.2.1. All the presented results highlight that

> *regularization can be performed thanks to a dimensionality reduction method such as Kernel-PCA.*

Consequently, the finite dimensional projection is a credible alternative to the Tikhonov's regularization used in the SVM algorithm. A nice property is that the training labels are used to select the "optimal" dimension: optimal means that the resulting D-dimensional representation of the data contains the right amount of information needed to classify the inputs. To sum up, the KPM can be seen as a dimensionality-reduction-based classification method that takes into account the labels for the dimensionality reduction step.

For the numerical results, two calibration methods of the penalty have been used in the KPM algorithm: the cross-validation and the slope heuristic. The slope heuristic allows to exploit the linear penalty obtained in Theorem 6.2.1 but it requires the tricky choice of a parameter $Dmax$. Its performances are comparable with those of cross-validation. Moreover, it saves computation time.

6.5.1 Comparison with Previous Works.

[Mas04] provides a precise mathematical comparison of finite dimensional and Tikhonov's regularization in the simpler framework of gaussian regression with fixed design. It shows that, in the case where the regularization constants C and λ are deterministic, from a minmax point of view, the squared norm regularization method is beaten by the so-called finite dimensional projection method (this gaussian intuition is summarized in chapter 5). Interestingly, [Mas04] provides a penalty of order of magnitude $\frac{D}{n}$ along with lower bounds witnessing its optimality. Consequently, except for the margin parameter h_0 witnessing the higher difficulty to deal with a classification problem rather than with a regression one, Theorem 6.2.1 provides the same order of magnitude.

[TG04] considers the same estimators $\hat{f}_D = \sum_{i=1}^{D} \alpha_i \phi_i$ with models S_D spanned by the D first functions of a basis of $L_2(Q)$. However, they use a L_1-penalty (depending on $\sum_{i=1}^{D} |\alpha_i|$) whereas we study a L_0-penalty (depending on the number of coefficients). They show that their procedure is adaptive under boundedness conditions on the involved basis: roughly speaking, we replace these assumptions by the use of the clip function. This leads to a theoretical result without any assumption on the infinity norm of the basis functions.

6.5.2 Discussion.

Theorem 6.2.1 offers a partial justification for the linear penalty used in the KPM algorithm: regarding min-max results in the framework of regression ([Mas04]) or prediction error of classification on a VC-class ([NM03]) convince us of the consistency of our result. However, we do not provide lower bounds ensuring optimality of the obtained oracle inequality (6.10).

Theorem 6.2.1 would specify the data dependent models used in the KPM algorithm: as far as we know, only few research have been undertaken on non-asymptotic studies of the closeness of empirical eigenfunctions and the true one ([Kol98] provides asymptotic results). We plan to extend results of chapter 4 to obtain oracle inequalities taking into account the specificities of the models used in the KPM algorithm.

The main drawback of Theorem 6.2.1 is the lack of adaptability to the margin: the penalty function involves the unknown margin parameter. This leads to high difficulties to calibrate the penalty in practice. We plan to study further the slope heuristic.

The main advantage of the presented finite dimensional projection with respect to SVM is that it allows to easily mix different kernels by considering finite dimensional spaces spanned by eigenfunctions of covariance operator associated to different kernels. Oracle inequalities can be obtained in this case using the same methodology (it suffices to change accordingly the weight x_D appearing in the proof of Theorem 6.2.1). To avoid additional technicalities, in this chapter, only the simplest version involving one model for each dimension is stated. However, this remark provides an interesting future direction.

The principle of finite dimensional regularization could be easily use to take into account the unlabeled data in a semi-supervised learning problem. We plan to investigate studies about this topic.

Acknowledgments

The authors are extremely grateful to P. Massart for giving us the original idea of the algorithm as well as for invaluable comments and to R. Vert and E. Lebarbier for their precious help.

Appendix

6.6 Appendix A: Risk Bounds for the Finite Dimensional Approach.

In this appendix, Theorem 6.2.1 is proved considering a more generic setting. The training data $((X_1, Y_1), \dots, (X_n, Y_n))$ belong to $(\mathcal{X} \times \mathcal{Y})^n$. Moreover, even in the classification setting, there are no special reason for clipping the empirical minimizer at 1 as stated for more clarity in Theorem 6.2.1. We thus define for $M > 0$:

$$\mathrm{clip}_M(g(x)) = \begin{cases} M & \text{if } g(x) \geq M \\ g(x) & \text{if } -M < g(x) < M \\ -M & \text{if } g(x) \leq -M \,. \end{cases}$$

This generality is motivated by the analysis of the influence of the clipped step on results of M-estimation given in [Mas00b] and [BBM03] and, particularly, how it allows to relax boundedness conditions. To do this, a simple risk bound for the clipped minimizer of the empirical risk over a fixed finite dimensional space is provided, together with an oracle inequality for model selection. The following notations will be used:

$$\mathrm{clip}_M(S_D) = \{\mathrm{clip}_M(f),\, f \in S_D\}\,,$$

and the star-hull of a class of functions $\mathcal{F}$ is defined by:

$$\mathrm{star}(\mathcal{F}) = \{\lambda f,\, 0 \leq \lambda \leq 1\,,\, f \in \mathcal{F}\}\,.$$

6.6.1 Clipped Empirical Risk Minimization on One Model.

In this section, we obtain a risk bound for a clipped empirical error minimizer $\widetilde{s}_D$ over a fixed vector space S_D of dimension at most D with a generic loss function γ:

$$\hat{s}_D = \arg\min_{t \in S_D} \frac{1}{n} \sum_{i=1}^{n} \gamma(t, (X_i, Y_i)) ,$$

and

$$\widetilde{s}_D = \mathrm{clip}_M(\hat{s}_D) .$$

The goal is to estimate the target function s minimizing the average loss over a "large" class $S \supset S_D$:

$$s = \arg\min_{t \in S} \mathbb{E}[\gamma(t, (X, Y))] .$$

The excess risk with respect to γ is:

$$L(g, s) = \mathbb{E}\left[\gamma(g, (X, Y))\right] - \mathbb{E}\left[\gamma(s, (X, Y))\right] .$$

In the sequel, s is supposed to take values in $[-M, M]$. If $M \geq 1$, it is true in the classification framework where $S = L_2(P)$ and $s = f^*$.

All the results will be stated under the following assumption on the loss which we will refer to as "assumption **(A)**" in the sequel:

(A1) *$\forall y \in \mathcal{Y}$, $\gamma(y, .)$ is B-Lipschitz.*

(A2) *$\forall t \in S$, $\forall y \in \mathcal{Y}$, $\forall x \in \mathcal{X}$,*

$$\gamma(y, t(x)) \geq \gamma(y, \mathrm{clip}_M(t(x))) .$$

The next result provides a risk bound when the vector space is fixed: its main advantage is the simplicity of the proof. With some additional technicalities, the same methodology offers a selection of model procedure (Theorem 6.6.2 below). Proofs of Theorems 6.6.1 and 6.6.2 use common material provided in the sequel.

Theorem 6.6.1. *Let S_D be a vector space of dimension at most D. We assume that*

- *The target function is bounded: $|s(x)| \leq M$.*
- *γ satisfies assumption (**A**).*
- *$\forall t \in \mathrm{clip_M(S_D)}$,*

$$\|\gamma(t) - \gamma(s)\|_2^2 \leq \kappa L(t, s) , \tag{6.22}$$

  where $\kappa = \kappa(M)$.

Then, for all $K > 1$, the following inequality holds:

$$\mathbb{E}\left[L(\widetilde{s}_D, s)\right] \leq \frac{K+1}{K-1} \left(\inf_{t \in S_D} L(t, s) + C_3 K \kappa' \frac{D}{n} \log n \right) + C_4 \frac{\kappa' K}{n} ,$$

where $\kappa' = \kappa \vee 1$, $C_3 = C_3(M, B)$ and $C_4 = C_4(M, B)$ depend only on M and B.

Proof of Theorem 6.6.1 . Let $s_D \in S_D$ be such that

$$s_D = \arg\min_{t \in S_D} P\gamma(f) ,$$

and $s_D^c = \mathrm{clip}_M(s_D)$. (If the infimum is not reached, one can use a dominated convergence argument to get the desired bound). To begin with,

$$P_n(\gamma(\widetilde{s}_D) - \gamma(s)) \leq P_n(\gamma(\hat{s}_D) - \gamma(s)) \leq P_n(\gamma(s_D) - \gamma(s)) ,$$

where the first inequality follows from the second condition of assumption (**A**) and the last one from the definition of the empirical minimizer. Thus

$$L(\widetilde{s}_D, s) \leq (P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s_D^c)) + (P - P_n)(\gamma(s_D^c) - \gamma(s)) + P_n(\gamma(s_D) - \gamma(s)) . \quad (6.23)$$

For the first term, we gather Theorems 6.7.1 and 6.8.3 below to the class of functions $\mathcal{F} = \mathrm{star}(\{\gamma(t) - \gamma(s_D^c), t \in \mathrm{clip}_M(S_D)\})$ to get that $\forall K > \frac{\kappa'}{7}$, $\forall \xi > 0$ with probability at least $1 - 3e^{-\xi}$, $\forall t \in \mathrm{clip}(S_D)$,

$$(P - P_n)(\gamma(t) - \gamma(s_D^c)) \leq \frac{1}{K}\|\gamma(t) - \gamma(s_D^c)\|_2^2 + \frac{E_1 K(2BM \vee 1)^2}{\kappa'^2}\hat{r}^*_{D,n} + \frac{E_2 K\xi(2BM \vee 1)^2}{n} ,$$

where $\kappa' = \kappa \vee 1$, $\hat{r}^*_{D,n}$ is defined in Theorem 6.8.3 and E_1, E_2 are universal constants. Since this bound holds simultaneously over $t \in \mathrm{clip}_M(S_D)$, we use it for the random function $t = \widetilde{s}_D$. Condition (6.22) will be useful to relate the squared norm back to the risk:

$$\begin{aligned}
\|\gamma(\widetilde{s}_D) - \gamma(s_D^c)\|_2^2 &\leq 2(\|\gamma(\widetilde{s}_D) - \gamma(s)\|_2^2 + \|\gamma(s) - \gamma(s_D^c)\|_2^2) \\
&\leq 2\kappa(L(\widetilde{s}_D, s) + L(s_D^c, s)) \\
&\leq 2\kappa(L(\widetilde{s}_D, s) + L(s_D, s)) ,
\end{aligned}$$

where the last inequality comes from assumption (**A2**). Finally, $\forall K > \frac{\kappa'}{7}$, $\forall \xi > 0$ with probability at least $1 - 3e^{-\xi}$,

$$\begin{aligned}
(P - P_n)(\gamma(\widetilde{s}_D) - \gamma(s_D^c)) \leq \frac{2\kappa'}{K}(L(\widetilde{s}_D, s) + L(s_D, s)) + \frac{E_1 K(2BM \vee 1)^2}{\kappa'^2}\hat{r}^*_{D,n} \\
+ \frac{E_2 K\xi(2BM \vee 1)^2}{n} .
\end{aligned}$$

Gathering this with inequality (6.23) yields that $\forall K > \frac{\kappa'}{7}$, with probability at least $1 - 3e^{-\xi}$,

$$\begin{aligned}
\left(1 - \frac{2\kappa'}{K}\right) L(\widetilde{s}_D, s) \leq \frac{2\kappa'}{K}L(s_D, s) + \frac{E_1 K(2BM \vee 1)^2}{\kappa'^2}\hat{r}^*_{D,n} + P_n(\gamma(s_D) - \gamma(s)) \\
+ (P - P_n)(\gamma(s_D^c) - \gamma(s)) + \frac{\xi E_2 K(2BM \vee 1)^2}{n} .
\end{aligned}$$

Integrating with respect to the sample yields that $\forall K > \frac{\kappa'}{7}$,

$$\begin{aligned}
(1 - \frac{2\kappa'}{K})\mathbb{E}[L(\widetilde{s}_D, s)] \leq \left(\frac{2\kappa'}{K} + 1\right) L(s_D, s) + \frac{E_1 K(2BM \vee 1)^2}{\kappa'^2}\mathbb{E}[\hat{r}^*_{D,n}] \\
+ \frac{3E_2 K(2BM \vee 1)^2}{n} ,
\end{aligned}$$

since $\mathbb{E}\left[(P-P_n)(\gamma(s_D^c)-\gamma(s))\right]=0$. Solving this inequality for $K>2\kappa'$ and setting $K'=\frac{K}{2\kappa'}$, for $K'>1$,

$$\mathbb{E}\left[L(\widetilde{s}_D,s)\right]\leq\frac{K'+1}{K'-1}\left(L(s_D,s)+\frac{2E_1K'(2BM\vee 1)^2}{\kappa'}\mathbb{E}[\hat{r}^*_{D,n}]\right)+C_4\frac{\kappa'K'}{n}.$$

This concludes the proof of Theorem 6.6.1 with $C_3=240000\times A_1(M,B)(2BM\vee 1)^2$ and $C_4=120672\times(2BM\vee 1)^2$. □

6.6.2 Model Selection.

A quite general result about penalized minimization of the clipped empirical loss over finite dimensional vector spaces is now presented.

Theorem 6.6.2. *Let $\{S_D\}_{D\geq 1}$ be a collection of vector spaces such that $\dim(S_D)\leq D$. We assume that*

- *The target function is bounded: $|s(x)|\leq M$.*
- *γ satisfies assumption (**A**).*
- *$\forall t\in\mathrm{clip_M(S_D)}$,*

$$\|\gamma(t)-\gamma(s)\|_2^2\leq\kappa L(t,s)\,, \tag{6.24}$$

where $\kappa=\kappa(M)$.

Let $K>1$. Choosing the dimension with the following penalized criterion

$$\widehat{D}=\arg\min_{D\geq 1}\left(\frac{1}{n}\sum_{i=1}^{n}\gamma(\widetilde{s}_D,(X_i,Y_i))+\mathrm{pen}_n(D)\right),$$

with a possibly data dependent penalty function pen_n such that

$$\forall\, D\geq 1\,,\,\mathrm{pen}_n(D)\geq C_5K\kappa'\frac{D}{n}\log n\,, \tag{6.25}$$

where $\kappa'=\kappa\vee 1$, the following inequality holds

$$\mathbb{E}[L(\widetilde{s}_{\widehat{D}},s)]\leq\frac{K}{K-1}\left(\inf_{D\geq 1}\left(\inf_{t\in S_D}L(t,s)+2\mathbb{E}[\mathrm{pen}_n(D)]\right)+\frac{C_6K\kappa'}{n}\right),$$

where $C_5=C_5(M,B)$ and $C_6=C_6(M,B)$ depend only on M and B.

Proof. Let $s_D=\arg\min_{t\in S_D}P\gamma(t)$. The definitions leads to the following chain of inequalities: $\forall D\geq 1$,

$$P_n\gamma(\widetilde{s}_{\widehat{D}})+\mathrm{pen}_n(\widehat{D})\leq P_n\gamma(\widetilde{s}_D)+\mathrm{pen}_n(D)\leq P_n\gamma(\hat{s}_D)+\mathrm{pen}_n(D)\leq P_n\gamma(s_D)+\mathrm{pen}_n(D)\,,$$

where the second inequality is due to assumption **(A2)**. Thus, $\forall D\geq 1$

$$\begin{aligned}L(\widetilde{s}_{\widehat{D}},s) &= (P-P_n)(\gamma(\widetilde{s}_{\widehat{D}})-\gamma(s))+P_n(\gamma(\widetilde{s}_{\widehat{D}})-\gamma(s))\\ &\leq (P-P_n)(\gamma(\widetilde{s}_{\widehat{D}})-\gamma(s))+P_n(\gamma(s_D)-\gamma(s))+\mathrm{pen}_n(D)-\mathrm{pen}_n(\widehat{D})\,.\end{aligned} \tag{6.26}$$

Let $D' \geq 1$ and $t_{D'} = \arg \min_{t \in \mathrm{clip}_M(S_{D'})} \|\gamma(t) - \gamma(s)\|^2$. (If the infimum is not reached, one can use a dominated convergence argument). Let us consider the following decomposition:

$$(P - P_n)(\gamma(\widetilde{s}_{D'}) - \gamma(s)) = (P - P_n)(\gamma(\widetilde{s}_{D'}) - \gamma(t_{D'})) + (P - P_n)(\gamma(t_{D'}) - \gamma(s)) \,. \quad (6.27)$$

Let x_D be such that $\sum_{D \geq 1} e^{-x_D} \leq 1$. Gathering Theorems 6.7.1 and 6.8.3 below with $\mathcal{F} = \mathrm{star}(\{\gamma(t) - \gamma(t_{D'}), t \in \mathrm{clip}_M(S_{D'})\})$ as in the proof of Theorem 6.6.1, we obtain that $\forall K > \frac{\kappa'}{7}$, with probability at least $1 - 3e^{-\xi - x_{D'}}$, $\forall t \in \mathrm{clip}(S_{D'})$,

$$\begin{aligned}(P - P_n)(\gamma(t) - \gamma(t_{D'})) \leq \frac{1}{K}\|\gamma(t) - \gamma(t_{D'})\|^2 + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{D',n} \\ + \frac{E_2 K (\xi + x_{D'})(2BM \vee 1)^2}{n} \,.\end{aligned}$$

where $\hat{r}^*_{D',n}$ is defined in Theorem 6.8.3 and E_1, E_2 are universal constants. Since the previous bound holds simultaneously for all $t \in \mathrm{clip}(S_{D'})$, we can use it for the random choice $t = \widetilde{s}_{D'}$. This leads to

$$\begin{aligned}(P - P_n)(\gamma(\widetilde{s}_{D'}) - \gamma(t_{D'})) \leq \frac{1}{K}\|\gamma(\widetilde{s}_{D'}) - \gamma(t_{D'})\|^2 + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{D',n} \\ + \frac{E_2 K (\xi + x_{D'})(2BM \vee 1)^2}{n} \,,\end{aligned}$$

with probability at least $1 - 3e^{-\xi - x_{D'}}$.

Condition (6.24) will be useful to relate the term involving the squared norm back to the risk:

$$\begin{aligned}\|\gamma(\widetilde{s}_{D'}) - \gamma(t_{D'})\|^2 &\leq 2(\|\gamma(\widetilde{s}_{D'}) - \gamma(s)\|^2 + \|\gamma(s) - \gamma(t_{D'})\|^2) \\ &\leq 4\|\gamma(\widetilde{s}_{D'}) - \gamma(s)\|^2 \\ &\leq 4\kappa L(\widetilde{s}_{D'}, s) \,.\end{aligned}$$

Thus with probability at least $1 - 3e^{-\xi - x_{D'}}$,

$$\begin{aligned}(P - P_n)(\gamma(\widetilde{s}_{D'}) - \gamma(t_{D'})) \leq \frac{4\kappa}{K} L(\widetilde{s}_{D'}, s) + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{D',n} \\ + \frac{E_2 K (\xi + x_{D'})(2BM \vee 1)^2}{n} \,. \quad (6.28)\end{aligned}$$

As in the proof of Theorem 6.6.1, Bernstein's inequality yields that with probability greater than $1 - e^{-\xi}$,

$$\begin{aligned}(P - P_n)(\gamma(t_{D'}) - \gamma(s)) &\leq \sqrt{\frac{2\xi \mathrm{Var}(\gamma(t_{D'}) - \gamma(s))}{n}} + \frac{2BM\xi}{3n} \\ &\leq \frac{1}{K}\|\gamma(t_{D'}) - \gamma(s))\|_2^2 + \left(\frac{2BM}{3} + \frac{K}{2}\right)\frac{\xi}{n} \,.\end{aligned}$$

The definition of $t_{D'}$ implies

$$(P - P_n)(\gamma(t_{D'})) - \gamma(s)) \leq \frac{1}{K}\|\gamma(\widetilde{s}_{D'}) - \gamma(s)\|_2^2 + \left(\frac{2BM}{3} + \frac{K}{2}\right)\frac{\xi}{n}.$$

Finally, condition (6.24) implies that with probability greater than $1-e^{-\xi}$,

$$(P-P_n)(\gamma(t_{D'}))-\gamma(s)) \leq \frac{\kappa}{K} L(\tilde{s}_{D'}, s) + \left(\frac{2BM}{3}+\frac{K}{2}\right)\frac{\xi}{n}. \tag{6.29}$$

Combining inequalities (6.29) and (6.28) in (6.27), yields that $\forall K > \frac{\kappa'}{7}$, $\forall D' \geq 1$, with probability at least $1-4e^{-\xi-x_{D'}}$,

$$(P-P_n)(\gamma(\tilde{s}_{D'})-\gamma(s)) \leq \frac{5\kappa}{K} L(\tilde{s}_{D'}, s) + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{D',n} + W_1 \frac{K(2BM\vee 1)^2(\xi + x_{D'})}{n},$$

where $W_1 = 20114$. We now use an union bound to obtain the previous inequality simultaneously for all $D' \geq 1$ and apply it with $D' = \widehat{D}$: $\forall K > \frac{\kappa'}{7}$, with probability $1-4e^{-\xi}$,

$$(P-P_n)(\gamma(\tilde{s}_{\widehat{D}})-\gamma(s)) \leq \frac{5\kappa}{K} L(\tilde{s}_{\widehat{D}}, s) + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{\widehat{D},n} + W_1 \frac{K(2BM\vee 1)^2(\xi + x_{\widehat{D}})}{n}.$$

Plugging this inequality into equation (6.26) yields that $\forall K > \frac{\kappa'}{7}$, with probability at least $1-4e^{-\xi}$, $\forall D \geq 1$,

$$\begin{aligned} L(\tilde{s}_{\widehat{D}}, s) \leq \frac{5\kappa'}{K} L(\tilde{s}_{\widehat{D}}, s) + \frac{E_1 K (2BM \vee 1)^2}{\kappa'^2} \hat{r}^*_{\widehat{D},n} + W_1 \frac{K(2BM\vee 1)^2(\xi + x_{\widehat{D}})}{n} \\ + P_n(\gamma(s_D)-\gamma(s)) + \mathrm{pen}_n(D) - \mathrm{pen}_n(\widehat{D}). \end{aligned}$$

Choosing $K' = \frac{K}{5\kappa'}$ leads to: $\forall K' > \frac{1}{35}$, with probability at least $1-4e^{-\xi}$, $\forall D \geq 1$,

$$\begin{aligned} L(\tilde{s}_{\widehat{D}}, s) \leq \frac{1}{K'} L(\tilde{s}_{\widehat{D}}, s) + \frac{5E_1 K' (2BM \vee 1)^2}{\kappa'} \hat{r}^*_{\widehat{D},n} + 5W_1 \frac{K'\kappa'(2BM\vee 1)^2(\xi + x_{\widehat{D}})}{n} \\ + P_n(\gamma(s_D)-\gamma(s)) + \mathrm{pen}_n(D) - \mathrm{pen}_n(\widehat{D}). \end{aligned}$$

Solving this inequality for $K' > 1$ leads to:

$$\begin{aligned} \frac{K'-1}{K'} L(\tilde{s}_{\widehat{D}}, s) \leq \frac{5E_1 K' (2BM \vee 1)^2}{\kappa'} \hat{r}^*_{\widehat{D},n} + \frac{5W_1 K'\kappa'(2BM\vee 1)^2(\xi + x_{\widehat{D}})}{n} \\ + P_n(\gamma(s_D)-\gamma(s)) + \mathrm{pen}_n(D) - \mathrm{pen}_n(\widehat{D}), \end{aligned}$$

By choosing $x_D = 2\log(D+1)$, condition (6.25) implies that, with probability greater than $1-4e^{-\xi}$, $\forall D \geq 1$,

$$L(\tilde{s}_{\widehat{D}}, s) \leq \frac{K'}{K'-1}\left(\frac{5W_1K'\kappa'(2BM\vee 1)^2}{n}\xi + P_n(\gamma(s_D)-\gamma(s)) + \mathrm{pen}_n(D)\right),$$

and

$$L(\tilde{s}_{\widehat{D}}, s) \leq \frac{K'}{K'-1}\left(\frac{5W_1K'\kappa'(2BM\vee 1)^2}{n}\xi + P_n(\gamma(s_D)-\gamma(s)) + 2\mathrm{pen}_n(D)\right).$$

Taking the infimum over $D \geq 1$ and integrating with respect to the sample entails:

$$\mathbb{E}L(\tilde{s}_{\widehat{D}}, s) \leq \frac{K'}{K'-1}\left(\mathbb{E}[\inf_{D\geq 1}(P_n(\gamma(s_D)-\gamma(s)) + 2\mathrm{pen}_n(D))] + \frac{20W_1K'\kappa'(2BM\vee 1)^2}{n}\right),$$

Finally,

$$\mathbb{E}L(\tilde{s}_{\widehat{D}}, s) \leq \frac{K'}{K'-1}\left(\inf_{D\geq 1} L(s_D, s) + 2\mathbb{E}[\mathrm{pen}_n(D))] + \frac{20W_1K'\kappa'(2BM\vee 1)^2}{n}\right).$$

This concludes the proof of Theorem 6.6.2 with $C_5(M,B) = 702256(2BM\vee 1)^2 A_1(M,B)$ and $C_6(M,B) = 402256(2BM\vee 1)^2$. □

6.6.3 Application to Classification.

As used all along the proofs, the clip function amounts to dealing with a complexity involving a bounded class of functions since it amounts to working with a bounded loss γ. This is explicit for the hinge loss γ_h since for $M \geq 1$,

$$\gamma_h(\text{clip}_M(g), (X_i, Y_i)) = \gamma_h(g, (X_i, Y_i)) \wedge (M+1)\,.$$

This last truncation has already been considered in [Bou02a, Theorem 8.3] and [BM02, Theorem 21] to obtain global risk bounds. Interestingly, this clipping has a clear interpretation in terms of classification: $\text{clip}(g)$ and g have the same sign and thus lead to the same classification rule.

Proof of Theorem 6.2.1. It is a simple consequence of Theorem 6.6.2. The assumptions of this result are met for the hinge loss considering $B = 1$, $S = L_2(Q)$ and $s = f^*$. Condition **(A2)** is satisfied only for $M \geq 1$ (the constants are minimal for $M = 1$). The lemma below ensures condition (6.24) with $\kappa = \frac{1}{h_0} \geq 1$.

Lemma 6.6.3. *Let $f : \mathcal{X} \to [-1, 1]$. We suppose that $|p(x) - \frac{1}{2}| \geq h_0$ where $p(x) = P[Y = 1|X = x]$. Then*

$$\|\gamma_h(f) - \gamma_h(f^*)\|_2^2 \leq \frac{1}{h_0} L_h(f, f^*)\,.$$

Proof of Lemma 6.6.3. It is a simple consequence of the reasoning of Lemma 4 of [BBM04] since a simple calculation shows that for $f : \mathcal{X} \to [-1, 1]$, $\|f - f^*\|_2^2 = \|\gamma_h(f) - \gamma_h(f^*)\|_2^2$. However, we provide a proof for the sake of completeness.

To begin with,

$$L_h(f, f^*) = \int p(x)[(1-f(x))_+ - (1-f^*(x))_+] + (1-p(x))[(1+f(x))_+ - (1+f^*(x))_+]\; dP(x)\,,$$

and

$$\|\gamma_h(f) - \gamma_h(f^*)\|_2^2 = \int p(x)[(1-f(x))_+ - (1-f^*(x))_+]^2 + (1-p(x))[(1+f(x))_+ - (1+f^*(x))_+]^2\; dP(x)\,.$$

Without loss of generality, we suppose that $f^*(x) = 1$ i.e. $p(x) \geq \frac{1}{2}$. Since $-1 \leq f(x) \leq 1$, $p(x)[(1-f(x))_+ - (1-f^*(x))_+]^2 + (1-p(x))[(1+f(x))_+ - (1+f^*(x))_+]^2 = (1-f(x))^2$. It yields $\frac{(f(x)-1)^2}{p(x)[(1-f(x))_+)] + (1-p(x))[(1+f(x))_+ - 2]} = \frac{1-f(x)}{2p(x)-1}$. We therefore obtain

$$\frac{p(x)[(1-f(x))_+)]^2 + (1-p(x))[(1+f(x))_+ - 2]^2}{p(x)[(1-f(x))_+)] + (1-p(x))[(1+f(x))_+ - 2]} \leq \frac{1}{h_0}\,,$$

and the statement follows by integrating with respect to x. □

6.7 Appendix B: Uniform Deviation Bound.

We need to introduce the notion of *sub-root* function introduced in [BBM03]. A function $\Psi : [0, \infty) \to [0, \infty)$ is sub-root if it is non-negative, non-decreasing and if $r \to \frac{\Psi(r)}{\sqrt{r}}$ is non-increasing for $r > 0$. The key quantity associated to such function is its fixed point: indeed, it can be shown that the fixed point equation $\Psi(r) = r$ has a unique positive solution (except for the trivial case $\Psi \equiv 0$). [BBM03] has shown that this solution r^* satisfies the following property:

$$r^* \le r \text{ if and only if } \Psi(r) \le r\,. \tag{6.30}$$

Let $\mathcal{F}$ be a set of functions. The following notation for the Rademacher average of $\mathcal{F}$ will be useful in the proofs:

$$\mathcal{R}_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\,,$$

where $(\varepsilon_i)_{i=1\cdots n}$ are independent and identically distributed Rademacher variables ($\mathbb{P}[\varepsilon_1 = 1] = \mathbb{P}[\varepsilon_1 = -1] = 1/2$). The notation $\mathbb{E}_\varepsilon$ means that the expectation is considered only with respect to ε: the variables $X_1, \dots, X_n$ are "fixed".

The main results of the chapter rely on the following theorem of concentration. It is an empirical version of Theorem 3 of [BBM04]: using the work of [BBM03], a control of the empirically localized Rademacher average is the key to control the deviations of an empirical process.

Theorem 6.7.1. *Let $\mathcal{F}$ be a star-shaped class of functions (i.e.* $\mathrm{star}(\mathcal{F}) = \mathcal{F}$*) containing* 0 *and with values in $[-b, b]$. Let A be a real number. Suppose that there exists a sub-root function $\hat\phi_n$ such that*

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}, P_n f^2 \le 2r} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i) \le \hat\phi_n(r)\,.$$

Let $\hat r_n^*$ be such that $\hat\phi_n(\hat r_n^*) = \frac{\hat r_n^*}{A'}$ where $A' = A \vee 1$.

Then, there exist universal constants E_1 and E_2 such that $\forall\, \xi > 0$, $K > \frac{A'}{7}$ with probability at least $1 - 3e^{-\xi}$, $\forall f \in \mathcal{F}$

$$(P - P_n)(f) \le \frac{1}{K}Pf^2 + E_1\frac{Kb'^2}{A'^2}\hat r_n^* + E_2\frac{K\xi b'^2}{n}\,,$$

where $b' = b \vee \frac{1}{5}$.

Proof of Theorem 6.7.1. Since $\mathrm{Var}(f) \le Pf^2$, Theorem 3 of [BBM04] states that if

$$\forall r \ge r^*,\ \mathbb{E} \sup_{f\in\mathcal{F}, Pf^2 \le r} (P - P_n)(f) \le \phi(r)\,,$$

with ϕ a sub-root function and r^* the unique solution of $A'\phi(r) = r$ then for all $K > \frac{A'}{7}$, with probability at least $1 - e^{-\xi}$, $\forall\, f \in \mathcal{F}$

$$(P - P_n)(f) \le \frac{1}{K}Pf^2 + \frac{50K}{A'^2}r^* + \frac{(K + 18b)\xi}{n}\,.$$

Moreover, the symmetrization lemma yields

$$\mathbb{E} \sup_{f\in\mathcal{F}, Pf^2\leq r} (P - P_n)(f) \leq 2\mathbb{E}\mathcal{R}_n\{f \in \mathcal{F}\,,\, Pf^2 \leq r\}$$

Consequently, we are allowed to consider $\phi(r) = 10b'\mathbb{E}\mathcal{R}_n\{f \in \mathcal{F}\,,\, Pf^2 \leq r\} + \frac{11b^2\xi}{n}$. Indeed, Lemma 3.4 of [BBM03] ensures that ϕ is a sub-root function since $\mathcal{F}$ is star-shaped.

We now control the fixed point r^* following the path of Theorem 4.1 of [BBM03]. The concentration result for sub-additive functionals of [BLM00] applied to Rademacher average yields that with probability at least $1 - e^{-\xi}$,

$$\phi(r) \leq 20b'\mathbb{E}_\varepsilon\mathcal{R}_n\{f \in \mathcal{F}\,,\, Pf^2 \leq r\} + \frac{11b^2\xi}{n} + \frac{20bb'\xi}{n}$$

and Corollary 2.2 of [BBM03] implies that if

$$r \geq \phi(r)\,,$$

then with probability at least $1 - 2e^{-\xi}$,

$$\phi(r) \leq 20b'\hat{\phi}_n(r) + \frac{11b^2\xi}{n} + \frac{20bb'\xi}{n}\,.$$

Using the previous inequality with $r = r^*$ ($r^* \geq \phi(r^*)$ since $A' \geq 1$) and combining the result with Lemma 3.7.2 with $K = \frac{3}{2}$, we obtain that with probability at least $1 - 2e^{-\xi}$,

$$r^* \leq \frac{3}{2}(20b')^2\hat{r}_n^* + 6A' \times 31b'^2\frac{\xi}{n}\,.$$

Finally, we obtain Theorem 6.7.1 with $E_1 = 30000$ and $E_2 = 20112$. □

6.8 Appendix C: Local Rademacher Complexity.

Let (T, d) be a metric space. The key quantity to control the involved Rademacher average is the covering number $\mathcal{N}(\varepsilon, T, d)$ of the set T. It is defined as the smallest number of balls of radius ε necessary to cover the space T. We have:

Theorem 6.8.1 ([Dud99] and [Bou02a]). *For every class $\mathcal{G}$ and every $X_1, \cdots, X_n$,*

$$\mathbb{E}_\varepsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(X_i) \leq \frac{4\sqrt{2}}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, L_2(P_n))}du\,.$$

It is often convenient to work with the packing number $\mathcal{M}(\varepsilon, T, d)$: this is the largest number of points of T which are at distance at least ε from each other. A classical result of [KT61] states that, up to some constants, these two measures of the size of a metric space are equivalent: precisely,

$$\mathcal{M}(2\varepsilon, T, d) \leq \mathcal{N}(\varepsilon, T, d) \leq \mathcal{M}(\varepsilon, T, d)\,.$$

In order to control the covering number of the involved class of functions clip(S_D), we use the following result which allows to upper-bound the packing number of a bounded-class of function by using its Vapnik-Chervonenkis dimension. We recall that if $\mathcal{A}$ denotes a family of subsets of $\mathcal{X}$, $V(\mathcal{A}) = \sup\{n, \sup_{x_1,\cdots,x_n\in\mathcal{X}} |\mathcal{A} \cap \{x_1, \cdots, x_n\}| = 2^n\}$. To the reader not familiar with these notions of complexity, we refer to [GKKW02] for a detailed survey.

Theorem 6.8.2 (Theorem 9.4 of [GKKW02]). *Let $\mathcal{G}$ be a class of functions $g : \mathcal{X} \longrightarrow [0, A]$ with $V(\mathcal{G}^+) \geq 2$. Let $p \geq 1$ and ν be a probability measure on $\mathcal{X}$. We consider ε such that $0 < \varepsilon < \frac{B}{4}$. Then the following holds*

$$\mathcal{M}(\varepsilon, \mathcal{G}, \|.\|_{L_p(\nu)}) \leq 3 \left(\frac{2eA^p}{\varepsilon^p} \log \frac{3eA^p}{\varepsilon^p} \right)^{V(\mathcal{G}^+)},$$

where $V(\mathcal{G}^+)$ is the VC-dimension of all subgraphs of functions of $\mathcal{G}$:

$$\mathcal{G}^+ = \{\{(z,t) \in \mathcal{X} \times \mathbb{R};\, t \leq g(z)\};\, g \in \mathcal{G}\}.$$

We now state the main result of this section.

Theorem 6.8.3. *Let $\mathcal{F} = \mathcal{F}_D(t_0) = \{(x,y) \to \gamma(t(x),y) - \gamma(t_0(x),y), t \in \mathrm{clip}_M(S_D)\}$ with $t_0 \in \mathrm{clip}_M(S_D)$. Let γ be such that*

$$\forall\, y, y_1, y_2\,, |\gamma(y,y_1) - \gamma(y,y_2)| \leq B|y_1 - y_2|\,. \tag{6.31}$$

Then, $\forall r \geq 0$,

$$\mathbb{E}_\varepsilon \sup_{f \in \mathrm{star}(\mathcal{F}),\, P_n f^2 \leq 2r} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \leq \frac{15}{\sqrt{n}} \sqrt{D+1} \sqrt{2r} \sqrt{\log \left(\frac{33(MB \vee 1)e^{\frac{1}{2}}}{\sqrt{2r} \wedge 2MB} \right)} = \hat{\phi}(r)\,,$$

and $\hat{\phi}$ is a sub-root function. Moreover, if $\hat{r}^*_{D,n}$ is such that $\hat{\phi}(\hat{r}^*_{D,n}) = \frac{\hat{r}^*_{D,n}}{\kappa'}$ with $\kappa' \geq 1$, then

$$\hat{r}^*_{D,n} \leq \kappa'^2 A_1(M,B) \frac{D+1}{n} \left(\left(\log \frac{n}{D} \right)_+ + 1 \right). \tag{6.32}$$

where $A_1(M,B) = 450 \left(\frac{33\sqrt{e}(MB \vee 1)}{(2MB) \wedge 1} \right)^2$. In the particular case of $M = B = 1$, we can improve the constant by taking $A_1(1,1) \leq 1487$.

Proof of Theorem 6.8.3. To begin with, we control the covering numbers. As noticed in [BBM03], it is easy to see that we can construct an ε-cover for $\mathrm{star}(\mathcal{F})$ using an $\frac{\varepsilon}{2}$-cover for $\mathcal{F}$ and an $\frac{\varepsilon}{2}$-cover for the interval $[0,1]$, which implies

$$\mathcal{N}(\varepsilon, \mathrm{star}(\mathcal{F}), L_2(P_n)) \leq \left(\left\lceil \frac{2}{\varepsilon} \right\rceil + 1 \right) \mathcal{N}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n) \right). \tag{6.33}$$

Moreover,

$$\begin{aligned} \mathcal{N}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n)\right) &\leq \mathcal{M}\left(\frac{\varepsilon}{2}, \mathcal{F}, L_2(P_n)\right) \\ &\leq \mathcal{M}\left(\frac{\varepsilon}{2B}, \{t - t_0, t \in \mathrm{clip}_M(S_D)\}, L_2(P_n)\right) \\ &= \mathcal{M}\left(\frac{\varepsilon}{2B}, \{t + M, t \in \mathrm{clip}_M(S_D)\}, L_2(P_n)\right), \end{aligned}$$

where the second inequality holds by using condition (6.31).

Theorem 6.8.2 with $A = 4M$ yields that for $0 < \varepsilon < 2MB$,

$$\begin{aligned}\mathcal{M}\left(\frac{\varepsilon}{2B}, \{t+M, t \in \mathrm{clip}_M(S_D)\}, L_2(P_n)\right) &\leq 3\left(\frac{128eM^2B^2}{\varepsilon^2}\log(\frac{192eM^2B^2}{\varepsilon^2})\right)^{V(\mathrm{clip}_M(S_D)+M)} \\ &\leq 3\left(\frac{11MBe^{\frac{1}{2}}}{\varepsilon}\right)^{4V(\mathrm{clip}_M(S_D)+M)}.\end{aligned}$$

where $V(\mathcal{G})$ denotes the Vapnik-Chervonenkis dimension of the class of functions $\mathcal{G}$. The last inequality is obtained by using $\log(x) \leq \frac{x}{2}$.

Moreover, $V(\mathrm{clip}_M(S_D)+M) = V(\mathrm{clip}_M(S_D)) \leq V(S_D) \leq D+1$, where the last inequality is a consequence of Lemma 2.6.15 of [vdVW96].

Finally, gathering inequality (6.33) and the previous ones, we get for $0 < \varepsilon < 2MB$,

$$\begin{aligned}\log \mathcal{N}(\varepsilon, \mathrm{star}(\mathcal{F}), L_2(P_n)) &\leq \log\left(\frac{2+2MB}{\varepsilon}\right) + 4(D+1)\log\left(\frac{33MBe^{\frac{1}{2}}}{\varepsilon}\right) \\ &\leq 5(D+1)\log\left(\frac{33(MB\vee 1)e^{\frac{1}{2}}}{\varepsilon}\right).\end{aligned}$$

Gathering Theorem 6.8.1 applied to $\mathcal{G} = \mathrm{star}(\mathcal{F})$ and the lemma below with $b = 2MB$, $\mu = 33(MB \vee 1)e^{\frac{1}{2}} > 32MB\sqrt{e}$ and $x = \sqrt{2r} \wedge (2MB)$ leads to the first inequality of Theorem 6.8.3.

Lemma 6.8.4 (inspired by Lemma 2.4 of [Men02]). *Let x, $0 \leq x \leq b$ and $\mu > \sqrt{e}b$:*

$$\int_0^x \sqrt{\log\left(\frac{\mu}{u}\right)}du \leq C(b)x\sqrt{\log\left(\frac{\mu}{x}\right)},$$

with $C = \frac{2\log(\frac{\mu}{b})}{2\log(\frac{\mu}{b})-1}$ an increasing function of b.

Proof. Let $f(x)$ be the left hand side of the inequality and $g(x)$ the right hand side. $f'(x) = \sqrt{\log(\frac{\mu}{x})}$ and $g'(x) = C(b)\left(\sqrt{\log\frac{\mu}{x}} - \frac{1}{2\sqrt{\log\frac{\mu}{x}}}\right)$. Consequently, $f'(x) \leq g'(x)$ is equivalent to $2\log\frac{\mu}{x} \leq C(b)\left(2\log\frac{\mu}{x} - 1\right)$ i.e. $C(b) \geq \frac{2\log(\frac{\mu}{x})}{2\log(\frac{\mu}{x})-1}$. Noticing that $f(0) = g(0)$, this concludes the proof of Lemma 6.8.4. □

We easily check that $\hat{\phi}$ is a sub-root function (it is increasing since $\mu > b\sqrt{e}$).
Due to property (6.30), the bound (6.32) on $\hat{r}^*_{D,n}$ proposed in Theorem 6.8.3 is equivalent to

$$\kappa'\hat{\phi}\left(\kappa'^2 A_1(M,B)\frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\right) \leq \kappa'^2 A_1(M,B)\frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right). \quad (6.34)$$

Since $\kappa' \geq 1$,

$$\begin{aligned}&\kappa'\hat{\phi}\left(\kappa'^2 A_1(M,B)\frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\right) \\ &\quad\leq \kappa'^2 15\frac{D+1}{n}\sqrt{2A_1(M,B)}\sqrt{\left(\left(\log\frac{n}{D}\right)_+ + 1\right)}\sqrt{\log(\frac{33(MB\vee 1)\sqrt{e}}{(2MB)\wedge\sqrt{2\frac{A_1(M,B)(D+1)}{n}}})}.\end{aligned}$$

If $A_1(M,B) \geq \left(\frac{33(MB\vee 1)}{\sqrt{2e}}\right)^2$, $\log(\frac{33(MB\vee 1)\sqrt{e}}{\sqrt{2\frac{A_1(M,B)(D+1)}{n}}}) \leq \left(\log\frac{n}{D}\right)_+ + 1$ (It is clear for $n \leq D$ and use $\sqrt{\frac{D+1}{n}} \geq \frac{D}{n}$ for $n > D$).

Thus, if $A_1(M,B) \geq \left(\frac{33(MB\vee 1)}{\sqrt{2e}}\right)^2$ then

$$\kappa'\hat{\phi}\left(\kappa'^2 A_1(M,B)\frac{D+1}{n}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\right)$$
$$\leq \kappa'^2 15\frac{D+1}{n}\sqrt{2A_1(M,B)}\left(\left(\log\frac{n}{D}\right)_+ + 1\right)\sqrt{\log\frac{33(MB\vee 1)\sqrt{e}}{2MB}}.$$

Finally, provided that $A_1(M,B) \geq 450\log\frac{33(MB\vee 1)\sqrt{e}}{b}$, we obtain inequality (6.34). This concludes the proof of Theorem 6.8.3. □

6.9 Appendix D: Proof of Inequality (6.11).

In this proof, with some abuse of notations, we denote

$$\{\phi_j(x)\}_{j\geq 1} = \{1, \sqrt{2}\sin(2\pi x), \sqrt{2}\cos(2\pi x), \cdots, \sqrt{2}\sin(2\pi jx), \sqrt{2}\cos(2\pi jx), \cdots\},$$

the orthonormal trigonometric basis of $L_2([0,1])$.

Let $f^* \in \mathcal{F}_\alpha$ and $\alpha_i^* = \langle f^*, \phi_i\rangle_{L_2(P)}$:

$$\inf_{g\in\mathcal{H}}\left(d(g,f^*) + \Lambda_n\|g\|^2_{\mathcal{H}}\right) = \inf_\alpha\left(\mu_\alpha + \Lambda_n\sum_{i\geq 1}\frac{\alpha_i^2}{\lambda_i}\right),$$

where $\mu_\alpha = \sqrt{\sum_{i\geq 1}(\alpha_i - \alpha_i^*)^2}$. The infimum is reached at

$$\alpha_i^\circ = \frac{\alpha_i^*\lambda_i}{\lambda_i + 2\Lambda_n\mu_{\alpha^\circ}},$$

and

$$\mu_{\alpha^\circ} = 2\sqrt{\sum_{i\geq 1}\left(\frac{\alpha_i^*\Lambda_n\mu_{\alpha^\circ}}{\lambda_i + 2\Lambda_n\mu_{\alpha^\circ}}\right)^2} \geq \frac{1}{2}\sqrt{\sum_{i\geq 1,\lambda_i\leq 2\Lambda_n\mu_{\alpha^\circ}}\alpha_i^{*2}}.$$

Using $\lambda_j \sim j^{-2\gamma}$, $\Lambda_n \sim n^{-\frac{4\gamma}{2(2\gamma+1)}}$ and the definition of $\mathcal{F}_\alpha$ yields:

$$\mu_{\alpha^\circ} \geq Cn^{-\frac{2\alpha-1}{4\gamma+2}}\mu_{\alpha^\circ}^{\frac{2\alpha-1}{4\gamma}}.$$

Solving this inequality for $2\alpha - 1 < 4\gamma$ entails

$$\mu_{\alpha^\circ} \geq Cn^{\frac{-4(2\alpha-1)\gamma}{(4\gamma+2)(4\gamma-2\alpha+1)}},$$

and finally

$$\inf_{f^*\in\mathcal{F}_\alpha}\inf_{g\in\mathcal{H}_k}\left(\|g-f^*\|_2 + \Lambda_n\|g\|^2_{\mathcal{H}}\right) \geq cn^{\frac{-4(2\alpha-1)\gamma}{2(2\gamma+1)(4\gamma-2\alpha+1)}}.$$

This concludes the proof of Inequality (6.11). □

6.10 Appendix E: Eigenfunctions in the Gaussian case.

We use the following normalization for the Hermite polynomials: H_n is an orthogonal system of $L_2(e^{-x^2})$ i.e. $e^{2\lambda x-\lambda^2} = \sum_{n\geq 0} H_n(x)\frac{\lambda^n}{n!}$. In this case, if $f_n(x) = H_n(x)e^{-\frac{x^2}{2}}$ then $\langle f_n, f_m\rangle_{L_2(\mathbb{R})} = \delta_{n,m}\sqrt{\pi}2^n n!$.

Theorem 6.10.1 ([ZWRM98] and [WS00]). *Let $d\mu(x) = \frac{1}{\sqrt{2\pi}}e^{-2ax^2}dx$ and T_k be the integral operator associated with the gaussian kernel $k(x,y) = e^{-b(x-y)^2}$.*

$$\begin{array}{cccc} T_k: & L_2(\mu) & \to & L_2(\mu) \\ & f & \to & \int_{\mathbb{R}} f(x)e^{-b(x-y)^2}d\mu(x) \end{array}$$

An explicit orthonormal basis of $L_2(\mu)$ of eigenvectors of T_k associated to $\lambda_j = \sqrt{\frac{1}{2A}}\left(\frac{b}{A}\right)^{j-1}$ is given by:

$$\Psi_j(x) = \frac{(4p)^{\frac{1}{4}}e^{-(p-a)x^2}H_{j-1}\left(\sqrt{2p}\,x\right)}{(2^{j-1}(j-1)!)^{\frac{1}{2}}}$$

where $p = \sqrt{a^2+2ab}$ and $A = a+b+p$ and $j \geq 1$.

Proof. Since the Hermite polynomials are an orthogonal system of $L_2(e^{-x^2})$, a simple calculation shows that $(\Psi_j)_{j\geq 1}$ is an orthogonal family of functions of $L_2(e^{-2ax^2})$. Moreover, since H_n is a polynomial function of degree n, a classical result of analysis implies that $(\Psi_j)_{j\geq 1}$ is an orthonormal basis of $L_2(\mu)$.

It can be proved by using equation 7.374.8 of [GR80] that:

$$\int_{\mathbb{R}} e^{-(x-y)^2}H_n(\alpha x)\,dx = \sqrt{\pi}(1-\alpha^2)^{n/2}H_n\left(\frac{\alpha y}{(1-\alpha^2)^{1/2}}\right). \tag{6.35}$$

Let $C_j = \frac{(4p)^{\frac{1}{4}}}{(2^{j-1}(j-1)!)^{\frac{1}{2}}}$. We have:

$$\begin{aligned}
T_k\Psi_j(y) &= C_j\int_{\mathbb{R}} e^{-(p-a)x^2}H_{j-1}\left(\sqrt{2p}\,x\right)e^{-b(x-y)^2}d\mu(x) \\
&= C_j\frac{e^{-(p-a)y^2}}{\sqrt{2\pi}}\int_{\mathbb{R}} H_{j-1}\left(\sqrt{2p}\,x\right)e^{-\left[b(x-y)^2+2ax^2+(p-a)x^2-(p-a)y^2\right]}dx \\
&= C_j\frac{e^{-(p-a)y^2}}{\sqrt{2\pi}}\int_{\mathbb{R}} H_{j-1}\left(\sqrt{2p}\,x\right)e^{-\left(x\sqrt{b+p+a}-y\sqrt{b-p+a}\right)^2}dx \\
&= C_j\frac{e^{-(p-a)y^2}}{\sqrt{2A\pi}}\int_{\mathbb{R}} H_{j-1}\left(\sqrt{\frac{2p}{A}}\,x'\right)e^{-\left(x'-y\sqrt{b-p+a}\right)^2}dx' \\
&= C_j\frac{e^{-(p-a)y^2}}{\sqrt{2A}}\left(\frac{b+a-p}{A}\right)^{(j-1)/2}H_{j-1}\left(\sqrt{2p}\,y\right),
\end{aligned}$$

where equality (6.35) is used in the last equality. We get:

$$T_k\Psi_j(y) = \frac{1}{\sqrt{2A}}\left(\frac{b+a-p}{A}\right)^{(j-1)/2}\Psi_j(y).$$

Since $\left(\frac{b+a-p}{A}\right)^{1/2} = \frac{b}{A}$, this concludes the proof of Theorem 6.10.1. $\square$

Part III

Open Problems and Bibliography

Chapter 7

Perspectives and Open Problems

Each topic tackled in this thesis raises new questions in several directions. To begin with, no lower bound is provided. It would warrant the relevance of the quantities involved in the upper bounds. The minmax approach commonly used in statistics is a way to express such lower bounds. Next, the concentration of eigenvalues of the kernel matrix around those of the kernel operator is proved in chapter 3 for the sums of largest and smallest eigenvalues. It is only natural to study how each eigenvalue concentrates. Finally, we are hopeful that the finite dimensional regularization, exploited for binary supervised classification in chapters 5 and 6, will be efficient in other frameworks.

7.1 Concentration of the Single Eigenvalues and Minmax Approach for KPCA.

This section proposes open problems related to the first part of this thesis.

- [STWCK05] provides concentration inequalities for the single eigenvalues of the kernel matrix. However, they are coarse since they do not depend on the eigenvalue itself. Moreover, they imply slow convergence rates. We plan to get more accurate bounds depending on the eigenvalue itself and corresponding to fast convergence rates.

- Chapter 3 provides upper bounds concerning the reconstruction error of KPCA. However, their optimality has not been studied. It seems natural to do it by providing an adapted minmax framework.

7.2 Finite Dimensional Projection.

We now provide open problems for classification and varied feasible use of the finite dimensional regularization principle for some learning tasks.

- The approximation error has been studied recently to provide rates of convergence for SVM (see [SS03] and [VV05]). We plan to get rates of convergence for KPM. The first step is to study the eigenfunctions of the kernel operator: this can be done by working on the links between differential and kernel operators.

- Optimality properties of these rates could be obtained with a minmax point of view. However, it is arduous to define a minmax framework for classification adapted to SVM and finite dimensional projection. Indeed, the convexification of the risk allows to work with models composed of smooth functions whereas the target (the Bayes classifier) is not smooth. Minmax results would get comparisons between different algorithms. For example, it would allow to compare the qualities of the ellipsoids involved in the SVM with the finite dimensional vector spaces.

- Tikhonov's and finite dimensional regularization can be combined by considering balls of the RKHS intersected with vector spaces as models. The minimization of the empirical risk over these models leads to an algorithm using a mixed strategy of regularization. Theorem 4.5.1 provides a first step of the analyze of such a regularization procedure since it ensures a complexity control of the involved random models. This idea deserves to be further investigated.
- As explained in appendix B of chapter 5, the finite dimensional regularization allows to take into account informations given by unlabeled data. Consequently, we plan to carry out experiments on some data sets and to undertake theoretical studies about semi-supervised learning.

Bibliography

[AB04] J.Y. Audibert and O. Bousquet. PAC-Bayesian generic chaining. *Advances in Neural Information Processing Systems 16.*, 2004.

[And63] T. W. Anderson. Asymptotic theory for Principal Component Analysis. *Ann. Math. Stat.*, 34:122–148, 1963.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[Bak77] Baker. *The numerical treatment of integral equations.* Oxford: Clarendon Press, 1977.

[Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist. 6 127–146*, 2002.

[Bax76] P. Baxendale. Gaussian measures on function spaces. *Amer. J. Math.*, 98:891–952, 1976.

[BBL02] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning 48 (1-3), 85-113*, 2002.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Proba. Theory Relat. Fields*, 113:301–413, 1999.

[BBM03] P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. 2003. Available at http://www.kyb.mpg.de/publications/pss/ps2000.ps, Submitted.

[BBM04] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of Support Vector Machines. *Manuscript*, 2004.

[BBZ04] G. Blanchard, O. Bousquet, and L. Zwald. Statistical Properties of Kernel Principal Component Analysis. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th. Conference on Learning Theory (COLT 2004)*, volume 3120 of *Lecture Notes in Artificial Intelligence*, pages 594–608. Springer, 2004. Extended technical report available at `http://www.math.u-psud.fr/~blanchard/publi/BlaBouZwa05.ps.gz`, submitted.

[BCS76] I.A. Ibragimov B.S. Cirel'son and V.N. Sudakov. Norm of gaussian sample function. *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory. Lecture Notes in Mathematics 550 20-41. Springer-Verlag, Berlin*, 1976.

[BE94] R. Bhatia and L. Elsner. The Hoffman-Wielandt inequality in infinite dimensions. *Proc.Indian Acad.Sci(Math. Sci.), Vol.104,No.3,pp.483-494.*, 1994.

[Bes79] P. Besse. *Etude descriptive d'un pocessus ; approximation, interpolation.* PhD thesis, Université de Toulouse, 1979.

[Bes91] P. Besse. Approximation spline de l'Analyse en Composantes Principales d'une variable aléatoire hilbertienne. *Ann. Fac. Sci. Toulouse (Math.)*, 12(5):329–349, 1991.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, 1984.

[BGV92] B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

[BJ02] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 2002.

[BJM03] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Technical Report 638, Department of Statistics, U.C. Berkeley. Journal of the American Statistical Association*, 2003.

[BLM00] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[BLV03] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research (Special issue on learning theory), 4(Oct):861-894*, 2003.

[BM97] L. Birgé and P. Massart. From model selection to adaptative estimation. *Festschrift for Lucien le Cam: research papers in probability and statistics, page 55 87, Springer New York*, 1997.

[BM98] L. Birge and P. Massart. Minimum contrast estimator on sieves: exponential bounds and rate of convergence. *Bernoulli*, 4, 329-375, 1998.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society 3 n3, 203-268.*, 2001.

[BM02] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BMP04] P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. *Proceedings of the 17th Annual Conference on Learning Theory, COLT*, 2004.

[BMVZ04] G. Blanchard, P. Massart, R. Vert, and L. Zwald. Kernel projection machine: a new tool for pattern recognition. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Proceedings of the 18th. Neural Information Processing System (NIPS 2004)*, pages 1649–1656. MIT Press, 2004.

[Bou02a] O. Bousquet. PhD thesis, Ecole Polytechnique, 2002.

[Bou02b] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I 334, 495-500*, 2002.

[Bou03] O. Bousquet. New approaches to statistical learning theory. *Ann. Inst. Statist. Math. Vol. 55, No. 2, 371-389*, 2003.

[BZ05] G. Blanchard and L. Zwald. On the convergence of eigenspaces in Kernel Principal Component Analysis. In *Proceedings of the 19th. Neural Information Processing System (NIPS 2005).* MIT Press, 2005.

[CG05] L. Cavalier and G.K. Golubev. Risk hull method and regularization by projections of ill-posed inverse problems. *Submitted*, 2005.

[CST00] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, UK, 2000.

[CT01] L. Cavalier and A.B. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods of Stat. 10 (2001), 247-282.*, 2001.

[CV95] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of mathematics.* Springer, New York, 1996.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer Series in Statistics. Springer Verlag, New York, 2001.

[DLM90] D.L. Donoho, R.C. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *Ann. Statist. 18,1416-1437*, 1990.

[dlPnG99] V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence.* Springer, 1999.

[DP76] J. Dauxois and A. Pousse. *Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique.* PhD thesis, Université de Toulouse, 1976.

[DPR82] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the Principal Component Analysis of a vector random function: some applications to statistical inference 136-154. *Journal of multivariate analysis*, 1982.

[DS63] N. Dunford and J. T. Schwartz. *Linear Operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space.* Number VII in Pure and Applied Mathematics. John Wiley & Sons, New York, 1963.

[Dud67] R.M. Dudley. The size of compact subsets of hilbert spaces and continuity of gaussian processes. *J. Funct. Anal. 1, 290-330*, 1967.

[Dud84] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

[Dud99] R. Dudley. *Uniform Central Limit Theorems.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

[EPP99] T. Evgenious, T. Poggio, and M. Pontil. A unified framework for regularisation networks and Support Vector Machine. Technical report, MIT, march 1999.

[EPP00] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and Support Vector Machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 171–203, Cambridge, MA, 2000. MIT Press.

[Fer75] X. Fermique. Régularité des trajectoires des fonctions aléatoires gaussiennes. *Lecture notes in mathématics (Springer) 480*, 1975.

[GK02] R. E. Greene and S. G. Krantz. *Function theory of one complex variable.* American Mathematical Society, 2002.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution–Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer Verlag, New York, 2002.

[GR80] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products.* Academic Press, New York, 1980. Corrected and enlarged edition prepared by A. Jeffrey.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933.

[Jan97] S. Janson. *Gaussian Hilbert Spaces.* Cambridge University Press, 1997.

[Kat66] T. Kato. *Perturbation Theory for Linear Operators.* New-York: Springer-Verlag, 1966.

[KG00] V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[Kle02] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris, Ser I 334*, 2002.

[Kol98] V. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. *Progress in Probability*, 43:191–227, 1998.

[Kol01] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory 47(5), 1902-1914*, 2001.

[Kol04] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *To appear*, 2004.

[KT61] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961.

[Leb02] Emilie Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini.* PhD thesis, University Paris XI, 2002.

[Led96] M. Ledoux. On Talagrand deviation inequalities for product measures. *ESAIM: Probability and Statistics 1,63-87*, 1996.

[Lin99] Y. Lin. Support Vector Machines and the Bayes rule in classification. *Technical Report TR 1014, University of Wisconsin, Statistics departement, 1999. To appear in Data Mining and Knowledge Discovery*, 1999.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach spaces.* Springer-Verlag, Berlin, 1991.

[Lug00] G. Lugosi. *Lecture notes on statistical learning theory.* Garchy seminar, 2000.

[Lug02] G. Lugosi. *Principles of Nonparametric Learning.* Springer, Wien, New York, pp. 1–56, 2002.

[LW03] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Ann. of Stat., to appear*, 2003.

[Mas00a] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. of Probability 28,2, 863-884.*, 2000.

[Mas00b] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.

[Mas04] P. Massart. *Concentration Inequalities and Model Selection.* Springer-Verlag, 2004. Probability summer school, Saint Flour 2004, to be published.

[McD98] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, New York, 1998.

[Men02] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(1):251-263, 2002.

[MT95] E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *The Annals of Statistics*, 23(2):502–524, 1995.

[NM03] E. Nedelec and P. Massart. Risk bounds for statistical learning. *Technical report, Université Paris-sud*, 2003.

[Pea01] K. Pearson. On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2:559–572, 1901.

[Rae99] G. Raetsch. http://ida.first.gmd.de/~raetsch/data/benchmarks.htm. 1999. Benchmark repository used in several Boosting, KFD and SVM papers.

[RD91] J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B*, 53(3):539–572, 1991.

[Rio01] E. Rio. Une inégalité de Benett pour les maxima de processus empiriques. 2001.

[Rud73] W. Rudin. *Functional Analysis*. McGraw-Hill, New York, 1973.

[Sch64a] I. J. Schoenberg. On interpolation by spline functions and its minimal properties. In P. L. Butzer and J. Korevaar, editors, *Über Approximationstheorie (On Approximation Theory)*, pages 109–129, Basel, 1964. International Series of Numerical Mathematics, Birkäuser Verlag.

[Sch64b] I. J. Schoenberg. Spline functions and the problem of graduation. *Proc. N. A. S.*, 52:947, 1964.

[SS98] A. J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

[SS03] C. Scovel and I. Steinwart. Fast rates for Support Vector Machines. *Submitted to Annals of Statistics on 12/24/03*, 2003.

[SSM96] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. Technical Report 44, Max-Planck-Institut für biologische Kybernetik, 1996.

[SSM98] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[SSM99] B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel Principal Component Analysis. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999. Short version appeared in *Neural Computation* 10:1299–1319, 1998.

[STC03] J. Shawe-Taylor and N. Cristianini. Estimating the moments of a random vector with applications. *Proceedings of GRETSI 2003 Conference, pages pp. 47-52*, 2003.

[STWCK02] J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. Eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory : 13th International Conference, ALT 2002*, volume 2533 of *Lecture Notes in Computer Science*, pages 23–40. Springer-Verlag, 2002. Extended version available at `http://www.support-vector.net/papers/eigenspectrum.pdf`.

[STWCK05] J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalisation error of Kernel PCA. *IEEE Transactions on Information Theory 51*, 2005. (To appear).

[Tal87] M. Talagrand. Regularity of gaussian processes. *Acta math.*, pages 159, 99–149, 1987.

[Tal96] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math. 126, 505-563*, 1996.

[Tal05] M. Talagrand. *The Generic Chaining.* Springer, 2005.

[TG04] B. Tarigan and S.A. Van De Geer. Adaptivity of Support Vector Machines with ℓ_1 penalty. *to appear*, 2004.

[Tor97] M. Torki. Etude de la sensibilité de toutes les valeurs propres non nulles d'un opérateur compact autoadjoint. Technical Report LAO97-05, Université Paul Sabatier, 1997. Available at http://mip.ups-tlse.fr/publi/rappLAO/97.05.ps.gz.

[Tsy04] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics, 32 (1)*, 2004.

[TvdG05] A. Tsybakov and S. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics Vol.33, No. 3 - June*, 2005.

[Val84] L. G. Valiant. A theory of learnable. *Proc. of the 1984 STOC*, pages 436–445, 1984.

[Vap82] V. Vapnik. *Estimation of dependencies based on empirical data.* Springer, New York, 1982.

[Vap95] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[Vap98] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[Vay00] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et mesures de complexité.* PhD thesis, Ecole Polytechnique, 2000.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996.

[vLBB04a] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. Technical Report 134, Max Planck Institute for Biological Cybernetics, 2004.

[vLBB04b] U. von Luxburg, O. Bousquet, and M. Belkin. On the convergence of spectral clustering on random samples: the normalized case. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory (COLT 2004)*, volume 3120 of *Lecture notes in Artificial Intelligence*, pages 457–471. Springer, 2004.

[VV05] R. Vert and J.-P. Vert. Consistency and convergence rate of one-class SVM and related algorithms. Technical report, LRI, 2005.

[Wah90] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1990.

[WS00] C. K. I. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In P. Langley, editor, *Proceedings of the 17th International Conference on Machine Learning*, pages 1159–1166, San Francisco, California, 2000. Morgan Kaufmann.

[WSTSS99] R. C. Williamson, J. Shawe-Taylor, B. Schölkopf, and A. J. Smola. Sample-based generalization bounds. *IEEE Transactions on Information Theory*, 1999. Submitted. Also: NeuroCOLT Technical Report NC-TR-99-055.

[Zha93] Ping Zhang. Model selection via multifold cross validation. *The Annnals of Statistics, Vol. 21, No. 1, 299-313*, 1993.

[ZWRM98] H. Zhu, C.K.I. Williams, R.J. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. *Neural networks and machine learning*, 1998.

Résumé. La thèse se place dans le cadre de l'apprentissage statistique. Elle apporte des contributions à la communauté du machine learning en utilisant des techniques de statistiques modernes basées sur des avancées dans l'étude des processus empiriques. Dans une première partie, les propriétés statistiques de l'analyse en composantes principales à noyau (KPCA) sont explorées. Le comportement de l'erreur de reconstruction est étudié avec un point de vue non-asymptotique et des inégalités de concentration des valeurs propres de la matrice de Gram sont données. Tous ces résultats impliquent des vitesses de convergence rapides. Des propriétés non-asymptotiques concernant les espaces propres de la KPCA eux-mêmes sont également proposées. Dans une deuxième partie, un nouvel algorithme de classification a été conçu : la Kernel Projection Machine (KPM). Tout en s'inspirant des Support Vector Machines (SVM), il met en lumière que la sélection d'un espace vectoriel par une méthode de réduction de la dimension telle que la KPCA régularise convenablement. Le choix de l'espace vectoriel utilisé par la KPM est guidé par des études statistiques de sélection de modèle par minimisation pénalisée de la perte empirique. Ce principe de régularisation est étroitement relié à la projection fini-dimensionnelle étudiée dans les travaux statistiques de Birgé et Massart. Les performances de la KPM et de la SVM sont ensuite comparées sur différents jeux de données. Chaque thème abordé dans cette thèse soulève de nouvelles questions d'ordre théorique et pratique.

Mots-clés. Apprentissage statistique, inégalité de concentration, processus empirique, minimisation empirique du risque, classification, réduction de dimension, régularisation, Support Vector Machines (SVM), sélection de modèle, inégalité oracle, vitesse rapide.

Summary. This thesis takes place within the framework of statistical learning. It brings contributions to the machine learning community using modern statistical techniques based on progress in the study of empirical processes. The first part investigates the statistical properties of Kernel Principal Component Analysis (KPCA). The behavior of the reconstruction error is studied with a non-asymptotic point of view and concentration inequalities of the eigenvalues of the kernel matrix are provided. All these results correspond to fast convergence rates. Non-asymptotic results concerning the eigenspaces of KPCA themselves are also provided. A new algorithm of classification has been designed in the second part: the Kernel Projection Machine (KPM). It is inspired by the Support Vector Machines (SVM). Besides, it highlights that the selection of a vector space by a dimensionality reduction method such as KPCA regularizes suitably. The choice of the vector space involved in the KPM is guided by statistical studies of model selection using the penalized minimization of the empirical loss. This regularization procedure is intimately connected with the finite dimensional projections studied in the statistical work of Birgé and Massart. The performances of KPM and SVM are then compared on some data sets. Each topic tackled in this thesis raises new questions.

Keywords. Statistical learning, concentration inequality, empirical process, empirical risk minimization, classification, dimensionality reduction, regularization, Support Vector Machines (SVM), model selection, oracle inequality, fast rate.

AMS classification: 62G05, 62H25, 62H30, 62P99, 68W01, 68W40, 68Q15.