Aggregation procedures: optimality and fast rates

Guillaume Lecué

Université Pierre et Marie Curie - Paris–VI

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS–VI

Spécialité : **Mathématiques**

soutenue par

Guillaume Lecué

Méthodes d'agrégation : optimalité et vitesses rapides.

Directeur de thèse : Alexandre TSYBAKOV

Rapporteurs : M. Vladimir KOLTCHINSKI Georgia Institute of Technology.
M. Gábor LUGOSI Universitat Pompeu Fabra.

Soutenue publiquement le **18 mai 2007** devant le jury composé de

M. Alexandre	**TSYBAKOV**	Université Paris–VI	Directeur
M. Vladimir	**KOLTCHINSKI**	Georgia Institute of Technology	Rapporteur
M. Gábor	**LUGOSI**	Universitat Pompeu Fabra	Rapporteur
M. Pascal	**MASSART**	Université Paris–XI	Président
M. Olivier	**CATONI**	Université Paris–VI	Examinateur
M. Arkadi	**NEMIROVSKI**	Georgia Institute of Technology	Examinateur
M. Stéphane	**BOUCHERON**	Université Paris–VI	Examinateur
M. Lucien	**BIRGÉ**	Université Paris–VI	Examinateur

Remerciements

Je tiens en premier lieu à remercier très sincèrement mon directeur de thèse Alexandre Tsybakov. Durant ces trois ans de thèse, j'ai toujours eu le sentiment de faire ma thèse avec Sacha à mes côtés. J'ai toujours été très demandeur de chacune de tes idées et très honoré d'avoir pu travailler avec toi (dans ton bureau et dans ton jardin à Bures). Je tiens aussi à te remercier pour m'avoir emmené à Berkeley. Aussi, je t'exprime ma profonde gratitude et je n'imagine pas mes recherches futures sans ton soutien et tes conseils avisés.

Je remercie Lucien Birgé, Stéphane Boucheron, Olivier Catoni, Vladimir Koltchinski, Gábor Lugosi, Pascal Massart et Arkadi Nemirovski pour m'avoir fait l'immense honneur de faire partie de mon jury de thèse.

Je remercie vivement Pascal Massart pour son cours de DEA qui fait partie de mon viatique de statisticien, mais aussi pour m'avoir orienté vers Sacha. Il est difficile en quelques mots de vous exprimer toute ma reconnaissance et mon estime.

Ces trois années de thèse passées à Chevaleret en général, et au LPMA en particulier, ont représenté pour moi une immense chance de rencontrer et de travailler avec des chercheurs dont la qualité est telle que la moindre de mes questions a toujours trouvé réponse. J'ai ainsi, à de nombreuses occasions, profité de cette opportunité auprès de chercheurs comme Lucien Birgé, Stéphane Boucheron, Albert Cohen, Arnak Dalalyan, Gérard Kerkyacharian, Erwann Le Pennec, Nicolas Vayatis. Je tiens ici à les remercier vivement pour leur patience, leur générosité et les connaissances qu'ils m'ont fait partager.

Je tiens à exprimer toute ma reconnaissance à Vladimir Koltchinski et Gábor Lugosi pour avoir accepté de rapporter ma thèse. Je remercie Vladimir Koltchinski pour son invitation à exposer dans son groupe de travail à Atlanta et pour son accueil chaleureux pendant mon séjour à Georgia Tech. Je remercie vivement Peter Bartlett, Peter Bickel et Bin Yu pour leur accueil à Berkeley. J'adresse aussi mes remerciements à Gábor Lugosi pour sa gentillesse et la simplicité qu'il a su gardé. Il est incontestablement un modèle très influent pour moi.

Je tiens aussi à remercier Patricia Reynaud-Bourret et Gilles Stoltz pour m'avoir permis à deux reprises d'exposer mes travaux dans leur groupe de travail. J'ai particulièrement apprécié ces expériences. Aussi, je tiens à remercier Gilles pour les nombreux et précieux conseils amicaux qu'il m'a donné pendant mes années de thèse.

Au cours de diverses conférences, j'ai eu la chance de rencontrer de remarquables chercheurs qui ont eu la gentillesse de discuter avec moi et dont les travaux m'ont inspiré. Je souhaite citer ici, entre autres : András Antos, Jean-Yves Audibert, Olivier Bousquet, Laurent Cavalier, Alexander Goldenschluger, Mattias Hein, Anatoli Iouditski, Oleg Lepski, Axel Munk, Ben Rubenstein, Csaba Szepesvári, Tong Zhang, Laurent Zwald...

Je remercie chaleureusement mes collaborateurs, Christophe Chesneau et Stéphane Gaïffas, avec lesquels j'espère concrétiser encore beaucoup d'autres projets.

Au cours de mes études ou de séminaires, j'ai eu le plaisir de rencontrer de nombreux thésards avec qui j'ai sympathisé et avec qui j'ai eu de nombreuses discussions fructueuses: Etienne, Thanh Mai, Claire, Sylvain, Thomas, Assane, Juan Carlos, Elie, Anne-Laure,

Fabien, Karine, Tu, Olivier, Karim, Mohammed, Nicolas, Fanny, Sébastien, Philippe, Nathalie, Katia, Alex, Mathieu, Pierre, Joseph, Alexis,...

Certaines lourdeurs administratives se sont très vite allégées grâce à l'efficacité de Salima, Josette, Nelly, Yves et Corentin. Je les remercie ici pour leur aide précieuse. De même, je dois remercier Jacques Portes pour son efficacité à résoudre les problèmes informatiques les plus ardus et pour ses recommandations technologiques.

Merci donc à tous mes collègues, chercheurs ou enseignants.

Je termine par un grand remerciement à ma famille, mes amis, et Céline ; leur soutien et leurs encouragements constants ont permis à cette thèse de voir le jour.

..., à ma mère,

à ma soeur, ...

Contents

CHAPITRE 1

Introduction et présentation des résultats

Au sein d'un problème statistique, le statisticien peut disposer d'une large batterie d'estimateurs (estimateurs à noyaux, estimateurs par projection, estimateurs par moindres carrés (pénalisés ou non), etc). Sous différentes hypothèses sur le modèle, l'une de ces procédures pourra être plus performante que les autres. Ces hypothèses, faites a priori, n'ont aucune raison d'être réellement vérifiées. Nous aimerions pouvoir profiter des qualités propres de ces estimateurs, tout en faisant le moins d'hypothèses possible sur le modèle. Ce genre de problèmes est connu sous le nom de *problème d'adaptation*. Les méthodes étudiées dans cette thèse peuvent être utilisées pour résoudre ce genre de problèmes. Pour éviter ces hypothèses, nous pouvons aussi changer de problématique en cherchant à construire une procédure faisant approximativement aussi bien que la meilleure parmi un ensemble de procédures de base donnée a priori. C'est le paradigme que nous nous proposons d'étudier ici.

Le principal travail de cette thèse porte sur **l'étude des méthodes d'agrégation sous l'hypothèse de marge** (cf. [83, 81, 80, 38]). Nous avons mis en avant que l'hypothèse de marge améliore les vitesses d'agrégation qui peuvent s'approcher de $1/n$, où n est la taille de l'échantillon.

Un autre résultat de cette thèse montre que certaines **méthodes de minimisation du risque empirique pénalisé sont sous-optimales** quand le risque est convexe, même sous l'hypothèse de marge (cf. [85, 84]). Contrairement aux procédures d'agrégation à poids exponentiels, ces méthodes n'arrivent pas à profiter de la marge du modèle.

Ensuite, nous avons appliqué les méthodes d'agrégation à la **résolution de quelques problèmes d'adaptation**. Dans une première application, nous construisons des **procédures à la fois adaptatives au paramètre de marge et au paramètre de complexité** par agrégation d'estimateurs à vecteurs de support (cf. [82]). Nous avons ensuite appliqué les méthodes d'agrégation dans les problèmes d'estimation de densités et de fonctions de régression. En agrégeant seulement $\log n$ estimateurs par ondelette seuillés, nous avons obtenu un estimateur **adaptatif sur tous les espaces de Besov sans perte de vitesse logarithmique** (cf.[38]). Une autre application des méthodes d'agrégation a été de répondre positivement à une conjecture de Stone dans le modèle du "single index" (cf. [56]). En adoptant un point de vue différent des méthodes habituellement utilisées dans ce modèle (c'est-à-dire en s'adaptant à l'index plutôt qu'en l'estimant), nous avons construit une **procédure atteignant la vitesse conjecturée par Stone** (sans perte de vitesse logarithmique telle qu'on l'observait chez les estimateurs construits jusqu'ici).

Une dernière contribution apportée par cette thèse a été de proposer une approche du **contrôle du biais en classification** par l'introduction d'espaces de règles de prédiction parcimonieuses (cf. [79]). Des vitesses minimax ont été obtenues sur ces modèles et une méthode d'agrégation a donné une version adaptative de ces procédures d'estimation.

Sommaire

1. Problème d'agrégation

1.1. Problématique de l'agrégation. Soit $(\mathcal{Z}, \mathcal{A})$ un espace probabilisable, $\mathcal{P}$ l'ensemble des mesures de probabilité sur cet espace et $F : \mathcal{P} \longmapsto \mathcal{F}$ une fonction sur $\mathcal{P}$ à valeurs dans un espace vectoriel $\mathcal{F}$. Considérons Z, une variable aléatoire à valeurs dans $(\mathcal{Z}, \mathcal{A})$ de mesure de probabilité π. Nous souhaitons estimer $F(\pi)$ à partir de n observations $Z_1, \ldots, Z_n$ de la variable Z. La qualité d'estimation d'un élément $f \in \mathcal{F}$ est mesurée par un risque de la forme :

$$A(f) = \mathbb{E}[Q(Z, f)],$$

où $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$ est une fonction de perte. Dans la majorité des cas, $F(\pi)$ minimise $A(\cdot)$ sur $\mathcal{F}$. Notons A^* le minimum $\min_{f \in \mathcal{F}} A(f)$. La différence $A(f) - A^*$ est appelée **l'excès de risque** de $f \in \mathcal{F}$. Pour un estimateur $\hat{f}_n$, la quantité $A(\hat{f}_n)$ est prise égale à $\mathbb{E}[Q(Z, \hat{f}_n)|Z_1, \ldots, Z_n]$.

Plusieurs problèmes de l'estimation non-paramétrique peuvent s'écrire dans ce cadre.

Exemple 1 : le problème de régression. Soit $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, où $(\mathcal{X}, \mathcal{T})$ est un espace mesurable, et $Z = (X, Y)$ un couple de variables aléatoires sur $\mathcal{Z}$, de distribution de probabilité π, tel que X prend ses valeurs dans $\mathcal{X}$ et Y prend ses valeurs dans $\mathbb{R}$. Supposons que l'espérance conditionnelle de Y par rapport à X existe. Nous souhaitons estimer la fonction de régression de Y en fonction de X :

$$f^*(x) = \mathbb{E}\left[Y|X = x\right], \ \forall x \in \mathcal{X}.$$

En général, la variable Y n'est pas une fonction exacte de X. Ce problème peut être considéré comme un problème d'estimation avec bruit. En tout point X, la sortie Y est concentrée autour de $\mathbb{E}\left[Y|X\right]$ à un bruit additif près ζ de moyenne nulle. Le modèle de régression peut alors s'écrire sous la forme :

$$Y = \mathbb{E}\left[Y|X\right] + \zeta.$$

Soit $\mathcal{F}$ l'ensemble de toutes les fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$. La norme d'une fonction f dans $L^2(\mathcal{X}, \mathcal{T}, P^X)$, où P^X est la distribution de la marginale X, est définie par $||f||^2_{L^2(P^X)} = \int_{\mathcal{X}} f^2(x) dP^X(x)$. Considérons la fonction de perte :

$$Q((x, y), f) = (y - f(x))^2,$$

définie pour tout $(x, y) \in \mathcal{X} \times \mathbb{R}$ et $f \in \mathcal{F}$. Le théorème de Pythagore donne

$$A(f) = \mathbb{E}\left[Q((X,Y),f)\right] = ||f^* - f||^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right].$$

La fonction de régression f^* minimise $A(\cdot)$ sur $\mathcal{F}$ et $A^* = \mathbb{E}[\zeta^2]$.

Exemple 2 : le problème d'estimation de densité. Notons π la mesure de probabilité de Z. Supposons π absolument continue par rapport à une mesure connue μ et notons f^* une version de la densité de π par rapport à cette mesure. Considérons $\mathcal{F}$ l'ensemble de toutes les fonctions de densité sur $(\mathcal{Z}, \mathcal{A}, \mu)$ et la fonction de perte

$$Q(z, f) = -\log f(z),$$

définie pour tout $z \in \mathcal{Z}$ et $f \in \mathcal{F}$. Nous avons

$$A(f) = \mathbb{E}\left[Q(Z,f)\right] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z)) d\pi(z),$$

où $K(f^*|f) = \int_{\mathcal{Z}} \log(f^*(z)/f(z)) d\pi(z)$ est la divergence de Kullback-Leibler entre f^* et f. La fonction de densité f^* minimise $A(\cdot)$ sur $\mathcal{F}$ et $A^* = -\int_{\mathcal{Z}} \log(f^*(z)) d\pi(z)$.

Prenons la distance quadratique pour fonction de perte. Dans ce cas $\mathcal{F}$ est l'ensemble de toutes les fonctions de carré intégrable $L^2(\mathcal{Z}, \mathcal{A}, \mu)$. Pour la fonction de perte

$$Q(z, f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z),$$

définie pour tout $z \in \mathcal{Z}$ et $f \in \mathcal{F}$, le risque d'un élément $f \in \mathcal{F}$ est donné par

$$A(f) = \mathbb{E}\left[Q(Z,f)\right] = ||f^* - f||^2_{L^2(\mu)} - \int_{\mathcal{Z}} (f^*(z))^2 d\mu(z).$$

La fonction de densité f^* minimise $A(\cdot)$ sur $\mathcal{F}$ et $A^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

Exemple 3 : le problème de classification. Soit $(\mathcal{X}, \mathcal{T})$ un espace mesurable. Supposons $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ muni d'une mesure de probabilité inconnue π. Considérons une variable aléatoire $Z = (X, Y)$ à valeurs dans $\mathcal{Z}$ de mesure de probabilité π. Notons par $\mathcal{F}$ l'ensemble des fonctions mesurables de $\mathcal{X}$ sur $\mathbb{R}$. Soit ϕ une fonction de $\mathbb{R}$ dans $\mathbb{R}$. Pour tout $f \in \mathcal{F}$ le $\phi-$risque de f est défini par $A^\phi(f) = \mathbb{E}[Q((X,Y),f)]$, où la fonction de perte est donnée par

$$Q((x, y), f) = \phi(yf(x))$$

pour tout $(x, y) \in \mathcal{X} \times \{-1, 1\}$. La plupart du temps, un minimiseur f^*_ϕ du $\phi-$risque A^ϕ sur $\mathcal{F}$ ou son signe est égal à la *règle de Bayes* (cf. [130]). C'est la règle de prédiction minimisant la fonction de perte $A_0 \overset{\text{def}}{=} A^{\phi_0}$ où $\phi_0(z) = \mathbb{1}_{(z \leq 0)}$ est la fonction de perte usuelle de classification (cf. [47]). La règle de Bayes est définie par

$$f^*_{\phi_0}(x) = \text{sign}(2\eta(x) - 1), \tag{1.1}$$

où $\eta(x) = \mathbb{P}[Y = 1|X = x], \forall x \in \mathcal{X}$ et $\text{sign}(z) = 2\mathbb{1}_{z \geq 0} - 1, \forall z \in \mathbb{R}$. Pour les estimateurs à vecteurs de support (SVM), la fonction de perte est la perte chanrière

$$\phi_1(z) = \max(0, 1 - z), \forall z \in \mathbb{R}.$$

Le risque associé est noté A_1. Certaines fonctions de perte utilisées en classification vérifient l'hypothèse de convexité suivante (cf. [75, 84]) :

DEFINITION 1.1. *Soit $\beta \geq 0$. Une fonction $\phi : \mathbb{R} \longmapsto \mathbb{R}$ deux fois différentiable est dite $\beta-$convexe sur $[-1, 1]$ si*

$$|\phi'(x)|^2 \leq \beta\phi''(x), \forall x \in [-1, 1]. \tag{1.2}$$

Nous présentons maintenant la problématique de l'agrégation de type "sélection de modèle" dans le cadre général.

> Etant donné $\mathcal{F}_0 = \{f_1, \cdots, f_M\}$ un dictionnaire de M éléments de $\mathcal{F}$ et n observations i.i.d. $Z_1, \ldots, Z_n$, nous souhaitons construire un estimateur $\tilde{f}_n$ dont l'excès de risque moyen $\mathbb{E}[A(\tilde{f}_n) - A^*]$ est aussi petit que celui de l'oracle $\min_{f \in \mathcal{F}_0} A(f) - A^*$ à un résidu près. De tels estimateurs sont appellées **agrégats** ou **méthodes d'agrégation**.

Les éléments $f_1, \ldots, f_M$ de $\mathcal{F}_0$ sont aussi appelés "estimateurs faibles". Ils peuvent, par exemple, être construits à partir d'un échantillon préliminaire (considéré gelé) ou être les éléments d'un réseau minimal du modèle, ou le début d'une base, ou des objets simples comme des indicateurs de demi-espace. Par exemple, pour le problème de sélection de modèle, nous disposons de M modèles. Pour chacun d'entre eux, nous construisons un estimateur. Au lieu de prendre toutes les observations pour la construction de ces estimateurs, nous utilisons seulement les m premières : $Z_1, \ldots, Z_m$. Cette phase d'estimation fournit M estimateurs $\hat{f}_m^{(1)}, \ldots, \hat{f}_m^{(M)}$. Passons ensuite à la phase d'apprentissage : les $(n-m)$ observations restantes $Z_{m+1}, \ldots, Z_n$ sont utilisées pour agréger ces estimateurs. Par indépendance des observations, nous pouvons supposer que l'échantillon utilisé lors de la phase d'estimation est gelé. Les estimateurs de base sont ainsi considérés comme des éléments non-aléatoires de $\mathcal{F}$ et plutôt que de travailler avec $(n-m)$ observations, nous supposons disposer de n observations.

Concrètement, nous souhaitons obtenir des **inégalités d'oracle**, c'est-à-dire des inégalités de la forme

$$\mathbb{E}[A(\tilde{f}_n) - A^*] \leq C \min_{f \in \mathcal{F}_0} A(f) - A^* + \gamma(n, M) \tag{1.3}$$

où $C \geq 1$ est une constante et $\gamma(n, M) \geq 0$ est appelé **vitesse d'agrégation**. Les applications statistiques de ce type d'inégalité sont par exemple :

i) Obtenir les vitesses de convergence de certains estimateurs.

ii) Résoudre des problèmes d'adaptation.

iii) Imiter la meilleure procédure parmi les M estimateurs de base aussi bien que possible.

Pour les deux premiers problèmes i) et ii), une inégalité d'oracle où $C > 1$ est suffisante. En revanche, pour le troisième problème, nous avons besoin de considérer des **inégalités d'oracle exactes** (cf. Définition 1.1 ci-dessous et la discussion dans [81]), c'est-à-dire des inégalités du type (1.3) où $C = 1$.

Le cadre d'estimation considéré permet d'avoir accès à un **risque empirique**, donné par la quantité

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, f).$$

C'est une mesure de l'erreur commise par l'estimateur f sur les observations $Z_1, \ldots, Z_n$. Ce critère empirique est à la base de la construction des méthodes d'agrégation. Dans le cadre de cette thèse, nous avons principalement travaillé sur les procédures d'agrégation ci-dessous.

La méthode d'agrégation la plus utilisée est appelée la procédure de **minimisation du risque empirique** (**MRE**) sur $\mathcal{F}_0$. Elle est définie par

$$\tilde{f}_n^{(MRE)} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0} A_n(f). \tag{1.4}$$

La méthode d'agrégation principalement étudiée dans cette thèse est celle d'**agrégation avec poids exponentiels** (**APE**) (cf. [87, 55, 10, 63, 119, 37]). Elle est définie par

$$\tilde{f}_{n,T}^{(APE)} \stackrel{\text{def}}{=} \sum_{f\in\mathcal{F}_0} w_T^{(n)}(f) f, \tag{1.5}$$

où les poids exponentiels $w_T^{(n)}(f)$ sont donnés par :

$$w_T^{(n)}(f) = \frac{\exp\left(-nTA_n(f)\right)}{\sum_{g\in\mathcal{F}_0}\exp\left(-nTA_n(g)\right)}, \quad \forall f\in\mathcal{F}_0, \tag{1.6}$$

où $T^{-1} > 0$ est un paramètre appelé "température" (en référence aux mesures de Gibbs).

Il existe une version récursive de la méthode précédente. Nous allons l'appeller procédure d'**agrégation cumulée avec poids exponentiels** (**ACPE**) (cf. [33, 34, 35, 125, 126, 127]). Elle est définie par :

$$\tilde{f}_{n,T}^{(ACPE)} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_{k,T}^{(APE)}, \tag{1.7}$$

où $\tilde{f}_{k,T}^{(APE)}$ est construit de la même manière que dans (1.5) à partir des k premières observations $Z_1,\ldots,Z_k$ pour le paramètre de température T^{-1} c'est-à-dire :

$$\tilde{f}_{k,T}^{(APE)} = \sum_{f\in\mathcal{F}_0} w_T^{(k)}(f) f, \text{ où } w_T^{(k)}(f) = \frac{\exp\left(-TkA_k(f)\right)}{\sum_{g\in\mathcal{F}_0}\exp\left(-TkA_k(g)\right)}, \quad \forall f\in\mathcal{F}_0.$$

A chacune de ces méthodes d'agrégation, nous pouvons associer associée une version pénalisée. Pour la méthode MRE, l'idée de la pénalisation est bien connue (cf. [11],[92], [93]). La méthode d'agrégation par **minimisation du risque empirique pénalisé** (**MREp**) est définie par :

$$\tilde{f}_n^{(MREp)} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0}\Big[A_n(f) + \operatorname{pen}(f)\Big], \tag{1.8}$$

où pen est une pénalité indépendante de l'échantillon. Pour un aperçu exhaustif des méthodes de ce genre, nous renvoyons le lecteur à [13, 22, 23, 90]. Des versions pénalisées des méthodes APE et ACPE peuvent être aussi proposées (cf. [87] et référence dans cet article).

Pour comparer ces procédures, [114] a introduit une notion d'optimalité pour les méthodes d'agrégation. Cette définition a été donnée en régression gaussienne. Elle se généralise de manière évidente aux autres modèles statistiques (voir [102] pour l'estimation de densité). Dans cette thèse, nous utilisons cette notion généralisée (cf. [38]) qui a la forme suivante.

Définition 1.1. *Nous appellons* **vitesse optimale d'agrégation** *une suite à deux indices $(\gamma(n,M) : n, M\in\mathbb{N})$, s'il existe deux constantes absolues C_1 et C_2 telles que les deux inégalités suivantes sont satisfaites.*

(1) Pour tout sous-ensemble fini $\mathcal{F}_0$ de $\mathcal{F}$ à M éléments, il existe une statistique $\tilde{f}_n$ telle que, quelle que soit la distribution de probabilité sous-jacente π, on a pour

tout $n \geq 1$,

$$\mathbb{E}[A(\tilde{f}_n) - A^*] \leq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_1 \gamma(n, M). \tag{1.9}$$

(2) Il existe un ensemble fini $\mathcal{F}_0$ *à* M *éléments dans* $\mathcal{F}$ *tel que, pour toute statistique* $\bar{f}_n$*, il existe une mesure de probabilité* π*, telle que pour tout* $n \geq 1$

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{F}_0} (A(f) - A^*) + C_2 \gamma(n, M).$$

De plus, quand ces deux inégalités sont satisfaites, on dit que la procédure $\tilde{f}_n$*, apparaissant dans (1.9), est une* **procédure optimale d'agrégation**.

1.2. Historique des principaux résultats obtenus en agrégation. Nemirovski (cf. [98]) a introduit le cadre général de l'étude des méthodes d'agrégation en statistique non-paramétrique. Il a formulé les trois problèmes d'agrégation : le problème d'agrégation de type "sélection de modèle" (MS), le problème d'agrégation convexe (C) et le problème d'agrégation linéaire (L). Étant donné un dictionnaire $\mathcal{F}_0$, l'objectif de (MS), comme nous l'avons déjà énoncé, est de construire une méthode d'agrégation qui a un risque proche de celui de l'oracle $\min_{f \in \mathcal{F}_0} A(f)$. L'objectif de (C) est de fournir une procédure ayant le risque proche de celui de l'oracle convexe $\min_{f \in \mathcal{C}_0} A(f)$, où $\mathcal{C}_0$ est l'enveloppe convexe de $\mathcal{F}_0$. Finalement, le problème (L) vise à produire des procédures atteignant le risque de l'oracle linéaire $\min_{f \in \mathcal{L}_0} A(f)$, où $\mathcal{L}_0$ est l'espace linéaire engendré par $\mathcal{F}_0$. La plus grande partie de la littérature sur les méthodes d'agrégation concerne le problème (MS) (cf. [125, 35, 126, 59, 120, 87, 17, 127, 27, 28, 129, 26, 75]) et le problème (C) (cf. [74, 98, 125, 126, 78, 127, 7, 102, 26, 73, 77, 101]). Quant au problème d'agrégation linéaire, il a principalement été étudié dans [98, 102, 26, 101].

Tsybakov (cf. [114]) a formalisé la notion de vitesse optimale d'agrégation pour les trois types d'agrégation dans l'esprit de la Définition 1.1 (qui traite seulement ici de l'agrégation (MS)). Cette notion fournit un cadre de comparaison des méthodes d'agrégation aussi utile que le cadre minimax pour la mise en compétition des estimateurs. Il a obtenu les vitesses d'agrégation optimales dans le modèle de régression gaussienne. Nous les rappelons dans le tableau suivant :

vitesse optimale (MS)	$(\log M)/n$
vitesse optimale (C)	M/n si $M \leq \sqrt{n}$ $\left(\frac{1}{n}\log[M/\sqrt{n}+1]\right)^{1/2}$ si $M > \sqrt{n}$
vitesse optimale (L)	M/n

La méthode d'agrégation ACPE atteint la vitesse d'agrégation (MS). Un agrégat, obtenu par projection sur l'espace linéaire engendré par le dictionnaire $\mathcal{F}_0$, atteint la vitesse optimale d'agrégation (L). Enfin, un agrégat composite des deux agrégats précédents atteint la vitesse optimale d'agrégation (C).

Dans [7] l'auteur étudie une méthode d'agrégation pour le problème d'agrégation convexe dans le modèle de régression. Cette étude se fait dans le cadre PAC-Bayésien ("PAC" vient de Probablement Approximativement Correct). D'autres procédures d'agrégation, comme la méthode MDL ("Minimum Description Length") de Barron et Cover (cf. [12] et [129]) ont été développées dans ce cadre.

Dans [87], les auteurs utilisent la totalité de l'échantillon pour construire plusieurs estimateurs par projection et des poids exponentiels qui leurs sont associés. L'agrégat ainsi obtenu satisfait une inégalité d'oracle avec une vitesse d'agrégation en $(\log M)/n$, où M est le nombre d'estimateurs par projection construits. Contrairement au protocole habituel, aucune découpe de l'échantillon n'est nécessaire pour obtenir ce résultat.

2. Vitesses rapides de classification sous l'hypothèse de marge

Dans [123, 124, 47], des résultats de borne inférieure ont fait apparaître la vitesse $n^{-1/2}$ comme une vitesse maximale de classification. C'est-à-dire une vitesse en dessous de laquelle on ne peut pas construire de classifieur plus rapide.

Néanmoins, Mammen et Tsybakov (cf. [91]), pour le problème d'analyse discriminante, ont proposé une hypothèse – autre qu'une hypothèse de complexité – qui permet d'améliorer les vitesses de convergence. Tsybakov (cf. [116]) a ensuite proposé une hypothèse similaire dans le cadre de la classification. Elle peut s'énoncer sous deux formes équivalentes données ici :

Hypothèse de marge en classification :

– Il existe un paramètre $\alpha > 0$ et une constante $c > 0$ tels que

$$\mathbb{P}[|2\eta(X) - 1| \leq t] \leq ct^{\alpha}, \forall 0 < t \leq 1/2.$$

– Il existe un paramètre $\kappa \geq 1$ et une constante $C > 0$ tels que pour toute fonction $f : \mathcal{X} \longmapsto \{-1, 1\}$, on a

$$\mathbb{E}[|f(X) - f^*(X)|] \leq C(A_0(f) - A_0^*)^{1/\kappa}. \qquad (\mathbf{HM})(\kappa) \tag{1.10}$$

Les paramètres α et κ vérifient la relation suivante :

$$\kappa = \frac{\alpha + 1}{\alpha} \quad (\kappa = 1 \text{ quand } \alpha = 0).$$

Sous cette hypothèse et une hypothèse de complexité, Tsybakov [116] a proposé des estimateurs atteignant des **vitesses rapides**, c'est-à-dire des vitesses de convergence au delà de $n^{-1/2}$.

Massart et Nédélec [94] ont étudié le comportement d'un estimateur obtenu par minimisation du risque empirique sur des classes de dimension de Vapnik Chervonenkis finie et sous l'hypothèse de marge introduite par Tsybakov. Ils ont obtenu la vitesse de classification suivante :

$$\Big(\frac{V(1 + \log(n/V))}{n}\Big)^{\frac{\kappa}{2\kappa - 1}}, \tag{1.11}$$

quand la règle de Bayes f^* appartient à une classe de VC-dimension finie V. Ils ont donné une borne inférieure, pour $\kappa = 1$, qui correspond à la vitesse (1.11), à un logarithme près.

Les résultats de convergence des estimateurs SVM obtenus par Scovel et Steinwart [109, 108] ont été donnés sous l'hypothèse de marge. En y ajoutant certaines hypothèses de complexité, ils ont obtenu des vitesses de convergence rapides pour les estimateurs SVM. Pour des classifieurs par substitution, Audibert et Tsybakov (cf. [114]) ont aussi obtenu des vitesses rapides minimax.

D'autres vitesses rapides de convergence ont été obtenues dans [112, 77, 8]. Dans [8], l'auteur a étudié la vitesse de convergence d'un estimateur de la forme (1.5) sous l'hypothèse de marge classique, ainsi que pour l'hypothèse de marge suivante :

$$c'(A_0(f) - A_0^*)^{1/\kappa} \leq \mathbb{P}[f(X) \neq f^*(X)] \leq C'(A_0(f) - A_0^*)^{1/\kappa}, \forall f : \mathcal{X} \longmapsto \{-1, 1\}.$$

L'hypothèse de marge a été introduite dans le cadre du problème de classification. Son extension au cadre plus général décrit au paragraphe 1.1 est la suivante (cf. [38]).

Hypothèse de marge (HM) : *La mesure de probabilité π vérifie l'hypothèse de marge (HM)$(\kappa, c, \mathcal{F}_0)$, pour $\kappa \geq 1, c > 0$ et $\mathcal{F}_0$ un sous-ensemble de $\mathcal{F}$ si*

$$\mathbb{E}[(Q(Z,f) - Q(Z,f^*))^2] \leq c(A(f) - A^*)^{1/\kappa}, \forall f \in \mathcal{F}_0. \tag{1.12}$$

Le modèle de régression pour la perte $L^2(P^X)$ et le modèle d'estimation de densité pour la perte Kullback-Leibler et la perte L^2 vérifient l'inégalité (1.12) avec $\kappa = 1$. Le modèle de classification pour des pertes non strictement convexes (comme la perte usuelle ou la perte charnière utilisée pour les SVM), ne vérifie pas cette inégalité. L'hypothèse (HM) doit donc être faite dans le modèle de classification si l'on souhaite pouvoir atteindre des vitesses de convergence ou d'agrégation aussi rapides que dans les modèles de régression ou d'estimation de densité.

3. Travaux de thèse

Un théorème classique (cf. le "no-free-lunch Theorem" du chapitre 7 de [47]) montre que, sans hypothèse de complexité, nous ne pouvons pas construire une règle de classification qui converge à une vitesse donnée vers la règle de Bayes quel que soit le modèle. Il faut alors faire recours à des mesures de la complexité d'un modèle qui sont, entre autres, la dimension de Vapnik Chervonenkis (cf. [47]), l'entropie (cf. [123, 124]), les complexités de Rademacher (cf. [77]). Nous pouvons aussi éviter ces hypothèses en changeant d'objectif. Pour cela, nous nous plaçons dans le cadre des méthodes d'agrégation. Dans cette problématique, aucune hypothèse de complexité n'est requise.

Parallèlement, les hypothèses de marge se sont développées en apprentissage statistique. Sous ces hypothèses, des règles de classification atteignent des vitesses rapides de convergence, comme on l'a déjà discuté dans le paragraphe 2. Les principaux problèmes liés à l'hypothèse de marge sont les suivants : le premier problème est l'adaptation à ce paramètre en simultané avec le paramètre de complexité. En effet, le paramètre de marge est aussi inconnu (au vu des données) que le paramètre de complexité du modèle. Le deuxième problème est d'étudier le comportement des méthodes d'agrégation sous l'hypothèse de marge. Concrètement, nous savons qu'il est plus facile d'estimer sous l'hypothèse de marge, la question est donc : *est-il plus facile d'agréger sous l'hypothèse de marge ?*

Dans cette thèse, une notion plus générale d'hypothèse de marge est proposée (cf. (1.12)). Elle permet de comprendre pourquoi la classification est, en un sens, plus difficile que l'estimation de densité ou la prédiction en régression. On verra plus tard que ceci est dû à la relation entre le biais et la variance décrite par le paramètre de marge. La valeur du paramètre κ détermine, dans certains modèles, la vitesse optimale d'agrégation, qui est parfois $(\log M)/n$ alors que, dans d'autres modèles, elle est $\sqrt{(\log M)/n}$. Le lien entre le paramètre de marge et la convexité de la perte est établi dans cette thèse (cf. [84]). En régression, les fonctions de perte sont généralement convexes, voire strictement convexes, alors qu'en classification, la fonction de perte la plus naturelle n'est pas continue. D'autre part, en classification pour des fonctions de perte $\beta-$convexes (cf. (1.2)), l'hypothèse de marge est naturellement satisfaite avec un paramètre de marge égal à 1 (le cas le plus favorable de l'hypothèse de marge), ce qui explique la vitesse d'agrégation rapide $(\log M)/n$ et les vitesses d'estimation paramétriques en $1/n$. **Le paramètre de marge fait alors le lien entre la vitesse minimax d'estimation (ou la vitesse optimale d'agrégation) et la convexité de la fonction de perte**.

Une autre contribution de cette thèse a été de démontrer que les méthodes classiques de sélection de modèle par minimisation du risque empirique pénalisé sont sous-optimales alors que les méthodes d'agrégation à poids exponentiels atteignent la vitesse optimale d'agrégation.

Nous avons ensuite utilisé les méthodes d'agrégation à poids exponentiels pour résoudre quelques problèmes d'adaptation. Le but d'une méthode d'agrégation est de faire aussi bien que le meilleur estimateur d'un ensemble d'estimateurs de base et cela, sans aucune hypothèse de complexité sur le modèle. Ensuite, pour le problème d'estimation, une hypothèse de complexité sur le modèle est nécessaire. Les techniques d'agrégation, étant libres de toute hypothèse de complexité, elles peuvent s'appliquer pour résoudre des problèmes d'adaptation. Pour cela, il suffit de prendre pour estimateurs faibles, des estimateurs construits en connaissant le paramètre de complexité, pour différentes valeurs de ce paramètre.

3.1. Vitesses optimales d'agrégation sous l'hypothèse de marge. Donnons d'abord quelques résultats principaux de cette thèse concernant les vitesses optimales d'agrégation sous l'hypothèse de marge et un résumé des chapitres traitant de ce sujet.

Dans le cadre général introduit dans le paragraphe 1.1, nous obtenons une inégalité d'oracle exacte de la forme (1.9), dont la vitesse d'agrégation est donnée par la quantité suivante :

(1.13)
$$\gamma(n,M)=\begin{cases}\left(\frac{\min_{f\in\mathcal{F}_0}(A(f)-A^*)^{\frac{1}{\kappa}}\log M}{n}\right)^{1/2} & \text{si } \min_{f\in\mathcal{F}_0}(A(f)-A^*)\geq\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\\ \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{sinon,}\end{cases}$$

où $\kappa\geq 1$ est le paramètre de marge (cf. [38]).

THÉORÈME 1.1. *Soit $\mathcal{F}_0=\{f_1,\dots,f_M\}$ un sous-ensemble de $\mathcal{F}$. Supposons que la probabilité sous-jacente π satisfait l'hypothèse (HM)$(\kappa,c,\mathcal{F}_0)$ pour un $\kappa\geq 1, c>0$ et que la perte vérifie $|Q(Z,f)-Q(Z,f^*)|\leq K$ p.s., pour tout $f\in\mathcal{F}_0$, où $K\geq 1$ est une constante. la procédure de minimisation du risque empirique $\tilde{f}_n=\tilde{f}_n^{(MRE)}$ satisfait*

$$\mathbb{E}[A(\tilde{f}_n)-A^*]\leq\min_{j=1,\dots,M}(A(f_j)-A^*)+C\gamma(n,M), \tag{1.14}$$

où $\gamma(n,M)$ est donné dans (1.13) et $C>0$ est une constante.

De plus, si $Q(z,\cdot)$ est convexe pour π-presque tout $z\in\mathcal{Z}$, alors la procédure avec poids exponentiels $\tilde{f}_n=\tilde{f}_n^{(APE)}$ satisfait l'inégalité d'oracle (1.14).

De cette inégalité d'oracle exacte, des inégalités d'oracle, dans les cadres usuels de l'estimation non-paramétriqueon, peuvent être déduites. En régression bornée, nous obtenons le corollaire suivant :

COROLLAIRE 1.1. *Soit $f_1,\dots,f_M$ des fonctions de $\mathcal{X}$ dans $[0,1]$. Les procédures $\tilde{f}_n=\tilde{f}_n^{(MRE)}$ et $\tilde{f}_n=\tilde{f}_n^{(APE)}$ vérifient, pour tout $\epsilon>0$,*

$$\mathbb{E}[||f^*-\tilde{f}_n||^2_{L^2(P^X)}]\leq(1+\epsilon)\min_{j=1,\dots,M}(||f^*-f_j||^2_{L^2(P^X)})+C\frac{\log M}{\epsilon n}.$$

Dans le modèle d'estimation de densité, nous obtenons le corollaire suivant :

COROLLAIRE 1.2. *Supposons que la fonction de densité à estimer f^* est bornée par $B\geq 1$. Soient $f_1,\dots,f_M$ des fonctions bornées par B. Considérons $\tilde{f}_n$ qui correspond*

indépendamment à la procédure MRE ou à la procédure APE. Pour tout $\epsilon > 0$, nous avons

$$\mathbb{E}[||f^* - \tilde{f}_n||^2_{L^2(\mu)}] \leq (1+\epsilon) \min_{j=1,\dots,M} (||f^* - f_j||^2_{L^2(\mu)}) + C\frac{\log M}{\epsilon n}.$$

En classification, pour la perte charnière ϕ_1 et la perte usuelle ϕ_0, le corollaire suivant est déduit du Théorème 9.1.

COROLLAIRE 1.3. *Soient $\kappa \geq 1$ et $\mathcal{F} = \{f_1, \dots, f_M\}$ une famille de fonctions à valeurs dans $[-1, 1]$. Notons $\mathcal{C}$ l'enveloppe convexe de $\mathcal{F}$. Supposons que π satisfait l'hypothèse de marge (HM)(κ) (cf. (1.10)). Les agrégats $\tilde{f}_n = \tilde{f}_n^{(APE)}$ ou $\tilde{f}_n = \tilde{f}_n^{(MRE)}$ satisfont pour tous entiers n, M et tout $a > 0$ les inégalités suivantes*

$$\mathbb{E}\left[A_1(\tilde{f}_n) - A_1^*\right] \leq (1+a) \min_{f \in \mathcal{C}} (A_1(f) - A_1^*) + C(a) \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}},$$

où A_1 est le risque correspondant à la perte charnière et $C(a) > 0$ est une constante.

Pour le risque de Bayes de classification A_0, la procédure MRE vérifie

$$\mathbb{E}\left[A_0(\tilde{f}_n^{(MRE)}) - A_0^*\right] \leq (1+a) \min_{f \in \mathcal{C}} (A_0(f) - A_0^*) + C(a) \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}.$$

Nous n'avons pas besoin d'hypothèse de marge dans le cas de l'estimation de densité et de régression (cf. corollaires 9.1 et 9.2), car pour ces deux modèles le paramètre de marge vaut naturellement $\kappa = 1$. En revanche, pour la classification, le vitesse d'agrégation dépend du paramètre de marge et varie entre

$$\sqrt{\frac{\log M}{n}} (\text{pour } \kappa = +\infty) \text{ et } \frac{\log M}{n} (\text{pour } \kappa = 1).$$

Le problème de classification permet de considérer plusieurs types de convexité pour la fonction de perte. Nous avons introduit une échelle continue de fonctions de perte pour étudier le comportement de la vitesse optimale d'agrégation en fonction de la perte. Considérons l'ensemble $\{\phi_h : h \geq 0\}$ de fonctions de perte données par

$$\phi_h(x) = \begin{cases} h\phi_1(x) + (1-h)\phi_0(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \tag{1.15}$$

pour tout $x \in \mathbb{R}$, où ϕ_0 est la fonction de perte $0-1$ et ϕ_1 est la perte charnière.

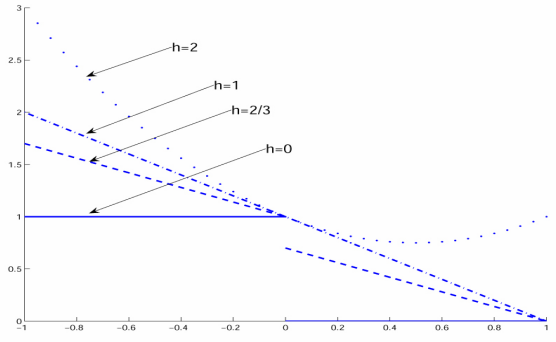

FIG. 1. Exemples de fonctions de perte de la famille $\{\phi_h : h \geq 0\}$ pour $h = 0$ (perte $0-1$), $h = 2/3$, $h = 1$ (perte charnière) et $h = 2$.

Nous avons choisi cet ensemble de fonctions de perte pour sa représentativité des différents types de convexité. Pour tout $h > 1$, ϕ_h est β_h-convexe sur $[-1, 1]$ pour un

$\beta_h \stackrel{\text{def}}{=} (2h-1)^2/(2(h-1)) \geq 2$, pour $h=1$ la fonction de perte est linéaire (c'est la perte charnière) et pour $h<1$, ϕ_h n'est pas convexe.

Pour le cas $h>1$, l'hypothèse de marge est naturellement satisfaite avec le paramètre de marge $\kappa=1$. Nous obtenons alors comme vitesse optimale d'agrégation pour les fonctions de perte ϕ_h :

$$\frac{\log M}{n} \quad (\text{ vitesse rapide d'agrégation}).$$

Pour le cas $h \leq 1$, l'hypothèse de marge n'est pas satisfaite, ce qui explique la faible vitesse d'agrégation

$$\sqrt{\frac{\log M}{n}} \quad (\text{ vitesse lente d'agrégation}).$$

Néanmoins, sous l'hypothèse de marge de paramètre de marge κ, nous obtenons la vitesse d'agrégation (1.13). Pour l'optimalité de cette vitesse d'agrégation, nous avons donné des théorèmes de borne inférieure (cf. [81, 84]). Cependant il reste toujours possible que la vitesse optimal d'agrégation sous l'hypothèse de marge soit

$$\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}}.$$

Si tel était le cas, la procédure MRE ne serait pas une procédure optimale d'agrégation (cf. [84]). Ce problème reste encore ouvert.

Donnons maintenant le descriptif des résultats sur ce sujet par chapitre.

Chapitre 2 : Dans ce chapitre (cf. [80]), nous montrons que la vitesse optimale d'agrégation dans le modèle de densité pour la divergence de Kullback-Leibler est

$$\frac{\log M}{n}.$$

La procédure d'agrégation atteignant cette vitesse est l'agrégat ACPE introduit en (1.7). Des inégalités de borne inférieure sont données pour la perte en variation totale et la perte Hellinger (elles sont probablement optimales).

Chapitre 3 : Ce travail [81] porte sur l'optimalité des procédures d'agrégation introduites en (1.4), (1.5) et (1.7) dans le modèle de classification, sous l'hypothèse de marge, pour la perte usuelle et la perte charnière.

Premièrement, sans hypothèse de marge, les trois procédures sont optimales et atteignent la vitesse optimale d'agrégation

$$\sqrt{\frac{\log M}{n}},$$

pour la perte charnière. Pour la perte usuelle, la vitesse optimale d'agrégation est aussi $\sqrt{(\log M)/n}$, et la procédure MRE (1.4) est une procédure optimale d'agrégation.

Sous l'hypothèse de marge, la vitesse d'agrégation est donnée en 1.13. Un résultat de type borne inférieure concernant l'optimalité de cette vitesse d'agrégation est donné.

Chapitre 5 : Ce travail porte sur le problème d'agrégation convexe en régression. Etant donné un dictionnaire de M fonctions, la vitesse optimale d'agrégation convexe est (cf. paragraphe 1.2)

$$M/n \text{ si } M \leq \sqrt{n} \text{ et } \Big(\frac{1}{n}\log[M/\sqrt{n}+1]\Big)^{1/2} \text{ si } M > \sqrt{n}.$$

Dans ce chapitre, nous montrons que, sous une hypothèse géométrique (disant que les estimateurs faibles sont dans un demi-cône), nous pouvons construire une procédure qui

imite la meilleure combinaison convexe à la vitesse

$$\frac{\log M}{n}.$$

Cette vitesse est habituellement la vitesse optimale d'agrégation pour le problème de (MS) agrégation (cf. paragraphe 1.2). Nous obtenons donc une amélioration de la vitesse due à la condition géométrique.

3.2. Sous-optimalité des méthodes de minimisation du risque empirique pénalisé. Dans les modèles à forte marge ($\kappa = 1$), les procédures de minimisation du risque empirique pénalisé n'arrivent pas à atteindre la vitesse optimale d'agrégation

$$\frac{\log M}{n}.$$

Concrètement, des exemples d'estimateurs faibles et de lois sous-jacentes sont exhibés pour lesquels ces procédures ne peuvent pas imiter l'oracle à la vitesse plus rapide que

$$\sqrt{\frac{\log M}{n}}.$$

Par conséquent, en estimation de densité, en régression et en classification (pour des fonctions de perte $\beta-$convexes), il est préférable d'utiliser des procédures d'agrégation à poids exponentiels plutôt que les méthodes usuelles de minimisation du risque empirique pénalisé pour construire des procédures adaptatives. Ces résultats sont donnés dans le chapitre 4 (cf. [84]).

Chapitre 4 : Dans [75], il est prouvé que la méthode d'agrégation ACPE (cf. 1.7) pour un paramètre de température T^{-1} convenablement choisi peut atteindre la vitesse d'agrégation

$$\frac{\log M}{n}$$

dans des modèles ayant une certaine propriété de convexité sur le risque. Dans ce chapitre nous montrons, sous certaines hypothèses sur la pénalité, que la méthode MREp (cf. 1.8) ne peut pas imiter l'oracle à la vitesse plus rapide que

$$\sqrt{\frac{\log M}{n}}.$$

Cette méthode n'est par conséquent pas optimale.

Nous avons fait ressortir le phénomène suivant : pour l'échelle de fonctions de perte (1.15), la vitesse optimale d'agrégation est $\sqrt{(\log M)/n}$ pour les pertes ϕ_h où $h \leq 1$. Elle est atteinte par la procédure MRE. Pour la perte charnière ($h = 1$), la vitesse optimale d'agrégation est atteinte par les trois procédures MRE, APE et ACPE. Pour les fonctions de perte ϕ_h avec $h > 1$, nous obtenons la vitesse rapide d'agrégation $(\log M)/n$, atteinte par la procédure ACPE. Dans ce cas, la procédure MREp (et donc aussi MRE) ne peut pas atteindre cette vitesse optimale d'agrégation, puisque nous avons exhibé un exemple pour lequel cette méthode ne peut pas imiter l'oracle à la vitesse plus rapide que $\sqrt{(\log M)/n}$.

Dans le chapitre 4, d'autres arguments concernant l'optimalité de la vitesse d'agrégation définie dans (1.13) sont donnés.

Ce chapitre met en exergue le lien étroit entre la convexité de la perte et la possibilité d'agréger à la vitesse rapide $(\log M)/n$, par l'intermédiaire du paramètre de marge.

Le tableau suivant rappelle les principaux résultats obtenus sur l'optimalité des méthodes d'agrégation dans le cadre de la classification sous l'hypothèse de marge. Pour

cela, l'échelle continue de fonctions de perte $\{\phi_h, h \geq 0\}$ (cf. (1.15)) est prise pour ensemble de fonctions tests.

Fonction de perte ϕ_h	$h=0$ perte 0-1	$0<h<1$	$h=1$ perte charnière	$h>1$ perte $\beta-$convexe
Hypothèse de marge	non automatiquement satisfaite ($\kappa=+\infty$)			automatiquement vérifiée ($\kappa=1$)
Vitesse optimale d'agrégation	(cf. 1.13) sous (HM)(κ) (conjecture)			$(\log M)/n$
Procédure optimale d'agrégation	ERM (conjecture)		ERM ou AEW (conjecture)	CAEW
MRE ou MREp	Optimale (conjecture)			Sous-optimale
APE	?		Optimale	optimale (conjecture)
ACPE	?			Optimale

3.3. Vitesses rapides de classification pour des règles de Bayes parcimonieuses. Le chapitre 6 rassemble des résultats d'approximation en classification. Une grande différence entre le modèle de classification et ceux de régression et d'estimation de densités réside dans le fait qu'en classification, le statisticien ne cherche pas à approcher le meilleur prédicteur possible (la règle de Bayes). Une telle approximation dans les classes habituelles de régularité, utilisées en régression et estimation de densités, n'est pas sensé dans le modèle de classification.

Nous avons proposé une autre approche en considérant des classes de fonctions à valeurs dans $\{-1,1\}$ pouvant être approchées en norme L^2 par des objets paramétriques dont les valeurs appartiennent également à $\{-1,1\}$. Sous des hypothèses sur la marge et le design du modèle, le risque de Bayes en classification est équivalent au risque L^2. Nous avons obtenu des vitesses minimax sur ces classes, atteintes par des "estimateurs par projection" sur ces espaces. Ces estimateurs se sont avérés être des arbres dyadiques.

Les classes de fonctions introduites sont dans le même esprit que les ellipsoïdes de Sobolev mis à part le fait qu'elles ne contiennent que des fonctions à valeurs dans $\{-1,1\}$ et sont plutôt à envisager comme une classe d'arbres dyadiques possédant une représentation parcimonieuse.

De plus, l'utilisation d'une méthode d'agrégation à poids exponentiel a donné une version adaptative de ces estimateurs. L'estimateur ainsi obtenu peut s'interpréter comme réalisant une procédure multi-échelle dans le même esprit que les estimateurs par projection en régression et en estimation de densité.

3.4. Applications aux modèles concrêts. Chapitre 7 : Dans ce chapitre (cf.[82]), nous utilisons la méthode d'agrégation introduite dans (1.5) pour construire des estimateurs implémentables et adaptatifs à la fois au paramètre de marge et à celui de complexité. Nous proposons une construction d'estimateurs SVM adaptatifs, en agrégeant des estimateurs faibles SVM par la méthode d'agrégation APE.

Les paramètres de marge et de complexité sont inconnus en pratique alors, pour profiter de grandes marges ou d'une faible complexité du modèle, nous devons être capable de construire des estimateurs indépendants de ces paramètres, apprenant aussi vite que des procédures ayant accès à ces paramètres. La procédure proposée dans ce chapitre est implémentable et réalise cet objectif.

Nous avons ensuite utilisé la méthode introduite en (1.5) pour l'agrégation d'estimateurs proposés par [116]. Elle fournit un estimateur adaptatif, plus simple que celui utilisé dans [116], atteignant la vitesse minimax d'estimation, (contrairement au résultat de [116], qui souffrait d'une perte de vitesse logarithmique lors de la phase d'adaptation). Un résultat similaire utilisant une autre méthode a été obtenu dans [77].

Chapitre 8 : Dans ce chapitre (cf. Chapter 8), nous montrons que la vitesse otpimale d'agrégation sous hypothèse de marge pour des inégalités d'oracle non exactes (cf. 1.3 pour $C > 1$) est

$$\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}}.$$

Nous avons ensuite utilisé cette inégalité d'oracle, satisfaite par la méthode APE (définie en 1.5), pour construire des estimateurs en agrégeant des classifieurs "par substitution", c'est-à-dire de la forme :

$$f_\eta(x) = \mathrm{sign}(2\eta(x) - 1),$$

où η varie dans un $\epsilon-$réseau de l'espace de Hölder pour la norme infinie et pour un ϵ convenablement choisi. L'estimateur ainsi construit est minimax. Nous avons utilisé une deuxième fois l'inégalité d'oracle de ce chapitre pour rendre adaptatif, en la complexité et la marge, ces estimateurs. Le désavantage de la méthode de construction de cet estimateur est qu'elle nécessite de connaître un réseau optimal des classes de Hölder pour la norme L^∞ et d'agréger un grand nombre d'estimateurs faibles (il y en a un nombre exponentiel en n). Cette méthode est donc difficilement implémentable dans la pratique. Pour résoudre ce problème, nous avons utilisé une troisième fois l'inégalité d'oracle démontrée dans ce chapitre pour agréger des classifieurs par substitution, où l'estimateur de la fonction de régression est un estimateur par polynômes locaux donc implémentable. Finalement, cette dernière procédure fournit un estimateur implémentable minimax et adaptatif en la marge et la régularité.

Chapitre 9 : Dans ce chapitre réalisé en collaboration avec Christophe Chesneau (cf.[38]), l'inégalité d'oracle du Théorème 9.1 et ses deux Corollaires (cf. Corollaires 9.1 et 9.2) sont prouvés. Ces résultats, en densité et régression, ont permis d'obtenir la vitesse de convergence d'un estimateur obtenu par agrégation d'estimateurs par ondelettes seuillés. Cet estimateur est minimax adaptatif en la régularité sur tous les espaces de Besov. De plus, cette procédure est implémentable car ne requiert que l'agrégation de $\log n$ estimateurs.

Chapitre 10 : Dans ce chapitre, nous répondons positivement à une conjecture de Stone, posée en 1982 dans [110]. C'est un travail commun avec Stéphane Gaïffas.

La problématique est la suivante. Plaçons nous dans le modèle de régression gaussienne

$$Y = f^*(X) + \sigma(X)\zeta,$$

où ζ est un bruit gaussien centré, réduit et indépendant de X, $f^* : \mathbb{R}^d \longmapsto \mathbb{R}$ est la fonction de régression introduite dans le paragraphe 1.1 et $\sigma : \mathbb{R}^d \longmapsto \mathbb{R}$ est une fonction vérifiant $0 < \sigma_0 \leq \sigma(X) \leq \sigma$ a.s.. La variable aléatoire Y est à valeurs réelles et X est une variable aléatoire à valeurs dans $\mathbb{R}^d$. L'hypothèse principale de ce chapitre est de supposer qu'il existe une direction $\theta \in \mathcal{S}_{d-1}$, où $\mathcal{S}_{d-1}$ est la sphère unité de $\mathbb{R}^d$, telle que

$$f^*(x) = g(\theta^t x), \forall x \in \mathbb{R}^d,$$

où la fonction $g : \mathbb{R} \longmapsto \mathbb{R}$, généralement appelée *fonction de lien*, est inconnue ainsi que le vecteur θ. Ceci est une hypothèse de réduction de dimension appelée hypothèse du "single index". Supposons ensuite, que g est $\alpha-$Höldérienne (ceci est l'hypothèse de complexité

du chapitre). La vitesse minimax dans ce modèle (sans l'hypothèse du "single-index") est $n^{-2\alpha/(2\alpha+d)}$ si f est $\alpha-$Höldérienne. Elle est donc d'autant moins rapide que la dimension d est grande. La conjecture de Stone consiste à prouver qu'il existe un estimateur $\hat{f}_n$ de f^* ayant une vitesse de convergence aussi rapide que dans un modéle uni-dimensionnel ($d = 1$), c'est-à-dire tel que son risque quadratique vérifie

$$\mathbb{E}[||\hat{f}_n - f^*||^2_{L^2(P^X)}] \leq Cn^{-\frac{2\alpha}{2\alpha+1}}. \tag{1.16}$$

La plupart des articles dans ce domaine proposent d'estimer le vecteur inconnu θ et d'utiliser la valeur estimée pour construire un estimateur uni-dimensionnel de g. L'approche que nous introduisons dans ce chapitre est de s'adapter en θ plutôt que de l'estimer. Pour cela nous avons agrégé des estimateurs par polynômes locaux unidimensionnels dans plusieures directions formant une grille de la sphère $\mathcal{S}_{d-1}$. Nous montrons une inégalité d'oracle satisfaite par une méthode d'agrégation avec poids exponentiels dépendant d'un paramètre de température. Cette inégalité d'oracle montre que cet agrégat s'adapte automatiquement en la direction. Parallèlement, pour une grille assez fine en la régularité α, cette procédure s'adapte aussi à la régularité. La vitesse obtenue par cet estimateur est la vitesse minimax prédite par Stone donnée en 1.16.

Des résultats de simulation montrent les meilleures performances des méthodes d'agrégation à poids exponentiels par rapport aux méthodes de minimisation du risque empirique. Dans le modèle du "single index", nous avons obtenu le graphique suivant, montrant l'évolution du risque quadratique de l'agrégat (en ordonnée) en fonction du paramètre de température. Pour une température proche de zéro, l'agrégat est une moyenne uniforme des estimateurs de base. Pour des grandes températures, l'agrégat est la méthode d'agrégation MRE.

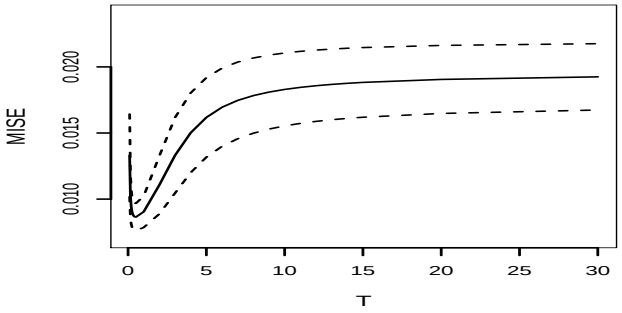

FIG. 2. Risque quadratique de l'agrégat en fonction de l'inverse de la température $T \geq 0$ (écart-type en ligne pointillée). Pout $T = 0$, nous avons le risque quadratique de l'agrégat à poids uniformes. Asymptotiquement, quand $T \longrightarrow +\infty$, on obtient le risque quadratique de l'ERM.

Le minimum atteint par cette fonction, nous permet de conjecturer l'existence d'une température optimale pour laquelle le risque quadratique de l'agrégat APE est minimale.

Part 1

Fast Rates and Optimality of Aggregation Procedures

CHAPTER 2

Lower Bounds and Aggregation in Density Estimation

In this chapter we prove the optimality of an aggregation procedure. We prove lower bounds for aggregation of model selection type of M density estimators for the Kullback-Leibler divergence (KL), the Hellinger's distance and the L_1-distance. The lower bound, with respect to the KL distance, can be achieved by the on-line type estimate suggested, among others, by [125]. Combining these results, we state that $\log M/n$ is an optimal rate of aggregation in the sense of [114], where n is the sample size.

Contents

The material of this chapter has been published in the *Journal of Machine Learning Research* (cf. [80]).

1. Introduction

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and ν be a σ-finite measure on $(\mathcal{X}, \mathcal{A})$. Let $D_n = (X_1, \dots, X_n)$ be a sample of n i.i.d. observations drawn from an unknown probability of density f on $\mathcal{X}$ with respect to ν. Consider the estimation of f from D_n.

Suppose that we have $M \geq 2$ different estimators $\hat{f}_1, \dots, \hat{f}_M$ of f. [33], [125], [98], [74], [114] and[35] have studied the problem of model selection type aggregation. It consists in construction of a new estimator $\tilde{f}_n$ (called *aggregate*) which is approximatively at least as good as the best among $\hat{f}_1, \dots, \hat{f}_M$. In most of these papers, this problem is solved by using a kind of cross-validation procedure. Namely, the aggregation is based on splitting the sample in two independent subsamples D_m^1 and D_l^2 of sizes m and l respectively, where $m \gg l$ and $m + l = n$. The size of the first subsample has to be greater than the one of the second because it is used for the true estimation, that is for the construction of the M estimators $\hat{f}_1, \dots, \hat{f}_M$. The second subsample is used for the adaptation step of the procedure, that is for the construction of an aggregate $\tilde{f}_n$, which has to mimic, in a certain sense, the behavior of the best among the estimators $\hat{f}_i$. Thus, $\tilde{f}_n$ is measurable w.r.t. the whole sample D_n unlike the first estimators $\hat{f}_1, \dots, \hat{f}_M$. Actually, [98] and [74] did not focus on model selection type aggregation. These papers give a bigger picture about the general topic of procedure aggregation and [125] complemented their results. [114] improved these results and formulated the three types of aggregation problems (cf. [114]).

One can suggest different aggregation procedures and the question is how to look for an optimal one. A way to define optimality in aggregation in a minimax sense for a regression problem is suggested in [114]. Based on the same principle we can define optimality for density aggregation. In this chapter we will not consider the sample splitting and concentrate only on the adaptation step, i.e. on the construction of aggregates (following [98], [74], [114]). Thus, the first subsample is fixed and instead of estimators $\hat{f}_1, \dots, \hat{f}_M$, we have fixed functions $f_1, \dots, f_M$. Rather than working with a part of the initial sample we will use, for notational simplicity, the whole sample D_n of size n instead of a subsample D_l^2.

The aim of this chapter is to prove the optimality, in the sense of [114], of the aggregation method proposed by Yang, for the estimation of a density on $(\mathbb{R}^d, \lambda)$ where λ is the Lebesgue measure on $\mathbb{R}^d$. This procedure is a convex aggregation with weights which can be seen in two different ways. Yang's point of view is to express these weights in function of the likelihood of the model, namely

$$\tilde{f}_n(x) = \sum_{j=1}^{M} \tilde{w}_j^{(n)} f_j(x), \quad \forall x \in \mathcal{X}, \tag{2.1}$$

where the weights are $\tilde{w}_j^{(n)} = (n+1)^{-1} \sum_{k=0}^{n} w_j^{(k)}$ and

$$w_j^{(k)} = \frac{f_j(X_1) \dots f_j(X_k)}{\sum_{l=1}^{M} f_l(X_1) \dots f_l(X_k)}, \ \forall k = 1, \dots, n \text{ and } w_j^{(0)} = \frac{1}{M}. \tag{2.2}$$

And the second point of view is to write these weights as exponential ones, as used in [10], [35], [63], [26], [72] and Chapter 7, for different statistical models. Define the empirical Kullback-Leibler loss $K_n(f) = -(1/n) \sum_{i=1}^{n} \log f(X_i)$ (keeping only the term independent of the underlying density to estimate) for all density f. We can rewrite these weights as exponential weights:

$$w_j^{(k)} = \frac{\exp(-kK_k(f_j))}{\sum_{l=1}^{M} \exp(-kK_k(f_l))}, \quad \forall k = 0, \dots, n.$$

Most of the results on convergence properties of aggregation methods are obtained for the regression and the gaussian white noise models. Nevertheless, [33, 35], [48], [125], [130] and [102] have explored the performances of aggregation procedures in the density estimation framework. Most of them have established upper bounds for some procedure and do not deal with the problem of optimality of their procedures. [98], [74] and [125] state lower bounds for aggregation procedure in the regression setup. To our knowledge, lower bounds for the performance of aggregation methods in density estimation are available only in [102]. Their results are obtained with respect to the mean squared risk. [33] and [125] construct procedures and give convergence rates w.r.t. the KL loss. One aim of this chapter is to prove optimality of one of these procedures w.r.t. the KL loss. Lower bounds w.r.t. the Hellinger's distance and L_1-distance (stated in Section 3) and some results of [17] and [48] (recalled in Section 4) suggest that the rates of convergence obtained in Theorem 2.2 and 2.4 are optimal in the sense given in Definition 2.1. In fact, an approximate bound can be achieved, if we allow the leading term in the RHS of the oracle inequality (i.e. in the upper bound) to be multiplied by a constant greater than one.

The chapter is organized as follows. In Section 2 we give a Definition of optimality, for a rate of aggregation and for an aggregation procedure, and our main results. Lower bounds, for different loss functions, are given in Section 3. In Section 4, we recall a result

of [125] about an exact oracle inequality satisfied by the aggregation procedure introduced in (2.1).

2. Main definition and main results

To evaluate the accuracy of a density estimator we use the Kullback-leibler (KL) divergence, the Hellinger's distance and the L_1-distance as loss functions. The *KL divergence* is defined for all densities f, g w.r.t. a $\sigma-$finite measure ν on a space $\mathcal{X}$, by

$$K(f|g) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{f}{g}\right) f d\nu & \text{if } P_f \ll P_g; \\ +\infty & \text{otherwise,} \end{cases}$$

where P_f (respectively P_g) denotes the probability distribution of density f (respectively g) w.r.t. ν. *Hellinger's distance* is defined for all non-negative measurable functions f and g by

$$H(f,g) = \left\| \sqrt{f} - \sqrt{g} \right\|_2,$$

where the L_2-norm is defined by $\|f\|_2 = \left(\int_{\mathcal{X}} f^2(x) d\nu(x)\right)^{1/2}$ for all functions $f \in L_2(\mathcal{X}, \nu)$. The L_1-*distance* is defined for all measurable functions f and g by

$$v(f,g) = \int_{\mathcal{X}} |f-g| d\nu.$$

The main goal of this chapter is to find optimal rate of aggregation in the sense of the definition given below. This definition is an analog, for the density estimation problem, of the one in [114] for the regression problem.

Definition 2.1. *Take $M \geq 2$ an integer, $\mathcal{F}$ a set of densities on $(\mathcal{X}, \mathcal{A}, \nu)$ and $\mathcal{F}_0$ a set of functions on $\mathcal{X}$ with values in $\mathbb{R}$ such that $\mathcal{F} \subseteq \mathcal{F}_0$. Let d be a loss function on the set $\mathcal{F}_0$. A sequence of positive numbers $(\psi_n(M))_{n\in\mathbb{N}^*}$ is called* ***optimal rate of aggregation of M functions in $(\mathcal{F}_0, \mathcal{F})$ w.r.t. the loss d*** *if :*

(i) There exists a constant $C < \infty$, depending only on $\mathcal{F}_0, \mathcal{F}$ and d, such that for all functions $f_1, \ldots, f_M$ in $\mathcal{F}_0$ there exists an estimator $\tilde{f}_n$ (aggregate) of f such that

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_f \left[d(f, \tilde{f}_n) \right] - \min_{i=1,\ldots,M} d(f, f_i) \right] \leq C\psi_n(M), \quad \forall n \in \mathbb{N}^*. \tag{2.3}$$

(ii) There exist some functions $f_1, \ldots, f_M$ in $\mathcal{F}_0$ and $c > 0$ a constant independent of M such that for all estimators $\hat{f}_n$ of f,

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_f \left[d(f, \hat{f}_n) \right] - \min_{i=1,\ldots,M} d(f, f_i) \right] \geq c\psi_n(M), \quad \forall n \in \mathbb{N}^*. \tag{2.4}$$

Moreover, when the inequalities (2.3) and (2.4) are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (2.3), is an **optimal aggregation procedure w.r.t. the loss d.**

Let $A > 1$ be a given number. In this chapter we are interested in the estimation of densities lying in

$$\mathcal{F}(A) = \{\text{densities bounded by } A\} \tag{2.5}$$

and, depending on the used loss function, we aggregate functions in $\mathcal{F}_0$ which can be:

(1) $\mathcal{F}_K(A) = \{\text{densities bounded by } A\}$ for KL divergence,
(2) $\mathcal{F}_H(A) = \{\text{non-negative measurable functions bounded by } A\}$ for Hellinger's distance,

(3) $\mathcal{F}_v(A) = \{\text{measurable functions bounded by } A\}$ for the L_1-distance.

The main result of this chapter, obtained by using Theorem 2.5 and assertion (2.6) of Theorem 2.3, is the following Theorem.

THEOREM 2.1. *Let $A > 1$. Let M and n be two integers such that $\log M \leq 16(\min(1, A-1))^2 n$. The sequence*

$$\psi_n(M) = \frac{\log M}{n}$$

is an optimal rate of aggregation of M functions in $(\mathcal{F}_K(A), \mathcal{F}(A))$ (introduced in (2.5)) w.r.t. the KL divergence loss. Moreover, the aggregation procedure with exponential weights, defined in (2.1), achieves this rate. So, this procedure is an optimal aggregation procedure w.r.t. the KL-loss.

Moreover, if we allow the leading term "$\min_{i=1,\dots,M} d(f, f_i)$", in the upper bound and the lower bound of Definition 2.1, to be multiplied by a constant greater than one, then the rate $(\psi_n(M))_{n\in\mathbb{N}^*}$ is said "near optimal rate of aggregation". Observing Theorem 2.6 and the result of [48] (recalled at the end of Section 4), the rates obtained in Theorems 2.2 and 2.4:

$$\left(\frac{\log M}{n}\right)^{\frac{q}{2}}$$

are near optimal rates of aggregation for the Hellinger's distance and the L_1-distance to the power q, where $q > 0$.

3. Lower bounds

To prove lower bounds of type (2.4) we use the following lemma on minimax lower bounds which can be obtained by combining Theorems 2.2 and 2.5 in [115]. We say that d is a semi-distance on Θ if d is symmetric, satisfies the triangle inequality and $d(\theta, \theta) = 0$.

LEMMA 2.1. *Let d be a semi-distance on the set of all densities on $(\mathcal{X}, \mathcal{A}, \nu)$ and w be a non-decreasing function defined on $\mathbb{R}_+$ which is not identically 0. Let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Let $\mathcal{C}$ be a finite set of densities on $(\mathcal{X}, \mathcal{A}, \nu)$ such that $card(\mathcal{C}) = M \geq 2$,*

$$\forall f, g \in \mathcal{C}, f \neq g \Longrightarrow d(f, g) \geq 4\psi_n > 0,$$

and the KL divergences $K(P_f^{\otimes n}|P_g^{\otimes n})$, between the product probability measures corresponding to densities f and g respectively, satisfy, for some $f_0 \in \mathcal{C}$,

$$\forall f \in \mathcal{C},\ K(P_f^{\otimes n}|P_{f_0}^{\otimes n}) \leq (1/16)\log(M).$$

Then,

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{C}} \mathbb{E}_f\left[w(\psi_n^{-1} d(\hat{f}_n, f))\right] \geq c_1,$$

where $\inf_{\hat{f}_n}$ denotes the infimum over all estimators based on a sample of size n from an unknown distribution with density f and $c_1 > 0$ is an absolute constant.

Now, we give a lower bound of the form (2.4) for the three different loss functions introduced in the beginning of the section. Lower bounds are given in the problem of estimation of a density on $\mathbb{R}^d$, namely we have $\mathcal{X} = \mathbb{R}^d$ and ν is the Lebesgue measure on $\mathbb{R}^d$.

THEOREM 2.2. *Let M be an integer greater than 2, $A > 1$ and $q > 0$ be two numbers. We have for all integers n such that $\log M \leq 16(\min(1, A-1))^2 n$,*

$$\sup_{f_1,\dots,f_M \in \mathcal{F}_H(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[H(\hat{f}_n, f)^q \right] - \min_{j=1,\dots,M} H(f_j, f)^q \right] \geq c \left(\frac{\log M}{n} \right)^{q/2},$$

where c is a positive constant which depends only on A and q. The sets $\mathcal{F}(A)$ and $\mathcal{F}_H(A)$ are defined in (2.5) when $\mathcal{X} = \mathbb{R}^d$ and the infimum is taken over all the estimators based on a sample of size n.

Proof : For all densities $f_1, \dots, f_M$ bounded by A we have,

$$\sup_{f_1,\dots,f_M \in \mathcal{F}_H(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[H(\hat{f}_n, f)^q \right] - \min_{j=1,\dots,M} H(f_j, f)^q \right] \geq \inf_{\hat{f}_n} \sup_{f \in \{f_1,\dots,f_M\}} \mathbb{E}_f \left[H(\hat{f}_n, f)^q \right].$$

Thus, to prove Theorem 1, it suffices to find M appropriate densities bounded by A and to apply Lemma 1 with a suitable rate.

We consider D the smallest integer such that $2^{D/8} \geq M$ and $\Delta = \{0,1\}^D$. We set $h_j(y) = h(y - (j-1)/D)$ for all $y \in \mathbb{R}$, where $h(y) = (L/D)g(Dy)$ and $g(y) = \mathbb{1}_{[0,1/2]}(y) - \mathbb{1}_{(1/2,1]}(y)$ for all $y \in \mathbb{R}$ and $L > 0$ will be chosen later. We consider

$$f_\delta(x) = \mathbb{1}_{[0,1]^d}(x) \left(1 + \sum_{j=1}^{D} \delta_j h_j(x_1) \right), \quad \forall x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

for all $\delta = (\delta_1, \dots, \delta_D) \in \Delta$. We take L such that $L \leq D \min(1, A-1)$ thus, for all $\delta \in \Delta$, f_δ is a density bounded by A. We choose our densities $f_1, \dots, f_M$ in $\mathcal{B} = \{f_\delta : \delta \in \Delta\}$, but we do not take all of the densities of $\mathcal{B}$ (because they are too close to each other), but only a subset of $\mathcal{B}$, indexed by a separated set (this is a set where all the points are separated from each other by a given distance) of Δ for the *Hamming distance* defined by $\rho(\delta^1, \delta^2) = \sum_{i=1}^{D} I(\delta_i^1 \neq \delta_i^2)$ for all $\delta^1 = (\delta_1^1, \dots \delta_D^1), \delta^2 = (\delta_1^2, \dots, \delta_D^2) \in \Delta$. Since $\int_{\mathbb{R}} h d\lambda = 0$, we have

$$\begin{aligned} H^2(f_{\delta^1}, f_{\delta^2}) &= \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} I(\delta_j^1 \neq \delta_j^2) \left(1 - \sqrt{1 + h_j(x)} \right)^2 dx \\ &= 2\rho(\delta^1, \delta^2) \int_0^{1/D} \left(1 - \sqrt{1 + h(x)} \right) dx, \end{aligned}$$

for all $\delta^1 = (\delta_1^1, \dots, \delta_D^1), \delta^2 = (\delta_1^2, \dots, \delta_D^2) \in \Delta$. On the other hand the function $\varphi(x) = 1 - \alpha x^2 - \sqrt{1+x}$, where $\alpha = 8^{-3/2}$, is convex on $[-1, 1]$ and we have $|h(x)| \leq L/D \leq 1$ so, according to Jensen, $\int_0^1 \varphi(h(x))dx \geq \varphi\left(\int_0^1 h(x)dx \right)$. Therefore $\int_0^{1/D} \left(1 - \sqrt{1 + h(x)} \right) dx \geq \alpha \int_0^{1/D} h^2(x)dx = (\alpha L^2)/D^3$, and we have

$$H^2(f_{\delta^1}, f_{\delta^2}) \geq \frac{2\alpha L^2}{D^3} \rho(\delta^1, \delta^2),$$

for all $\delta^1, \delta^2 \in \Delta$. According to Varshamov-Gilbert, cf. [115, p. 89] or [68], there exists a $D/8$-separated set, called $N_{D/8}$, on Δ for the Hamming distance such that its cardinal is higher than $2^{D/8}$ and $(0, \dots, 0) \in N_{D/8}$. On the separated set $N_{D/8}$ we have,

$$\forall \delta^1, \delta^2 \in N_{D/8}\,,\, H^2(f_{\delta^1}, f_{\delta^2}) \geq \frac{\alpha L^2}{4D^2}.$$

In order to apply Lemma 2.1, we need to control the KL divergences too. Since we have taken $N_{D/8}$ such that $(0,\ldots,0) \in N_{D/8}$, we can control the KL divergences w.r.t. P_0, the Lebesgue measure on $[0,1]^d$. We denote by P_δ the probability of density f_δ w.r.t. the Lebesgue's measure on $\mathbb{R}^d$, for all $\delta \in \Delta$. We have,

$$\begin{aligned} K(P_\delta^{\otimes n}|P_0^{\otimes n}) &= n\int_{[0,1]^d} \log\left(f_\delta(x)\right) f_\delta(x)dx \\ &= n\sum_{j=1}^{D}\int_{\frac{j-1}{D}}^{j/D} \log\left(1+\delta_j h_j(x)\right)\left(1+\delta_j h_j(x)\right)dx \\ &= n\left(\sum_{j=1}^{D}\delta_j\right)\int_0^{1/D}\log(1+h(x))(1+h(x))dx, \end{aligned}$$

for all $\delta=(\delta_1,\ldots,\delta_D)\in N_{D/8}$. Since $\forall u>-1, \log(1+u)\le u$, we have,

$$K(P_\delta^{\otimes n}|P_0^{\otimes n}) \le n\left(\sum_{j=1}^{D}\delta_j\right)\int_0^{1/D}(1+h(x))h(x)dx \le nD\int_0^{1/D}h^2(x)dx = \frac{nL^2}{D^2}.$$

Since $\log M \le 16(\min(1,A-1))^2 n$, we can take L such that $(nL^2)/D^2 = \log(M)/16$ and still having $L \le D\min(1,A-1)$. Thus, for $L=(D/4)\sqrt{\log(M)/n}$, we have for all elements δ^1,δ^2 in $N_{D/8}$, $H^2(f_{\delta^1},f_{\delta^2}) \ge (\alpha/64)(\log(M)/n)$ and $\forall \delta\in N_{D/8}$, $K(P_\delta^{\otimes n}|P_0^{\otimes n}) \le (1/16)\log(M)$.

Applying Lemma 1 when d is H, the Hellinger's distance, with M density functions $f_1,\ldots,f_M$ in $\left\{f_\delta : \delta\in N_{D/8}\right\}$ where $f_1=\mathbb{1}_{[0,1]^d}$ and the increasing function $w(u)=u^q$, we get the result.

■

REMARK 2.1. *The construction of the family of densities $\left\{f_\delta : \delta\in N_{D/8}\right\}$ is in the same spirit as the lower bound of [114], [102]. But, as compared to [102], we consider a different problem (model selection aggregation) and as compared to [114], we study in a different context (density estimation). Also, our risk function is different from those considered in these papers.*

Now, we give a lower bound for KL divergence. We have the same result as for square of Hellinger's distance.

THEOREM 2.3. *Let $M\ge 2$ be an integer, $A>1$ and $q>0$. We have, for any integer n such that $\log M \le 16(\min(1,A-1))^2 n$,*

$$\sup_{f_1,\ldots,f_M\in\mathcal{F}_K(A)}\ \inf_{\hat f_n}\ \sup_{f\in\mathcal{F}(A)}\left[\mathbb{E}_f\left[(K(f|\hat f_n))^q\right] - \min_{j=1,\ldots,M}(K(f|f_j))^q\right] \ge c\left(\frac{\log M}{n}\right)^q, \tag{2.6}$$

and

$$\sup_{f_1,\ldots,f_M\in\mathcal{F}_K(A)}\ \inf_{\hat f_n}\ \sup_{f\in\mathcal{F}(A)}\left[\mathbb{E}_f\left[(K(\hat f_n|f))^q\right] - \min_{j=1,\ldots,M}(K(f_j|f))^q\right] \ge c\left(\frac{\log M}{n}\right)^q, \tag{2.7}$$

where c is a positive constant which depends only on A. The sets $\mathcal{F}(A)$ and $\mathcal{F}_K(A)$ are defined in (2.5) for $\mathcal{X}=\mathbb{R}^d$.

Proof : Proof of the inequality (2.7) of Theorem 2.3 is similar to the one for (2.6). Since we have for all densities f and g,

$$K(f|g) \ge H^2(f,g),$$

[a proof is given in 115, p. 73], it suffices to note that, if $f_1, \ldots, f_M$ are densities bounded by A then,

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_K(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[(K(f|\hat{f}_n))^q \right] - \min_{j=1,\ldots,M} (K(f|f_i))^q \right]$$
$$\geq \inf_{\hat{f}_n} \sup_{f \in \{f_1,\ldots,f_M\}} \left[\mathbb{E}_f \left[(K(f|\hat{f}_n))^q \right] \right] \geq \inf_{\hat{f}_n} \sup_{f \in \{f_1,\ldots,f_M\}} \left[\mathbb{E}_f \left[H^{2q}(f, \hat{f}_n) \right] \right],$$

to get the result by applying Theorem 2.2.

■

With the same method as Theorem 1, we get the result below for the L_1-distance.

THEOREM 2.4. *Let $M \geq 2$ be an integer, $A > 1$ and $q > 0$. We have for any integers n such that $\log M \leq 16(\min(1, A-1))^2 n$,*

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_v(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[v(f, \hat{f}_n)^q \right] - \min_{j=1,\ldots,M} v(f, f_i)^q \right] \geq c \left(\frac{\log M}{n} \right)^{q/2}$$

where c is a positive constant which depends only on A. The sets $\mathcal{F}(A)$ and $\mathcal{F}_v(A)$ are defined in (2.5) for $\mathcal{X} = \mathbb{R}^d$.

Proof : The only difference with Theorem 2.2 is in the control of the distances. With the same notations as the proof of Theorem 2.2, we have,

$$v(f_{\delta^1}, f_{\delta^2}) = \int_{[0,1]^d} |f_{\delta^1}(x) - f_{\delta^2}(x)| dx = \rho(\delta^1, \delta^2) \int_0^{1/D} |h(x)| dx = \frac{L}{D^2} \rho(\delta^1, \delta^2),$$

for all $\delta^1, \delta^2 \in \Delta$. Thus, for $L = (D/4)\sqrt{\log(M)/n}$ and $N_{D/8}$, the $D/8$-separated set of Δ introduced in the proof of Theorem 2.2, we have,

$$v(f_{\delta^1}, f_{\delta^2}) \geq \frac{1}{32} \sqrt{\frac{\log(M)}{n}}, \quad \forall \delta^1, \delta^2 \in N_{D/8} \text{ and } K(P_\delta^{\otimes n} | P_0^{\otimes n}) \leq \frac{1}{16} \log(M), \quad \forall \delta \in \Delta.$$

Therefore, by applying Lemma 1 to the L_1-distance with M densities $f_1, \ldots, f_M$ in $\left\{ f_\delta : \delta \in N_{D/8} \right\}$ where $f_1 = \mathbb{1}_{[0,1]^d}$ and the increasing function $w(u) = u^q$, we get the result.

■

4. Upper bounds

In this section we use an argument in [125] (see also [35]) to show that the rate of the lower bound of Theorem 2.3 is an optimal rate of aggregation with respect to the KL loss. We use an aggregate constructed by Yang (defined in (2.1)) to attain this rate. An upper bound of the type (2.3) is stated in the following Theorem. Remark that Theorem 2.5 holds in a general framework of a measurable space $(\mathcal{X}, \mathcal{A})$ endowed with a σ-finite measure ν.

THEOREM 2.5 (Yang). *Let $X_1, \ldots, X_n$ be n observations of a probability measure on $(\mathcal{X}, \mathcal{A})$ of density f with respect to ν. Let $f_1, \ldots, f_M$ be M densities on $(\mathcal{X}, \mathcal{A}, \nu)$. The aggregate $\tilde{f}_n$, introduced in (2.1), satisfies, for any underlying density f,*

$$\mathbb{E}_f \left[K(f|\tilde{f}_n) \right] \leq \min_{j=1,\ldots,M} K(f|f_j) + \frac{\log(M)}{n+1}. \tag{2.8}$$

Proof : Proof follows the line of [125], although he does not state the result in the form (2.3), for convenience we reproduce the argument here. We define $\hat{f}_k(x; X^{(k)}) =$

$\sum_{j=1}^{M} w_j^{(k)} f_j(x)$, $\forall k = 1, \ldots, n$ (where $w_j^{(k)}$ is defined in (2.2) and $x^{(k)} = (x_1, \ldots, x_k)$ for all $k \in \mathbb{N}$ and $x_1, \ldots, x_k \in \mathcal{X}$) and $\hat{f}_0(x; X^{(0)}) = (1/M) \sum_{j=1}^{M} f_j(x)$ for all $x \in \mathcal{X}$. Thus, we have

$$\tilde{f}_n(x; X^{(n)}) = \frac{1}{n+1} \sum_{k=0}^{n} \hat{f}_k(x; X^{(k)}).$$

Let f be a density on $(\mathcal{X}, \mathcal{A}, \nu)$. We have

$$\begin{aligned}
\sum_{k=0}^{n} \mathbb{E}_f \left[K(f|\hat{f}_k) \right] &= \sum_{k=0}^{n} \int_{\mathcal{X}^{k+1}} \log \left(\frac{f(x_{k+1})}{\hat{f}_k(x_{k+1}; x^{(k)})} \right) \prod_{i=1}^{k+1} f(x_i) d\nu^{\otimes(k+1)}(x_1, \ldots, x_{k+1}) \\
&= \int_{\mathcal{X}^{n+1}} \left(\sum_{k=0}^{n} \log \left(\frac{f(x_{k+1})}{\hat{f}_k(x_{k+1}; x^{(k)})} \right) \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}) \\
&= \int_{\mathcal{X}^{n+1}} \log \left(\frac{f(x_1) \ldots f(x_{n+1})}{\prod_{k=0}^{n} \hat{f}_k(x_{k+1}; x^{(k)})} \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}),
\end{aligned}$$

but $\prod_{k=0}^{n} \hat{f}_k(x_{k+1}; x^{(k)}) = (1/M) \sum_{j=1}^{M} f_j(x_1) \ldots f_j(x_{n+1}), \forall x_1, \ldots, x_{n+1} \in \mathcal{X}$ thus,

$$\sum_{k=0}^{n} \mathbb{E}_f \left[K(f|\hat{f}_k) \right] = \int_{\mathcal{X}^{n+1}} \log \left(\frac{f(x_1) \ldots f(x_{n+1})}{\frac{1}{M} \sum_{j=1}^{M} f_j(x_1) \ldots f_j(x_{n+1})} \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}),$$

moreover $x \longmapsto \log(1/x)$ is a decreasing function so,

$$\begin{aligned}
&\sum_{k=0}^{n} \mathbb{E}_f \left[K(f|\hat{f}_k) \right] \\
&\leq \min_{j=1,\ldots,M} \left\{ \int_{\mathcal{X}^{n+1}} \log \left(\frac{f(x_1) \ldots f(x_{n+1})}{\frac{1}{M} f_j(x_1) \ldots f_j(x_{n+1})} \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}) \right\} \\
&\leq \log M + \min_{j=1,\ldots,M} \left\{ \int_{\mathcal{X}^{n+1}} \log \left(\frac{f(x_1) \ldots f(x_{n+1})}{f_j(x_1) \ldots f_j(x_{n+1})} \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}) \right\},
\end{aligned}$$

finally we have,

$$\sum_{k=0}^{n} \mathbb{E}_f \left[K(f|\hat{f}_k) \right] \leq \log M + (n+1) \inf_{j=1,\ldots,M} K(f|f_j). \tag{2.9}$$

On the other hand we have,

$$\mathbb{E}_f \left[K(f|\tilde{f}_n) \right] = \int_{\mathcal{X}^{n+1}} \log \left(\frac{f(x_{n+1})}{\frac{1}{n+1} \sum_{k=0}^{n} \hat{f}_k(x_{n+1}; x^{(k)})} \right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1, \ldots, x_{n+1}),$$

and $x \longmapsto \log(1/x)$ is convex, thus,

$$\mathbb{E}_f \left[K(f|\tilde{f}_n) \right] \leq \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{E}_f \left[K(f|\hat{f}_k) \right]. \tag{2.10}$$

Theorem 2.5 follows by combining (2.9) and (2.10).

■

Birgé constructs estimators, called *T-estimators* (the "T" is for "test"), which are adaptive in aggregation selection model of M estimators with a residual proportional at $(\log M/n)^{q/2}$ when Hellinger and L_1-distances are used to evaluate the quality of estimation (cf. [17]). But it does not give an optimal result as Yang, because there is a constant greater

than 1 in front of the main term $\min_{i=1,\dots,M} d^q(f,f_i)$ where d is the Hellinger distance or the L_1 distance. Nevertheless, observing the proof of Theorem 2.2 and 2.4, we can obtain

$$\sup_{f_1,\dots,f_M\in\mathcal{F}(A)} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}(A)} \left[\mathbb{E}_f\left[d(f,\hat{f}_n)^q\right] - C(q) \min_{i=1,\dots,M} d(f,f_i)^q\right] \geq c\left(\frac{\log M}{n}\right)^{q/2},$$

where d is the Hellinger or L_1-distance, $q>0$ and $A>1$. The constant $C(q)$ can be chosen equal to the one appearing in the following Theorem. The same residual appears in this lower bound and in the upper bounds of Theorem 2.6, so we can say that

$$\left(\frac{\log M}{n}\right)^{q/2}$$

is near optimal rate of aggregation w.r.t. the Hellinger distance or the L_1-distance to the power q, in the sense given at the end of Section 2. We recall Birgé's results in the following Theorem.

THEOREM 2.6 (Birgé). *If we have n observations of a probability measure of density f w.r.t. ν and $f_1,\dots,f_M$ densities on $(\mathcal{X},\mathcal{A},\nu)$, then there exists an estimator $\tilde{f}_n$ (T-estimator) such that for any underlying density f and $q>0$, we have*

$$\mathbb{E}_f\left[H(f,\tilde{f}_n)^q\right] \leq C(q)\left(\min_{j=1,\dots,M} H(f,f_j)^q + \left(\frac{\log M}{n}\right)^{q/2}\right),$$

and for the L_1-distance we can construct an estimator $\tilde{f}_n$ which satisfies :

$$\mathbb{E}_f\left[v(f,\tilde{f}_n)^q\right] \leq C(q)\left(\min_{j=1,\dots,M} v(f,f_j)^q + \left(\frac{\log M}{n}\right)^{q/2}\right),$$

where $C(q)>0$ is a constant depending only on q.

Another result, which can be found in [48], states that the minimum distance estimate proposed by Yatracos (1985) (cf. [48, p. 59]) achieves the same aggregation rate as in Theorem 2.6 for the L_1-distance with $q=1$. Namely, for all $f,f_1,\dots,f_M\in\mathcal{F}(A)$,

$$\mathbb{E}_f\left[v(f,\breve{f}_n)\right] \leq 3\min_{j=1,\dots,M} v(f,f_j) + \sqrt{\frac{\log M}{n}},$$

where $\breve{f}_n$ is the estimator of Yatracos defined by

$$\breve{f}_n = \arg\min_{f\in\{f_1,\dots,f_M\}} \sup_{A\in\mathcal{A}} \left|\int_A f - \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i\in A\}}\right|,$$

and $\mathcal{A}=\{\{x : f_i(x)>f_j(x)\} : 1\leq i,j\leq M\}$.

CHAPTER 3

Optimal Rates of Aggregation in Classification

In the same spirit as [114], we define the optimality of an aggregation procedure in the problem of classification. Using an aggregate with exponential weights, we obtain an optimal rate of convex aggregation for the hinge risk under the margin assumption. Moreover we obtain an optimal rate of model selection aggregation under the margin assumption for the excess Bayes risk.

Contents

The material of this chapter is an article accepted for publication in the journal *Bernoulli* (cf. [81]).

1. Introduction.

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We consider a random variable (X, Y) on $\mathcal{X} \times \{-1, 1\}$ with probability distribution denoted by π. Denote by P^X the marginal of π on $\mathcal{X}$ and by $\eta(x) \stackrel{\text{def}}{=} \mathbb{P}(Y = 1 | X = x)$ the conditional probability function of $Y = 1$ knowing that $X = x$. We have n i.i.d. observations of the couple (X, Y) denoted by $D_n = ((X_i, Y_i))_{i=1,\dots,n}$. The aim is to predict the output label Y for any input X in $\mathcal{X}$ from the observations D_n.

We recall some usual notation introduced for the classification framework. A **prediction rule** is a measurable function $f : \mathcal{X} \longmapsto \{-1, 1\}$. The **misclassification error** associated with f is

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

It is well known (see, e.g., [47]) that

$$\min_{f:\mathcal{X} \longmapsto \{-1,1\}} R(f) = R(f^*) \stackrel{\text{def}}{=} R^*,$$

where the prediction rule f^*, called the **Bayes rule**, is defined by

$$f^*(x) \stackrel{\text{def}}{=} \text{sign}(2\eta(x) - 1), \forall x \in \mathcal{X}.$$

The minimal risk R^* is called the **Bayes risk**. A **classifier** is a function, $\hat{f}_n = \hat{f}_n(X, D_n)$, measurable with respect to D_n and X with values in $\{-1, 1\}$, that assigns to the sample D_n a prediction rule $\hat{f}_n(., D_n) : \mathcal{X} \longmapsto \{-1, 1\}$. A key characteristic of $\hat{f}_n$ is the **generalization error** $\mathbb{E}[R(\hat{f}_n)]$, where

$$R(\hat{f}_n) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq \hat{f}_n(X)|D_n).$$

The aim of statistical learning is to construct a classifier $\hat{f}_n$ such that $\mathbb{E}[R(\hat{f}_n)]$ is as close to R^* as possible. Accuracy of a classifier $\hat{f}_n$ is measured by the value $\mathbb{E}[R(\hat{f}_n) - R^*]$ called **excess Bayes risk** of $\hat{f}_n$. We say that the classifier $\hat{f}_n$ learns with the convergence rate $\psi(n)$, where $(\psi(n))_{n\in\mathbb{N}}$ is a decreasing sequence, if there exists an absolute constant $C > 0$ such that for any integer n, $\mathbb{E}[R(\hat{f}_n) - R^*] \leq C\psi(n)$.

Theorem 7.2 of [47] shows that no classifier can learn with a given convergence rate for arbitrary underlying probability distribution π. To achieve rates of convergence, we need a complexity assumption on the set which the Bayes rule f^* belongs to. For instance [123, 124] provide examples of classifiers learning with a given convergence rate under complexity assumptions. These rates can not be faster than $n^{-1/2}$ (cf. [47]). Nevertheless, they can be as fast as n^{-1} if we add a control on the behavior of the conditional probability function η at the level 1/2 (the distance $|\eta(\cdot) - 1/2|$ is sometimes called the margin). The papers [91], for the problem of discriminant analysis, which is close to our classification problem, and [116] have introduced the following assumption

(MA) Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies MA(κ) with $1 \leq \kappa < +\infty$ if there exists $c_0 > 0$ such that,*

$$\mathbb{E}[|f(X) - f^*(X)|] \leq c_0 \left(R(f) - R^*\right)^{1/\kappa}, \tag{3.1}$$

for any measurable function f with values in $\{-1, 1\}$.

According to [116] and [22], this assumption is equivalent to a control on the margin given by

$$\mathbb{P}[|2\eta(X) - 1| \leq t] \leq ct^{\alpha}, \forall 0 \leq t < 1.$$

Several example of **fast rates**, i.e. rates faster than $n^{-1/2}$, can be found in [19, 109, 108, 92, 94, 93] and [9].

The aim of this chapter is the following:

(1) We define a concept of optimality for aggregation procedures in classification.
(2) We introduce several aggregation procedures in classification and obtain exact oracle inequalities for their risks.
(3) We prove lower bounds and show optimality of the suggested procedures and derive optimal rates of aggregation under the margin assumption.

The chapter is organized as follows. In Section 2 we introduce definitions and the procedures which are used throughout the chapter. Section 3 contains oracle inequalities for our aggregation procedures w.r.t. the excess hinge risk. Section 4 contains similar results for the excess Bayes risk. Proofs are postponed in Section 5.

2. Definitions and Procedures.

2.1. Loss functions. The quality of a classifier is often measured by a convex surrogate ϕ for the classification loss ([41, 54, 89, 55, 25, 14, 15]). Let us introduce some notations. Take ϕ a measurable function from $\mathbb{R}$ to $\mathbb{R}$. The risk associated with the loss

function ϕ is called the $\phi-$**risk** and is defined by

$$A^{(\phi)}(f) \stackrel{\text{def}}{=} \mathbb{E}[\phi(Yf(X))],$$

where $f : \mathcal{X} \longmapsto \mathbb{R}$ is a measurable function. The **empirical** $\phi-$**risk** is defined by

$$A_n^{(\phi)}(f) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$$

and we denote by $A^{(\phi)*}$ the infimum over all real valued functions $\inf_{f:\mathcal{X}\longmapsto\mathbb{R}} A^{(\phi)}(f)$.

Classifiers obtained by minimization of the empirical $\phi-$risk, for different convex losses, have been proved to have very good statistical properties (cf. [89, 20, 130, 109, 108] and [15]). A wide variety of classification methods in machine learning are based on this idea, in particular, on using the convex loss $\phi(x) \stackrel{\text{def}}{=} \max(1-x, 0)$, associated with support vector machines ([41, 104]), called the **hinge-loss**. The corresponding risk is called the **hinge risk** and is defined by

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}[\max(1 - Yf(X), 0)],$$

for any measurable function $f : \mathcal{X} \longmapsto \mathbb{R}$ and the **optimal hinge risk** is defined by

$$A^* \stackrel{\text{def}}{=} \inf_{f:\mathcal{X}\longmapsto\mathbb{R}} A(f). \tag{3.2}$$

It is easy to check that the Bayes rule f^* attains the infimum in (3.2) and

$$R(f) - R^* \leq A(f) - A^*, \tag{3.3}$$

for any measurable function f with values in $\mathbb{R}$ (cf. [88] and generalizations in [130] and [15]), where we extend the definition of R to the class of real valued functions by $R(f) = R(\text{sign(f)})$. Thus, minimization of the **excess hinge risk**, $A(f) - A^*$, provides a reasonable alternative for minimization of the excess Bayes risk, $R(f) - R^*$.

2.2. Aggregation Procedures. Now, we introduce the problem of aggregation and the aggregation procedures which will be studied in this chapter.

Suppose that we have $M \geq 2$ different classifiers $\hat{f}_1, \ldots, \hat{f}_M$ taking values in $\{-1, 1\}$. The problem of model selection type aggregation, as studied in [98, 125, 34, 35, 114], consists in construction of a new classifier $\tilde{f}_n$ (called *aggregate*) which mimics approximatively the best classifier among $\hat{f}_1, \ldots, \hat{f}_M$. In most of these papers the aggregation is based on splitting of the sample in two independent subsamples D_m^1 and D_l^2 of sizes m and l respectively, where $m + l = n$. The first subsample D_m^1 is used to construct the classifiers $\hat{f}_1, \ldots, \hat{f}_M$ and the second subsample D_l^2 is used to aggregate them, i.e., to construct a new classifier that mimics in a certain sense the behavior of the best among the classifiers $\hat{f}_j, j = 1, \ldots, M$.

In this chapter we will not consider the sample splitting and concentrate only on the construction of aggregates (following [74, 114, 17, 27]). Thus, the first subsample is fixed and instead of classifiers $\hat{f}_1, \ldots, \hat{f}_M$, we have fixed prediction rules $f_1, \ldots, f_M$. Rather than working with a part of the initial sample we will suppose, for notational simplicity, that the whole sample D_n of size n is used for the aggregation step instead of a subsample D_l^2.

Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite set of real-valued functions, where $M \geq 2$. An **aggregate** is a real valued statistic of the form

$$\tilde{f}_n = \sum_{f\in\mathcal{F}} w^{(n)}(f) f,$$

where the weights $(w^{(n)}(f))_{f\in\mathcal{F}}$ satisfy

$$w^{(n)}(f) \geq 0 \text{ and } \sum_{f\in\mathcal{F}} w^{(n)}(f) = 1.$$

Let ϕ be a convex loss for classification. The Empirical Risk Minimization aggregate (**ERM**) is defined by the weights

$$w^{(n)}(f) = \begin{cases} 1 & \text{for one } f \in \mathcal{F} \text{ such that } A_n^{(\phi)}(f) = \min_{g\in\mathcal{F}} A_n^{(\phi)}(g), \\ 0 & \text{for other } f \in \mathcal{F}. \end{cases} \quad , \quad \forall f \in \mathcal{F}$$

The ERM aggregate is denoted by $\tilde{f}_n^{(ERM)}$.

The **averaged ERM** aggregate is defined by the weights

$$w^{(n)}(f) = \begin{cases} 1/N & \text{if } A_n^{(\phi)}(f) = \min_{g\in\mathcal{F}} A_n^{(\phi)}(g), \\ 0 & otherwise, \end{cases} \quad , \quad \forall f \in \mathcal{F},$$

where N is the number of functions in $\mathcal{F}$ minimizing the empirical $\phi-$risk. The averaged ERM aggregate is denoted by $\tilde{f}_n^{(AERM)}$.

The Aggregation with Exponential Weights aggregate (**AEW**) is defined by the weights

$$w^{(n)}(f) = \frac{\exp\left(-nA_n^{(\phi)}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-nA_n^{(\phi)}(g)\right)}, \quad \forall f \in \mathcal{F}. \tag{3.4}$$

The AEW aggregate is denoted by $\tilde{f}_n^{(AEW)}$.

The **cumulative AEW** aggregate is an on-line procedure defined by the weights

$$w^{(n)}(f) = \frac{1}{n}\sum_{k=1}^{n} \frac{\exp\left(-kA_k^{(\phi)}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-kA_k^{(\phi)}(g)\right)}, \quad \forall f \in \mathcal{F}.$$

The cumulative AEW aggregate is denoted by $\tilde{f}_n^{(CAEW)}$.

When $\mathcal{F}$ is a class of prediction rules, intuitively, the AEW aggregate is more robust than the ERM aggregate w.r.t. the problem of overfitting. If the classifier with smallest empirical risk is overfitted, i.e., it fits too much to the observations, then the ERM aggregate will be overfitted. But, if other classifiers in $\mathcal{F}$ are good classifiers, the aggregate with exponential weights will consider their "opinions" in the final decision procedure and these opinions can balance with the opinion of the overfitted classifier in $\mathcal{F}$ which can be false because of its overfitting property. The ERM only considers the "opinion" of the classifier with the smallest risk, whereas the AEW takes into account all the opinions of the classifiers in the set $\mathcal{F}$. Moreover, the AEW aggregate does not need any minimization algorithm contrarily to the ERM aggregate.

The exponential weights, defined in (3.4), can be found in several situations. First, one can check that the solution of the following minimization problem

$$\min\Big(\sum_{j=1}^{M} \lambda_j A_n^{(\phi)}(f_j) + \epsilon \sum_{j=1}^{M} \lambda_j \log\lambda_j : \sum_{j=1}^{M} \lambda_j \leq 1, \lambda_j \geq 0, j = 1, \ldots, M\Big), \tag{3.5}$$

for all $\epsilon > 0$, is

$$\lambda_j = \frac{\exp\left(-\frac{A_n^{(\phi)}(f_j)}{\epsilon}\right)}{\sum_{k=1}^M \exp\left(-\frac{A_n^{(\phi)}(f_k)}{\epsilon}\right)}, \forall j = 1, \dots, M.$$

Thus, for $\epsilon = 1/n$, we find the exponential weights used for the AEW aggregate. Second, these weights can also be found in the theory of prediction of individual sequences, cf. [119].

2.3. Optimal Rates of Aggregation. In the same spirit as in [114], where the regression problem is treated, we introduce a concept of optimality for an aggregation procedure and for rates of aggregation, in the classification framework. Our aim is to prove that the aggregates introduced above are optimal in the following sense. All the results are given under the margin assumption. We denote by $\mathcal{P}_\kappa$ the set of all probability measures π on $\mathcal{X} \times \{-1, 1\}$ satisfying MA(κ).

DEFINITION 3.1. *Let ϕ be a loss function. The remainder term $\gamma(n, M, \kappa, \mathcal{F}, \pi)$ is called* **optimal rate of model selection type aggregation (MS-aggregation) for the $\phi-$risk**, *if the two following inequalities hold:*

(i) $\forall \mathcal{F} = \{f_1, \dots, f_M\}$, there exists a statistic $\tilde{f}_n$, depending on $\mathcal{F}$, such that $\forall \pi \in \mathcal{P}_\kappa$, $\forall n \geq 1$,

$$\mathbb{E}\left[A^{(\phi)}(\tilde{f}_n) - A^{(\phi)*}\right] \leq \min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right) + C_1 \gamma(n, M, \kappa, \mathcal{F}, \pi). \tag{3.6}$$

(ii) $\exists \mathcal{F} = \{f_1, \dots, f_M\}$ such that for any statistic $\bar{f}_n$, $\exists \pi \in \mathcal{P}_\kappa$, $\forall n \geq 1$

$$\mathbb{E}\left[A^{(\phi)}(\bar{f}_n) - A^{(\phi)*}\right] \geq \min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right) + C_2 \gamma(n, M, \kappa, \mathcal{F}, \pi). \tag{3.7}$$

Here, C_1 and C_2 are positive constants which may depend on κ. Moreover, when these two inequalities are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (3.6), is an **optimal MS-aggregate for the $\phi-$risk**. *If $\mathcal{C}$ denotes the convex hull of $\mathcal{F}$ and if (3.6) and (3.7) are satisfied with $\min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right)$ replaced by $\min_{f \in \mathcal{C}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right)$ then, we say that $\gamma(n, M, \kappa, \mathcal{F}, \pi)$ is an* **optimal rate of convex aggregation type for the $\phi-$risk** *and $\tilde{f}_n$ is an* **optimal convex aggregation procedure for the $\phi-$risk**.

In [114], the optimal rate of aggregation depends only on M and n. In our case the residual term may be a function of the underlying probability measure π, of the class $\mathcal{F}$ and of the margin parameter κ. Remark that, without any margin assumption, we obtain $\sqrt{(\log M)/n}$ for residual, which is free from π and $\mathcal{F}$. Under the margin assumption we got a residual term dependent of π and $\mathcal{F}$ and it should be interpreted as a normalizing factor in the ratio

$$\frac{\mathbb{E}\left[A^{(\phi)}(\bar{f}_n) - A^{(\phi)*}\right] - \min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right)}{\gamma(n, M, \kappa, \mathcal{F}, \pi)},$$

and in that case our definition does not imply the uniqueness of the residual.

3. Optimal Rates of Convex Aggregation for the Hinge Risk.

Take M functions $f_1, \dots, f_M$ with values in $[-1, 1]$. Consider the convex hull $\mathcal{C} = \text{Conv}(f_1, \dots, f_M)$. We want to mimic the best function in $\mathcal{C}$ using the hinge risk and working under the margin assumption. We first introduce a margin assumption w.r.t. the hinge loss.

(MAH) Margin (or low noise) assumption for hinge-risk. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the margin assumption for hinge-risk MAH(κ) with parameter $1 \leq \kappa < +\infty$ if there exists $c > 0$ such that,*

$$\mathbb{E}\left[|f(X) - f^*(X)|\right] \leq c\left(A(f) - A^*\right)^{1/\kappa}, \tag{3.8}$$

for any function f on $\mathcal{X}$ with values in $[-1, 1]$.

PROPOSITION 3.1. *The assumption MAH(κ) is equivalent to the margin assumption MA(κ).*

In what follows we will assume that MA(κ) holds and thus also MAH(κ) holds.

The AEW aggregate of M functions $f_1, \ldots, f_M$ with values in $[-1, 1]$, introduced in (3.4) for a general loss, has a simple form, for the case of the hinge loss, given by

$$\tilde{f}_n = \sum_{j=1}^{M} w^{(n)}(f_j) f_j, \text{ where } w^{(n)}(f_j) = \frac{\exp\left(\sum_{i=1}^{n} Y_i f_j(X_i)\right)}{\sum_{k=1}^{M} \exp\left(\sum_{i=1}^{n} Y_i f_k(X_i)\right)}, \quad \forall j = 1, \ldots, M. \tag{3.9}$$

In Theorems 3.1 and 3.2, we state the optimality of our aggregates in the sense of Definition 3.1.

THEOREM 3.1 (**Oracle inequality**). *Let $\kappa \geq 1$. We assume that π satisfies MA(κ). We denote by $\mathcal{C}$ the convex hull of a finite set $\mathcal{F}$ of functions $f_1, \ldots, f_M$ with values in $[-1, 1]$. Let $\tilde{f}_n$ be either of the four aggregates introduced in Section 2.2. Then, for any integers $M \geq 3, n \geq 1$, $\tilde{f}_n$ satisfies the following inequality*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{C}} (A(f) - A^*) + C\left(\sqrt{\frac{\min_{f \in \mathcal{C}} (A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}\right),$$

where $C = 32(6 \vee 537c \vee 16(2c + 1/3))$ for the ERM, AERM and AEW aggregates with $\kappa \geq 1$ and $c > 0$ is the constant in (3.8) and $C = 32(6 \vee 537c \vee 16(2c + 1/3))(2 \vee (2\kappa - 1)/(\kappa - 1)$ for the CAEW aggregate with $\kappa > 1$. For $\kappa = 1$ the CAEW aggregate satisfies

$$\mathbb{E}\left[A(\tilde{f}_n^{(CAEW)}) - A^*\right] \leq \min_{f \in \mathcal{C}} (A(f) - A^*) + 2C\left(\sqrt{\frac{\min_{f \in \mathcal{C}} (A(f) - A^*) \log M}{n}} + \frac{(\log M) \log n}{n}\right).$$

REMARK 3.1. *The hinge loss is linear on $[-1, 1]$, thus, MS-aggregation or convex aggregation of functions with values in $[-1, 1]$ are identical problems. Namely, we have*

$$\min_{f \in \mathcal{F}} A(f) = \min_{f \in \mathcal{C}} A(f). \tag{3.10}$$

THEOREM 3.2 (**Lower bound**). *Let $\kappa \geq 1$, M, n be two integers such that $2 \log_2 M \leq n$. We assume that the input space $\mathcal{X}$ is infinite. There exists an absolute constant $C > 0$, depending only on κ and c, and a set of prediction rules $\mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any real-valued procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ) for which*

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{C}} (A(f) - A^*) + C\left(\sqrt{\frac{(\min_{f \in \mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}\right),$$

where $C = c^{\kappa}(4e)^{-1} 2^{-2\kappa(\kappa-1)/(2\kappa-1)} (\log 2)^{-\kappa/(2\kappa-1)}$ and $c > 0$ is the constant in (3.8).

Combining the exact oracle inequality of Theorem 3.1 and the lower bound of Theorem 3.2, we see that the residual

$$\sqrt{\frac{(\min_{f\in\mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}, \tag{3.11}$$

is an optimal rate of convex aggregation of M functions with values in $[-1,1]$ for the hinge-loss. Moreover, for any real valued function f, we have $\max(1-y\psi(f(x)),0) \leq \max(1-yf(x),0)$ for all $y\in\{-1,1\}$ and $x\in\mathcal{X}$, thus

$$A(\psi(f)) - A^* \leq A(f) - A^*, \text{ where } \psi(x) = \max(-1,\min(x,1)), \quad \forall x\in\mathbb{R}. \tag{3.12}$$

Thus, by aggregating $\psi(f_1),\dots,\psi(f_M)$, it is easy to check that

$$\sqrt{\frac{(\min_{f\in\mathcal{F}} A(\psi(f)) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

is an optimal rate of model-selection aggregation of M real valued functions $f_1,\dots,f_M$ w.r.t. the hinge loss. In both cases, the aggregate with exponential weights as well as ERM and AERM attain these optimal rates and the CAEW aggregate attains the optimal rate if $\kappa > 1$. Applications and learning properties of the AEW procedure can be found in Chapters 7 and 8 (in particular, adaptive SVM classifiers are constructed by aggregating only $(\log n)^2$ SVM estimators). In Theorem 3.1, the AEW procedure satisfies an exact oracle inequality with an optimal residual term whereas in Chapters 7 and 8 the oracle inequalities satisfied by the AEW procedure are not exact (there is a multiplying factor greater than 1 in front of the bias term) and in Chapter 7 the residual is not optimal. In Chapter 8, it is proved that for any finite set $\mathcal{F}$ of functions $f_1,\dots,f_M$ with values in $[-1,1]$ and any $\epsilon>0$, there exists an absolute constant $C(\epsilon)>0$, such that, for $\mathcal{C}$ the convex hull of $\mathcal{F}$,

$$\mathbb{E}\left[A(\tilde{f}_n^{(AEW)}) - A^*\right] \leq (1+\epsilon)\min_{f\in\mathcal{C}}(A(f)-A^*) + C(\epsilon)\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}. \tag{3.13}$$

This oracle inequality is good enough for several applications (see the examples in Chapter 8). Nevertheless, (3.13) can be easily deduced from Theorem 3.1 using Lemma 3.1 and may be inefficient to construct adaptive estimators with exact constant (because of the factor greater than 1 in front of $\min_{f\in\mathcal{C}}(A(f)-A^*)$). Moreover, oracle inequalities with a factor greater than 1 in front of the oracle $\min_{f\in\mathcal{C}}(A(f)-A^*)$ do not characterize the real behavior of the used technique of aggregation. For instance, for any strictly convex loss ϕ, the ERM procedure satisfies, (cf. Chapter 9),

$$\mathbb{E}\left[A^{(\phi)}(\tilde{f}_n^{(ERM)}) - A^{(\phi)*}\right] \leq (1+\epsilon)\min_{f\in\mathcal{F}}(A^{(\phi)}(f) - A^{(\phi)*}) + C(\epsilon)\frac{\log M}{n}. \tag{3.14}$$

But, it has been recently proved in [85], that the ERM procedure can not mimic the oracle faster than $\sqrt{(\log M)/n}$, whereas, for strictly convex losses, the CAEW procedure can mimic the oracle at the rate $(\log M)/n$ (cf. [75]). Thus, for strictly convex losses, it is better to use aggregation procedure with exponential weights than ERM (or even penalized ERM procedures (cf. in Chapter 4)) to mimic the oracle. Non-exact oracle inequalities of the form (3.14) cannot tell us which procedure is better to use, since, both ERM and CAEW procedures satisfy this inequality.

It is interesting to note that the rate of aggregation (3.11) depends on both the class $\mathcal{F}$ and π through the term $\min_{f\in\mathcal{C}} A(f) - A^*$. This is different from the regression problem

(cf. [114]), where the optimal aggregation rates depends only on M and n. Three cases can be considered, where $\mathcal{M}(\mathcal{F},\pi)$ denotes $\min_{f\in\mathcal{C}}(A(f)-A^*)$ and M may depend on n:

(1) If $\mathcal{M}(\mathcal{F},\pi)\le a\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, for an absolute constant $a>0$, then the hinge risk of our aggregates attains $\min_{f\in\mathcal{C}}A(f)-A^*$ with the rate $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, which can be $\log M/n$ in the case $k=1$.

(2) If $a\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\le\mathcal{M}(\mathcal{F},\pi)\le b$, for some constants $a,b>0$, then our aggregates mimic the best prediction rule in $\mathcal{C}$ with a rate slower than $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$ but faster than $((\log M)/n)^{1/2}$.

(3) If $\mathcal{M}(\mathcal{F},\pi)\ge a>0$, where $a>0$ is a constant, then the rate of aggregation is $\sqrt{\frac{\log M}{n}}$, as in the case of no margin assumption.

We can explain this behavior by the fact that not only κ but also $\min_{f\in\mathcal{C}}A(f)-A^*$ measures the difficulty of classification. For instance, in the extreme case where $\min_{f\in\mathcal{C}}A(f)-A^*=0$, which means that $\mathcal{C}$ contains the Bayes rule, we have the fastest rate $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$. In the worst cases, which are realized when κ tends to ∞ or $\min_{f\in\mathcal{C}}(A(f)-A^*)\ge a>0$, where $a>0$ is an absolute constant, the optimal rate of aggregation is the slow rate $\sqrt{\frac{\log M}{n}}$.

4. Optimal Rates of MS-Aggregation for the Excess Risk.

Now, we provide oracle inequalities and lower bounds for the excess Bayes risk. First, we can deduce from Theorem 3.1 and 3.2, 'almost optimal rates of aggregation' for the excess Bayes risk achieved by the AEW aggregate. Second, using the ERM aggregate, we obtain optimal rates of model selection aggregation for the excess Bayes risk.

Using inequality (3.3), we can derive from Theorem 3.1, an oracle inequality for the excess Bayes risk. The lower bound is obtained using the same proof as in Theorem 3.2.

COROLLARY 3.1. *Let $\mathcal{F}=\{f_1,\dots,f_M\}$ be a finite set of prediction rules for an integer $M\ge 3$ and $\kappa\ge 1$. We assume that π satisfies MA(κ). Denote by $\tilde{f}_n$ either the ERM or the AERM or the AEW aggregate. Then, $\tilde{f}_n$ satisfies for any number $a>0$ and any integer n*

(3.15)

$$\mathbb{E}\left[R(\tilde{f}_n)-R^*\right]\le 2(1+a)\min_{j=1,\dots,M}(R(f_j)-R^*)+\left[C+(C^{2\kappa}/a)^{1/(2\kappa-1)}\right]\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

where $C=32(6\vee 537c\vee 16(2c+1/3))$. The CAEW aggregate satisfies the same inequality with $C=32(6\vee 537c\vee 16(2c+1/3))(2\vee(2\kappa-1)/(\kappa-1))$ when $\kappa>1$. For $\kappa=1$ the CAEW aggregate satisfies (3.15) where we need to multiply the residual by $\log n$.

Moreover there exists a finite set of prediction rules $\mathcal{F}=\{f_1,\dots,f_M\}$ such that for any classifier $\bar{f}_n$, there exists a probability measure π on $\mathcal{X}\times\{-1,1\}$ satisfying MA(κ), such that for any $n\ge 1, a>0$,

$$\mathbb{E}\left[R(\bar{f}_n)-R^*\right]\ge 2(1+a)\min_{f\in\mathcal{F}}(R(f)-R^*)+C(a)\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

where $C(a)>0$ is a constant depending only on a.

Due to Corollary 3.1,

$$\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$$

is an almost optimal rate of MS-aggregation for the excess risk and the AEW aggregate achieves this rate. The word "almost" is here because $\min_{f\in\mathcal{F}}(R(f)-R^*)$ is multiplied by a constant greater than 1.

Oracle inequality (3.15) is not exact since the minimal excess risk over $\mathcal{F}$ is multiplied by the constant $2(1+a)>1$. This is not the case while using the ERM aggregate as explained in the following Theorem.

THEOREM 3.3. *Let $\kappa\geq 1$. We assume that π satisfies MA(κ). We denote by $\mathcal{F}=\{f_1,\dots,f_M\}$ a set of prediction rules. The ERM aggregate over $\mathcal{F}$ satisfies for any integer $n\geq 1$*

$$\mathbb{E}\left[R(\tilde{f}_n^{(ERM)})-R^*\right]\leq \min_{f\in\mathcal{F}}(R(f)-R^*)+C\left(\sqrt{\frac{\min_{f\in\mathcal{F}}(R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

where $C=32(6\vee 537c_0\vee 16(2c_0+1/3))$ and c_0 is the constant appearing in MA(κ).

Using Lemma 3.1, we can deduce the results of [64] from Theorem 3.3. Oracle inequalities under MA(κ) have already been stated in [93] (cf. [22]), but the obtained remainder term is worse than the one obtained in Theorem 3.3.

According to Definition 3.1, combining Theorem 3.3 and the following Theorem, the rate

$$\sqrt{\frac{\min_{f\in\mathcal{F}}(R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$$

is an optimal rate of MS-aggregation w.r.t. the excess Bayes risk. The ERM aggregate achieves this rate.

THEOREM 3.4 (**Lower bound**). *Let $M\geq 3$ and n be two integers such that $2\log_2 M\leq n$ and $\kappa\geq 1$. Assume that $\mathcal{X}$ is infinite. There exists an absolute constant $C>0$ and a set of prediction rules $\mathcal{F}=\{f_1,\dots,f_M\}$ such that for any procedure $\bar{f}_n$ with values in $\mathbb{R}$, there exists a probability measure π satisfying MA(κ) for which*

$$\mathbb{E}\left[R(\bar{f}_n)-R^*\right]\geq \min_{f\in\mathcal{F}}(R(f)-R^*)+C\left(\sqrt{\frac{(\min_{f\in\mathcal{F}}R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

where $C=c_0{}^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$ and c_0 is the constant appearing in MA(κ).

5. Proofs.

Proof of Proposition 3.1. Since for any function f from $\mathcal{X}$ to $\{-1,1\}$ we have $2(R(f)-R^*)=A(f)-A^*$, then, MA(κ) is implied by MAH(κ).

Assume that MA(κ) holds. We first explore the case $\kappa>1$, then, MA(κ) implies that there exists a constant $c_1>0$ such that $\mathbb{P}\left(|2\eta(X)-1|\leq t\right)\leq c_1 t^{1/(\kappa-1)}$ for any $t>0$ (cf. [22]). Let f from $\mathcal{X}$ to $[-1,1]$. We have, for any $t>0$,

$$A(f)-A^*=\mathbb{E}\left[|2\eta(X)-1||f(X)-f^*(X)|\right]\geq t\mathbb{E}\left[|f(X)-f^*(X)|\mathbb{I}_{|2\eta(X)-1|\geq t}\right]$$

$$\geq t\left(\mathbb{E}\left[|f(X)-f^*(X)|\right]-2\mathbb{P}\left(|2\eta(X)-1|\leq t\right)\right)\geq t\left(\mathbb{E}\left[|f(X)-f^*(X)|\right]-2c_1t^{1/(\kappa-1)}\right).$$

For $t_0 = ((\kappa-1)/(2c_1\kappa))^{\kappa-1}\mathbb{E}\left[|f(X)-f^*(X)|\right]^{\kappa-1}$, we obtain

$$A(f) - A^* \geq ((\kappa-1)/(2c_1\kappa))^{\kappa-1}\kappa^{-1}\mathbb{E}\left[|f(X)-f^*(X)|\right]^{\kappa}.$$

For the case $\kappa = 1$, MA(1) implies that there exists $h > 0$ such that $|2\eta(X)-1| \geq h$ a.s.. Indeed, if for any $N \in \mathbb{N}^*$, there exists $A_N \in \mathcal{A}$ such that $P^X(A_N) > 0$ and $|2\eta(x)-1| \leq N^{-1}, \forall x \in A_N$, then, for

$$f_N(x) = \begin{cases} -f^*(x) & \text{if } x \in A_N \\ f^*(x) & \text{otherwise,} \end{cases}$$

we obtain $R(f_N) - R^* \leq 2P^X(A_N)/N$ and $\mathbb{E}\left[|f_N(X)-f^*(X)|\right] = 2P^X(A_N)$, and there is no constant $c_0 > 0$ such that $P^X(A_N) \leq c_0 P^X(A_N)/N$ for all $N \in \mathbb{N}^*$. So, assumption MA(1) does not hold if no $h > 0$ satisfies $|2\eta(X)-1| \geq h$ a.s.. Thus, for any f from $\mathcal{X}$ to $[-1,1]$, we have $A(f) - A^* = \mathbb{E}\left[|2\eta(X)-1||f(X)-f^*(X)|\right] \geq h\mathbb{E}\left[|f(X)-f^*(X)|\right]$.

Proof of Theorem 3.1: Cf. proof of Theorem 9.1 in Chapter 9.

Proof of Theorem 3.2. Let a be a positive number and $f_1, \ldots, f_M$ be M prediction rules. Using (3.10), we have, for any finite set $\mathcal{F}$ of M real valued functions,

(3.16)

$$\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\kappa}\left(\mathbb{E}\left[A(\hat{f}_n) - A^*\right] - (1+a)\min_{f\in \mathrm{Conv}(\mathcal{F})}(A(f)-A^*)\right) \geq \inf_{\hat{f}_n}\sup_{\substack{\pi\in\mathcal{P}_\kappa \\ f^*\in\{f_1,\ldots,f_M\}}}\mathbb{E}\left[A(\hat{f}_n) - A^*\right],$$

where $\mathrm{Conv}(\mathcal{F})$ is the set made of all the convex combinations of elements in $\mathcal{F}$. Let N be an integer such that $2^{N-1} \leq M$, $x_1, \ldots, x_N$ be N distinct points of $\mathcal{X}$ and w be a positive number satisfying $(N-1)w \leq 1$. Denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\}) = w$, for $j = 1, \ldots, N-1$ and $P^X(\{x_N\}) = 1-(N-1)w$. We consider the cube $\Omega = \{-1,1\}^{N-1}$. Let $0 < h < 1$. For all $\sigma = (\sigma_1, \ldots, \sigma_{N-1}) \in \Omega$ we consider

$$\eta_\sigma(x) = \begin{cases} (1+\sigma_j h)/2 & \text{if } x = x_1, \ldots, x_{N-1}, \\ 1 & \text{if } x = x_N. \end{cases}$$

For all $\sigma \in \Omega$ we denote by π_σ the probability measure on $\mathcal{X}\times\{-1,1\}$ having P^X for marginal on $\mathcal{X}$ and η_σ for conditional probability function.

Assume that $\kappa > 1$. We have $\mathbb{P}\left(|2\eta_\sigma(X)-1| \leq t\right) = (N-1)w\mathbb{1}_{h\leq t}$ for any $0 \leq t < 1$. Thus, if we assume that $(N-1)w \leq h^{1/(\kappa-1)}$ then $\mathbb{P}\left(|2\eta_\sigma(X)-1| \leq t\right) \leq t^{1/(\kappa-1)}$ for all $0 \leq t < 1$. Thus, according to [116], π_σ belongs to $\mathcal{P}_\kappa$.

We denote by ρ the Hamming distance on Ω. Let $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma,\sigma') = 1$. Denote by H the Hellinger's distance. Since $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\Big(1-\Big(1-H^2(\pi_\sigma,\pi_{\sigma'})/2\Big)^n\Big)$ and

$$H^2(\pi_\sigma,\pi_{\sigma'}) = w\sum_{j=1}^{N-1}\Big(\sqrt{\eta_\sigma(x_j)}-\sqrt{\eta_{\sigma'}(x_j)}\Big)^2+\Big(\sqrt{1-\eta_\sigma(x_j)}-\sqrt{1-\eta_{\sigma'}(x_j)}\Big)^2 = 2w(1-\sqrt{1-h^2}),$$

then, the Hellinger's distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\left(1-(1-w(1-\sqrt{1-h^2}))^n\right).$$

Take w and h such that $w(1-\sqrt{1-h^2}) \leq n^{-1}$. Then, $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) \leq \beta = 2(1-e^{-1}) < 2$ for any integer n.

Let $\sigma \in \Omega$ and $\hat{f}_n$ be an estimator with values in $[-1,1]$ (according to (3.12), we consider only estimators in $[-1,1]$). Using MA(κ), we have, conditionally to the observations D_n

and for $\pi = \pi_\sigma$,

$$A(\hat{f}_n) - A^* \geq \left(c\mathbb{E}_{\pi_\sigma}\left[|\hat{f}_n(X) - f^*(X)| \right] \right)^\kappa \geq (cw)^\kappa \left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j| \right)^\kappa .$$

Taking here the expectations, we find $\mathbb{E}_{\pi_\sigma}\left[A(\hat{f}_n) - A^* \right] \geq (cw)^\kappa \mathbb{E}_{\pi_\sigma}\left[\left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j| \right)^\kappa \right]$. Using Jensen's inequality and Lemma 3.3, we obtain

$$\inf_{\hat{f}_n} \sup_{\sigma \in \Omega} \left(\mathbb{E}_{\pi_\sigma}\left[A(\hat{f}_n) - A^* \right] \right) \geq (cw)^\kappa \left(\frac{N-1}{4e^2} \right)^\kappa . \tag{3.17}$$

Take now $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$, $h = \left(n^{-1} \lceil \log M / \log 2 \rceil \right)^{(\kappa-1)/(2\kappa-1)}$. Replace w and N in (3.17) by there values, thus, from (3.16), there exist $f_1, \ldots, f_M$ (the 2^{N-1} first ones are $\text{sign}(2\eta_\sigma - 1)$ for $\sigma \in \Omega$ and any choice for the $M - 2^{N-1}$ remaining ones) such that for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}\left[A(\hat{f}_n) - A^* \right] - (1+a)\min_{j=1,\ldots,M}(A(f_j) - A^*) \geq C_0 \left(\frac{\log M}{n} \right)^{\frac{\kappa}{2\kappa-1}}$, where $C_0 = c^\kappa (4e)^{-1} 2^{-2\kappa(\kappa-1)/(2\kappa-1)} (\log 2)^{-\kappa/(2\kappa-1)}$.
Moreover, according to Lemma 3.1, we have

$$a \min_{f \in \mathcal{C}}(A(f) - A^*) + \frac{C_0}{2} \left(\frac{\log M}{n} \right)^{\frac{\kappa}{2\kappa-1}} \geq \sqrt{2^{-1} a^{1/\kappa} C_0} \sqrt{\frac{(\min_{f \in \mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}}.$$

Thus,

$$\mathbb{E}\left[A(\hat{f}_n) - A^* \right] \geq \min_{f \in \mathcal{C}}(A(f) - A^*) + \frac{C_0}{2} \left(\frac{\log M}{n} \right)^{\frac{\kappa}{2\kappa-1}} + \sqrt{2^{-1} a^{1/\kappa} C_0} \sqrt{\frac{(A_{\mathcal{C}} - A^*)^{\frac{1}{\kappa}} \log M}{n}}.$$

For $\kappa = 1$, we take $h = 1/2$. Then $|2\eta_\sigma(X) - 1| \geq 1/2$ a.s. so $\pi_\sigma \in$MA(1). It suffices then to take $w = 4/n$ and $N = \lceil \log M / \log 2 \rceil$ to obtain the result.

Proof of Corollary 3.1. The result follows from Theorems 3.1 and 3.2. Using the fact that for any prediction rule f we have $A(f) - A^* = 2(R(f) - R^*)$, inequality (3.3) and Lemma 3.1, for any $a > 0$, with $t = a(A_{\mathcal{C}} - A^*)$ and $v = (C^2 (\log M)/n)^{\kappa/(2\kappa-1)} a^{-1/(2\kappa-1)}$ we obtain the result.

Proof of Theorem 3.3: Cf. proof of Theorem 9.1 in Chapter 9.

Proof of Theorem 3.4: For all prediction rules $f_1, \ldots, f_M$, we have

$$\sup_{g_1,\ldots,g_M} \inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\kappa} \left(\mathbb{E}\left[R(\hat{f}_n) - R^* \right] - (1+a) \min_{j=1,\ldots,M} (R(g_j) - R^*) \right) \geq \inf_{\hat{f}_n} \sup_{\substack{\pi \in \mathcal{P}_\kappa \\ f^* \in \{f_1,\ldots,f_M\}}} \mathbb{E}\left[R(\hat{f}_n) - R^* \right].$$

Consider the set of probability measures $\{\pi_\sigma, \sigma \in \Omega\}$ introduced in the proof of Theorem 3.2. Assume that $\kappa > 1$. Since for any $\sigma \in \Omega$ and any classifier $\hat{f}_n$, we have, by using MA(κ),

$$\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^* \right] \geq (c_0 w)^\kappa \mathbb{E}_{\pi_\sigma} \left[\left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j| \right)^\kappa \right],$$

using Jensen's inequality and Lemma 3.3, we obtain

$$\inf_{\hat{f}_n} \sup_{\sigma \in \Omega} \left(\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^* \right] \right) \geq (c_0 w)^\kappa \left(\frac{N-1}{4e^2} \right)^\kappa .$$

By taking $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$, $h = \left(n^{-1}\lceil \log M/\log 2\rceil\right)^{\frac{\kappa-1}{2\kappa-1}}$, there exist $f_1, \dots, f_M$ (the 2^{N-1} first ones are $\mathrm{sign}(2\eta_\sigma - 1)$ for $\sigma \in \Omega$ and any choice for the $M - 2^{N-1}$ remaining ones) such that for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}\left[R(\hat{f}_n) - R^*\right] - (1+a)\min_{j=1,\dots,M}(R(f_j) - R^*) \geq C_0 \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, where $C_0 = c_0{}^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$. Moreover, according to Lemma 3.1, we have

$$a\min_{f\in\mathcal{F}}(R(f) - R^*) + \frac{C_0}{2}\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} \geq \sqrt{a^{1/\kappa}C_0/2}\sqrt{\frac{(\min_{f\in\mathcal{F}} R(f) - R^*)^{\frac{1}{\kappa}}\log M}{n}}.$$

The case $\kappa = 1$ is treated in the same way as in the proof of Theorem 3.2.

LEMMA 3.1. *Let $v, t > 0$ and $\kappa \geq 1$. The concavity of the logarithm yields*

$$t + v \geq t^{\frac{1}{2\kappa}} v^{\frac{2\kappa-1}{2\kappa}}.$$

LEMMA 3.2. *Let f be a function from $\mathcal{X}$ to $[-1,1]$ and π a probability measure on $\mathcal{X}\times\{-1,1\}$ satisfying MA(κ), for a $\kappa \geq 1$. Denote by $\mathbb{V}$ the symbol of variance. We have $\mathbb{V}\Big(Y(f(X) - f^*(X))\Big) \leq c(A(f) - A^*)^{1/\kappa}$ and $\mathbb{V}\Big(\mathbb{1}_{Yf(X)\leq 0} - \mathbb{1}_{Yf^*(X)\leq 0}\Big) \leq c(R(f) - R^*)^{1/\kappa}$.*

LEMMA 3.3. *Let $\{P_\omega / \omega \in \Omega\}$ be a set of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, indexed by the cube $\Omega = \{0,1\}^m$. Denote by $\mathbb{E}_\omega$ the expectation under P_ω and by ρ the Hamming distance on Ω. Assume that*

$$\forall \omega, \omega' \in \Omega / \rho(\omega, \omega') = 1,\ H^2(P_\omega, P_{\omega'}) \leq \alpha < 2,$$

then

$$\inf_{\hat{w}\in[0,1]^m} \max_{\omega\in\Omega} \mathbb{E}_\omega \left[\sum_{j=1}^m |\hat{w}_j - w_j|\right] \geq \frac{m}{4}\left(1 - \frac{\alpha}{2}\right)^2.$$

Proof. Obviously, we can replace $\inf_{\hat{w}\in[0,1]^m}$ by $(1/2)\inf_{\hat{w}\in\{0,1\}^m}$ since for all $w \in \{0,1\}$ and $\hat{w} \in [0,1]$ there exists $\tilde{w} \in \{0,1\}$ (for instance the projection of $\hat{w}$ on $\{0,1\}$) such that $|\hat{w} - w| \geq (1/2)|\tilde{w} - w|$. Then, we use Theorem 2.10 p.103 of [114].

CHAPTER 4

Suboptimality of Penalized Empirical Risk Minimization

Let $f \in \mathcal{F}$ be an object to estimate and $\mathcal{F}_0 \subset \mathcal{F}$ be a subset with cardinality M. For instance the elements in $\mathcal{F}_0$ may have been constructed with preliminary observations which are throughout the chapter assumed to be frozen. The elements in $\mathcal{F}_0$ are considered as non-random. Given a loss function, we want to construct a procedure which mimics at the best possible rate the best procedure in $\mathcal{F}_0$. This fastest rate is called optimal rate of aggregation. In this chapter, we prove that, in several estimation problems (classification under margin assumption for different losses, density estimation and regression), the usual penalized (or structural) Empirical Risk Minimization (ERM) procedures cannot achieve this optimal rate of aggregation. On the other hand, in those cases, aggregation procedures with exponential weights attain the optimal rate of aggregation. Moreover, we prove that quality of aggregation of the ERM procedure depends on both the margin and approximation quality of the model.

Contents

The material of this chapter combines a paper accepted for publication in *COLT07* and a paper submitted for publication.

1. Introduction

1.1. Framework. Let $(\mathcal{Z}, \mathcal{T})$ a measurable space. Denote by $\mathcal{P}$ the set of all probability measures on $(\mathcal{Z}, \mathcal{T})$. Let F be a function on $\mathcal{P}$ with values in an algebra $\mathcal{F}$. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability measure. Let D_n be a sample of n i.i.d. observations $Z_1, \ldots, Z_n$ having the common probability measure π.

The probability measure π is unknown. Our aim is to estimate $F(\pi)$ from the observations D_n. Consider a loss function $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}_+$ and the corresponding average loss

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}\left[Q(Z, f)\right],$$

where $\mathbb{E}$ denotes the expectation. If the minimum over all f in $\mathcal{F}$

$$A^* \stackrel{\text{def}}{=} \min_{f \in \mathcal{F}} A(f)$$

is achieved by at least one function, we denote by f^* a minimizer in $\mathcal{F}$. In this chapter we will assume that $\min_{f \in \mathcal{F}} A(f)$ is achievable.

In most of the cases f^* will be equal to our target $F(\pi)$. We don't know the risk A, since π is not available to the statistician. Thus, we minimize the empirical version of A constructed from the observations D_n, i.e.

$$A_n(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, f). \tag{4.1}$$

Now, we introduce an assumption which improves the quality of estimation and of aggregation in our framework. This assumption has been first introduced by [91], for the problem of discriminant analysis, and [116], for the classification problem. With this assumption, fast rates of convergence can be achieved, for instance, in classification problem (cf. [116], [109]).

Margin Assumption(MA): *The probability measure π satisfies the margin assumption MA($\kappa, c, \mathcal{F}_0$), where $\kappa \geq 1, c > 0$ and $\mathcal{F}_0$ is a subset of $\mathcal{F}$ if*

$$\mathbb{E}[(Q(Z, f) - Q(Z, f^*))^2] \leq c(A(f) - A^*)^{1/\kappa}, \forall f \in \mathcal{F}_0. \tag{4.2}$$

In the regression setup on $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a measurable space, with the L^2 risk w.r.t. the probability measure of the design on $\mathcal{X}$ (cf. Example 1, Section 1.1 of Chapter 1), it is easy to see that any probability distribution π on $\mathcal{X} \times \mathbb{R}$ satisfies the margin assumption $\mathrm{MA}(1, 1, \mathcal{F})$, where $\mathcal{F}$ is the set of all square integrable functions from $\mathcal{X}$ to $\mathbb{R}$. In density estimation with the integrated square risk (cf. Example 2, Section 1.1 of Chapter 1) with the densities a.s. bounded by a constant $B \geq 1$, satisfy the margin assumption $\mathrm{MA}(1, 16B^2, \mathcal{F}_B)$ where $\mathcal{F}_B$ is the set of all non-negative functions $f \in L^2(\mathcal{Z}, \mathcal{T}, \mu)$ bounded by B.

The margin assumption is linked to the convexity of the underlying loss. In density and regression estimation it is naturally satisfied with the best margin parameter $\kappa = 1$, but, for non-convex loss (for instance in classification) this assumption is an additional restriction.

1.2. Aggregation Procedures and Optimality. We work with the notation introduced at the beginning of the previous subsection. Our framework is the same as the one considered, among others, in [98, 35, 75, 125, 126, 127]. We have a family $\mathcal{F}_0$ of M "weak estimators" $f_1, \ldots, f_M \in \mathcal{F}$ and the goal is to construct an estimator, based on a sample D_n of n i.i.d. observations $Z_1, \ldots, Z_n$ of Z, which has a risk close to the one of the oracle, that is $\min_{f \in \mathcal{F}_0}(A(f) - A^*)$. Those weak estimators could have been constructed from a preliminary set of observations or they can be the first M functions of a basis or simple objects like decision stumps. The problem is to find a strategy which mimics as fast as we can the best element in $\mathcal{F}_0$. Such a strategy can then be used to construct efficient

adaptive estimators (cf. [98] and Chapters 7,8, 9 and 10). In this chapter we consider four different aggregation strategies.

The most well known one is the Empirical Risk Minimization (**ERM**) procedure over $\mathcal{F}_0$, defined by

$$\tilde{f}_n^{(ERM)} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0} A_n(f), \tag{4.3}$$

and the penalized Empirical Risk Minimization (**pERM**) procedures given by

$$\tilde{f}_n^{(pERM)} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0} \Big(A_n(f) + \operatorname{pen}(f)\Big), \tag{4.4}$$

where $\operatorname{pen}(\cdot)$ is some penalty function (cf.,e.g., [92],[93]).

A **selector** is an aggregate with values in the family $\mathcal{F}_0$. Penalized ERM and ERM procedures are examples of selectors.

Aggregation with Exponential Weights (**AEW**) procedure over $\mathcal{F}_0$ is defined by

$$\tilde{f}_{n,\beta}^{(AEW)} \stackrel{\text{def}}{=} \sum_{f\in\mathcal{F}_0} w_\beta^{(n)}(f) f, \tag{4.5}$$

where $\beta > 0$ is a parameter called the *temperature* and the exponential weights $w_\beta^{(n)}(f)$ are defined by

$$w_\beta^{(n)}(f) = \frac{\exp\left(-n\beta^{-1}A_n(f)\right)}{\sum_{g\in\mathcal{F}_0}\exp\left(-n\beta^{-1}A_n(g)\right)}, \quad \forall f \in \mathcal{F}_0. \tag{4.6}$$

Cumulative Aggregation with Exponential Weights (**CAEW**) procedure is defined by

$$\tilde{f}_{n,\beta}^{(CAEW)} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_{k,\beta}^{(AEW)}, \tag{4.7}$$

where $\tilde{f}_{k,\beta}^{(AEW)}$ is constructed as in (4.5) with the sample $Z_1, \ldots, Z_k$ of size k and with the temperature parameter $\beta > 0$. Namely,

$$\tilde{f}_{k,\beta}^{(AEW)} = \sum_{f\in\mathcal{F}} w_\beta^{(k)}(f) f, \text{ where } w_\beta^{(k)}(f) = \frac{\exp\left(-\beta^{-1}kA_k(f)\right)}{\sum_{g\in\mathcal{F}}\exp\left(-\beta^{-1}kA_k(g)\right)}, \quad \forall f \in \mathcal{F}.$$

Since there are many different ways to combine the weak estimators, we consider the following definition, which is inspired by the one given in [114] for the regression model. This definition provides a way to compare aggregation strategies.

Definition 4.1. *The remainder term $\gamma(n, M)$ is called* **optimal rate of aggregation**, *if the two following inequalities hold.*

(1) For any finite set $\mathcal{F}_0$ of M elements in $\mathcal{F}$, there exists a statistic $\tilde{f}_n$ such that for any underlying probability measure π and any integer $n \geq 1$,

$$\mathbb{E}[A(\tilde{f}_n) - A^*] \leq \min_{f\in\mathcal{F}_0}\left(A(f) - A^*\right) + C_1\gamma(n, M). \tag{4.8}$$

(2) There exists a finite set $\mathcal{F}_0$ of M elements in $\mathcal{F}$ such that for any statistic $\bar{f}_n$, there exists a probability distribution π, such that for any $n \geq 1$

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f\in\mathcal{F}_0}\left(A(f) - A^*\right) + C_2\gamma(n, M). \tag{4.9}$$

Here, C_1 and C_2 are absolute positive constants. Moreover, when these two inequalities are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (4.8), is an **optimal aggregation procedure**.

The aim of this chapter is to obtain the optimal rate of aggregation in several situations and to prove that the ERM and certain penalized ERM procedures cannot achieve the optimal rate when the loss function has some convexity property.

The chapter is organized as follows. In the three following sections, we explore, respectively, the classification under the margin assumption setup for different loss functions, the gaussian regression and the density estimation frameworks. In Section 6, we discuss the results. All the proofs are postponed to Section 7.

2. Classification Under Margin Assumption.

Consider the problem of binary classification. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Consider a couple (X, Y) of random variables where X takes its values in $\mathcal{X}$ and Y is a random label taking values in $\{-1, 1\}$. We denote by π the probability distribution of (X, Y). For any function $\phi : \mathbb{R} \longmapsto \mathbb{R}$, define the $\phi-$risk of a real valued classifier f on $\mathcal{X}$ by

$$A^{\phi}(f) = \mathbb{E}[\phi(Yf(X))].$$

Comparing with the notation of the previous section we have $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ and $Q((x, y), f) = \phi(yf(x))$.

Many different losses have been discussed in the literature along the last decade (cf. [41, 54, 89, 55, 25]), for instance:

$\phi_0(x) = \mathbb{1}_{(x \leq 0)}$	classical loss or $0-1$ loss
$\phi_1(x) = \max(0, 1-x)$	hinge loss (SVM loss)
$x \longmapsto \log_2(1 + \exp(-x))$	logit-boosting loss
$x \longmapsto \exp(-x)$	exponential boosting loss
$x \longmapsto (1-x)^2$	squared loss
$x \longmapsto \max(0, 1-x)^2$	2-norm soft margin loss.

In particular, we are interested in losses having the following convexity property (cf. [75] for examples).

Definition 4.2. *Let $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a function and β be a non-negative number. We say that ϕ is $\beta-$***convex on** $[-1, 1]$ *when*

$$[\phi'(x)]^2 \leq \beta\phi''(x), \quad \forall |x| \leq 1.$$

For example, logit-boosting loss is $(e/\log 2)-$convex, exponential boosting loss is $e-$convex, squared and $2-$norm soft margin losses are $2-$convex.

There are some links with the usual concepts of convexity. We recall the definition of these concepts (cf. [103]). Let $\phi : [-1, 1] \longmapsto \mathbb{R}$ be a function. If

$$\phi(\alpha x_1 + (1-\alpha)x_2) < \alpha\phi(x_1) + (1-\alpha)\phi(x_2),$$

for all $x_1 \neq x_2$ in $[-1, 1]$, then, ϕ is called a *strictly convex* function on $[-1, 1]$. If there is a constant $c > 0$ such that for any $x_1, x_2 \in [-1, 1]$,

$$\phi(\alpha x_1 + (1-\alpha)x_2) < \alpha\phi(x_1) + (1-\alpha)\phi(x_2) - \frac{1}{2}c\alpha(1-\alpha)|x_1 - x_2|^2,$$

then, f is called a *strongly convex* function on $[-1, 1]$.

Proposition 4.1. *Let $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a twice differentiable function. If ϕ is strongly convex then, there exists $\beta > 0$, such that ϕ is β-convex. Moreover, the constant function is 0-convex but not strictly convex and the function $x \longmapsto (x+1)^3/3 - (x+1)$ is strictly convex on $[-1, 1]$ but not $\beta-$convex for any $\beta \geq 0$.*

We denote by f_ϕ^* a function from $\mathcal{X}$ to $\mathbb{R}$ which minimizes $A^\phi(\cdot)$ over the set of real-valued functions. We denote by $A^{\phi *} \stackrel{\text{def}}{=} A^\phi(f_\phi^*)$ the minimal $\phi-$risk. In many interesting cases studied in the literature, either f_ϕ^* or its sign are equal to the Bayes classifier

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

where η is the conditional probability function $x \longmapsto \mathbb{P}(Y = 1|X = x)$ defined on $\mathcal{X}$. The Bayes classifier f^* is a minimizer of the ϕ_0-risk (cf. [47]) and is the best classifier that we want to mimic.

To understand how behaves the optimal rate of aggregation depending on the loss function we introduce a "continuous scale" of loss functions indexed by a non-negative number h:

$$\phi_h(x) = \begin{cases} h\phi_1(x) + (1-h)\phi_0(x) & \text{if } 0 \le h \le 1, \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases} \qquad \forall x \in \mathbb{R}, \tag{4.10}$$

where ϕ_0 is the $0-1$ loss and ϕ_1 is the Hinge loss.

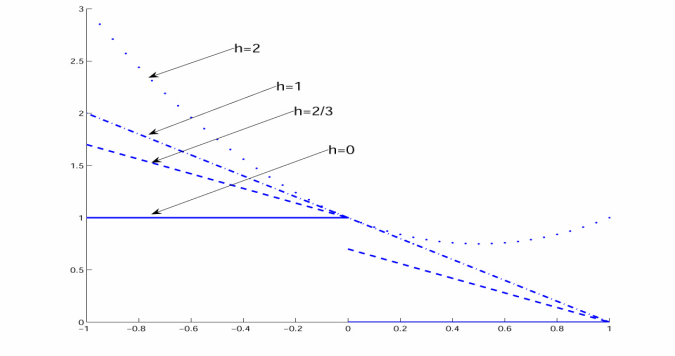

FIGURE 1. Examples of loss functions. The solid line is for $h = 0$ (the $0-1$ loss), the dashed line is for $h = 2/3$, the dashed-dotted line is for $h = 1$ (the hinge loss), the dotted line is for $h = 2$.

This set of losses is representative enough since it describes different types of convexity: for any $h > 1$, ϕ_h is $\beta-$convex on $[-1, 1]$ with $\beta \ge \beta_h \stackrel{\text{def}}{=} (2h-1)^2/(2(h-1)) \ge 2$, for $h = 1$ the loss is linear and for $h < 1$, ϕ_h is non-convex. For $h \ge 0$, we consider

$$A_h(f) \stackrel{\text{def}}{=} A^{\phi_h}(f), f_h^* \stackrel{\text{def}}{=} f_{\phi_h}^* \text{ and } A_h^* \stackrel{\text{def}}{=} A_*^{\phi_h} = A^{\phi_h}(f_h^*).$$

We have

$$f_h^*(x) = \begin{cases} f^*(x) & \text{if } 0 \le h \le 1 \\ \frac{2\eta(x)-1}{2(h-1)} & h > 1, \end{cases} \qquad \forall x \in \mathbb{R}. \tag{4.11}$$

2.1. Optimal Rates of Aggregation Under the Margin Assumption. In the classification setup the margin assumption (cf. (4.2)) has the following form.

($\phi-$MA) $\phi-$Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the $\phi-$margin assumption ($\phi-$MA)(κ) with margin parameter $1 \le \kappa < +\infty$ if there exists $c_\phi > 0$ such that,*

$$\mathbb{E}\left[(\phi(Yf(X)) - \phi(Yf_\phi^*(X)))^2\right] \le c_\phi \left(A^\phi(f) - A^{\phi *}\right)^{1/\kappa}, \tag{4.12}$$

for all measurable functions f with values in $[-1, 1]$.

We first start with a proposition dealing with the $\phi-$margin assumption.

PROPOSITION 4.2. *For any $0 \leq h \leq 1$ and $\kappa \geq 1$, ($\phi_h-MA)(\kappa)$ is equivalent to ($\phi_0-MA)(\kappa)$. For any $h > 1$, (ϕ_h-MA)(1) is satisfied.*

We denote by $\mathcal{P}_\kappa$ the set of all probability distributions π on $\mathcal{X} \times \{-1, 1\}$ satisfying the usual margin assumption (ϕ_0-MA)(κ) of [116].

THEOREM 4.1. *Let $h \geq 0$, $\kappa \geq 1$ be two numbers and $M \geq 2$ be an integer. We assume that $\mathcal{X}$ is infinite.*

If $h \leq 1$, then there exists a family $\mathcal{F}_0$ of M classifiers $f_1, \ldots, f_M$ with values in $\{-1, 1\}$ such that for any statistic $\bar{f}_n$ there exists a probability distribution $\pi \in \mathcal{P}_\kappa$ such that $\min_{f \in \mathcal{F}_0} (A_h(f) - A_h^*) = 0$ and

$$\mathbb{E}\left[A_h(\bar{f}_n) - A_h^*\right] \geq \min_{f \in \mathcal{F}_0} (A_h(f) - A_h^*) + C_2\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}},$$

for any integer $n \geq 1$.

If $h \leq 1$ and $\kappa > 1$, then there exists a family $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ of M classifiers with values in $\{-1, 1\}$ such that for any statistic $\bar{f}_n$ there exists a probability distribution $\pi \in \mathcal{P}_\kappa$ such that $\min_{f \in \mathcal{F}_0} (A_h(f) - A_h^*) \geq C\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}}$ and

$$\mathbb{E}\left[A_h(\bar{f}_n) - A_h^*\right] \geq \min_{f \in \mathcal{F}_0} (A_h(f) - A_h^*) + \left(\frac{(\min_{f \in \mathcal{F}}(A_h(f) - A_h^*))^{\frac{1}{\kappa}} \log M}{n}\right)^{1/2}, \tag{4.13}$$

for any integer $n \geq 1$.

If $h > 1$, there exists a family $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ of M classifiers with values in $\{-1, 1\}$ such that for any statistic $\bar{f}_n$ there exists a probability distribution π on $\mathcal{X} \times \{-1, 1\}$ such that

$$\mathbb{E}\left[A_h(\bar{f}_n) - A_h^*\right] \geq \min_{f \in \mathcal{F}_0} (A_h(f) - A_h^*) + C\frac{\log M}{n},$$

for any integer $n \geq 1$.

For any probability measure π on $\mathcal{X} \times \{-1, 1\}$, any loss function ϕ, any set $\mathcal{F}_0$ of functions from $\mathcal{X}$ to $[-1, 1]$ with cardinality M and any margin parameter $\kappa \geq 1$, consider the rate of aggregation

$$\gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi) = \begin{cases} \left(\frac{\mathcal{B}(\mathcal{F}_0, \pi, \phi)^{\frac{1}{\kappa}} \log M}{\beta_1 n}\right)^{1/2} & \text{if } \mathcal{B}(\mathcal{F}_0, \pi, \phi) \geq \left(\frac{\log M}{\beta_1 n}\right)^{\frac{\kappa}{2\kappa-1}} \\ \left(\frac{\log M}{\beta_2 n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\mathcal{F}_0, \pi, \phi)$ denotes the bias term $\min_{f \in \mathcal{F}_0} (A(f) - A^*)$ and β_1 and β_2 are positive constants depending only on ϕ. It is proved, in Chapter 9, that, if ϕ is a bounded function from $[-1, 1]$ to $\mathbb{R}$ and if the underlying probability measure π satisfies $\phi-$MA(κ), then the Empirical Risk Minimization procedure $\tilde{f}_n = \tilde{f}_n^{ERM}$ satisfies, for any family $\mathcal{F}_0$ of functions $f_1, \ldots, f_M$ with values in $[-1, 1]$,

$$\mathbb{E}[A^\phi(\tilde{f}_n) - A^{\phi *}] \leq \min_{f \in \mathcal{F}_0} (A^\phi(f) - A^{\phi *}) + \gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi). \tag{4.14}$$

Moreover, it is proved in Chapter 9 that if ϕ is convex, then the CAEW procedure $\tilde{f}_n = \tilde{f}_{n,\beta}^{CAEW}$ with temperature parameter $\beta = 1$ and the AEW procedure $\tilde{f}_n = \tilde{f}_n^{AEW}$ satisfy (4.14). Besides, corollary 4.4 of [75] provides the following result. If ϕ is $\beta-$convex

for a positive number β, then the CAEW procedure with temperature parameter β, satisfies

$$\mathbb{E}[A^{\phi}(\tilde{f}_{n,\beta}^{CAEW}) - A^{\phi *}] \leq \min_{f \in \mathcal{F}_0}(A^{\phi}(f) - A^{\phi *}) + \beta \frac{\log M}{n}. \tag{4.15}$$

Remark that the last result, does not require a margin assumption. This can be explained by the fact that, for $h > 1$, assumption ϕ_h-MA(1) is automatically satisfied.

Thus, if we allow the residual term of aggregation to depend on the bias term $\mathcal{B}(\mathcal{F}_0, \pi, \phi)$, in the same spirit as in Chapter 3, we find that $h \longmapsto \mathcal{R}(n, M, \kappa, \mathcal{F}_0, \pi, \phi_h)$, where

$$\mathcal{R}(n, M, \kappa, \mathcal{F}_0, \pi, \phi) = \begin{cases} \beta \frac{\log M}{n} & \text{if } \phi \text{ is } \beta-\text{convex} \\ \gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi) & \text{otherwise,} \end{cases} \tag{4.16}$$

is an optimal rate of aggregation for the scale of loss functions $(\phi_h)_{h \geq 0}$. Nevertheless, the lower bound construction worked out in Theorem 4.1 cannot guarantee that the optimal rate of aggregation for $0 \leq h \leq 1$ is actually not

$$\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa - 1}}.$$

Indeed, the lower bound obtained in (4.13), is constructed on the distribution π such that the bias term $\mathcal{B}(\mathcal{F}_0, \pi, \phi)$ equals to $((\log M)/n)^{\kappa/(2\kappa-1)}$, so, the residual term defined in (4.16) is, up to a constant factor, equal to $((\log M)/n)^{\kappa/(2\kappa-1)}$. Nevertheless, if $\gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi_h)$ is not the optimal rate of aggregation for $0 \leq h \leq 1$, then the ERM procedure cannot be the optimal aggregation procedure (cf. Theorem 4.2 below).

2.2. Suboptimality of Penalized ERM Procedures in Classification under Margin Assumption. In this Section we prove a lower bound under the margin assumption for any selector and we give a more precise lower bound for penalized ERM procedures.

THEOREM 4.2. *Let $M \geq 2$ be an integer, $\kappa \geq 1$ be a real number, $\mathcal{X}$ be infinite and $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a loss function such that $a_{\phi} \stackrel{\text{def}}{=} \phi(-1) - \phi(1) > 0$. There exists a family $\mathcal{F}_0$ of M classifiers with values in $\{-1, 1\}$ satisfying the following.*

Let $\tilde{f}_n$ be a selector with values in $\mathcal{F}_0$. Assume that $\sqrt{(\log M)/n} \leq 1/2$. There exists a probability measure $\pi \in \mathcal{P}_{\kappa}$ and an absolute constant $C_3 > 0$ such that $\tilde{f}_n$ satisfies

$$\mathbb{E}\left[A^{\phi}(\tilde{f}_n) - A_*^{\phi}\right] \geq \min_{f \in \mathcal{F}}\left(A^{\phi}(f) - A_*^{\phi}\right) + C_3\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa - 1}}. \tag{4.17}$$

Consider the penalized ERM procedure $\tilde{f}_n^{pERM}$ associated with $\mathcal{F}$, defined by

$$\tilde{f}_n^{pERM} \in \operatorname{Arg\,min}_{f \in \mathcal{F}}(A_n^{\phi}(f) + \operatorname{pen}(f))$$

where the penalty function $\operatorname{pen}(\cdot)$ satisfies $|\operatorname{pen}(f)| \leq C\sqrt{(\log M)/n}, \forall f \in \mathcal{F}$, with $0 \leq C < \sqrt{2}/3$. Assume that $1188\pi C^2 M^{9C^2} \log M \leq n$. If $\kappa > 1$ then, there exists a probability measure $\pi \in \mathcal{P}_{\kappa}$ and an absolute constant $C_4 > 0$ such that the penalized ERM procedure $\tilde{f}_n^{pERM}$ satisfies

$$\mathbb{E}\left[A^{\phi}(\tilde{f}_n^{pERM}) - A_*^{\phi}\right] \geq \min_{f \in \mathcal{F}}\left(A^{\phi}(f) - A_*^{\phi}\right) + C_4\sqrt{\frac{\log M}{n}}. \tag{4.18}$$

REMARK 4.1. *Inspection of the proof shows that Theorem 4.2 is valid for any family $\mathcal{F}_0$ of classifiers $f_1, \ldots, f_M$, with values in $\{-1, 1\}$, such that there exist points $x_1, \ldots, x_{2^M}$ in $\mathcal{X}$ satisfying $\big\{(f_1(x_j), \ldots, f_M(x_j)) : j = 1, \ldots, 2^M\big\} = \{-1, 1\}^M$.*

REMARK 4.2. *If we use a penalty function such that* $|\mathrm{pen}(f)| \leq \gamma n^{-1/2}, \forall f \in \mathcal{F}_0$, *where* $\gamma > 0$ *is an absolute constant (i.e.* $0 \leq C \leq \gamma(\log M)^{-1/2}$*), then the condition "*$(3376C)^2(2\pi M^{36C^2}\log M) \leq n$*" of Theorem 4.2 is equivalent to "n greater than a constant".*

REMARK 4.3. *It has been observed in [22] that several model selection methods can be described in the following way: for each pair of functions* $(f_k, f_{k'})$*, a threshold* $\tau(k, k', D_n)$ *is built and the function* f_k *is favored with respect to the function* $f_{k'}$ *if*

$$A_n^\phi(f_k) - A_n^\phi(f_{k'}) \leq \tau(k, k', D_n).$$

In the penalization setting, the threshold $\tau(k, k', D_n)$ *is given by* $\tau(k, k', D_n) = \mathrm{pen}(f_{k'}) - \mathrm{pen}(f_k)$*. In the setting of Theorem 4.2, if we select* $\tilde{f}_n = f_{\hat{k}}$ *such that for any* $k \in \{1, \dots, M\}$,

$$A_n^\phi(\tilde{f}_n) - A_n^\phi(f_k) \leq \tau(\hat{k}, k, D_n),$$

and if the threshold satisfies $\tau(k, k', D_n) \leq C\sqrt{(\log M)/n}, \forall k, k' \in \{1, \dots, M\}$*, then* $\tilde{f}_n$ *satisfies (4.18) for the same class* $\mathcal{F}_0$ *and probability measure* π *as in Theorem 4.2.*

REMARK 4.4. *It has been proved in chapter 14 of [47] that no selector (that is a statistic with values in* $\mathcal{F}_0$*) can mimic the oracle with rates faster than* $((\log M)/n)^{1/2}$ *for the* $0-1$ *loss function. Similar bounds in a less general form have been obtained earlier in [118, 106]. In Theorem 4.2, we show that this is still the case for the pERM procedure even if we work under the* ϕ_0-*margin assumption with margin parameter* $\kappa > 1$*.*

Theorem 4.2 states that the ERM procedure (and even penalized ERM procedures) cannot mimic the best classifier in $\mathcal{F}_0$ with rate faster than $((\log M)/n)^{1/2}$ if the basis classifiers in $\mathcal{F}_0$ are different enough and under a very mild assumption on the loss. If there is no margin assumption (which corresponds to the case $\kappa = +\infty$), the result of Theorem 4.2 can be easily deduced from the lower bound in Chapter 7 of [47]. The main message of Theorem 4.2 is that such a negative statement remains true even under the margin assumption MA(κ). Selectors aggregates cannot mimic the oracle faster than $((\log M)/n)^{1/2}$ in general. Under MA(κ), they cannot mimic the best classifier in $\mathcal{F}_0$ with rates faster than $((\log M)/n)^{\kappa/(2\kappa-1)}$ (which is greater than $(\log M)/n$ when $\kappa > 1$).

We know, according to [75], that the CAEW procedure mimics the best classifier in $\mathcal{F}_0$ at the rate $(\log M)/n$ if the loss is $\beta-$convex (cf. 4.15). This and Theorem 4.2 show that penalized ERM procedures are suboptimal aggregation procedures when the loss function is $\beta-$convex. In particular, if one wants to construct adaptive classifiers, then it is better in the rate to consider aggregation procedures with exponential weights.

Remark that, when $h \leq 1$, even if we assume that $(\phi_h-\mathrm{MA})(\kappa)$ holds then, the ERM procedure cannot achieve the rate $((\log M)/n)^{\frac{\kappa}{2\kappa-1}}$ in general. To achieve such a rate, we need to assume that the bias $\min_{f\in\mathcal{F}_0}(A_h(f) - A_h^*)$ is not greater than $((\log M)/n)^{\frac{\kappa}{2\kappa-1}}$. Thus, the behavior of the ERM depends on both the margin and the approximation of the model.

3. Gaussian Regression Framework.

Take $\mathcal{Z} = \mathbb{R}^2$ and let $Z = (X, Y)$ be a couple of random variables on $\mathcal{Z} = \mathbb{R} \times \mathbb{R}$ such that

$$Y = f^*(X) + \sigma\zeta,$$

where ζ is a standard gaussian random variable independent of X and $\sigma > 0$. We consider the prediction of Y given X. The best prediction using the quadratic loss is the regression

function

$$f^*(X) = \mathbb{E}[Y|X].$$

We want to estimate f^* w.r.t. the $L^2(P^X)-$risk, where P^X is the marginal probability distribution of X. Recall that the norm in $L^2(P^X)$ is defined by $||f||_{L^2(P^X)} = (\int f^2 dP^X)^{1/2}$. The loss function is defined by

$$Q((x,y),f) = (y - f(x))^2, \tag{4.19}$$

for any $(x,y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$. Pythagora's theorem yields

$$A(f) = \mathbb{E}[Q((X,Y),f)] = ||f^* - f||^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right].$$

Hence, f^* is a minimizer of $A(f)$ and $A^* = \mathbb{E}[\zeta^2]$.

According to [114], the optimal rate of aggregation in our gaussian regression setup is

$$\frac{\log M}{n}.$$

This rate is achieved by the CAEW procedure with suitably chosen temperature parameter β (cf.[75]). This fast rate of aggregation can be explained by the fact that the intrinsic margin parameter in the gaussian regression setup under the $L^2(P^X)-$risk is equal to 1, which is the best case for the margin parameter.

In the following theorem we prove that selectors (like usual penalized ERM procedures) cannot achieve this rate and thus are suboptimal aggregation procedures, as compared to the aggregation methods with exponential weights.

THEOREM 4.3. *Let $M \geq 2$ be an integer. In the gaussian regression model describe above with $\mathcal{X} = [0,1]$, there exists a family $\mathcal{F}_0$ of M functions $f_1, \ldots, f_M$ such that for any selector $\tilde{f}_n$, there exists a probability measure π of (X,Y) on $[0,1] \times \mathbb{R}$ with regression function f^* of Y given X satisfying*

$$\mathbb{E}\left[||\tilde{f}_n - f^*||^2_{L^2(P^X)}\right] \geq \min_{f \in \mathcal{F}_0}\left(||f - f^*||^2_{L^2(P^X)}\right) + C_3\sqrt{\frac{\log M}{n}},$$

for any integer $n \geq 1$ and where $C_3 > 0$ is an absolute constant.

A similar result is given in [86] for the bounded regression framework. The authors proved that a selector cannot mimic the oracle faster than $n^{-1/2}$. Here, our bound is sharp, since there is the factor $\sqrt{\log M}$ in the bound. The same factor appears in the upper bound for ERM.

4. Density Estimation Framework.

Let $(\mathcal{Z}, \mathcal{T}, \mu)$ be a measurable space. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability distribution. We assume that π is absolutely continuous w.r.t. to μ and denote by f^* a version of the density of π w.r.t. μ. Consider the set $\mathcal{F}$ of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$ and the loss function

$$Q(z,f) = -\log f(z),$$

defined for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have

$$A(f) = \mathbb{E}[Q(Z,f)] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z)) d\pi(z),$$

where $K(f^*|f)$ is the Kullback-Leibler divergence between f^* and f. Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} \log(f^*(z)) d\pi(z)$.

Instead of using the Kullback-Leibler loss, one can use the quadratic loss. For this setup, consider $\mathcal{F} = L^2(\mu) \stackrel{\text{def}}{=} L^2(\mathcal{Z}, \mathcal{T}, \mu)$. Define the loss function

$$Q(z, f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z), \tag{4.20}$$

for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have, for any $f \in \mathcal{F}$,

$$A(f) = \mathbb{E}\left[Q(Z, f)\right] = ||f^* - f||^2_{L^2(\mu)} - \int_{\mathcal{Z}} (f^*(z))^2 d\mu(z).$$

Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

THEOREM 4.4. *Let $M \geq 2$ be an integer. For the setup of density estimation problem with $\mathcal{Z} = [0, 1]$, there exists a family $\mathcal{F}_0$ of M functions $f_1, \ldots, f_M$ such that for any selector $\tilde{f}_n$ there exists a probability measure π on $[0, 1]$ with density function f^* w.r.t. the Lebesgue measure on $[0, 1]$ satisfying*

$$\mathbb{E}\left[||\tilde{f}_n - f^*||_2^2\right] \geq \min_{f \in \mathcal{F}_0} \left(||f - f^*||_2^2\right) + C_3 \sqrt{\frac{\log M}{n}},$$

and

$$\mathbb{E}\left[K(\tilde{f}_n | f^*)\right] \geq \min_{f \in \mathcal{F}_0} \left(K(f | f^*)\right) + C_3 \sqrt{\frac{\log M}{n}},$$

for any integer $n \geq 1$ such that $\sqrt{(\log M)/(2n)} \leq 2$.

Combining [75] and the result of Chapter 2, the optimal rate of aggregation for this estimation problem is

$$\frac{\log M}{n}$$

and the CAEW procedure with suitable choice of temperature parameter β attains this rate of aggregation. Theorem 4.4 shows that this rate cannot be achieved by the penalized ERM procedure.

5. Direct suboptimality of pERM in regression and density estimation.

The following theorems can be deduced from Theorem 4.3 and 4.4. However, they can be proven directly without using results from the minimax theory due to the special form of the ERM and pERM procedures. We give the corresponding proofs in Section 7.

THEOREM 4.5. *Let $M \geq 2$ be an integer. In the gaussian regression model described above with $\mathcal{X} = [0, 1]$, there exists a family $\mathcal{F}_0$ of M functions $f_1, \ldots, f_M$ and a probability measure π such that the penalized ERM procedure*

$$\tilde{f}_n^{pERM} \in \operatorname{Arg\,min}_{f \in \mathcal{F}_0} (A_n(f) + \operatorname{pen}(f)),$$

where $|\operatorname{pen}(f)| \leq C\sqrt{(\log M)/n}, \forall f \in \mathcal{F}_0$, and $0 \leq C < \sigma/(4\sqrt{2}c^*)$ is an absolute constant, satisfies

$$\mathbb{E}\left[||\tilde{f}_n^{pERM} - f^*||^2_{L^2(P^X)}\right] \geq \min_{f \in \mathcal{F}_0} \left(||f - f^*||^2_{L^2(P^X)}\right) + C_3 \sqrt{\frac{\log M}{n}},$$

for any integer $n \geq 1$ such that $2n^{-1} \log[(M-1)(M-2)] \leq 1/4$ where C_3 is an absolute constant and c^* is the constant in Sudakov's minoration (cf. Theorem 4.8 below).

THEOREM 4.6. *Let $M \geq 2$ be an integer. For the setup of density estimation problem with $\mathcal{Z} = [0, 1]$, there exists a family $\mathcal{F}_0$ of M functions $f_1, \ldots, f_M$ and a probability*

measure π such that the penalized ERM procedure w.r.t. the L^2 loss,

$$\tilde{f}_n^{pERM} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0}\Big(\int_{\mathbb{R}} f^2(x)dx - \frac{2}{n}\sum_{i=1}^n f(X_i) + \operatorname{pen}(f)\Big)$$

where $|\operatorname{pen}(f)| \leq C\sqrt{(\log M)/n}, \forall f \in \mathcal{F}_0$, and $0 < C < \sqrt{2}/3$ is an absolute constant, satisfies

$$\mathbb{E}\left[||\tilde{f}_n^{pERM} - f^*||_2^2\right] \geq \min_{f\in\mathcal{F}_0}\left(||f - f^*||_2^2\right) + C_3\sqrt{\frac{\log M}{n}},$$

and for the penalized ERM procedure w.r.t. the Kullback-Leibler loss:

$$\tilde{f}_n^{pERM} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0}\Big(\int_{\mathbb{R}} -\frac{1}{n}\sum_{i=1}^n \log f(X_i) + \operatorname{pen}(f)\Big)$$

we have

$$\mathbb{E}\left[K(\tilde{f}_n^{pERM}|f^*)\right] \geq \min_{f\in\mathcal{F}_0}\left(K(f|f^*)\right) + C_3\sqrt{\frac{\log M}{n}},$$

for any integer $n \geq 1$ such that $C\sqrt{(\log M)/n} \leq 1/2$ where C_3 is an absolute constant.

6. Discussion and Open Problems.

Here we discuss the results of this chapter concerning classification.
We recall the following definition

$$\gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi) = \begin{cases} \left(\frac{\mathcal{B}(\mathcal{F}_0,\phi,\pi)^{\frac{1}{\kappa}}\log M}{n}\right)^{1/2} & \text{if } \mathcal{B}(\mathcal{F}_0,\phi,\pi) \geq \left(\beta_1\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} \\ \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\mathcal{F}_0, \phi, \pi)$ denotes the bias $\min_{f\in\mathcal{F}_0}\left(A^\phi(f) - A^{\phi*}\right)$. The following table summarizes results on optimal rates of aggregation in classification.

Loss function ϕ_h	$h=0$ $0-1$loss	$0<h<1$	$h=1$ Hinge loss	$h>1$ $\beta-$ convex losses
Margin Assumption	Not automatically satisfied ($\kappa = +\infty$)			Automatically satisfied with $\kappa = 1$
Optimal rate of aggregation	$\gamma(n, M, \kappa, \mathcal{F}_0, \pi, \phi_h)$ (conjecture)			$(\log M)/n$
Optimal aggregation procedure	ERM (conjecture)		ERM or AEW (conjecture)	CAEW
ERM or pERM	Optimal (conjecture)			Suboptimal
AEW	?		Optimal (conjecture)	Optimal (conjecture)
CAEW	?			Optimal

Table 1. Optimal rate of aggregation, optimal aggregation procedures and margin assumption for the continuous scale of loss functions of Figure 2.

It is easy to see that CAEW is optimal when ϕ_h is the Hinge loss and when the margin parameter κ is strictly greater than 1 and in the case where $\kappa = 1$ the CAEW procedure achieves the optimal rate of aggregation up to a logarithmic factor (cf. Chapter 3). In the case $h > 1$, the loss function is convex, so that

$$\frac{1}{n}\sum_{i=1}^{n} A_h(\tilde{f}_{k,\beta}^{(AEW)}) \leq A_h(\tilde{f}_{n,\beta}^{(CAEW)})$$

and less observations are used for the construction of $\tilde{f}_{k,\beta}^{(AEW)}, 1 \leq k \leq n-1$, than for the construction of $\tilde{f}_{n,\beta}^{(AEW)}$. We can therefore expect the ϕ_h-risk of $\tilde{f}_{n,\beta}^{(AEW)}$ to be smaller than the ϕ_h-risk of $\tilde{f}_{k,\beta}^{(AEW)}$ for all $1 \leq k \leq n-1$ and hence smaller than the ϕ_h-risk of $\tilde{f}_{n,\beta}^{(CAEW)}$. Moreover, according to (4.14), AEW is optimal when $h > 1$ and when the bias is smaller than $(\log M)/n$. Next, it is easy to get from (4.14) that, we have for any convex loss ϕ and all $\epsilon > 0$

$$\mathbb{E}[A^{\phi}(\tilde{f}_n^{(AEW)}) - A^{\phi *}] \leq (1+\epsilon)\min_{f\in\mathcal{F}}(A^{\phi}(f) - A^{\phi *}) + \frac{C}{\epsilon}\frac{\log M}{n}.$$

Thus, the AEW procedure is likely to be optimal for loss functions ϕ_h, with $h > 1$.

We just proved that the ERM procedure is optimal only for non-convex losses (except for the borderline case of the hinge loss). But, in those cases, the implementation of the ERM procedure requires the minimization of a function which is not convex, thus this procedure is computationally hard and is sometimes not efficient from a practical point of view. Actually, convex surrogate for the $0-1$ loss have been introduced to avoid the minimization of non-convex functionals. Thus, the ERM procedure is theoretically optimal only for non-convex losses but in that case it is practically inefficient and it is practically efficient only for the cases where ERM is theoretically suboptimal.

If we assume that the conjectures of Table 1 are true, the Hinge loss is really hinge for three different reasons. For losses "between" the hinge loss and the $0-1$ loss, we have:

- the intrinsic margin parameter is $\kappa = +\infty$,
- an optimal aggregation procedure is the ERM,
- the optimal rate of aggregation depends both on the margin parameter and on the approximation property of the class $\mathcal{F}_0$ (through its bias).

For losses " over" the Hinge loss ($h > 1$), we have:

- the intrinsic margin parameter is $\kappa = 1$,
- an optimal aggregation procedure is CAEW and the ERM is suboptimal
- the optimal rate of aggregation is the fast aggregation rate $(\log M)/n$.

Moreover for the hinge loss we get, by linearity

$$\min_{f\in\mathcal{C}} A_1(f) - A_1^* = \min_{f\in\mathcal{F}} A_1(f) - A_1^*,$$

where $\mathcal{C}$ is the convex hull of $\mathcal{F}$. Thus, for the particular case of the hinge loss, "model selection" aggregation and "convex" aggregation are identical problems (cf. Chapter 3 for more details).

The intrinsic margin parameter is a very good characterization of the "difficulty" of a model. For model with an intrinsic margin parameter $\kappa = 1$ (density estimation, regression, classification w.r.t. $\beta-$convex losses), we can, under a complexity assumption achieve rates of convergence as fast as approximately $1/n$, and we can aggregate as fast as $(\log M)/n$. But, for models with an intrinsic margin parameter $\kappa = +\infty$ (classification w.r.t. a non

$\beta-$convex loss), we cannot expect convergence rates faster than $n^{-1/2}$ and aggregation rates faster than $\sqrt{(\log M)/n}$. Nevertheless, for models with "bad" intrinsic margin, we can assume to work under an additional margin assumption with a margin parameter $1 < \kappa \leq +\infty$. Under this assumption we can achieve, in these models, the fast convergence rate approaching n^{-1} (under an additive complexity assumption) and $\gamma(n, M, \kappa, \mathcal{F}, \pi, \phi)$ for aggregation rate. In the aggregation case we can see that a complexity assumption is needed if we want to benefit from the margin assumption. Otherwise the bias term is greater than an absolute constant in general and thus, the aggregation rate is $\sqrt{(\log M)/n}$ like in the case of no margin assumption. Finally, we can see that the margin parameter is strongly related to the convexity of the loss function of the model (cf. Proposition 4.2). This may give an explanation why convexity is so important here.

7. Proofs.

Proof of Proposition 4.1: If ϕ is strongly convex then, there exists $a > 0$ such that $\phi''(x) \geq a$. To complete the proof, it suffices to remark that ϕ' is bounded on $[-1, 1]$.

LEMMA 4.1. *Let $\phi : \mathbb{R} \longmapsto \mathbb{R}_+$ be a loss function. For any f, g from $\mathcal{X}$ to $\{-1, 1\}$, we have*

$$A^\phi(f) - A^\phi(g) = a_\phi(A_0(f) - A_0(g)) \text{ where } a_\phi = \phi(-1) - \phi(1).$$

Proof: We have

$$\begin{aligned}\mathbb{E}[\phi(Yf(X))|X] &= \mathbb{E}[\phi(Y)|X]\mathbb{1}_{f(X)=1} + \mathbb{E}[\phi(-Y)|X]\mathbb{1}_{f(X)=-1} \\ &= [\phi(1)\eta(X) + \phi(-1)(1-\eta(X))]\mathbb{1}_{f(X)=1} + [\phi(-1)\eta(X) + \phi(1)(1-\eta(X))]\mathbb{1}_{f(X)=-1}.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}[\phi(Yf(X))|X] - \mathbb{E}[\phi(Yg(X))|X] \\ &\quad=[\phi(1)\eta(X) + \phi(-1)(1-\eta(X))](\mathbb{1}_{f(X)=1} - \mathbb{1}_{g(X)=1}) \\ &\qquad + [\phi(-1)\eta(X) + \phi(1)(1-\eta(X))](\mathbb{1}_{f(X)=-1} - \mathbb{1}_{g(X)=-1}) \\ &\quad=(\phi(1) - \phi(-1))(1 - 2\eta(X))\frac{g(X) - f(X)}{2}.\end{aligned}$$

■

Proof of Proposition 4.2: Let $0 \leq h \leq 1$ and $\kappa \geq 1$. Assume that $(\phi_0-\text{MA})(\kappa)$ holds. Let f be a function defined on $\mathcal{X}$ with values in $[-1, 1]$. By convexity and (4.11), we have

$$\begin{aligned}\mathbb{E}[(\phi_h(Yf(X)) - \phi_h(Yf_h^*(X)))^2] \quad\leq\quad & h\mathbb{E}[(\phi_1(Yf(X)) - \phi_1(Yf^*(X)))^2] \\ & +(1-h)\mathbb{E}[(\phi_0(Yf(X)) - \phi_0(Yf^*(X)))^2].\end{aligned}$$

According to Proposition 3.1 of Chapter 3, $(\phi_1-\text{MA})(\kappa)$ is satisfied. So, using $(\phi_0-\text{MA})(\kappa)$, $(\phi_1-\text{MA})(\kappa)$ and concavity of $x \longmapsto x^{1/\kappa}$ we obtain

$$\begin{aligned}\mathbb{E}[(\phi_h(Yf(X)) - \phi_h(Yf_h^*(X)))^2] \quad&\leq\quad hc_1(A_1(f) - A_1^*)^{1/\kappa} + (1-h)c_0(A_0(f) - A_0^*)^{1/\kappa} \\ &\leq\quad \max(c_0, c_1)(A_h(f) - A_h^*)^{1/\kappa}.\end{aligned}$$

Thus, $(\phi_h-\text{MA})(\kappa)$ holds.

Assume that $(\phi_h-\text{MA})(\kappa)$ holds. Let f be a function defined on $\mathcal{X}$ with values in $[-1, 1]$. We want to prove that $(\phi_0-\text{MA})(\kappa)$ holds. Taking $g = \text{sign}(f)$ we have

$\phi_0(Yg(X)) = \phi_0(Yf(X))$ thus we can assume that f takes its values in $\{-1,1\}$. We have

$$A_h(f) - A_h^* = (1+h)(A_0(f) - A_0^*)$$

and

$$\mathbb{E}\Big[(\phi_h(Yf(X)) - \phi_h(Yf_h^*(X)))^2\Big] = (1+h)^2\mathbb{E}\Big[(\phi_0(Yf(X)) - \phi_0(Yf_h^*(X)))^2\Big].$$

So, $(\phi_0-\mathrm{MA})(\kappa)$ holds.

Let $h > 1$ be a real number and f be a real valued function defined on $\mathcal{X}$. We have

$$A_h(f) - A_h^* = (h-1)\mathbb{E}[(f(X) - f_h^*(X))^2]$$

and

$$|\phi_h(x) - \phi_h(y)| \leq (2h+1)|x-y|, \quad \forall |x|,|y| \leq \max(1, 1/(2(h-1))).$$

So, we have

$$\begin{aligned}\mathbb{E}\left[(\phi_h(Yf(X)) - \phi_h(Yf_h^*(X)))^2\right] &\leq (2h+1)^2\mathbb{E}\left[(f(X) - f_h^*(X))^2\right] \\ &\leq \frac{(2h+1)^2}{h-1}\Big(A_h(f) - A_h^*\Big).\end{aligned}$$

Thus, $(\phi_h-\mathrm{MA})(1)$ is satisfied.

■

Proof of Theorem 4.1: Let $0 \leq h \leq 1$ and $\kappa \geq 1$.

For any real valued function f we have $A_1(f) - A_1^* \geq A_0(f) - A_0^*$ (cf. [130]) and for any prediction rule f we have $A_1(f) - A_1^* = 2(A_0(f) - A_0^*)$. Hence, we have

$$\begin{aligned}&\sup_{f_1,\ldots,f_M}\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\kappa}\left(\mathbb{E}\left[A_h(\hat{f}_n) - A_h^*\right] - \min_{j=1,\ldots,M}(A_h(f_j) - A_h^*)\right) \\ &\geq \sup_{f_1,\ldots,f_M}\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\kappa}\Big(\mathbb{E}\left[A_p(\hat{f}_n) - A_p^*\right] - \min_{j=1,\ldots,M}(A_p(f_j) - A_p^*)\Big),\end{aligned}$$

where $\sup_{f_1,\ldots,f_M}$ denotes the supremum over all prediction rules $f_1,\ldots,f_M$ and $\inf_{\hat{f}_n}$ is the infimum over all statistics constructed with n observations in our model.

We consider an integer N such that $2^{N-1} \leq M$, $2N-1$ different points of $\mathcal{X}$ denoted by $x_1,\ldots,x_N$, $y_1,\ldots,y_{N-1}$ and a positive number w such that $1 \geq 2(N-1)w$. We denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\}) = P^X(\{y_j\}) = w$ for $j = 1,\ldots,N-1$ and $P^X(\{x_N\}) = 1 - 2(N-1)w$. We consider the cube $\Omega = \{-1,1\}^{N-1}$ and a number $0 < \mathfrak{h} < 1$. For all $\sigma \in \Omega$ we consider

$$\eta_\sigma(x) = \begin{cases} (1+\sigma_j\mathfrak{h})/2 & \text{if } x = x_1,\ldots,x_{N-1},y_1,\ldots,y_{N-1} \\ 1 & \text{if } x = x_N. \end{cases}.$$

For all $\sigma \in \Omega$ we denote by π_σ the probability measure on $\mathcal{X}\times\{-1,1\}$ where P^X is the marginal on $\mathcal{X}$ and η_σ the a conditional probability function of Y knowing X. The Bayes rules associated to π_σ is

$$f_\sigma^*(x) = \begin{cases} \sigma_j & \text{if } x = x_1,\ldots,x_{N-1},y_1,\ldots,y_{N-1} \\ 1 & \text{if } x = x_N. \end{cases}.$$

Assume that $\kappa > 1$. We have $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) = 2(N-1)w\mathbb{1}_{\mathfrak{h}\leq t}$ for any $0 \leq t < 1$. Thus, if we assume that $2(N-1)w \leq \theta\mathfrak{h}^{1/(\kappa-1)}$, where θ is a positive number, then $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) \leq \theta t^{1/(\kappa-1)}$ for all $0 \leq t < 1$. Thus, according to [116] and Chapter 3,

π_σ belongs to $\mathcal{P}_\kappa$ with

$$c_0 \stackrel{\text{def}}{=} c_{\phi_0} = \frac{2}{(\kappa-1)\theta^{\kappa-1}}\Big(\frac{\kappa}{\kappa-1}\Big)^\kappa.$$

We denote by ρ the Hamming distance on Ω. Let $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma,\sigma') = 1$. Then, the Hellinger's distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\left(1 - (1 - 2w(1-\sqrt{1-\mathfrak{h}^2}))^n\right).$$

Take w and $\mathfrak{h}$ such that $2w(1-\sqrt{1-\mathfrak{h}^2}) \le n^{-1}$. Then, $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) \le 2(1-e^{-1}) < 2$ for any integer n.

Let $\hat{f}_n$ be a real-valued statistic. Since only the sign of a classifier is taken into account in the ϕ_0-risk, w.l.o.g. we assume that $\hat{f}_n$ takes its values in $\{-1,1\}$.

Let σ be in Ω. Assume that the underlying probability distribution π of the i.i.d. observations D_n is π_σ. Since π_σ belongs to $\mathcal{P}_\kappa$, we have, conditionally to the observations D_n,

$$A_0(\hat{f}_n) - A_0^* \ge \Big(c_0^{-1}\mathbb{E}_{\pi_\sigma}\Big[|\hat{f}_n(X) - f_\sigma^*(X)|\Big]\Big)^\kappa \ge (c_0^{-1}w)^\kappa \left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j|\right)^\kappa.$$

Taking here the expectation, we obtain

$$\mathbb{E}_{\pi_\sigma}\Big[A_0(\hat{f}_n) - A_0^*\Big] \ge (c_0^{-1}w)^\kappa \mathbb{E}_{\pi_\sigma}\left[\left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j|\right)^\kappa\right].$$

Using Jensen's inequality and Lemma 4.2 (p. 77), we obtain:

$$\inf_{\hat{f}_n}\sup_{\pi\in\{\pi_\sigma:\sigma\in\Omega\}}\Big(\mathbb{E}_{\pi_\sigma}\Big[A_0(\hat{f}_n) - A_0^*\Big]\Big) \ge \Big(\frac{(N-1)w}{2c_0e^2}\Big)^\kappa.$$

Take now $N = \lceil \log M/\log 2\rceil$, $h = ((N-1)/(n\theta))^{(\kappa-1)/(2\kappa-1)}$ and $w = (2nh^2)^{-1}$. We have

$$\inf_{\hat{f}_n}\sup_{\pi\in\{\pi_\sigma:\sigma\in\Omega\}}\Big(\mathbb{E}_{\pi_\sigma}\Big[A_0(\hat{f}_n) - A_0^*\Big]\Big) \ge C_0(\kappa,\theta)\Big(\frac{N-1}{n}\Big)^{\kappa/(2\kappa-1)},$$

where

$$C_0(\kappa,\theta) = \Big(\frac{\theta^{\frac{2\kappa-2}{2\kappa-1}}}{4c_0e^2}\Big)^\kappa.$$

For $\kappa = 1$, we take $h = 1/2$, then $|2\eta_\sigma(X) - 1| \ge 1/2$ a.s. so $\pi_\sigma \in \mathcal{P}_1$. It suffices then to take $w = 2/n$ and $N = \lceil \log M/\log 2\rceil$ to get

$$\inf_{\hat{f}_n}\sup_{\pi\in\{\pi_\sigma:\sigma\in\Omega\}}\Big(\mathbb{E}_{\pi_\sigma}\Big[A_0(\hat{f}_n) - A_0^*\Big]\Big) \ge C_0(1,1)\frac{N-1}{n},$$

for $C_0(1,1) = (2e^2)^{-1}$.

For the case $\min_{f\in\mathcal{F}_0}(A_0(f) - A_0^*) = 0$ we take $\mathcal{F}_0 = \{f_\sigma^* : \sigma\in\Omega\}$ and $\theta = 1$, then,

$$\sup_{\mathcal{F}_0=\{f_1,\dots,f_M\}}\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\kappa}\left(\mathbb{E}\Big[A_0(\hat{f}_n) - A_0^*\Big] - \min_{f\in\mathcal{F}_0}(A_0(f) - A_0^*)\right)$$
$$\ge \inf_{\hat{f}_n}\sup_{\pi\in\{\pi_\sigma:\sigma\in\Omega\}}\mathbb{E}\Big[A_0(\hat{f}_n) - A_0^*\Big] \ge C_0(\kappa,1)\Big(\frac{\log M}{n}\Big)^{\kappa/(2\kappa-1)}.$$

For the case $\min_{f\in\mathcal{F}_0}(A_0(f)-A_0^*)\geq C\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}}$ and $\kappa>1$, we consider for any $\sigma\in\Omega$,

$$f_\sigma(x)=\begin{cases}\sigma_j & \text{if } x=x_1,\dots,x_{N-1},\\ -\sigma_j & \text{if } x=y_1,\dots,y_{N-1}\\ 1 & \text{if } x=x_N.\end{cases}$$

For any $\sigma^{(1)},\sigma^{(2)}\in\Omega$, we have under $\pi_{\sigma^{(2)}}$,

$$A_0(f_{\sigma^{(1)}})-A_0^*=2(N-1)\hbar w=\theta^{\frac{\kappa-1}{2\kappa-1}}\Big(\frac{N-1}{n}\Big)^{\kappa/(2\kappa-1)}.$$

Thus, for $\mathcal{F}_0=\{f_\sigma:\sigma\in\Omega\}$, we have

$$\begin{aligned}&\sup_{\mathcal{F}_0=\{f_1,\dots,f_M\}}\inf_{\hat f_n}\sup_{\pi\in\mathcal{P}_\kappa}\Big(\mathbb{E}\Big[A_0(\hat f_n)-A_0^*\Big]-\min_{f\in\mathcal{F}_0}(A_0(f)-A_0^*)\Big)\\ &\quad\geq\inf_{\hat f_n}\sup_{\pi\in\{\pi_\sigma:\sigma\in\Omega\}}\Big(\mathbb{E}\Big[A_0(\hat f_n)-A_0^*\Big]-\min_{\sigma\in\Omega}(A_0(f_\sigma)-A_0^*)\Big)\\ &\quad\geq(C_0(\kappa,\theta)-\theta^{\frac{\kappa-1}{2\kappa-1}})\Big(\frac{N-1}{n}\Big)^{\frac{\kappa}{2\kappa-1}}\\ &\quad=\Big(\frac{\min_{f\in\mathcal{F}}(A_0(f)-A_0^*)^{1/\kappa}(N-1)}{n}\Big)^{1/2},\end{aligned}$$

where we chose

$$\theta=\theta_0\stackrel{\text{def}}{=}\Big[(2e^2)^\kappa\Big[\frac{1}{\kappa}\Big(\frac{\kappa-1}{2\kappa}\Big)^{-1/\kappa}\Big]\Big]^{\frac{2\kappa-1}{(\kappa-1)(\kappa^2+(\kappa-1)^2-1)}}.$$

We have $\min_{f\in\mathcal{F}_0}(A_0(f)-A_0^*)=\theta_0^{\frac{\kappa-1}{2\kappa-1}}\Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa-1}}$ and

$$\begin{aligned}&\sup_{\mathcal{F}_0=\{f_1,\dots,f_M\}}\inf_{\hat f_n}\sup_{\pi\in\mathcal{P}_\kappa}\Big(\mathbb{E}\Big[A_0(\hat f_n)-A_0^*\Big]-\min_{f\in\mathcal{F}_0}(A_0(f)-A_0^*)\Big)\\ &\quad\geq\Big(\frac{\min_{f\in\mathcal{F}_0}(A_0(f)-A_0^*)^{1/\kappa}\log M}{n}\Big)^{1/2}\end{aligned}$$

For the case $h>1$, we consider an integer N such that $2^{N-1}\leq M$, $N-1$ different points $x_1,\dots,x_N$ of $\mathcal{X}$ and a positive number w such that $(N-1)w\leq 1$. We denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\})=w$ for $j=1,\dots,N-1$ and $P^X(\{x_N\})=1-(N-1)w$. Denote by Ω the cube $\{-1,1\}^{N-1}$. For any $\sigma\in\Omega$ and $h>1$, we consider the conditional probability function η_σ in two different cases. If $2(h-1)\leq 1$ we take

$$\eta_\sigma(x)=\begin{cases}(1+2\sigma_j(h-1))/2 & \text{if } x=x_1,\dots,x_{N-1}\\ 2(h-1) & \text{if } x=x_N,\end{cases}$$

and if $2(h-1)>1$ we take

$$\eta_\sigma(x)=\begin{cases}(1+\sigma_j)/2 & \text{if } x=x_1,\dots,x_{N-1}\\ 1 & \text{if } x=x_N.\end{cases}$$

For all $\sigma\in\Omega$ we denote by π_σ the probability measure on $\mathcal{X}\times\{-1,1\}$ with the marginal P^X on $\mathcal{X}$ and the conditional probability function η_σ of Y knowing X.

Consider

$$\rho(h)=\begin{cases}1 & \text{if } 2(h-1)\leq 1\\ (4(h-1))^{-1} & \text{if } 2(h-1)>1\end{cases}\quad\text{and } g_\sigma^*(x)=\begin{cases}\sigma_j & \text{if } x=x_1,\dots,x_{N-1}\\ 1 & \text{if } x=x_N.\end{cases}$$

A minimizer of the ϕ_h-risk when the underlying distribution is π_σ is given by

$$f^*_{h,\sigma} \stackrel{\text{def}}{=} \frac{2\eta_\sigma(x)-1}{2(h-1)} = \rho(h) g^*_\sigma(x), \quad \forall x \in \mathcal{X},$$

for any $h > 1$ and $\sigma \in \Omega$.

When we choose $\{f^*_{h,\sigma} : \sigma \in \Omega\}$ for the set $\mathcal{F} = \{f_1, \ldots, f_M\}$ of basis functions, we obtain

$$\sup_{\{f_1,\ldots,f_M\}} \inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A^*_h\right] - \min_{j=1,\ldots,M} (A_h(f_j) - A^*_h) \right)$$
$$\geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P}: \\ f^*_h \in \{f^*_{h,\sigma}:\sigma\in\Omega\}}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A^*_h\right] \right).$$

Let σ be an element of Ω. Under the probability distribution π_σ, we have $A_h(f) - A^*_h = (h-1)\mathbb{E}[(f(X) - f^*_{h,\sigma}(X))^2]$, for any real-valued function f on $\mathcal{X}$. Thus, for a real valued estimator $\hat{f}_n$ based on D_n, we have

$$A_h(\hat{f}_n) - A^*_h \geq (h-1)w \sum_{j=1}^{N-1} (\hat{f}_n(x_j) - \rho(h)\sigma_j)^2.$$

We consider the projection function $\psi_h(x) = \psi(x/\rho(h))$ for any $x \in \mathcal{X}$, where $\psi(y) = \max(-1, \min(1, y)), \forall y \in \mathbb{R}$. We have

$$\begin{aligned}
\mathbb{E}_\sigma[A_h(\hat{f}_n) - A^*_h] &\geq w(h-1) \sum_{j=1}^{N-1} \mathbb{E}_\sigma(\psi_h(\hat{f}_n(x_j)) - \rho(h)\sigma_j)^2 \\
&\geq w(h-1)(\rho(h))^2 \sum_{j=1}^{N-1} \mathbb{E}_\sigma(\psi(\hat{f}_n(x_j)) - \sigma_j)^2 \\
&\geq 4w(h-1)(\rho(h))^2 \inf_{\hat{\sigma}\in[0,1]^{N-1}} \max_{\sigma\in\Omega} \mathbb{E}_\sigma \left[\sum_{j=1}^{N-1} |\hat{\sigma}_j - \sigma_j|^2 \right],
\end{aligned}$$

where the infimum $\inf_{\hat{\sigma}\in[0,1]^{N-1}}$ is taken over all estimators $\hat{\sigma}$ based on one observation from the statistical experience $\{\pi^{\otimes n}_\sigma | \sigma \in \Omega\}$ and with values in $[0,1]^{N-1}$.

For any $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma, \sigma') = 1$, the Hellinger's distance between the measures $\pi^{\otimes n}_\sigma$ and $\pi^{\otimes n}_{\sigma'}$ satisfies

$$H^2\left(\pi^{\otimes n}_\sigma, \pi^{\otimes n}_{\sigma'}\right) = \begin{cases} 2\left(1 - (1 - 2w(1 - \sqrt{1-h^2}))^n\right) & \text{if } 2(h-1) < 1 \\ 2\left(1 - (1 - 2w(1 - \sqrt{3/4}))^n\right) & \text{if } 2(h-1) \geq 1 \end{cases}.$$

We take

$$w = \begin{cases} (2n(h-1)^2) & \text{if } 2(h-1) < 1 \\ 8n^{-1} & \text{if } 2(h-1) \geq 1. \end{cases}$$

Thus, we have for any $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma, \sigma') = 1$,

$$H^2\left(\pi^{\otimes n}_\sigma, \pi^{\otimes n}_{\sigma'}\right) \leq 2(1 - e^{-1}).$$

To complete the proof we apply Lemma 4.2 (p. 77) with $N = \lceil (\log M)/n \rceil$.

■

Proof of Theorem 4.2: Consider $\mathcal{F}_0$ a family of classifiers $f_1, \ldots, f_M$, with values in $\{-1,1\}$, such that there exist 2^M points $x_1, \ldots, x_{2^M}$ in $\mathcal{X}$ satisfying $\big\{(f_1(x_j), \ldots, f_M(x_j)) : j = 1, \ldots, 2^M\big\} = \{-1,1\}^M \stackrel{\text{def}}{=} \mathcal{S}_M$.

Consider the lexicographic order on $\mathcal{S}_M$:

$$(-1, \ldots, -1) \preccurlyeq (-1, \ldots, -1, 1) \preccurlyeq (-1, \ldots, -1, 1, -1) \preccurlyeq \ldots \preccurlyeq (1, \ldots, 1).$$

Take j in $\{1, \ldots, 2^M\}$ and denote by x'_j the element in $\{x_1, \ldots, x_{2^M}\}$ such that the vector $(f_1(x'_j), \ldots, f_M(x'_j))$ is the $j-$th element of $\mathcal{S}_M$ for the lexicographic order. We denote by φ the bijection between $\mathcal{S}_M$ and $\{x_1, \ldots, x_{2^M}\}$ such that the value of φ at the $j-$th element of $\mathcal{S}_M$ is x'_j. By using the bijection φ we can work independently either on the set $\mathcal{S}_M$ or on $\{x_1, \ldots, x_{2^M}\}$. Without any assumption on the space $\mathcal{X}$, we consider, in what follows, functions and probability measures on $\mathcal{S}_M$. Remark that for the bijection φ we have

$$f_j(\varphi(x)) = x^j, \quad \forall x = (x^1, \ldots, x^M) \in \mathcal{S}_M, \forall j \in \{1, \ldots, M\}.$$

With a slight abuse of notation, we still denote by $\mathcal{F}$ the set of functions $f_1, \ldots, f_M$ defined by $f_j(x) = x^j$, for any $j = 1, \ldots, M$.

First remark that for any f, g from $\mathcal{X}$ to $\{-1,1\}$, using

$$\mathbb{E}[\phi(Yf(X))|X] = \mathbb{E}[\phi(Y)|X]\mathbb{1}_{(f(X)=1)} + \mathbb{E}[\phi(-Y)|X]\mathbb{1}_{(f(X)=-1)},$$

we have

$$\mathbb{E}[\phi(Yf(X))|X] - \mathbb{E}[\phi(Yg(X))|X] = a_\phi(1/2 - \eta(X))(f(X) - g(X)).$$

Hence, we obtain $A^\phi(f) - A^\phi(g) = a_\phi(A_0(f) - A_0(g))$. So, we have for any $j = 1, \ldots, M$,

$$A^\phi(f_j) - A^\phi(f^*) = a_\phi(A_0(f_j) - A_0^*).$$

Moreover, for any $f : \mathcal{S}_M \longmapsto \{-1,1\}$ we have $A_n^\phi(f) = \phi(1) + a_\phi A_n^{\phi_0}(f)$ and $a_\phi > 0$ by assumption, hence,

$$\tilde{f}_n^{pERM} \in \operatorname{Arg}\min_{f \in \mathcal{F}}(A_n^{\phi_0}(f) + \operatorname{pen}(f)).$$

Thus, it suffices to prove Theorem 4.2, when the loss function ϕ is the classical $0-1$ loss function ϕ_0.

We denote by $\mathcal{S}_{M+1}$ the set $\{-1,1\}^{M+1}$ and by $X^0, \ldots, X^M$, $M+1$ independent random variables with values in $\{-1,1\}$ such that X^0 is distributed according to a Bernoulli $\mathcal{B}(w,1)$ with parameter w (that is $\mathbb{P}(X^0 = 1) = w$ and $\mathbb{P}(X^0 = -1) = 1 - w$) and the M other variables $X^1, \ldots, X^M$ are distributed according to a Bernoulli $\mathcal{B}(1/2, 1)$. The parameter $0 \leq w \leq 1$ will be chosen wisely in what follows.

For any $j \in \{1, \ldots, M\}$, we consider the probability distribution $\pi_j = (P^X, \eta^{(j)})$ of a couple of random variables (X, Y) with values in $S_{M+1} \times \{-1,1\}$, where P^X is the probability distribution on S_{M+1} of $X = (X^0, \ldots, X^M)$ and $\eta^{(j)}(x)$ is the regression function at the point $x \in \mathcal{S}_{M+1}$, of $Y = 1$ knowing that $X = x$, given by

$$\eta^{(j)}(x) = \begin{cases} 1 & \text{if } x^0 = 1 \\ 1/2 + h/2 & \text{if } x^0 = -1, x^j = -1 \\ 1/2 + h & \text{if } x^0 = -1, x^j = 1 \end{cases}, \quad \forall x = (x^0, x^1, \ldots, x^M) \in \mathcal{S}_{M+1},$$

where $h > 0$ is a parameter chosen wisely in what follows. The Bayes rule f^*, associated with the distribution $\pi_j = (P^X, \eta^{(j)})$, is identically equal to 1 on $\mathcal{S}_{M+1}$.

If the probability distribution of (X,Y) is π_j for a $j \in \{1,\dots,M\}$ then, for any $0<t<1$, we have $\mathbb{P}[|2\eta(X)-1| \leq t] \leq (1-w)\mathbb{1}_{h\leq t}$. Now, we take

$$1-w = h^{\frac{1}{\kappa-1}},$$

then, we have $\mathbb{P}[|2\eta(X)-1| \leq t] \leq t^{\frac{1}{\kappa-1}}$ and so $\pi_j \in \mathcal{P}_\kappa$.

We extend the definition of the f_j's to the set $\mathcal{S}_{M+1}$ by $f_j(x) = x^j$ for any $x = (x^0,\dots,x^M) \in \mathcal{S}_{M+1}$ and $j = 1,\dots,M$. Consider $\mathcal{F} = \{f_1,\dots,f_M\}$. Assume that (X,Y) is distributed according to π_j for a $j \in \{1,\dots,M\}$. For any $k \in \{1,\dots,M\}$ and $k \neq j$, we have

$$A_0(f_k) - A_0^* = \sum_{x\in\mathcal{S}_{M+1}} |\eta(x)-1/2||f_k(x)-1|\mathbb{P}[X=x] = \frac{3h(1-w)}{8} + \frac{w}{2}$$

and the excess risk of f_j is given by $A_0(f_j) - A_0^* = (1-w)h/4 + w/2$. Thus, we have

$$\min_{f\in\mathcal{F}} A_0(f) - A_0^* = A_0(f_j) - A_0^* = (1-w)h/4 + w/2.$$

First, we prove the lower bound for any selector. Let $\tilde{f}_n$ be a selector with values in $\mathcal{F}$. If the underlying probability measure is π_j for a $j \in \{1,\dots,M\}$ then,

$$\begin{aligned}\mathbb{E}_n^{(j)}[A_0(\tilde{f}_n) - A_0^*] &= \sum_{k=1}^{M}(A_0(f_k) - A_0^*)\pi_j^{\otimes n}[\tilde{f}_n = f_k] \\ &= \min_{f\in\mathcal{F}}(A_0(f) - A_0^*) + \frac{h(1-w)}{8}\pi_j^{\otimes n}[\tilde{f}_n \neq f_j],\end{aligned}$$

where $\mathbb{E}_n^{(j)}$ denotes the expectation w.r.t. the observations D_n when (X,Y) is distributed according to π_j. Hence, we have

$$\max_{1\leq j\leq M}\{\mathbb{E}_n^{(j)}[A_0(\tilde{f}_n) - A_0^*] - \min_{f\in\mathcal{F}}(A_0(f) - A_0^*)\} \geq \frac{h(1-w)}{8}\inf_{\hat{\phi}_n}\max_{1\leq j\leq M}\pi_j^{\otimes n}[\hat{\phi}_n \neq j],$$

where the infimum $\inf_{\hat{\phi}_n}$ is taken over all tests valued in $\{1,\dots,M\}$ constructed from one observation in the model $(\mathcal{S}_{M+1}\times\{-1,1\}, \mathcal{A}\times\mathcal{T}, \{\pi_1,\dots,\pi_M\})^{\otimes n}$, where $\mathcal{T}$ is the natural $\sigma-$algebra on $\{-1,1\}$. Moreover, for any $j \in \{1,\dots,M\}$, we have

$$K(\pi_j^{\otimes n}|\pi_1^{\otimes n}) \leq \frac{nh^2}{4(1-h-2h^2)},$$

where $K(P|Q)$ is the Kullback-Leibler divergence between P and Q (that is $\int \log(dP/dQ)dP$ if $P << Q$ and $+\infty$ otherwise). Thus, if we apply Lemma 4.3 (p. 78) with $h = ((\log M)/n)^{(\kappa-1)/(2\kappa-1)}$, we obtain the result.

Second, we prove the lower bound for the pERM procedure $\hat{f}_n = \tilde{f}_n^{pERM}$. Now, we assume that the probability distribution of (X,Y) is π_M and we take

$$h = \left(C^2\frac{\log M}{n}\right)^{\frac{\kappa-1}{2\kappa}}. \tag{4.21}$$

We have $\mathbb{E}[A_0(\hat{f}_n) - A_0^*] = \min_{f\in\mathcal{F}}(A_0(f) - A_0^*) + \frac{h(1-w)}{8}\mathbb{P}[\hat{f}_n \neq f_M]$. Now, we upper bound $\mathbb{P}[\hat{f}_n = f_M]$, conditionally to $\mathcal{Y} = (Y_1,\dots,Y_n)$. We have

$$\begin{aligned}&\mathbb{P}[\hat{f}_n = f_M|\mathcal{Y}] \\ &= \mathbb{P}[\forall j = 1,\dots,M-1, A_n^{\phi_0}(f_M) + \mathrm{pen}(f_M) \leq A_n^{\phi_0}(f_j) + \mathrm{pen}(f_j)|\mathcal{Y}] \\ &= \mathbb{P}[\forall j = 1,\dots,M-1, \nu_M \leq \nu_j + n(\mathrm{pen}(f_j) - \mathrm{pen}(f_M))|\mathcal{Y}],\end{aligned}$$

where $\nu_j = \sum_{i=1}^n \mathbb{1}_{(Y_i X_i^j \le 0)}, \forall j = 1, \dots, M$ and $X_i = (X_i^j)_{j=0,\dots,M} \in \mathcal{S}_{M+1}, \forall i = 1, \dots, n$. Moreover, the coordinates $X_i^j, i = 1, \dots, n; j = 0, \dots, M$ are independent, $Y_1, \dots, Y_n$ are independent of $X_i^j, i = 1, \dots, n; j = 1, \dots, M-1$ and $|\mathrm{pen}(f_j)| \le h^{\kappa/(\kappa-1)}, \forall j = 1, \dots, M$. So, we have

$$\begin{aligned}
\mathbb{P}[\hat{f}_n = f_M|\mathcal{Y}] &= \sum_{k=0}^n \mathbb{P}[\nu_M = k|\mathcal{Y}] \prod_{j=1}^{M-1} \mathbb{P}[k \le \nu_j + n(\mathrm{pen}(f_j) - \mathrm{pen}(f_M))|\mathcal{Y}] \\
&\le \sum_{k=0}^n \mathbb{P}[\nu_M = k|\mathcal{Y}] \Big(\mathbb{P}[k \le \nu_1 + 2nh^{\kappa/(\kappa-1)}|\mathcal{Y}]\Big)^{M-1} \\
&\le \mathbb{P}[\nu_M \le \bar{k}|\mathcal{Y}] + \big(\mathbb{P}[\bar{k} \le \nu_1 + 2nh^{\kappa/(\kappa-1)}|\mathcal{Y}]\big)^{M-1},
\end{aligned}$$

where

$$\begin{aligned}
\bar{k} &= \mathbb{E}[\nu_M|\mathcal{Y}] - 2nh^{\kappa/(\kappa-1)} \\
&= \frac{1}{2}\sum_{i=1}^n \Big(\frac{2-4h}{2-3h}\mathbb{1}_{(Y_i=-1)} + \frac{1+h^{1/(\kappa-1)}(h/2-1/2)}{1+h^{1/(\kappa-1)}(3h/4-1/2)}\mathbb{1}_{(Y_i=1)}\Big) - 2nh^{\kappa/(\kappa-1)}.
\end{aligned}$$

Using Einmahl and Masson's concentration inequality (cf. [53]), we obtain

$$\mathbb{P}[\nu_M \le \bar{k}|\mathcal{Y}] \le \exp(-2nh^{2\kappa/(\kappa-1)}).$$

Using Berry-Esséen's theorem (cf. p.471 in [16]), the fact that $\mathcal{Y}$ is independent of $(X_i^j; 1 \le i \le n, 1 \le j \le M-1)$ and $\bar{k} \ge n/2 - 9nh^{\kappa/(\kappa-1)}/4$, we get

$$\mathbb{P}[\bar{k} \le \nu_1 + 2nh^{\frac{\kappa}{\kappa-1}}|\mathcal{Y}] \le \mathbb{P}\left[\frac{n/2-\nu_1}{\sqrt{n}/2} \le 6h^{\frac{\kappa}{\kappa-1}}\sqrt{n}\right] \le \Phi(6h^{\frac{\kappa}{\kappa-1}}\sqrt{n}) + \frac{66}{\sqrt{n}},$$

where Φ stands for the standard normal distribution function. Thus, we have

$$\begin{aligned}
(4.22)\ \mathbb{E}[A_0(\hat{f}_n) - A_0^*] &\ge \min_{f\in\mathcal{F}}(A_0(f) - A_0^*) \\
&+ \frac{(1-w)h}{8}\Big(1 - \exp(-2nh^{2\kappa/(\kappa-1)}) - \Big(\Phi(6h^{\kappa/(\kappa-1)}\sqrt{n}) + 66/\sqrt{n}\Big)^{M-1}\Big).
\end{aligned}$$

Next, for any $a > 0$, by the elementary properties of the tails of normal distribution, we have

$$(4.23)\qquad 1 - \Phi(a) = \frac{1}{\sqrt{2\pi}}\int_a^{+\infty} \exp(-t^2/2)dt \ge \frac{a}{\sqrt{2\pi}(a^2+1)}e^{-a^2/2}.$$

Besides, we have for $0 < C < \sqrt{2}/6$ (a modification for $C = 0$ is obvious) and the condition $(3376C)^2(2\pi M^{36C^2}\log M) \le n$, thus, if we replace h by its value given in (4.21) and if we apply (4.23) with $a = 16C\sqrt{\log M}$, then we obtain

$$(4.24)\qquad \Big(\Phi(6h^{\kappa/(\kappa-1)}\sqrt{n}) + 66/\sqrt{n}\Big)^{M-1} \le \exp\Big[-\frac{M^{1-18C^2}}{18C\sqrt{2\pi\log M}} + \frac{66(M-1)}{\sqrt{n}}\Big].$$

Combining (4.22) and (4.24), we obtain the result with $C_4 = (C/4)\Big(1 - \exp(-8C^2) - \exp(-1/(36C\sqrt{2\pi\log 2}))\Big) > 0$.

■

Proof of Theorem 4.3: We consider a random variable X uniformly distributed on $[0,1]$ and its dyadic representation:

$$X = \sum_{k=1}^{+\infty} X^{(k)} 2^{-k}, \tag{4.25}$$

where $(X^{(k)} : k \geq 1)$ is a sequence of i.i.d. random variables following a Bernoulli $\mathcal{B}(1/2, 1)$ with parameter $1/2$. The random variable X is the design of the regression model worked out here.

We consider $h > 0$, which will be chosen wisely in what follows, and the following regression functions

$$f^*_{(j)}(x) = \left\{ \begin{array}{ll} 2h & \text{if } x^{(j)} = 1 \\ h & \text{if } x^{(j)} = 0, \end{array} \right.$$

for any $j = 1, \ldots, M$ and where x has $\sum_{k=1}^{+\infty} x^{(j)} 2^{-j}$, with $x^{(j)} \in \{0, 1\}$, for dyadic decomposition.

We consider the following dictionary $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ where

$$f_j(x) = 2x^{(j)} - 1, \forall j = 1, \ldots, M.$$

We denote by $\mathbb{P}_j$ the probability measure of (X, Y) taking values in $[0,1] \times \mathbb{R}$, such that X is uniformly distributed on $[0,1]$ and $Y = f^*_{(j)}(X) + \epsilon$, where ϵ is a standard gaussian random variable independent of X.

For any $k, j = 1, \ldots, M$, we have

$$||f_k - f^*_{(j)}||^2_{L^2(P^X)} = \left\{ \begin{array}{ll} \frac{1}{2}[5h^2 + 2] & k \neq j \\ \frac{1}{2}[5h^2 - 2h + 2] & k = j. \end{array} \right.$$

Let $\tilde{f}_n$ be a selector with values in the dictionary $\mathcal{F}_0$ constructed from the sample D_n made of n i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). We denote by $\mathbb{E}^{(j)}_n$ the expectation w.r.t. the sample D_n when (X, Y) is distributed according to $\mathbb{P}_j$. Let $j \in \{1, \ldots, M\}$, we have

$$\begin{aligned} \mathbb{E}^{(j)}_n[||\tilde{f}_n - f^*_{(j)}||^2_{L^2(P^X)}] - \min_{1 \leq k \leq M} ||f_k - f^*_{(j)}||^2_{L^2(P^X)} & \\ = \sum_{k=1}^{M} ||f_k - f^*_{(j)}||^2_{L^2(P^X)} \mathbb{P}_j^{\otimes n}[\tilde{f}_n = f_k] - (1/2)[5h^2 - 2h + 2] & \\ = (1/2)[5h^2 - 2h + 2][1 - \mathbb{P}_j^{\otimes n}[\tilde{f}_n = f_j]] + (1/2)[5h^2 + 2]\mathbb{P}_j^{\otimes n}[\tilde{f}_n \neq f_j] & \\ = \frac{h}{2} \mathbb{P}_j^{\otimes n}[\tilde{f}_n \neq f_j] & \end{aligned}$$

On the other side, the Kullback-Leibler divergence between $\mathbb{P}_j^{\otimes n}$, for a $j \in \{2, \ldots, M\}$, and $\mathbb{P}_1^{\otimes n}$ satisfies

$$K(\mathbb{P}_j^{\otimes n} | \mathbb{P}_1^{\otimes n}) = \frac{n}{2} ||f^*_{(j)} - f^*_{(1)}||^2_{L^2(P^X)} = \frac{nh^2}{2}.$$

Thus, if we take

$$h = \sqrt{\frac{\log M}{8n}},$$

according to Lemma 4.3 (p. 78), we have

$$\inf_{\tilde{f}_n} \max_{1 \leq j \leq M} \mathbb{P}_j^{\otimes n}[\tilde{f}_n \neq f_j] \geq 1/8,$$

where $\inf_{\tilde{f}_n}$ denotes the infimum over all selectors with values in $\mathcal{F}_0$. Then, for a selector $\tilde{f}_n$ there exists $j \in \{1, \dots, M\}$ such that

$$\mathbb{E}_n^{(j)}[||\tilde{f}_n - f^*_{(j)}||^2_{L^2(P^X)}] - \min_{1\le k\le M} ||f_k - f^*_{(j)}||^2_{L^2(P^X)} \ge \frac{1}{16}\sqrt{\frac{\log M}{8n}}.$$

Proof of Theorem 4.4: We consider M density functions on $[0,1]$ given by

$$f^*_{(j)}(x) = \begin{cases} 3/2 & \text{if } x^{(j)} = 1 \\ 1/2 & \text{if } x^{(j)} = 0 \end{cases}$$

for any $x \in [0,1]$ having $\sum_{k=1}^{+\infty} x^{(j)} 2^{-j}$, with $x^{(j)} \in \{0,1\}$, for dyadic decomposition, and we denote by $\mathbb{P}_j$ the probability measure on $[0,1]$ with density function $f^*_{(j)}$ w.r.t. the Lebesgue measure on $[0,1]$. We consider the dictionary $\mathcal{F}_0 = \{f_1, \dots, f_M\}$ such that

$$f_k(x) = \begin{cases} 1+h & \text{if } x^{(j)} = 1 \\ 1-h & \text{if } x^{(j)} = 0, \end{cases} \quad \forall k \in \{1, \dots, M\}$$

for any $x = \sum_{k=1}^{+\infty} x^{(j)} 2^{-j} \in [0,1]$ and for a $h > 0$ chosen wisely in what follows.

For any $k, j \in \{1, \dots, M\}$, we have

$$||f_k - f^*_{(j)}||_2^2 = \begin{cases} \frac{1}{2}\Big[(1/2-h)^2 + (1/2+h)^2\Big] & \text{if } k \neq j \\ (1/2-h)^2 & \text{if } k = j. \end{cases}$$

We denote by $\inf_{\tilde{f}_n}$ the infimum over all selector with values in $\mathcal{F}_0$. We have

$$\inf_{\tilde{f}_n} \max_{1\le j\le M} \Big[\mathbb{E}_n^{(j)}[||\tilde{f}_n - f^*_{(j)}||_2^2] - \min_{1\le k\le M} ||f_k - f^*_{(j)}||_2^2\Big] = h \inf_{\tilde{f}_n} \max_{1\le j\le M} \mathbb{P}_j^{\otimes n}[\tilde{f}_n \neq f_j] \tag{4.26}$$

Moreover, the Kullback-Leibler divergence between $\mathbb{P}_j^{\otimes n}$ and $\mathbb{P}_1^{\otimes n}$, for a $j \in \{2, \dots, M\}$, satisfies

$$K(\mathbb{P}_j^{\otimes n} | \mathbb{P}_1^{\otimes n}) = nK(\mathbb{P}_j^{\otimes n} | \mathbb{P}_1^{\otimes n}) \le \frac{nh^2}{1-h^2}.$$

Thus, taking $h = (1/4)\sqrt{(\log M)/(2n)} \le 1/2$ and applying Lemma 4.3 (p. 78) in (4.26) for the set $\{\mathbb{P}_j^{\otimes n} : j = 1, \dots, M\}$, we complete the proof for the case of the L^2 loss.

For the case of the Kullback-Leibler loss, we have for any $k, j \in \{1, \dots, M\}$,

$$K(f_k | f^*_{(j)}) = \begin{cases} (1/2)[\log(4/3) + 3h^2] & \text{if } k \neq j \\ (1/2)[\log(4/3) - h\log 3 + 2h^2] & \text{if } k = j. \end{cases}$$

Then, using the same arguments than previously, we complete the proof for this case.

Proof of Theorem 4.5: We consider a random variable X uniformly distributed on $[0,1]$ and its dyadic representation:

$$X = \sum_{k=1}^{+\infty} X^{(k)} 2^{-k}, \tag{4.27}$$

where $(X^{(k)} : k \ge 1)$ is a sequence of i.i.d. random variables following a Bernoulli $\mathcal{B}(1/2, 1)$ with parameter $1/2$. The random variable X is the design of the regression model worked out here. For the regression function we take

$$f^*(x) = \begin{cases} 2h & \text{if } x^{(M)} = 1 \\ h & \text{if } x^{(M)} = 0, \end{cases} \tag{4.28}$$

where x has the dyadic decomposition $x = \sum_{k=1}^{+\infty} x^{(k)} 2^{-k}$ and

$$h = \frac{C}{4}\sqrt{\frac{\log M}{n}}.$$

We consider the set $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ of basis functions

$$f_j(x) = 2x^{(j)} - 1, \quad \forall j \in \{1, \ldots, M\}, \tag{4.29}$$

where we consider the dyadic decomposition of $x \in [0,1]$ given by $x = \sum_{k=1}^{+\infty} x^{(k)} 2^{-k}$, where $x^{(k)} \in \{0,1\}, \forall k \geq 1$.

For any $j = 1, \ldots, M-1$,

$$A(f_j) - A^* = ||f_j - f^*||^2_{L^2([0,1])} = \frac{1}{2}[5h^2 + 2]$$

and

$$A(f_M) - A^* = ||f_M - f^*||^2_{L^2([0,1])} = \frac{1}{2}[5h^2 - 2h + 2].$$

Thus we have

$$\min_{j=1,\ldots,M} A(f_j) - A^* = A(f_M) - A^* = (1/2)[5h^2 - 2h + 2].$$

For

$$\tilde{f}_n^{pERM} \in \text{Arg}\min_{f \in \mathcal{F}_0} \Big(\frac{1}{n}\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \text{pen}(f) \Big),$$

we have

$$\mathbb{E}[||\hat{f}_n - f^*||_{L^2([0,1])}] = \min_{j=1,\ldots,M} ||f_j - f^*||_{L^2([0,1])} + 2h\mathbb{P}[\hat{f}_n \neq f_M] \tag{4.30}$$

Now, we upper bound $\mathbb{P}[\hat{f}_n = f_M]$. We have

$$\begin{aligned}
\mathbb{P}[\hat{f}_n = f_M] &= \mathbb{P}[\forall j = 1, \ldots, M-1, A_n(f_M) + \text{pen}(f_M) \leq A_n(f_j) + \text{pen}(f_j)] \\
&= \mathbb{P}[\forall j = 1, \ldots, M-1, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - f_M(X_i))^2 \leq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - f_j(X_i))^2 \\
&\quad + \sqrt{n}(\text{pen}(f_j) - \text{pen}(f_M))] \\
&\leq \mathbb{P}[\forall j = 1, \ldots, M-1, N_M \geq N_j \\
&\quad + \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\frac{h}{2}(\epsilon_i^{(M)}\epsilon_i^{(j)} - 1) + \frac{3h}{2}(\epsilon_i^j - 1) - \frac{C}{\sigma}\sqrt{\log M}],
\end{aligned}$$

where for any $j = 1, \ldots, M$,

$$N_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i\epsilon_i^{(j)} \text{ and } \epsilon_i^{(j)} = 2X_i^{(j)} - 1.$$

It is easy to check that $N_1, \ldots, N_M$ are M normalized standard gaussian random variables uncorrelated (but non independent).

Next, denote by ϵ the family of Rademacher variables $(\epsilon_i^{(j)} : i = 1, \ldots, n, j = 1, \ldots, M)$. We have for any $2C/\sigma < \gamma < (2\sqrt{2}c^*)$ (c^* is given in Theorem 4.8),

$$\begin{aligned}
\mathbb{P}[\hat{f}_n = f_M] &\leq \mathbb{E}\Big[\mathbb{P}[N_M \geq \max_{j=1,\ldots,M-1} N_j - 2C/\sigma\sqrt{\log M}|\epsilon]\Big] \\
&\leq \mathbb{P}[N_M \geq -\gamma\sqrt{\log M} + \mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon]] \qquad (4.31) \\
&\quad + \mathbb{E}\Big[\mathbb{P}[\mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon] - \max_{j=1,\ldots,M-1} N_j \geq (\gamma - 2C/\sigma)\sqrt{\log M}|\epsilon]\Big].
\end{aligned}$$

Remark that, conditionally to ϵ, the vector $(N_1, \ldots, N_{M-1})$ is a linear transform of the gaussian vector $(\zeta_1, \ldots, \zeta_n)$. Hence, $(N_1, \ldots, N_{M-1})$ is a gaussian vector (conditionally to ϵ). Then, we can use the gaussian concentration Theorem (cf. [93]), to obtain the following inequality for the second term of the RHS in (4.31):

(4.32)
$$\mathbb{P}[\mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon] - \max_{j=1,\ldots,M-1} N_j \geq (\gamma - 2C/\sigma)\sqrt{\log M}|\epsilon] \leq \exp(-(C/\sigma - \gamma/2)^2 \log M).$$

Remark that we used $\mathbb{E}[N_j^2|\epsilon] = 1$ for any $j = 1, \ldots, M-1$.

For the first term in the RHS of (4.31), we have

$$\begin{aligned}
&\mathbb{P}\Big[N_M \geq -\gamma\sqrt{\log M} + \mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon]\Big] \\
&\leq \mathbb{P}\Big[N_M \geq -2\gamma\sqrt{\log M} + \mathbb{E}[\max_{j=1,\ldots,M-1} N_j]\Big] \\
&\quad + \mathbb{P}\Big[-\gamma\sqrt{\log M} + \mathbb{E}[\max_{j=1,\ldots,M-1} N_j] \geq \mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon]\Big]
\end{aligned} \tag{4.33}$$

Next, we lower bound $\mathbb{E}[\max_{j=1,\ldots,M-1} N_j]$. Since $(N_1, \ldots, N_{M-1})$ is a gaussian vector (conditionally to ϵ) and for any $k \neq j \in \{1, \ldots, M\}$, we have

$$\mathbb{E}[(N_k - N_j)^2|\epsilon] = \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^{(k)} - \epsilon_i^{(j)})^2$$

then, according to Sudakov's lower bound Theorem (cf. Theorem 4.8 in Section 8), there exits an absolute constant $c^* > 0$, such that

$$c^*\mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon] \geq \min_{k\neq j\in\{1,\ldots,M-1\}} \Big(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^{(k)} - \epsilon_i^{(j)})^2\Big)^{1/2}\sqrt{\log M}. \tag{4.34}$$

Thus, we have

$$c^*\mathbb{E}[\max_{j=1,\ldots,M-1} N_j] \geq \mathbb{E}\Big[\min_{k\neq j\in\{1,\ldots,M-1\}} \Big(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^{(k)} - \epsilon_i^{(j)})^2\Big)^{1/2}\Big]\sqrt{\log M}. \tag{4.35}$$

Moreover, using that $\sqrt{x} \geq x/\sqrt{2}, \forall x \in [0, 2]$, we have

$$\mathbb{E}\Big[\min_{k\neq j\in\{1,\ldots,M-1\}} \Big(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^{(k)} - \epsilon_i^{(j)})^2\Big)^{1/2}\Big] \geq \sqrt{2}\Big(1 - \mathbb{E}\Big[\max_{j\neq k}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^{(k)}\epsilon_i^{(j)}\Big]\Big). \tag{4.36}$$

Besides, using a maximal inequality (cf. Theorem 4.9 in Section 8) and $2n^{-1}\log[(M-1)(M-2)] \leq 1/4$, we have

$$\mathbb{E}\Big[\max_{j\neq k}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^{(k)}\epsilon_i^{(j)}\Big] \leq \Big(\frac{2}{n}\log[(M-1)(M-2)]\Big)^{1/2} \leq \frac{1}{2}. \tag{4.37}$$

Remark that we used Hoeffding's inequality to obtain $\mathbb{E}[\exp(s\xi^{(j,k)})] \leq \exp(s^2/(4n)), \forall s > 0$, where $\xi^{(j,k)} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^{(k)}\epsilon_i^{(j)}$. Then, combining equations (4.35), (4.36) and (4.37), we obtain

$$c^*\mathbb{E}[\max_{j=1,\ldots,M-1} N_j] \geq ((\log M)/2)^{1/2}.$$

Then, using this inequality in the first RHS of (4.33) and the usual inequality on the tail of a gaussian random variable (remark that N_M is a standard gaussian variable), we obtain:

$$\mathbb{P}\Big[N_M \geq -2\gamma\sqrt{\log M} \quad + \quad \mathbb{E}[\max_{j=1,\ldots,M-1} N_j]\Big] \leq \mathbb{P}\Big[N_M \geq ((c^*\sqrt{2})^{-1} - 2\gamma)\sqrt{\log M}\Big]$$

$$\leq \mathbb{P}\Big[N_M \geq ((c^*\sqrt{2})^{-1} - 2\gamma)\sqrt{\log M}\Big] \tag{4.38}$$
$$\leq \exp\Big(-((c^*\sqrt{2})^{-1} - 2\gamma)^2(\log M)/2\Big).$$

Remark that, we used $2\sqrt{2}c^*\gamma \leq 1$. For the second term in (4.33), we use the concentration inequality of Theorem 4.7 of Section 8, to obtain

$$\mathbb{E}\Big[\mathbb{P}\Big[-\gamma\sqrt{\log M} + \mathbb{E}[\max_{j=1,\ldots,M-1} N_j] \geq \mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon]|\epsilon\Big]\Big] \leq \exp(-\gamma^2/4), \tag{4.39}$$

where we used the concentration inequality of Theorem 4.9 and inequality (4.34) to obtain $0 \leq \mathbb{E}[\max_{j=1,\ldots,M-1} N_j|\epsilon] \leq \sqrt{2\log M}$.

Finally, combining (4.31), (4.38), (4.33), (4.32) in the initial inequality (4.31), we obtain

$$\begin{aligned}\mathbb{P}[\hat{f}_n = f_M] &\leq \exp(-2(2C-\gamma)^2\log M) \\ &\quad + \exp\Big(-((c^*\sqrt{2})^{-1} - 2\gamma)^2(\log M)/2\Big) + \exp(-\gamma^2/4)\end{aligned}$$

We complete the proof by using this inequality in (4.30).

■

Proof of Theorem 4.6: We consider a random variable X uniformly distributed on $[0,1]$ and its dyadic representation:

$$X = \sum_{k=1}^{+\infty} X^{(k)} 2^{-k}, \tag{4.40}$$

where $(X^{(k)} : k \geq 1)$ is a sequence of i.i.d. random variables following a Bernoulli $\mathcal{B}(1/2, 1)$ with parameter $1/2$. In the density estimation setup we have n i.i.d. observations of the random variable X. Hence, the density function to estimate is

$$f^*(x) \stackrel{\text{def}}{=} 1, \forall x \in [0,1]. \tag{4.41}$$

We consider the set of basis density functions $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$, where, for any $j \in \{1, \ldots, M-1\}$,

$$f_j(x) = \begin{cases} 3/2 & \text{if } x^{(j)} = 1 \\ 1/2 & \text{if } x^{(j)} = 0 \end{cases} \tag{4.42}$$

where we consider the dyadic decomposition of $x \in [0,1]$ given by $x = \sum_{k=1}^{+\infty} x^{(k)} 2^{-k}$, where $x^{(k)} \in \{0,1\}, \forall k \geq 1$. For $j = M$, we consider

$$f_M(x) = \begin{cases} 3/2 - h & \text{if } x^{(M)} = 1 \\ 1/2 + h & \text{if } x^{(M)} = 0 \end{cases} \tag{4.43}$$

where x has the dyadic decomposition $x = \sum_{k=1}^{+\infty} x^{(k)} 2^{-k}$ and

$$h = \frac{1}{2}\sqrt{\frac{\log M}{n}}.$$

First, we explore the case when the loss is the L^2-norm. For any $j = 1, \ldots, M-1$,

$$A(f_j) - A^* = ||f_j - f^*||^2_{L^2([0,1])} = 1/4$$

and

$$A(f_M) - A^* = ||f_M - f^*||^2_{L^2([0,1])} = (1/2 - h)^2.$$

Thus we have

$$\min_{j=1,\ldots,M} A(f_j) - A^* = A(f_M) - A^* = (1/2 - h)^2.$$

For

$$\tilde{f}_n^{pERM} \in \text{Arg}\min_{f\in\mathcal{F}_0}\Big(\int_{\mathbb{R}} f^2(x)dx - \frac{2}{n}\sum_{i=1}^n f(X_i) + \text{pen}(f)\Big),$$

we have

$$\mathbb{E}[||\hat{f}_n - f^*||_{L^2([0,1])}] = \min_{j=1,\dots,M} ||f_j - f^*||_{L^2([0,1])} + (h-h^2)\mathbb{P}[\hat{f}_n \neq f_M]$$

Now, we upper bound $\mathbb{P}[\hat{f}_n = f_M]$. We have

$$\begin{aligned}\mathbb{P}[\hat{f}_n = f_M] &= \mathbb{P}[\forall j = 1,\dots,M-1, A_n(f_M) + \text{pen}(f_M) \leq A_n(f_j) + \text{pen}(f_j)] \\ &= \mathbb{P}[\forall j = 1,\dots,M-1, \nu_M \leq \nu_j + n(\text{pen}(f_j) - \text{pen}(f_M))],\end{aligned}$$

where $\nu_j = \sum_{i=1}^n X_i^{(j)}, \forall j = 1,\dots,M$. Moreover, $(X_i^{(j)})_{j=1,\dots,M;i=1,\dots,n}$ are i.i.d. $\mathcal{B}(1/2,1)$ and $|\text{pen}(f_j)| \leq h, , \forall j = 1,\dots,M$. So using the same arguments as in the proof of Theorem 4.2, we have

$$\mathbb{P}[\hat{f}_n = f_M] = \mathbb{P}[\nu_M > \bar{k}] + \big(\mathbb{P}[\nu_1 \leq (1-2h)\bar{k} + 2nh]\big)^{M-1},$$

where we choose $\bar{k} = 2nh$. Using Hoeffding's inequality for binomial variables, we have

$$\mathbb{P}[\nu_M \geq 2nh] \leq \exp(-2nh^2).$$

Using the normal approximation Theorem (cf. Theorem 4.10 in Section 8), we get

$$\mathbb{P}[\nu_1 \leq (1-2h)\bar{k} + 2nh] \leq \Phi(8h\sqrt{n}) + \frac{264}{2\sqrt{n}},$$

where Φ is the standard normal distribution function. We complete the proof with the same arguments as in the proof of Theorem 4.2.

Second, with the same notation as in the L^2 case, we have for the Kullback-Leibler divergence case, for any $j = 1,\dots,M-1$,

$$A(f_j) - A^* = K(f_j|f^*) = (1/2)\log(4/3)$$

and

$$A(f_M) - A^* = K(f_M|f^*) = (1/2)\log(4/3) + \epsilon_h,$$

where $\epsilon_h = (1/2)\log\Big(1 - (h(4+4h))/[(3-2h)(1+2h)]\Big)$. Thus we have

$$\min_{j=1,\dots,M} A(f_j) - A^* = A(f_M) - A^* = (1/2)\log(4/3) + \epsilon_h.$$

For

$$\tilde{f}_n^{pERM} \in \text{Arg}\min_{f\in\mathcal{F}_0}\Big(\int_{\mathbb{R}} -\frac{1}{n}\sum_{i=1}^n \log f(X_i) + \text{pen}(f)\Big),$$

we have

$$\mathbb{E}[K(\hat{f}_n|f^*)] = \min_{j=1,\dots,M} K(f_j|f^*) - \epsilon_h\mathbb{P}[\hat{f}_n \neq f_M]$$

Now, we upper bound $\mathbb{P}[\hat{f}_n = f_M]$. We have

$$\begin{aligned}\mathbb{P}[\hat{f}_n = f_M] &= \mathbb{P}[\forall j = 1,\dots,M-1, A_n(f_M) + \text{pen}(f_M) \leq A_n(f_j) + \text{pen}(f_j)] \\ &= \mathbb{P}[\forall j = 1,\dots,M-1, \nu_M \leq \nu_j + n(\text{pen}(f_j) - \text{pen}(f_M))] \\ &\leq \mathbb{P}\Big[\sum_{i=1}^n X_i^{(1)} \leq \sum_{i=1}^n X_i^{(M)} + \frac{20nh}{3\log 3}\Big]\end{aligned}$$

where $\nu_j = -\sum_{i=1}^{n} \log[(1/2-h)X_i^{(j)}+1], \forall j = 1,\dots,M$. Moreover, $(X_i^{(j)})_{j=1,\dots,M;i=1,\dots,n}$ are i.i.d. $\mathcal{B}(1/2,1)$ and $|\text{pen}(f_j)| \leq h, , \forall j = 1,\dots,M$. So using the same arguments as in the proof of Theorem 4.2, we have

$$\mathbb{P}[\hat{f}_n = f_M] = \mathbb{P}[\nu_M > \bar{k}] + \left(\mathbb{P}[\nu_1 \leq (1-2h)\bar{k} + 2nh]\right)^{M-1},$$

where we take $\bar{k} = nh$. We complete the proof with the same arguments as in the previous case.

■

8. Appendix.

The following Lemma is used to establish the lower bounds. It is a slightly different version of the Assouad's Lemma (cf. [115]).

LEMMA 4.2. *Let $(\mathcal{X},\mathcal{A})$ be a measurable space. Consider a set of probability $\{P_\omega/\omega\in\Omega\}$ indexed by the cube $\Omega = \{0,1\}^m$. Denote by $\mathbb{E}_\omega$ the expectation under P_ω. Let $\theta \geq 1$ be a number. Assume that:*

$$\forall \omega,\omega' \in \Omega/\rho(\omega,\omega') = 1,\ H^2(P_\omega, P_{\omega'}) \leq \alpha < 2,$$

then we have

$$\inf_{\hat{w}\in[0,1]^m} \max_{\omega\in\Omega} \mathbb{E}_\omega \left[\sum_{j=1}^{m} |\hat{w}_j - w_j|^\theta\right] \geq m2^{-3-\theta}(2-\alpha)^2$$

where the infimum $\inf_{\hat{w}\in[0,1]^m}$ is taken over all estimator based on an observation from the statistical experience $\{P_\omega/\omega\in\Omega\}$ and with values in $[0,1]^m$.

Proof: Let $\hat{\omega} = (\hat{\omega}_1,\dots,\hat{\omega}_m)$ be an estimator with values in $[0,1]^m$, we have:

$$\begin{aligned}
\max_{\omega\in\Omega} \mathbb{E}_\omega \left[\sum_{j=1}^{m} |\hat{\omega}_j - \omega_j|^\theta\right] &\geq \frac{1}{2^m}\sum_{\omega\in\Omega} \mathbb{E}_\omega \left[\sum_{j=1}^{m} |\hat{\omega}_j - \omega_j|^\theta\right] \\
&\geq \frac{1}{2^m}\sum_{j=1}^{m} \left(\sum_{\omega\in\Omega:\omega_j=1} + \sum_{\omega\in\Omega:\omega_j=0}\right) \mathbb{E}_\omega\left[|\hat{\omega}_j - \omega_j|^\theta\right].
\end{aligned}$$

Each term of the sum over j are lower bounded in the same way. Let see the case of the term $j = m$. We have

$$\begin{aligned}
&\left(\sum_{\omega\in\Omega:\omega_m=1} + \sum_{\omega\in\Omega:\omega_m=0}\right) \mathbb{E}_\omega\left[|\hat{\omega}_m - \omega_m|^\theta\right] \\
&= \sum_{(\omega_1,\dots,\omega_{m-1})\in\{0,1\}^{m-1}} \mathbb{E}_{(\omega_1,\dots,\omega_{m-1},1)}\left[|\hat{\omega}_m - 1|^\theta\right] + \mathbb{E}_{(\omega_1,\dots,\omega_{m-1},0)}\left[|\hat{\omega}_m|^\theta\right] \\
&= \sum_{(\omega_1,\dots,\omega_{m-1})\in\{0,1\}^{m-1}} \int_{\mathcal{X}} (1-\hat{\omega}_m(x))^\theta dP_{(\omega_1,\dots,\omega_{m-1},1)}(x) + \int_{\mathcal{X}} \hat{\omega}_m(x)^\theta dP_{(\omega_1,\dots,\omega_{m-1},0)}(x).
\end{aligned}$$

Thus, if μ is a measure on $(\mathcal{X},\mathcal{A})$ which dominates $P_{(\omega_1,\dots,\omega_{m-1},1)}$ and $P_{(\omega_1,\dots,\omega_{m-1},0)}$ then we obtain

$$\left(\sum_{\omega\in\Omega:\omega_m=1} + \sum_{\omega\in\Omega:\omega_m=0}\right) \mathbb{E}_\omega\left[|\hat{\omega}_m - \omega_m|^\theta\right]$$

$$= \sum_{(\omega_1,\dots,\omega_{m-1})\in\{0,1\}^{m-1}} \int_{\mathcal{X}} \Big[(1-\hat{\omega}_m(x))^\theta f_{(\omega_1,\dots,\omega_{m-1},1)}(x) + \hat{\omega}_m(x)^\theta f_{(\omega_1,\dots,\omega_{m-1},0)}(x)\Big]\, d\mu(x),$$

where $f_{(\omega_1,\dots,\omega_{m-1},1)}$ and $f_{(\omega_1,\dots,\omega_{m-1},0)}$ are one version of the density of the probability measures $P_{(\omega_1,\dots,\omega_{m-1},1)}$ and $P_{(\omega_1,\dots,\omega_{m-1},0)}$ with respect to μ. Since for all $\alpha\in[0,1], a, b\in\mathbb{R}$ we have $(1-\alpha)^\theta a + \alpha^\theta b \geq 2^{1-\theta}\min(a,b)$, we get:

$$\begin{aligned}
&\left(\sum_{\omega\in\Omega:\omega_m=1} + \sum_{\omega\in\Omega:\omega_m=0}\right) \mathbb{E}_\omega\left[|\hat{\omega}_m - \omega_m|^\theta\right] \\
&\geq\ 2^{1-\theta}2^{m-1}\int_{\mathcal{X}} \min\left(f_{(\omega_1,\dots,\omega_{m-1},1)}(x) + f_{(\omega_1,\dots,\omega_{m-1},0)}(x)\right) d\mu(x) \\
&=\ 2^{1-\theta}2^{m-1}\int \min(dP_{(\omega_1,\dots,\omega_{m-1},1)}, dP_{(\omega_1,\dots,\omega_{m-1},0)}) \\
&\geq\ 2^{1-\theta}2^{m-1}\min_{\substack{w,w'\in\Omega:\\ \rho(\omega,\omega')=1}} \int \min(dP_\omega, dP_{\omega'}).
\end{aligned}$$

We complete the proof with Le Cam's inequality (cf., for instance, [115] p.73) which states that for all probabilities P and Q, we have

$$\int \min(dP, dQ) \geq \frac{1}{2}\left(1 - \frac{H^2(P,Q)}{2}\right)^2.$$

■

THEOREM 4.7 (Einmahl and Mason (cf. [53])). *Let $Z_1,\dots,Z_n$ be n independent positive random variables such that $\mathbb{E}[Z_i^2]\leq\sigma^2, \forall i=1,\dots,n$. Then, we have, for any $\delta>0$,*

$$\mathbb{P}\Big[\sum_{i=1}^n Z_i - \mathbb{E}[Z_i] \leq -n\delta\Big] \leq \exp(-n\delta^2/(2\sigma^2)).$$

THEOREM 4.8 (Sudakov (cf. [93])). *There exists an absolute constant $c^*>0$ such that for any integer M, any centered gaussian vector $X=(X_1,\dots,X_M)$ in $\mathbb{R}^M$, we have,*

$$c^*\mathbb{E}[\max_{1\leq j\leq M} X_j] \geq \epsilon\sqrt{\log M},$$

where $\epsilon = \min\Big[\mathbb{E}[(X_i-X_j)^2]^{1/2}; i\neq j\in\{1,\dots,M\}\Big]$.

THEOREM 4.9 (Maximal inequality (cf. [93])). *Let $Y_1,\dots,Y_M$ be M random variables satisfying $\mathbb{E}[\exp(sY_j)]\leq\exp((s^2\sigma^2)/2)$ for any integer j and any $s>0$. Then, we have*

$$\mathbb{E}[\max_{1\leq j\leq M} Y_j] \leq \sigma\sqrt{\log M}.$$

THEOREM 4.10 (Berry-Esséen (cf. page 471 in [16])). *Suppose that $(X_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. random variables with mean μ and variance $\sigma^2>0$. Then, for all n,*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\Big(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq t\Big) - \Phi(t)\right| \leq \frac{33}{4}\frac{\mathbb{E}|X_1-\mu|^3}{\sigma^3\sqrt{n}}.$$

We use the following lemma to prove the weakness of selector aggregates. A proof can be found p. 84 in [115].

LEMMA 4.3. *Let $\mathbb{P}_1, \ldots, \mathbb{P}_M$ be M probability measures on a measurable space $(\mathcal{Z}, \mathcal{T})$ satisfying $\frac{1}{M}\sum_{j=1}^{M} K(\mathbb{P}_j|\mathbb{P}_1) \leq \alpha \log M$, where $0 < \alpha < 1/8$. We have*

$$\inf_{\hat{\phi}} \max_{1\leq j\leq M} \mathbb{P}_j(\hat{\phi} \neq j) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\Big(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log 2}}\Big),$$

where the infimum $\inf_{\hat{\phi}}$ is taken over all tests $\hat{\phi}$ with values in $\{1, \ldots, M\}$ constructed from one observation in the statistical model $(\mathcal{Z}, \mathcal{T}, \{\mathbb{P}_1, \ldots, \mathbb{P}_M\})$.

CHAPTER 5

Convex Aggregation under Positive Covariance Assumption

We prove a convex oracle inequality with a rate of aggregation equals to $\frac{\log M}{n}$. This rate is faster than the standard optimal rate of convex aggregation which is (cf. [114]) M/n if $M < \sqrt{n}$ and $\sqrt{(\log(1 + M/\sqrt{n})/n)}$ otherwise. Here, we obtain the optimal rate of Model Selection aggregation $\frac{\log M}{n}$. This result is obtained under a positive covariance assumption of the estimators to aggregate. It means that, when estimators are positively correlated, it is as easy to mimic the best convex combination of these estimators as to mimic the best of them.

Contents

1. Introduction

Let (X, Y) be a random variable on $\mathcal{X} \times \mathbb{R}$. Denote by π the probability distribution of (X, Y) and by P^X the marginal of X. Consider the norm

$$||f||_{L^2(P^X)} = \left(\int_{\mathcal{X}} |f(x)|^2 dP^X(x) \right)^{1/2},$$

defined for any $f \in L^2(P^X)$. In the regression framework, we want to estimate the regression function

$$\eta(x) = \mathbb{E}\left[Y|X = x\right], \forall x \in \mathcal{X},$$

from a sample of n i.i.d. observations of the couple (X, Y). We denote these observations by $D_n = ((X_i, Y_i))_{1 \leq i \leq n}$. Usually, the variable Y is not an exact function of X. Given is an input $X \in \mathcal{X}$, we are not able to predict the exact value of the output $Y \in \mathbb{R}$. This issue can be seen in the regression framework as a noised estimation. It means that in each spot X of the input set, the predicted label Y is concentrated around $\mathbb{E}\left[Y|X\right]$ up to an additional noise with null mean. Denote this noise by ζ. It is equal to the real random variable $Y - \mathbb{E}\left[Y|X\right]$. The regression model can be written as

$$Y = \mathbb{E}\left[Y|X\right] + \zeta.$$

In this chapter we study a convex aggregation procedure under an geometric assumption. For this problem, we consider M measurable functions $\eta_1, \ldots, \eta_M$ from $\mathcal{X}$ to $\mathbb{R}$, usually called *weak estimators*. Our aim is to mimic the best combination of them where coefficients of this combination are taken in a bounded subset H^M of $\mathbb{R}^M$. For instance, if $H^M = \Lambda^M$,

where

$$\Lambda^M = \left\{ (\lambda_1, \dots, \lambda_M) / \lambda_j \geq 0, \sum_{j=1}^{M} \lambda_j = 1 \right\},$$

we speak about convex aggregation problem.

2. Convex Aggregation Oracle Inequality.

We introduce the following notation

$$\eta_\lambda = \sum_{j=1}^{M} \lambda_j \eta_j, \quad \forall \lambda \in \mathbb{R}^M.$$

We consider the following aggregation procedure $\eta_{\hat{\lambda}_n}$ defined by the weights

$$\hat{\lambda}_n \in \operatorname{Arg}\min_{\lambda \in H^M} R_n(\eta_\lambda) \tag{5.1}$$

where

$$R_n(\eta_0) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \eta_0(X_i))^2$$

is the empirical risk of η_0, for any measurable function η_0 from $\mathcal{X}$ to $\mathbb{R}$.

We give an oracle inequality satisfied by the procedure $\eta_{\hat{\lambda}_n}$

THEOREM 5.1. *In the regression framework $Y = \eta(X) + \sigma(X)\zeta$, where $X \in \mathcal{X}$ and ζ are independent variables. Let $\mathcal{F} = \{\eta_1, \dots, \eta_M\}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that*

- *There exists $B > 0$ such that $||\eta||_\infty, ||\eta_j||_\infty \leq B$ for all $1 \leq j \leq M$.*

Then the aggregation procedure defined in (5.1) satisfies for any $a > 0$ and for any bounded subset H^M of $\mathbb{R}^M$:

$$\mathbb{E}[||\eta_{\hat{\lambda}_n} - \eta||^2_{L^2(P^X)}] \leq (1+a) \min_{\lambda \in H^M} ||\eta_\lambda - \eta||^2_{L^2(P^X)} + \frac{2(1+a)^2}{na} \mathbb{E}\left[\max_{i=1,\dots,n} |\sigma(X_i)\zeta_i|^2 \right] + C_0 \sqrt{\frac{\log M}{n}}.$$

Moreover, if we have the positive covariance assumption

- *For any $j, k = 1, \dots, M$, $\quad \mathbb{E}[(\eta_j(X) - \eta(X))(\eta_k(X) - \eta(X))] \geq 0.$*

and if we only consider subsets H^M of $(\mathbb{R}_+)^M$ (that is for positive coefficients), then the aggregation procedure defined in (5.1) satisfies for any $a > 0$ and for any bounded subset H^M of $(\mathbb{R}_+)^M$:

$$\mathbb{E}[||\eta_{\hat{\lambda}_n} - \eta||^2_{L^2(P^X)}] \leq (1+a) \min_{\lambda \in H^M} ||\eta_\lambda - \eta||^2_{L^2(P^X)} + \frac{2(1+a)^2}{na} \mathbb{E}\left[\max_{i=1,\dots,n} |\sigma(X_i)\zeta_i|^2 \right] + C_0 \frac{\log M}{n}.$$

If ζ is gaussian centered with square deviation equals to 1 and if there exists σ^2 such that $\sigma(X)^2 \leq \sigma^2$ a.s. then $\mathbb{E}\left[\max_{i=1\dots,n} |\sigma(X_i)\zeta_i|^2\right] \leq 2\sigma^2 \log n$.

If ζ is bounded by $L > 0$ (cf. this is called the bounded regression) and if there exists σ^2 such that $\sigma(X)^2 \leq \sigma^2$ a.s., we have $\mathbb{E}\left[\max_{i=1\dots,n} |\sigma(X_i)\zeta_i|^2\right] \leq (\sigma L)^2$.

REMARK 5.1. *Assumption on the covariance of estimators can be replace by*

$$\mathbb{E}_{P^X}[(\eta_j - \eta)(\eta_k - \eta)] \mathbb{E}_{P^X}[(\eta_{j'} - \eta)(\eta_{k'} - \eta)] \geq 0, \forall j, k, j', k' = 1, \dots, M$$

which means that $\eta_1, \dots, \eta_M$ are on the same "side" w.r.t. η. Rudely speaking, η_j's belong to half a cone with vertex η in $L^2(P^X)$.

Proof of Theorem 5.1: For any measurable real-valued functions f, g from $\mathcal{X}$, consider

$$||f||_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(X_i) \text{ and } < f, g >_n = \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$$

and for any real vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$, consider

$$< \epsilon, f >_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i).$$

For any measurable real-valued function f from $\mathcal{X}$ we have

$$||f - \eta||_n^2 = R_n(f) - R_n(\eta) + 2 < \mathbf{N}, f - \eta >_n,$$

where $\mathbf{N}$ denote the random vector $(\sigma(X_1)\zeta_1, \ldots, \sigma(X_n)\zeta_n)$ of the noises.

Denote by $\mathcal{C}$ the set of all η_λ where $\lambda \in H^M$. Let $\bar{\eta}$ be in $\mathcal{C}$. We have for any $a > 0$,

$$\begin{aligned}
&||\eta_{\hat{\lambda}_n} - \eta||^2_{L^2(P^X)} \\
=&||\eta_{\hat{\lambda}_n} - \eta||^2_{L^2(P^X)} + (1+a)\left[2 < \mathbf{N}, \eta_{\hat{\lambda}_n} - \eta >_n + R_n(\eta_{\hat{\lambda}_n}) - R_n(\eta) - ||\eta_{\hat{\lambda}_n} - \eta||_n^2\right] \\
\leq&(1+a)(R_n(\bar{\eta}) - R_n(\eta)) \\
&+ \sup_{\eta_0 \in \mathcal{C}}\left[||\eta_0 - \eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - (1+a)||\eta_0 - \eta||_n^2\right].
\end{aligned}$$

Moreover, for any measurable real valued function $\bar{\eta}$ on $\mathcal{X}$, we have $\mathbb{E}\left[R_n(\bar{\eta}) - R_n(\eta)\right] = ||\bar{\eta} - \eta||^2_{L^2(P^X)}$. Thus,

$$\begin{aligned}
\mathbb{E}||\eta_{\hat{\lambda}_n} - \eta||^2_{L^2(P^X)} \leq& (1+a)\min_{\lambda \in H^M}||\eta_\lambda - \eta||^2_{L^2(P^X)} \\
&+ \quad \mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}}\left[||\eta_0 - \eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - (1+a)||\eta_0 - \eta||_n^2\right]\right].
\end{aligned}$$

Lemma 5.1. *Under the positive covariance assumption we have*

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}}\left[||\eta_0 - \eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - (1+a)||\eta_0 - \eta||_n^2\right]\right] \\
&\qquad \leq \frac{2(1+a)^2}{na}\mathbb{E}\left[\max_{i=1,\ldots,n}|\sigma(X_i)\zeta_i|\right] + C_1\frac{\log M}{n}
\end{aligned}$$

Proof: We have

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\eta_0 \in \mathcal{C}}\left[||\eta_0 - \eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - (1+a)||\eta_0 - \eta||_n^2\right]\Big] \\
&\qquad \leq \mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}} 2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - \frac{a}{2}||\eta_0 - \eta||_n^2\right] \\
&\qquad\quad + \mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}}||\eta_0 - \eta||^2_{L^2(P^X)} - \frac{2+a}{2}||\eta_0 - \eta||_n^2\right]
\end{aligned}$$

Moreover for any $\eta_0 \in \mathcal{C}$,

$$2(1+a) < \mathbf{N}, \eta_0 - \eta >_n - \frac{a}{2}||\eta_0 - \eta||_n^2 \leq 2(1+a)2||\eta_0 - \eta||_n| < \mathbf{N}, h_n(\eta_0) >_n | - \frac{a}{2}||\eta_0 - \eta||_n^2,$$

where

$$h_n(\eta_0) = \begin{cases} \frac{\eta_0-\eta}{||\eta_0-\eta||_n} & \text{if } ||\eta_0-\eta||_n \neq 0 \\ 0 & \text{otherwise} \end{cases} .$$

Using inequality $2|xy| \leq (a/2)x^2 + (2/a)y^2$ which holds for any scalar x, y, we get:

$$2(1+a)||\eta_0-\eta||_n| < \mathbf{N}, h_n(\eta_0) >_n | - \frac{a}{2}||\eta_0-\eta||_n^2 \leq \frac{2(1+a)^2}{a} < \mathbf{N}, h_n(\eta_0) >_n^2 .$$

Since $||(h_n(\eta_0)(X_1), \ldots, h_n(\eta_0)(X_n))||_2^2 = n$, we have

$$\mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}} < \mathbf{N}, h_n(\eta_0) >_n^2\right] \leq \frac{1}{n}\mathbb{E}\left[\sup_{\theta \in \mathcal{S}_n^2} < \mathbf{N}, \theta >^2\right] = \frac{1}{n}\mathbb{E}\left[\max_{i=1,\ldots,n} |\sigma(X_i)\zeta_i|^2\right],$$

where $\mathcal{S}_n^2$ is the set of all unit vectors of $(\mathbb{R}^n, ||.||_2)$ and $< ., . >$ is the usual inner product of $(\mathbb{R}^n, ||.||_2)$.

Let $W_0 = (w_1, \ldots, w_M)$ be a vector in H^M and denote by η_0 the combination $\sum_{j=1}^M w_j\eta_j$. Since $w_j \geq 0, \forall j = 1, \ldots, M$, we have

$$||\eta_0-\eta||_{L^2(P^X)}^2 - \frac{2+a}{2}||\eta_0-\eta||_n^2 = W_0^t Z W_0 \leq \sup_{1\leq k,j\leq M} Z_{k,j}||W_0||_2 \leq C_0 \sup_{1\leq k,j\leq M} Z_{k,j},$$

where C_0 is a bound for H^M and $(Z_{k,j})_{1\leq k,j\leq M}$ is given by

$$Z_{k,j} = \int_{\mathbb{R}} g_k(x)g_j(x)P^X(dx) - \frac{2+a}{2n}\sum_{i=1}^n g_k(X_i)g_j(X_i)$$

and $g_k = \eta_k - \eta$ for any $k = 1, \ldots, M$. Hence,

$$\mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}} ||\eta_0-\eta||_{L^2(P^X)}^2 - \frac{2+a}{2}||\eta_0-\eta||_n^2\right] \leq C_0\mathbb{E}[\sup_{1\leq k,j\leq M} Z_{k,j}].$$

Denote by P^X the distribution of X. For any real-valued function g defined on $\mathcal{X}$, denote by $P^X g$ the expectation $\mathbb{E}(g(X))$ and $P_n^X g$ its empirical version. Since for all $j, k = 1, \ldots, M$, $P^X g_k g_j \geq 0$ we have for any $\delta > 0$

$$\mathbb{P}\left[P^X g_k g_j - \frac{2+a}{2}P_n^X g_k g_j \geq \delta\right] \leq \mathbb{P}\left[P^X g_k g_j - P_n^X g_k g_j \geq \frac{2\delta + aP^X g_k g_j}{2+a}\right].$$

We apply Bernstein's concentration inequality to obtain

$$\begin{aligned}\mathbb{P}&\left[P^X g_k g_j - P_n^X g_k g_j \geq \frac{2\delta + aP^X g_k g_j}{2+a}\right] \\ &\leq \exp\left(-\frac{3n(2\delta + aP^X g_k g_j)^2}{6(2+a)^2 P^X g_k^2 g_j^2 + 8(2+a)B^2(2\delta + aP^X g_k g_j)}\right).\end{aligned}$$

There exists a constant $C_1 > 0$ depending only on a, B such that for all $0 < \delta \leq 2(4+a)B^2$ and all $0 \leq j, k \leq M$, we have

$$\frac{3n(2\delta + aP^X g_k g_j)^2}{6(2+a)^2 P^X g_k^2 g_j^2 + 8(2+a)B^2(2\delta + aP^X g_k g_j)} \geq C_1\delta.$$

Using the union bound, for all positive number u, we have

$$\mathbb{E}\left[\sup_{1\leq k,j\leq M} Z_{k,j}\right] \leq \mathbb{E}\left[\sup_{1\leq k,j\leq M} Z_{k,j}(\mathbb{1}_{\sup_{1\leq k,j\leq M} Z_{k,j}\leq u} + \mathbb{1}_{\sup_{1\leq k,j\leq M} Z_{k,j}\geq u})\right]$$

$$\leq \quad 2u + \int_u^{+\infty} \mathbb{P}\left[\sup_{1\leq k,j\leq M} Z_{k,j} \geq \delta\right] d\delta \leq 2u + \frac{M^2}{C_1}\exp(-C_1 u)$$

Denote by $\mu(M)$ the unique solution of $X = (M^2/2)\exp(-X)$, we get $\log M \leq \mu(M) \leq 2\log M$. Take u such that $nC_1 u = \mu(M)$ then we obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} ||f-\eta||^2 - \frac{2+a}{2}||f-\eta||_n^2\right] \leq \frac{4}{C_1}\frac{\log M}{n}.$$

■

LEMMA 5.2. *Without any covariance assumption we have*

$$\mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}}\left[||\eta_0-\eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0-\eta >_n -(1+a)||\eta_0-\eta||_n^2\right]\right]$$
$$\leq \frac{(1+a)^2}{na}\mathbb{E}\left[\max_{i=1,\ldots,n}|\sigma(X_i)\zeta_i|\right] + C_2\sqrt{\frac{\log M}{n}}$$

Proof: We have

$$\mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}}\left[||\eta_0-\eta||^2_{L^2(P^X)} + 2(1+a) < \mathbf{N}, \eta_0-\eta >_n -(1+a)||\eta_0-\eta||_n^2\right]\right]$$
$$\leq \mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}} 2(1+a) < \mathbf{N}, \eta_0-\eta >_n -a||\eta_0-\eta||_n^2\right]$$
$$+ \mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}} ||\eta_0-\eta||^2_{L^2(P^X)} - ||\eta_0-\eta||_n^2\right]$$

Like in Lemma 5.1, we have

$$2(1+a) < \mathbf{N}, \eta_0-\eta >_n -a||\eta_0-\eta||_n^2 \leq \frac{(1+a)^2}{a} < \mathbf{N}, h_n(\eta_0) >_n^2,$$

thus,

$$\mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}} 2(1+a) < \mathbf{N}, \eta_0-\eta >_n -a||\eta_0-\eta||_n^2\right] \leq \frac{(1+a)^2}{an}\mathbb{E}\left[\max_{i=1,\ldots,n}|\sigma(X_i)\zeta_i|^2\right].$$

Let $W_0 = (w_1,\ldots,w_M)$ be a vector in H^M and denote by η_0 the combination $\sum_{j=1}^M w_j\eta_j$. We have

$$||\eta_0-\eta||^2_{L^2(P^X)} - ||\eta_0-\eta||_n^2 = W_0^t Z W_0 \leq ||Z||_\infty ||W_0||_2 \leq C_0||Z||_\infty,$$

where C_0 is a bound for H^M, Z is the random matrix $(Z_{k,j})_{1\leq k,j\leq M}$ with

$$Z_{k,j} = \int_{\mathbb{R}} g_k(x)g_j(x)P^X(dx) - \frac{1}{n}\sum_{i=1}^n g_k(X_i)g_j(X_i)$$

and $g_k = \eta_k - \eta$ for any $k = 1,\ldots,M$. Hence,

$$\mathbb{E}\left[\sup_{\eta_0\in\mathcal{C}} ||\eta_0-\eta||^2_{L^2(P^X)} - ||\eta_0-\eta||_n^2\right] \leq C_0\mathbb{E}[||Z||_\infty].$$

Denote by P^X the distribution of X. For any real-valued function g defined on $\mathcal{X}$, denote by $P^X g$ the expectation $\mathbb{E}(g(X))$ and $P_n^X g$ its empirical version. Using Bernstein's

concentration inequality, we have for any $0 < \delta < 32B^2$

$$\mathbb{P}\left[|P^X g_k g_j - P_n^X g_k g_j| \geq \delta\right] \leq 2\exp\left(-\frac{n\delta^2}{2P^X g_k^2 g_j^2 + 2B^2\delta/3}\right) \leq 2\exp\left(-\frac{n\delta^2}{54B^4}\right).$$

Using the union bound, for all positive number u, we have

$$\begin{aligned}\mathbb{E}\left[||Z||_\infty\right] &\leq \mathbb{E}\left[||Z||_\infty(\mathbb{1}_{||Z||_\infty \leq u} + \mathbb{1}_{||Z||_\infty \geq u})\right] \\ &\leq 2u + \int_u^{+\infty} \mathbb{P}\left[||Z||_\infty \geq \delta\right] d\delta \leq 2u + \frac{108B^2M^2}{nu}\exp\left(-\frac{nu^2}{54B^2}\right)\end{aligned}$$

Denote by $\mu(M)$ the unique solution of $X = M^2\exp(-X)$, we get $\log M \leq \mu(M) \leq 2\log M$. Take u such that $nu^2 = 54B^2\mu(M)$ then we obtain

$$\mathbb{E}\left[\sup_{\eta_0 \in \mathcal{C}} ||\eta_0 - \eta||^2_{L^2(P^X)} - ||\eta_0 - \eta||_n^2\right] \leq 8B\sqrt{54}\sqrt{\frac{\log M}{n}}.$$

■

In the gaussian case, we have for any $\delta > 0$

$$\mathbb{P}\left[\max_{i=1\ldots,n} |\zeta_i| \geq \delta\right] \leq n\mathbb{P}[|\zeta_1| \geq \delta] \leq n\sqrt{\frac{2}{\pi}}\frac{\exp(-\delta^2/2)}{\delta}.$$

Thus, for any $u > 0$,

$$\begin{aligned}\mathbb{E}\left[\max_{i=1,\ldots,n} |\zeta_i|^2\right] &\leq \mathbb{E}\left[\max_{i=1,\ldots,n} |\zeta_i|^2(\mathbb{1}_{\max_{i=1,\ldots,n}|\zeta_i|^2 \leq u} + \mathbb{1}_{\max_{i=1,\ldots,n}|\zeta_i|^2 \geq u})\right] \\ &\leq 2u + \frac{2n}{\pi}\int_u^{+\infty}\frac{\exp(-\delta/2)}{\sqrt{\delta}} \leq 2u + \frac{4n}{\pi}\frac{\exp(-u/2)}{\sqrt{u}}\end{aligned}$$

Denote by $\mu(M)$ the unique solution of $X^3 = \frac{n}{2\sqrt{\pi}}\exp(-X^2)$, we get $\sqrt{(\log n)/2} \leq \mu(M) \leq \sqrt{\log n}$. Take u such that $(u/2)^{1/2} = \mu(M)$ then we obtain

$$\mathbb{E}\left[\max_{i=1,\ldots,n} |\sigma(X_i)\zeta_i|^2\right] \leq 2\sigma^2\log n.$$

■

Part 2

Fast Rates for Sparse Bayes Rules in Classification

CHAPTER 6

Classification with Minimax Fast Rates for Classes of Bayes Rules with Sparse Representation

We study the behavior of an adaptive estimator for classification problem on $[0,1]^d$, considering piecewise constant classifiers on a dyadic, regular grid. We consider classes of classifier functions that satisfy certain conditions regarding their coefficients when developed over the (overcomplete) basis of indicator functions of dyadic cubes of $[0,1]^d$ and these coefficients are restricted to values in $\{-1,0,1\}$. Lower bounds on the minimax rates of convergence over these classes are established when the underlying marginal of the design is comparable to the Lebesgue measure. An upper bound for the performance of the estimator is derived, which is shown to match the lower bound (up to a logarithmic factor).

Contents

The material of this chapter is an article accepted for publication in the journal *Electronic Journal of Statistics* (cf. [79]).

1. Introduction

Denote by $D_n = (X_i, Y_i)_{1 \le i \le n}$ n i.i.d. observations of a couple (X, Y) of random variables with values in $[0,1]^d \times \{-1,1\}$. Denote by π the probability distribution of (X, Y). We want to construct measurable functions which associate a label $y \in \{-1,1\}$ to each point x of $[0,1]^d$. Such functions are called *prediction rules*. The quality of a prediction rule f is given by the value

$$R(f) = \mathbb{P}(f(X) \neq Y)$$

called *misclassification error of* f. It is well known (e.g. [47]) that there exists an optimal prediction rule which attains the minimum of R over all measurable functions with values in $\{-1, 1\}$. It is called the *Bayes rule* and it is defined by

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

where η is the *conditional probability function of* $Y = 1$ *knowing* X defined by

$$\eta(x) = \mathbb{P}(Y = 1|X = x).$$

The value

$$R^* = R(f^*) = \min_f R(f)$$

is known as the *Bayes risk*. The aim of classification is to construct a prediction rule, using the observations D_n, with a risk as close to R^* as possible. Such a construction is called a *classifier*. Performance of a classifier $\hat{f}_n$ is measured by the value

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi[R(\hat{f}_n) - R^*]$$

called *excess risk of* $\hat{f}_n$. In this case $R(\hat{f}_n) = \mathbb{P}(\hat{f}_n(X) \neq Y|D_n)$ and $\mathbb{E}_\pi$ denotes the expectation w.r.t. D_n when the probability distribution of (X_i, Y_i) is π for any $i = 1, \ldots, n$. Consider $(\phi(n))_{n\in\mathbb{N}}$ a decreasing sequence of positive numbers. We say that a classifier $\hat{f}_n$ *learns at the convergence rate* $\phi(n)$, if there exists an absolute constant $C > 0$ such that for any integer n,

$$\mathbb{E}_\pi[R(\hat{f}_n) - R^*] \leq C\phi(n).$$

We introduce a loss function on the set of all prediction rules:

$$d_\pi(f, g) = |R(f) - R(g)|.$$

This loss function is a *semi-distance* (it is symmetric, satisfies the triangle inequality and $d_\pi(f, f) = 0$). For all classifiers $\hat{f}_n$, it is linked to the excess risk by

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi[d_\pi(\hat{f}_n, f^*)],$$

where the RHS is the risk of $\hat{f}_n$ associated with the loss d_π.

Theorem 7.2 of [47] shows that no classifier can learn with a given convergence rate for arbitrary underlying probability distribution π. To achieve rates of convergence, we need a complexity assumption on the set which the Bayes rule f^* belongs to. For instance, [123, 124] provide examples of classifiers learning, with a given convergence rate, under complexity assumptions on the set of conditional probability functions. Other rates of convergence have been obtained under the assumption that the Bayes rule belongs to a class of prediction rules with a finite dimension of Vapnik and Chervonenkis (cf.[47]). In both cases, the problem of a direct approximation of f^* is not treated. In the first case, the problem of approximation of f^* is shifted to the problem of approximation of the regression function η. In fact, if $\bar{f}$ denote the plug-in rule $\mathbb{1}_{\bar{\eta}\geq 1/2}$, where $\bar{\eta}$ is a function with values in $[0, 1]$ then, we have

$$d_\pi(\bar{f}, f^*) \leq 2\mathbb{E}[|\bar{\eta}(X) - \eta(X)|] \tag{6.1}$$

Thus, under smoothness assumption on the conditional function η, we can control the approximation term. However, global smoothness assumptions on η are somehow too restrictive for the estimation of f^* since the behavior of η away from the decision boundary $\{x \in [0, 1]^d : \eta(x) = 1/2\}$ has no effect on the estimation of f^*. In the second case, the approximation term equals to zero, since it is assumed that the Bayes rule belongs to a

class with a finite VC dimension and so we don't need to approach the Bayes rule by a simpler object.

Many authors pointed out the need for developing a suitable approximation theory for classification. Given a model $\mathcal{C}$ of prediction rules, it is written in p.34 in [22]: "estimating the model bias $\min_{f\in\mathcal{C}}(R(f) - R^*)$ seems to be beyond the reach of our understanding. In fact, estimating R^* is known to be a difficult statistical problem, see [47] and [5]." In [20], question on the control of the approximation error for a class of models in the boosting framework is asked. In this chapter, it is assumed that the Bayes rule belongs to the model and form of distribution satisfying such condition is explored. Another related work is [89], where, under general conditions, it can be guaranteed that the approximation error converges to zero for some specific models. In [116], the author examines classes that are indexed by a complexity exponent that reflects the smoothness of the Bayes decision boundary. An argument of entropy is then used to upper bound the bias term. A generalization of these classes is given in [105]. Finally, on the general topic of approximation theory in classification we want to mention the recent work of [107].

The main difficulty of a direct approximation of f^* is the dependence of the loss d_π on π. Given a model $\mathcal{P}$ (a set of probability measures on $[0,1]^d \times \{-1,1\}$) with a known complexity, we want to be able to construct a decreasing family $(\mathcal{F}_\epsilon)_{\epsilon>0}$ of classes of prediction rules, such that we have an approximation result of the form:

$$\forall \pi = (P^X, \eta) \in \mathcal{P}, \forall \epsilon > 0, \exists f_\epsilon \in \mathcal{F}_\epsilon : d_\pi(f_\epsilon, f^*) \leq \epsilon, \tag{6.2}$$

where P^X is the marginal distribution of π on $[0,1]^d$ and $f^* = \text{Sign}(2\eta - 1)$ is the Bayes rule, associated with the regression function η of π. In fact, we want the classes $\mathcal{F}_\epsilon$ to be parametric, such that, for the estimation problem, we just have to estimate a parametric object in a class $\mathcal{F}_{\epsilon_n}$, for a well chosen ϵ_n (generally obtained by a trade-off between the bias/approximation term and the variance term, coming from the estimation of the best parametric object in $\mathcal{F}_{\epsilon_n}$ approaching f^*).

We upper bound the loss d_π, but, we still work directly with the approximation of f^*. For a prediction rule f we have

$$d_\pi(f, f^*) = \mathbb{E}[|2\eta(X) - 1| \mathbb{1}_{f(X)\neq f^*(X)}] \leq (1/2)||f - f^*||_{L^1(P^X)}. \tag{6.3}$$

In order to get a distribution-free loss function, we assume that the following assumption holds. This assumption is close to assuming that the marginal distribution of X is the Lebesgue measure on $[0,1]^d$.

(A1) *The marginal P^X is absolutely continuous w.r.t. the Lebesgue measure λ_d and there exist two constants $0 < a < A < +\infty$ such that $a \leq dP^X(x)/d\lambda_d \leq A, \quad \forall x \in [0,1]^d$.*

The behavior of the regression function η near the level $1/2$ is a key characteristic of the classification's quality (cf. e.g. [116]). In fact, the closest is η to $1/2$, the more difficult is the classification problem. Here, we work under the following assumption introduced by [94].

Strong Margin Assumption (SMA): There exists an absolute constant $0 < h \leq 1$ such that:

$$\mathbb{P}\left(|2\eta(X) - 1| > h\right) = 1.$$

Under assumptions (A1) and (SMA) we have, for any prediction rule f,

$$\frac{ah}{2}||f - f^*||_{L_1(\lambda_d)} \leq d_\pi(f, f^*) \leq \frac{A}{2}||f - f^*||_{L_1(\lambda_d)}.$$

Thus, estimation of f^* w.r.t. the loss d_π is the same as estimation w.r.t. the $L_1(\lambda_d)-$norm, where λ_d is the Lebesgue measure on $[0,1]^d$.

The chapter is organized as follows. In the next section, we introduce a class of functions, with values in $\{-1,1\}$, developed in a fundamental system of $L^2([0,1]^d)$. Section 3is devoted to the approximation and the estimation of Bayes rules having a sparse representation in this system. In Section 4, we discuss this approach. Proofs are postponed to Section 5.

2. Classes of Bayes Rules with Sparse Representation.

In this section, we introduce a class of prediction rules. For that, we consider two different representations of prediction rules.

The first way is to represent a prediction rule as an infinite dyadic tree. An *infinite dyadic decision tree* is defined as a partitioning of the hypercube $[0,1]^d$ obtained by cutting in half perpendicular to one of the axis coordinates, then cutting recursively the two pieces obtained in half again, and so on. Most of the time, finite dyadic trees are considered (cf. [21] and [105]). It means that the previous constructions stop at an arbitrary point along every branches. For a survey on decision trees we refer to [96]. Here, we consider also infinite dyadic trees.

The other way is more "analytic". Namely, we consider the representation of prediction rules in a fundamental system of $L^2([0,1]^d,\lambda_d)$ (that is a countable family of functions such that all their finite linear combinations is dense in $L^2([0,1]^d,\lambda_d)$), inherited from the Haar basis, and control the number of non-zero coefficients (which can take values $-1,0,1$ in this case).

2.1. Analytic representation of decision trees. First we consider a fundamental system of $L^2([0,1]^d,\lambda_d)$. We consider a sequence of partitions of $[0,1]^d$ by setting for any integer j,

$$\mathcal{I}_{\mathbf{k}}^{(j)} = E_{k_1}^{(j)} \times \ldots \times E_{k_d}^{(j)},$$

where $\mathbf{k}$ is the multi-index

$$\mathbf{k} = (k_1,\ldots,k_d) \in I_d(j) = \{0,1,\ldots,2^j-1\}^d,$$

and for any integer j and any $k \in \{1,\ldots,2^j-1\}$,

$$E_k^{(j)} = \begin{cases} \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right) & \text{if } k = 0,\ldots,2^j-2 \\ \left[\frac{2^j-1}{2^j}, 1\right] & \text{if } k = 2^j-1 \end{cases}.$$

We consider the family $\mathcal{S} = \left(\phi_{\mathbf{k}}^{(j)} : j \in \mathbb{N}, \mathbf{k} \in I_d(j)\right)$ where

$$\phi_{\mathbf{k}}^{(j)} = \mathbb{1}_{\mathcal{I}_{\mathbf{k}}^{(j)}}, \quad \forall j \in \mathbb{N}, \mathbf{k} \in I_d(j),$$

where $\mathbb{1}_A$ denotes the indicator of a set A. The set $\mathcal{S}$ is a fundamental system of $L^2([0,1]^d,\lambda_d)$. This is the class of indicators of the dyadic sets of $[0,1]^d$.

Formal definition of the classes $\mathcal{F}^{(d)}$: We consider the class $\mathcal{F}^{(d)}$ of functions $f : [0,1]^d \longmapsto \{-1,1\}$ defined by

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, \lambda_d - a.s., \text{ where } a_{\mathbf{k}}^{(j)} \in \{-1,0,1\}.$$

In what follows, we use the vocabulary appearing in the wavelet literature. The index "j" of $a_{\mathbf{k}}^{(j)}$ and $\phi_{\mathbf{k}}^{(j)}$ is called "level of frequency".

Writing convention (W): Since $\mathcal{S}$ is not an orthogonal basis of $L^2([0,1]^d, \lambda_d)$, the expansion of f w.r.t. this system is not unique. Therefore, to avoid any ambiguity, we define an unique writing for any mapping f in $\mathcal{F}^{(d)}$ by taking $a_{\mathbf{k}}^{(j)} \in \{-1, 1\}$ with preferences for low frequencies when it is possible. Roughly speaking, for $f \in \mathcal{F}^{(d)}$, denoted by $f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, \lambda_d - a.s.$ where $a_{\mathbf{k}}^{(j)} \in \{-1, 0, 1\}$. This convention means that, we construct $A_{\mathbf{k}}^{(j)} \in \{-1, 0, 1\}, j \in \mathbb{N}, \mathbf{k} \in I_d(j)$, such that: if there exists $J \in \mathbb{N}$ and $\mathbf{k} \in I_d(J)$ such that for all $\mathbf{k}' \in I_d(J+1)$ satisfying $\phi_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}'}^{(J+1)} \neq 0$ we have $a_{\mathbf{k}'}^{(J+1)} = 1$, then we take $A_{\mathbf{k}}^{(J)} = 1$ and the other 2^d coefficients of higher frequency $A_{\mathbf{k}'}^{(J+1)} = 0$, instead of having these 2^d coefficients equal to 1, and the same convention holds for -1. Moreover, if we have $A_{\mathbf{k}}^{(J_0)} \neq 0$ then $A_{\mathbf{k}'}^{(J)} = 0$ for all $J > J_0$ and $\mathbf{k}' \in I_d(J)$ satisfying $\phi_{\mathbf{k}}^{(J_0)} \phi_{\mathbf{k}'}^{(J)} \neq 0$.

We can describe a mapping $f \in \mathcal{F}^{(d)}$ satisfying this convention by using an infinite dyadic decision tree. Each node corresponds to a coefficient $A_{\mathbf{k}}^{(J)}$. The root is $A_{(0,\ldots,0)}^{(0)}$. If a node, describing the coefficient $A_{\mathbf{k}}^{(J)}$, equals to 1 or -1 then, it has no branches, otherwise it has 2^d branches, corresponding to the 2^d coefficients at the following frequency, describing the coefficients $A_{\mathbf{k}'}^{(J+1)}$ satisfying $\phi_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}'}^{(J+1)} \neq 0$. At the end, all the leaves of the tree equal to 1 or -1, and the depth of a leaf is the frequency of the associated coefficient. The writing convention says that a node cannot have all his leaves equal to 1 together (or -1). In this case we write this mapping by putting a 1 at the node (or -1). In what follows we say that a function $f \in \mathcal{F}^{(d)}$ satisfies the writing convention (W) when f is written in $\mathcal{S}$ using the writing convention described in this paragraph. Remark that, this writing convention is not an assumption on the function since we can write all $f \in \mathcal{F}^{(d)}$ using this convention.

We can avoid the problem of the non-uniqueness of the expansion of a function in the overcomplete system $\mathcal{S}$. For instance, by using the wavelet tensor product of the Haar basis (cf. [95]), we obtain an orthonormal wavelet basis of $L^2([0,1]^d)$. In that case the link with dyadic decision trees is much more complicated and the obtained results are not easily interpretable.

It is easy to see that all measurable functions from $[0,1]^d$ to $\{-1, 1\}$ cannot be represented in this way. A simple example is given by the following construction. Consider $(q_k)_{k \geq 1}$ an enumeration of the rational numbers of $(0, 1)$. Denote by A the union, over $k \in \mathbb{N}$, of the open balls $\mathcal{B}(q_k, 2^{-(k+1)})$. This is a dense open set of Lebesgue measure bounded by $1/2$. The prediction rule $f = 2\mathbb{1}_A - 1$ cannot be written in the fundamental system $\mathcal{S}$ using coefficients with values in $\{-1, 0, 1\}$ ($f \notin \mathcal{F}^{(1)}$). Nevertheless, under a mild assumption (cf. the following definition) a prediction rule belongs to $\mathcal{F}^{(d)}$.

Definition 6.1. *Let A be a Borel subset of $[0,1]^d$. We say that A is* **almost everywhere open** *if there exists an open subset $\mathcal{O}$ of $[0,1]^d$ such that $\lambda_d(A \Delta \mathcal{O}) = 0$, where λ_d is the Lebesgue measure on $[0,1]^d$ and $A\Delta\mathcal{O}$ is the symmetric difference.*

Theorem 6.1. *Let η be a function from $[0,1]^d$ to $[0,1]$. We consider*

$$f_\eta(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ -1 & \text{otherwise.} \end{cases}$$

We assume that $\{\eta \geq 1/2\}$ and $\{\eta < 1/2\}$ are almost everywhere open. Thus, there exists $g \in \mathcal{F}^{(d)}$ such that $g = f_\eta, \lambda_d - a.s.$.

For instance, if $\lambda_d(\partial\{\eta = 1/2\}) = 0$ and, either η is λ_d-almost everywhere continuous (it means that there exists an open subset of $[0,1]^d$ with a Lebesgue measure equals to

1 such that η is continuous on this open subset) or if η is λ_d-almost everywhere equal to a continuous function, then $f_\eta \in \mathcal{F}^{(d)}$. Moreover, the Lebesgue measure satisfies the property of regularity, which says that for any Borel $B \in [0,1]^d$ and any $\epsilon > 0$, there exists a compact subset K and an open subset $\mathcal{O}$ such that $K \subseteq A \subseteq \mathcal{O}$ and $\lambda_d(\mathcal{O} - K) \leq \epsilon$. Hence, one can easily check that for any measurable function f from $[0,1]^d$ to $\{-1,1\}$ and any $\epsilon > 0$, there exists a function $g \in \mathcal{F}^{(d)}$ such that $\lambda_d(\{x \in [0,1]^d : f(x) \neq g(x)\}) \leq \epsilon$. Thus, $\mathcal{F}^{(d)}$ is dense in $L^2(\lambda_d)$ intersected with the set of all measurable functions from $[0,1]^d$ to $\{-1,1\}$.

2.2. Related works and main results. The best known decision tree algorithms are CART (cf. [24]) and C4.5 (cf. [100]). These methods use a growing and pruning algorithm. First, a large tree is grown by splitting recursively nodes along coordinates axes according to an "impurity" criterion. Next, this tree is pruned using a penalty function. Penalties are usually based on standard complexity regularization like the square root of the size of the tree. Spatially adaptive penalties depend not only on the complexity of the tree, but also on the spatial distribution of training samples. More recent constructions of decision trees have been proposed in [105] and [21]. In [105], the authors consider, in the multi-class framework, dyadic decision trees and exhibit near-minimax rates of convergence by considering spatial adaptive penalties. They obtained rates of convergence over classes of prediction functions having a complexity defined in the same spirit as [91] and [116]. In [21], a general framework is worked out including classification for different loss functions. The authors select among a set of dyadic trees having a finite depth, the best tree realizing an optimal trade-off between the empirical risk and a penalty term. Here, the penalty term is proportional to the number of leaves in the tree. They obtained oracle inequalities and derived rates of convergence in the regression setup under a regularity assumption on the underlying regression function to estimate. Rates of convergence, for the classification problem, are not derived from these oracle inequalities, since, they do not treat the bias term.

Our estimation procedure does not provide an algorithm in the same spirit as these previous works. The main reason is that, we obtain results under the assumption on the marginal distribution given by (A1). This assumption allows us to work at a given "frequency" and we do not need a multi-scale construction of the dyadic tree as in the previous related work. Once the optimal frequency obtained (by trade off), the estimation procedure is a regular histogram rule as considered in Chapter 9 of [47].

The present work focuses on the control of the approximation term and the introduction of classes of prediction rules having different complexities and approximation qualities. As we shall see, one crucial difference of our estimator is that it is able to deal with infinite trees. Such infinite trees can be considered since we control the bias term. Nevertheless, when the complexity parameter α (associated with the concept of complexity that we consider), is unknown we use a multi-scale approach to construct an adaptive procedure. This procedure learns with the rate

$$\Big(\frac{\log n}{n}\Big)^{1-\alpha},$$

for any complexity parameter α. This multi-scale classifier is the following: we split the sample in two subsamples $D_m^{(1)}$, containing the first m observations, and $D_l^{(2)}$, the $l(=n-m)$ last ones. We use $D_m^{(1)}$ to construct a family of classifiers $\hat{f}_m^{(J)}$ for different frequency levels $J \in [0, J^{(n)}]$, for an integer $J^{(n)}$ chosen later. For instance $\hat{f}_m^{(0)}$ is the classifier which makes

a majority vote in the cell $\mathcal{I}_0^{(0)}$, $\hat{f}_m^{(1)}$ is the classifier making a majority vote in each cell $\mathcal{I}_{\mathbf{k}}^{(1)}$, for $\mathbf{k} \in I_d(1)$ of the partition $\mathcal{S}^{(1)} = \{\mathcal{I}_{\mathbf{k}}^{(1)}, \mathbf{k} \in I_d(1)\}$ of $[0,1]^d$, etc.. Subsample $D_l^{(2)}$ is used to construct exponential weights $w_J^{(l)}$ (cf. Chapter 8). The weight $w_J^{(l)}$ is associated with the basic classifier $\hat{f}_m^{(J)}$, for any $J \in [0, J^{(n)}]$. Finally, the procedure that we propose is the sign of the convex combination

$$\tilde{f}_n = \sum_{J=1}^{J^{(n)}} w_J^{(l)} \hat{f}_m^{(J)}. \tag{6.4}$$

An interesting fact is that, we can consider the set $\mathcal{S}$, introduced in Subsection 2.1, as a dictionary of basic functions. Considering prediction rules as linear combinations of the functions in this dictionary with coefficients in $\{-1, 0, 1\}$ (using the convention of writing (W)), we obtain that, the LASSO estimator (cf. [113]) is given, in this framework, by

$$\text{Arg} \max_{f \in \mathcal{F}^{(d)}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{f(X_i) \neq Y_i} + \gamma \sum_{j,\mathbf{k}} |a_{\mathbf{k}}^{(j)}|,$$

where $f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, \lambda_d - a.s.$ Since the coefficients $a_{\mathbf{k}}^{(j)}$ take their values in $\{-1, 0, 1\}$, the l_1-type penalty $\sum_{j,\mathbf{k}} |a_{\mathbf{k}}^{(j)}|$ is exactly the number of leaves of the dyadic tree associated with the prediction rule f. Thus, LASSO estimator, in this framework and for the dictionary $\mathcal{S}$, is the same as the estimator considered in [21].

2.3. Class of Bayes rules. Now, we define a model for the Bayes rule by taking a subset of $\mathcal{F}^{(d)}$. For all functions w defined on $\mathbb{N}$ and with values in $\mathbb{N}$, we consider $\mathcal{F}_w^{(d)}$, the class for Bayes rules, composed of all prediction rules f which can be written, using the previous writing convention (W), by

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)},$$

where $a_{\mathbf{k}}^{(j)} \in \{-1, 0, 1\}$ and

$$\text{card}\left\{\mathbf{k} \in \mathrm{I_d(j)} : \mathrm{a_{\mathbf{k}}^{(j)}} \neq 0\right\} \leq \mathrm{w(j)}, \quad \forall \mathrm{j} \in \mathbb{N}.$$

The class $\mathcal{F}_w^{(d)}$ depends on the choice of the function w. If w is too small then the class $\mathcal{F}_w^{(d)}$ is poor. That is the subject of the following Proposition 6.1.

PROPOSITION 6.1. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{N}$ such that $w(0) \geq 1$. The two following assertions are equivalent:*

(i) $\mathcal{F}_w^{(d)} \neq \{\mathbb{I}_{[0,1]^d}\}$.
(ii) $\sum_{j=1}^{+\infty} 2^{-dj} w(j) \geq 1$.

This proposition is strongly connected to the Kraft inequality from coding theory (see e.g. [42]).

If w is too large then, the approximation of the model $\mathcal{F}_w^{(d)}$, by a parametric model will be impossible. That is why we give a particular look on the class of functions introduced in the following Definition 6.2.

DEFINITION 6.2. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{N}$. If w satisfies*

$$\sum_{j=0}^{+\infty} \frac{w(j)}{2^{dj}} < +\infty \tag{6.5}$$

then, we say that $\mathcal{F}_w^{(d)}$ is a **$\mathbf{L^1}-$ellipsoid of prediction rules**.

We say that $\mathcal{F}_w^{(d)}$ is a "L^1-ellipsoid" for a function w satisfying (6.5), because , the sequence $(w(j))_{j\in\mathbb{N}}$ belongs to a L^1-ellipsoid of $\mathbb{N}^{\mathbb{N}}$, with sequence of radius $(2^{dj})_{j\in\mathbb{N}}$. Moreover, Definition 6.2 can be linked to the definition of a L^1-ellipsoid for real valued functions, since we have a kind of basis, given by $\mathcal{S}$, and we have a control on coefficients which increases with the frequency. Control on coefficients, given by (6.5), is close to the one for coefficients of a real valued function in a L^1-ellipsoid of Sobolev, since it deals with the quality of approximation of the class $\mathcal{F}_w^{(d)}$ by a parametric model.

REMARK 6.1. *A L^1-ellipsoid of prediction rules is made of "sparse" prediction rules. In fact, for $f \in \mathcal{F}_w^{(d)}$ with w satisfying (6.5), the number of non-zero coefficients in the decomposition of f (using the writing convention (W)), at a given frequency, becomes small as the frequency grows. That is the reason why $\mathcal{F}_w^{(d)}$ can be called a* **sparse class of prediction rules**.

Next, we provide examples of functions satisfying (6.5). Classes $\mathcal{F}_w^{(d)}$ associated with these functions are used in what follows as statistical models. We first define the minimal infinite class of prediction rules $\mathcal{F}_0^{(d)}$ which is the class $\mathcal{F}_w^{(d)}$ when $w = w_0^{(d)}$ where $w_0^{(d)}(0) = 1$ and $w_0^{(d)}(j) = 2^d - 1$, for all $j \geq 1$. To understand why this class is important we introduce a concept of local oscillation of a prediction rule. This concept defines a kind of "regularity" for functions with values in $\{-1, 1\}$. For f a function from $[0,1]^d$ to $\{-1,1\}$ in $\mathcal{F}^{(d)}$, we consider the writing of f in the fundamental system introduced in Section 3.1 with writing convention (W):

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k}\in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, \lambda_d - a.s..$$

Let $J \in \mathbb{N}$ and $\mathbf{k} \in I_d(J)$. We say that $\mathcal{I}_{\mathbf{k}}^{(J)}$ is a **low oscillating block** of f when f has exactly $2^d - 1$ non-zero coefficients, in this block, at each level of frequencies greater than $J + 1$. In this case we say that f **has a low oscillating block of frequency** J. Remark that, if f has an oscillating block of frequency J, then f has an oscillating block of frequency J', for all $J' \geq J$. The function class $\mathcal{F}_0^{(d)}$ is made of all prediction rules with one oscillating block at level 1 and of the indicator function $\mathbb{1}_{[0,1]^d}$. If we have $w(j_0) < w_0^{(d)}(j_0)$ for one $j_0 \geq 1$ and $w(j) = w_0^{(d)}(j)$ for $j \neq j_0$ then the associated class $\mathcal{F}_w^{(d)}$ contains only the indicator function $\mathbb{1}_{[0,1]^d}$, that is the reason why we say that $\mathcal{F}_0^{(d)}$ is "minimal".

Nevertheless, the following proposition shows that $\mathcal{F}_0^{(d)}$ is a rich class of prediction rules from a combinatorial point of view. We recall some quantities which measure a combinatorial richness of a class of prediction rules (cf. [47]). For any class $\mathcal{F}$ of prediction rules from $[0,1]^d$ to $\{-1,1\}$, we consider

$$N(\mathcal{F}, (x_1, \ldots, x_m)) = card\left(\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}\right)$$

where $x_1, \ldots, x_m \in [0,1]^d$ and $m \in \mathbb{N}$,

$$S(\mathcal{F}, m) = \max\left(N(\mathcal{F}, (x_1, \ldots, x_m)) : x_1, \ldots, x_m \in [0,1]^d\right)$$

and the VC-dimension of $\mathcal{F}$ is

$$VC(\mathcal{F}) = \min\left(m \in \mathbb{N} : S(\mathcal{F}, m) \neq 2^m\right).$$

Consider $x_j = \left(\frac{2^j+1}{2^{j+1}}, \frac{1}{2^{j+1}}, \ldots, \frac{1}{2^{j+1}}\right)$, for any $j \in \mathbb{N}$. For any integer m, we have $N(\mathcal{F}_0^{(d)}, (x_1, \ldots, x_m)) = 2^m$. Hence, the following proposition holds.

PROPOSITION 6.2. *The class of prediction rules $\mathcal{F}_0^{(d)}$ has an infinite VC-dimension.*

Every class $\mathcal{F}_w^{(d)}$ such that $w \geq w_0^{(d)}$ has an infinite VC-dimension (since $\mathcal{F}_w^{(d)} \subseteq \mathcal{F}_{w'}^{(d)}$ when $w \leq w'$), which is the case for the following classes.

We denote by $\mathcal{F}_K^{(d)}$, for a $K \in \mathbb{N}^*$, the class $\mathcal{F}_w^{(d)}$ of prediction rules where w is equal to the function

$$w_K^{(d)}(j) = \begin{cases} 2^{dj} & \text{if } j \leq K, \\ 2^{dK} & \text{otherwise.} \end{cases}$$

This class is called the **truncated class of level K**.

We consider **exponential classes**. These sets of prediction rules are denoted by $\mathcal{F}_\alpha^{(d)}$, where $0 < \alpha < 1$, and are equal to $\mathcal{F}_w^{(d)}$ when $w = w_\alpha^{(d)}$ and

$$w_\alpha^{(d)}(j) = \begin{cases} 2^{dj} & \text{if } j \leq N^{(d)}(\alpha) \\ \lceil 2^{d\alpha j} \rceil & \text{otherwise} \end{cases},$$

where $N^{(d)}(\alpha) = \inf\left(N \in \mathbb{N} : \lceil 2^{d\alpha N} \rceil \geq 2^d - 1\right)$, that is for $N^{(d)}(\alpha) = \lceil \log(2^d - 1)/(d\alpha \log 2) \rceil$.

The classes $\mathcal{F}_0^{(d)}$, $\mathcal{F}_K^{(d)}$ and $\mathcal{F}_\alpha^{(d)}$ are examples of L^1-ellipsoid of prediction rules.

REMARK 6.2. *Other sets of prediction rules are described by the classes $\mathcal{F}_w^{(d)}$ where w is from $\mathbb{N}$ to $\mathbb{N}$ and satisfies*

$$\sum_{j \geq 1} a_j \frac{w(j)}{2^{dj}} \leq L,$$

where $(a_j)_{j \geq 1}$ is an increasing sequence of positive numbers.

3. Rates of Convergence over $\mathcal{F}_w^{(d)}$ under (SMA)

3.1. Approximation Result. Let w be a function from $\mathbb{N}$ to $\mathbb{N}$ and $A > 1$. We denote by $\mathcal{P}_{w,A}$ the set of all probability measures π on $[0,1]^d \times \{-1, 1\}$ such that the Bayes rules f^*, associated with π, belongs to $\mathcal{F}_w^{(d)}$ and the marginal of π on $[0,1]^d$ is absolutely continuous and a version of its Lebesgue density is upper bounded by A. The following theorem can be seen as an approximation theorem for the Bayes rules w.r.t. the loss d_π uniformly in $\pi \in \mathcal{P}_{w,A}$.

THEOREM 6.2 (Approximation theorem). *Let $\mathcal{F}_w^{(d)}$ be a L^1-ellipsoid of prediction rules. We have:*

$$\forall \epsilon > 0, \exists J_\epsilon \in \mathbb{N} : \forall \pi \in \mathcal{P}_{w,A}, \exists f_\epsilon = \sum_{\mathbf{k} \in I_d(J_\epsilon)} B_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)}$$

where $B_{\mathbf{k}}^{(J_\epsilon)} \in \{-1, 1\}$ and

$$d_\pi(f^*, f_\epsilon) \leq \epsilon,$$

where f^* is the Bayes rule associated to π. For example, J_ϵ can be the smallest integer J satisfying $\sum_{j=J+1}^{+\infty} 2^{-dj} w(j) < \epsilon/A$.

Theorem 6.2 is the first step to prove an estimation theorem using a trade-off between a bias term and a variance term. We write

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi[d_\pi(\hat{f}_n, f^*)] \leq \mathbb{E}_\pi[d_\pi(\hat{f}_n, f_\epsilon)] + d_\pi(f_\epsilon, f^*).$$

Since f_ϵ belongs to a parametric model we expect to have a control of the variance term $\mathbb{E}_\pi[d_\pi(\hat{f}_n, f_\epsilon)]$, depending on the dimension of the parametric model which is linked to the quality of the approximation in the bias term. Remark that, no assumption on the quality of the classification problem (like an assumption on the margin) is required to obtain Theorem 6.2. Only assumption on the "number of oscillations" of f^* is used. Theorem 6.2 deals with approximation of functions in the L^1-ellipsoid $\mathcal{F}_w^{(d)}$ by functions with values in $\{-1, 1\}$ and no estimation issues are considered.

3.2. Estimation Result. We consider the following class of estimators indexed by the frequency rank $J \in \mathbb{N}$:

$$\hat{f}_n^{(J)} = \sum_{\mathbf{k} \in I_d(J)} \hat{A}_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}}^{(J)}, \tag{6.6}$$

where coefficients are defined by

$$\hat{A}_{\mathbf{k}}^{(J)} = \begin{cases} 1 & \text{if } \exists X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } N_{\mathbf{k}}^{(J)+} > N_{\mathbf{k}}^{(J)-} \\ -1 & \text{otherwise,} \end{cases}$$

where, for any $\mathbf{k} \in I_d^{(J)}$, we consider $N_{\mathbf{k}}^{(J)+} = \text{Card}\left\{i : X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } Y_i = 1\right\}$ and $N_{\mathbf{k}}^{(J)-} = \text{Card}\left\{i : X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } Y_i = -1\right\}$.

To obtain a good control of the variance term, we need to assure a good quality of the estimation problem. Therefore, estimation results are obtained in Theorem 6.3 under (SMA) assumption. Nevertheless, (SMA) assumption is not enough to assure any rate of convergence (cf. chapter 7 of [47] or corollary 6.1 at the end of section 3.3). We have to define a model for η or f^* with a finite complexity. Here we assume that the underlying Bayes rule f^*, associated with π, belongs to a L^1-ellipsoid of prediction rules.

THEOREM 6.3 (Estimation theorem). *Let $\mathcal{F}_w^{(d)}$ be a L^1-ellipsoid of prediction rules. Let π be a probability measure on $[0,1]^d \times \{-1,1\}$ satisfying assumptions (A1) and (SMA), and such that the Bayes rule belongs to $\mathcal{F}_w^{(d)}$. The excess risk of the classifier $\hat{f}_n^{(J_\epsilon)}$ satisfies:*

$$\forall \epsilon > 0, \quad \mathcal{E}_\pi(\hat{f}_n^{(J_\epsilon)}) = \mathbb{E}_\pi[d_\pi(\hat{f}_n^{(J_\epsilon)}, f^*)] \leq (1+A)\epsilon + \exp\left(-na(1-\exp(-h^2/2))2^{-dJ_\epsilon}\right),$$

where J_ϵ is the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w(j) < \epsilon/A$. Parameters a, A appear in Assumption (A1) and h is used in (SMA).

3.3. Optimality. This section is devoted to the optimality, in a minimax sense, of estimation in the classification models $\mathcal{F}_w^{(d)}$. Let $0 < h < 1$, $0 < a \leq 1 \leq A < +\infty$ and w a mapping from $\mathbb{N}$ to $\mathbb{N}$. We denote by $\mathcal{P}_{w,h,a,A}$ the set of all probability measures $\pi = (P^X, \eta)$ on $[0,1]^d \times \{-1,1\}$ such that

(1) The marginal P^X satisfies (A1).
(2) The Assumption (SMA) is satisfied.
(3) The Bayes rule f^*, associated with π, belongs to $\mathcal{F}_w^{(d)}$.

We apply a version of the Assouad Lemma to lower bound the risk over $\mathcal{P}_{w,h,a,A}$.

THEOREM 6.4. *Let w be a function from $\mathbb{N}$ to $\mathbb{N}$ such that*

(i) $w(0) \geq 1$ *and* $\forall j \geq 1, \quad w(j) \geq 2^d - 1$
(ii) $\forall j \geq 1, \quad w(j-1) \geq 2^{-d} w(j)$.
We have for any $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1} \left(w\left(\lfloor \log n/(d \log 2) \rfloor + 1 \right) - (2^d - 1) \right),$$

where $C_0 = (h/8) \exp\left(-(1 - \sqrt{1-h^2}) \right)$. *Moreover, if* $w(j) \geq 2^d, \forall j \geq 1$ *then*

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1}.$$

REMARK 6.3. *For a function* w *satisfying assumptions of Theorem 6.4 and under (SMA), we cannot expect a convergence rate faster than* $1/n$, *which is the usual lower bound for the classification problem under (SMA).*

We can deduce Theorem 7.1 of [47] from our Theorem 6.4. We denote by $\mathcal{P}_1$ the class of all probability measures on $[0,1]^d \times \{-1, 1\}$ such that the marginal distribution P^X is λ_d (the Lebesgue probability distribution on $[0,1]^d$) and (SMA) is satisfied with the margin $h = 1$. The case "$h = 1$" is equivalent to $R^* = 0$.

COROLLARY 6.1. *For any integer* n *we have*

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_1} \mathcal{E}(\hat{f}_n) \geq \frac{1}{8e}.$$

It means that no classifier can achieve any rate of convergence in the classification models $\mathcal{P}_1$.

3.4. Rates of Convergence for Different Classes of Prediction Rules. In this section we apply results stated in Theorem 6.3 and Theorem 6.4 to different L^1-ellipsoid classes $\mathcal{F}_w^{(d)}$ introduced at the end of Section 2. We give rates of convergence and lower bounds for these models. Using notation introduced in Section 2 and Subsection 3.3, we consider the following models. For $w = w_K^{(d)}$ we denote by $\mathcal{P}_K^{(d)}$ the set of probability measures $\mathcal{P}_{w_K^{(d)},h,a,A}$ and by $\mathcal{P}_\alpha^{(d)}$ for the exponential case $w = w_\alpha^{(d)}$.

THEOREM 6.5. *For the truncated class* $\mathcal{F}_K^{(d)}$, *we have*

$$\sup_{\pi \in \mathcal{P}_K^{(d)}} \mathcal{E}_\pi(\hat{f}_n^{(J_n(K))}) \leq C_{K,h,a,A} \frac{\log n}{n},$$

where $C_{K,h,a,A} > 0$ *is depending only on* K, h, a, A. *For the lower bound, there exists* $C_{0,K,h,a,A} > 0$ *depending only on* K, h, a, A *such that, for all* $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_K^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C_{0,K,h,a,A} n^{-1}.$$

For the exponential class $\mathcal{F}_\alpha^{(d)}$ *where* $0 < \alpha < 1$, *we have for any integer* n

$$\sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n^{(J_n(\alpha))}) \leq C'_{\alpha,h,a,A} \left(\frac{\log n}{n} \right)^{1-\alpha}, \tag{6.7}$$

where $C'_{\alpha,h,a,A} > 0$. *For the lower bound, there exists* $C'_{0,\alpha,h,a,A} > 0$ *depending only on* α, h, a, A *such that, for all* $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C'_{0,\alpha,h,a,A} n^{-1+\alpha}.$$

The order of $J_n(\alpha)$ and $J_n(K)$ is $\lceil \log\left(an/(2^d \log n)\right)/(d \log 2)\rceil$, up to a multiplying constant.

A remarkable point is that the class $\mathcal{F}_K^{(d)}$ has an infinite VC-dimension (cf. Proposition 6.2). Nevertheless, the rate $\log n/n$ is achieved in this model. Existence of classes of rules with infinite VC dimension that are consistent when the marginal distribution of the design X is without atoms has been remarked in [47].

3.5. Adaptation to the complexity. In this section we provide an adaptive estimator for the exponential classes. The estimator $\hat{f}_n^{(J_n(\alpha))}$, appearing in (6.7), depends on the complexity parameter α, since

$$J_n(\alpha) = \left\lceil \frac{\log(A/(\epsilon_n(2^{d(1-\alpha)}-1)))}{d(1-\alpha)\log 2} \right\rceil$$

and $\epsilon_n = (\log n/(nC))^{1-\alpha}$, where $C = a(1-e^{-h^2/2})2^{-d}(A^{-1}(2^{d(1-\alpha)}-1))^{1/(1-\alpha)}$. In practice, we do not have access to this parameter. Thus, it is important to construct an estimator free from this parameter and which can learn at the near-optimal rate $((\log n)/n)^{1-\alpha}$ if the underlying probability distribution belongs to $\mathcal{F}_\alpha^{(d)}$ for any α. This is the problem of adaptation to the complexity parameter α.

To construct an adaptive estimator, we use an aggregation procedure. We split the sample in two parts. Denote by $D_m^{(1)}$ the subsample containing the first m observations and $D_l^{(2)}$ the one containing the $l(=n-m)$ last ones. Subsample $D_m^{(1)}$ is used to construct classifiers $\hat{f}_m^{(J)}$ for different frequency levels $J \in [0, J^{(n)}]$, for an integer $J^{(n)}$ chosen later. Subsample $D_l^{(2)}$ is used to construct the exponential weights of our aggregation procedure (cf. Chapter 8). We aggregate the basis classifiers $\hat{f}_m^{(J)}$, $J \in [1, J^{(n)}]$, by the procedure

$$\tilde{f}_n = \sum_{J=1}^{J^{(n)}} w_J^{(l)} \hat{f}_m^{(J)}, \tag{6.8}$$

where

$$w_J^{(l)} = \frac{\exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(J)}(X_i)\right)}{\sum_{J'=1}^{J^{(n)}} \exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(J')}(X_i)\right)}, \quad \forall J = 1, \dots, J^{(n)}. \tag{6.9}$$

The classifier that we propose is

$$\hat{f}_n = \mathrm{Sign}(\tilde{f}_n). \tag{6.10}$$

THEOREM 6.6. *Assume that $J^{(n)}$ is greater than $(\log n)^2$ and choose $l = \lceil n/\log n \rceil$ for the learning sample size. For any $\alpha \in (0,1)$, we have, for n large enough,*

$$\sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \le 6C'_{\alpha,h,a,A} \left(\frac{\log n}{n}\right)^{1-\alpha}, \tag{6.11}$$

where $C'_{\alpha,h,a,A} > 0$ has been introduced in Theorem 6.5.

The classifier $\hat{f}_n$ does not assume the knowledge of the parameter α neither of a, A, h. Thus, it is also adaptive to the parameters a, A and h.

REMARK 6.4. *We may compare our method with the ERM type aggregate defined by*

$$\bar{f}_n \in \mathrm{Arg} \min_{f \in \{\hat{f}_m^{(0)}, \dots, \hat{f}_m^{(J^{(n)})}\}} \sum_{i=m+1}^n \mathbb{1}_{(f(X_i) \neq Y_i)}.$$

This aggregate also satisfies (6.11), if we replace $\hat{f}_n$ by $\bar{f}_n$ (cf. Chapter 8). The difference is that the aggregate (6.8) uses a multi-scale approach (it associates a weight to each frequency), whereas the adaptive classifier $\bar{f}_n$ selects the best "empirical frequency".

The other way to extend our approach deals with the problem of choice of the geometry by taking $\mathcal{S}$ as fundamental system. One possible solution is to consider classifiers "adaptive to the geometry". Using an adaptive procedure, for instance the same as in (6.8), we can construct classifiers adaptive to the "rotation" and "translation". Consider, for example, the dyadic partition of $[0,1]^2$ at the frequency level J_n. We can construct classifiers using the same procedure as (6.6) but for partitions obtained by translation of the dyadic partition by the vector $(n_1/(2^{J_n}\log n), n_2/(2^{J_n}\log n))$, where $n_1, n_2 = 0, \ldots, \lceil \log n \rceil$. We can do the same thing by aggregating classifiers obtained by the procedure (6.6) for partitions obtained by rotation of center $(1/2, 1/2)$ with angle $n_3\pi/(2\log n)$, where $n_3 = 0, \ldots, \lceil \log n \rceil$, of the initial dyadic partition. In this heuristic we don't discuss about the way to solve problems near the boundary of $[0,1]^2$.

4. Discussion

In this chapter we start by considering a model of prediction rules. Then, we provide an approximation theorem for these models. The form of object approaching the Bayes rule in these models leads to a particular form of estimators (here the histogram estimators). Finally, the way the estimator depends on the complexity of the underlying model (here the level of frequency) impose a way to construct adaptive estimators. As we can see everything depends on the starting model we consider. In this section we discuss the representation and the estimation of prediction rules lying in these models in simple cases.

For the one-dimensional case, another point of view is to consider $f^* \in L^2([0,1])$ and to develop f^* in an orthonormal wavelet basis of $L^2([0,1])$. Namely,

$$f^* = \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} a_k^{(j)} \psi_k^{(j)},$$

where $a_k^{(j)} = \int_0^1 f^*(x)\psi_k^{(j)}(x)dx$ for any $j \in \mathbb{N}$ and $k = 0, \ldots, 2^j - 1$. For the control of the bias term we assume that the family of coefficients $(a_k^{(j)}, j \in \mathbb{N}, k = 0, \ldots, 2^j - 1)$ belongs to our L^1-ellipsoid. But this point of view leads to functional analysis and estimation issues. First problem: which functions with values in $\{-1, 1\}$ have wavelet coefficients in our L^1-ellipsoid and which wavelet basis is more adapted to our problem (maybe the Haar basis)? Second problem: which kind of estimators could be used for the estimation of these coefficients? As we can see, the main problem is that there is no approximation theory for functions with values in $\{-1, 1\}$. We do not know how to approach, in $L^2([0,1])$, measurable functions with values in $\{-1, 1\}$ by "parametric" functions with values in $\{-1, 1\}$. Methods developed in this chapter may be seen as a first step in this direction. We can generalize this approach to functions with values in $\mathbb{Z}$. When functions take values in $\mathbb{R}$, for instance in the regression problem, usual approximation theory is used to obtain a control on the bias term. Finally, remark that functions with values in $\{-1, 1\}$ can be approximated by real-valued (possibly smooth) functions; this is for example what is used for SVM or boosting. In those cases, control of the approximation term is still an open question (cf. [109] and [89]).

In considering the classification problem over the square $[0,1]^2$, a classifier has to be able to approach, for instance, the "simple" Bayes rule $f_{\mathcal{C}}^*$ which is equal to 1 inside $\mathcal{C}$,

where $\mathcal{C}$ is a disc inside $[0,1]^2$, and -1 outside $\mathcal{C}$. In our framework, two questions need to be considered:

- What is the representation of a simple function $f_{\mathcal{C}}^*$ in our fundamental system, using only coefficients with values in $\{-1,0,1\}$ and with the writing convention (W)?
- Is the estimate $\hat{f}_n^{(J_n)}$, where $J_n = \lceil \log\left(an/(2^d \log n)\right)/(d \log 2) \rceil$ is the frequency rank appearing in Theorem 6.5, a good classifier when the underlying probability measure yields $f_{\mathcal{C}}^*$ as Bayes rule?

At a first glance, our point of view is not the right way to estimate $f_{\mathcal{C}}^*$. In this regular case (the boundary is an infinite differentiable curve), the direct estimation of the boundary is a better approach. The main reason is that a 2-dimensional estimation problem becomes a 1-dimensional problem. Such reduction of dimension makes the estimation easier (in passing, our approach is specifically good in the 1-dimensional case, since the notion of boundary does not exist in this case). Nevertheless, our approach is applicable for the estimation of such functions (cf. Theorem 6.7). Actually, a direct estimation of the boundary reduces the dimension but there is a loss of observations since observations far from the boundary are not used by this estimation point of view. This may explain why our approach is applicable. Denote by

$$\mathcal{N}(A,\epsilon,||.||_\infty) = \min\left(N : \exists x_1,\ldots,x_N \in \mathbb{R}^2 : A \subseteq \cup_{j=1}^N B_\infty(x_j,\epsilon)\right)$$

the $\epsilon-$covering number of a subset A of $[0,1]^2$, w.r.t. the infinity norm of $\mathbb{R}^2$. For example, the circle $\mathcal{C} = \{(x,y) \in \mathbb{R}^2 : (x-1/2)^2 + (y-1/2)^2 = (1/4)^2\}$ satisfies $\mathcal{N}(\mathcal{C},\epsilon,||.||_\infty) \leq (\pi/4)\epsilon^{-1}$. For any set A of $[0,1]^2$, denote by ∂A the boundary of A.

THEOREM 6.7. *Let A be a subset of $[0,1]^2$ such that $\mathcal{N}(\partial A,\epsilon,||.||_\infty) \leq \delta(\epsilon)$, for any $\epsilon > 0$, where δ is a decreasing function on $\mathbb{R}_+^*$ with values in $\mathbb{R}^+$ satisfying $\epsilon^2\delta(\epsilon) \longrightarrow 0$ when $\epsilon \longrightarrow 0$. Consider the prediction rule $f_A = 2\mathbb{1}_A - 1$. For any $\epsilon > 0$, denote by ϵ_0 the greatest positive number satisfying $\delta(\epsilon_0)\epsilon_0^2 \leq \epsilon$. There exists a prediction rule constructed in the fundamental system $\mathcal{S}$ at the frequency rank J_{ϵ_0} with coefficients in $\{-1,1\}$ denoted by*

$$f_{\epsilon_0} = \sum_{\mathbf{k} \in I_2(J_{\epsilon_0})} a_{\mathbf{k}}^{(J_{\epsilon_0})} \phi_{\mathbf{k}}^{(J_{\epsilon_0})},$$

with $J_{\epsilon_0} = \lfloor \log(1/\epsilon_0)/\log 2 \rfloor$ such that

$$||f_{\epsilon_0} - f_A||_{L^1(\lambda_2)} \leq 36\epsilon.$$

For instance, there exists a function f_n, written in the fundamental system $\mathcal{S}$ at the frequency level $J_n = \lfloor \log(4n/(\pi \log n))/\log 2 \rfloor$, which approaches the prediction rule $f_{\mathcal{C}}^*$ with a $L^1(\lambda_2)$ error upper bounded by $36(\log n)/n$. This frequency level is, up to a constant factor, the same as the one appearing in Theorem 6.5. In a more general way, any prediction rule with a boundary having a finite perimeter (for instance polygons) is approached by a function written in the fundamental system at the same frequency rank J_n and the same order of $L^1(\lambda_2)$ error $(\log n)/n$. Remark that for this frequency level J_n, we have to estimate $n/\log n$ coefficients. Estimations of one coefficient $a_{\mathbf{k}}^{(J_n)}$, for $\mathbf{k} \in I_2(J_n)$, depends on the number of observation in the square $\mathcal{I}_{\mathbf{k}}^{(J_n)}$. The probability that no observation "falls" in $\mathcal{I}_{\mathbf{k}}^{(J_n)}$ is smaller than n^{-1}. Thus, number of coefficient estimated with no observations is small compared to the order of approximation $(\log n)/n$ and is taken into account in the variance term. Now, the problem is about finding an L^1-ellipsoid of prediction rules

such that for any integer n the approximation function f_n belongs to such a ball. This problem depends on the geometry of the boundary set ∂A. It arises naturally since we chose a particular geometry for our partition: dyadic partitions of the space $[0,1]^d$, and we have to pay a price for this choice which has been made independently of the type of functions to estimate. But, this choice of geometry is, in our case, the same as the choice of a wavelet basis, for instance, in the density estimation problem. Depending on the type of Bayes rules we have to estimate, a special partition can be considered. For example our "dyadic approach" is very well adapted for the estimation of Bayes rules associated with chessboard (with the value 1 for black square and -1 for white square). This kind of Bayes rules are very badly estimated by classification procedures estimating the boundary since most of these procedures require regularity assumptions which are not fulfilled in the case of chessboards. In the general case, the ideal choice of the geometry is adapted to the particular geometry induced by the measure μ on $[0,1]^d$, defined by

$$\mu(A) = \int_A |2\eta(x) - 1| P^X(dx),$$

for any measurable set $A \subseteq [0,1]^d$. Namely, we do not need a good resolution of the partition for the regions of $[0,1]^d$ with a low $\mu-$probability. However, we need a sharper resolution for regions with a high $\mu-$probability. In our case (under assumptions (A1) and (SMA)), the measure μ is equivalent to the Lebesgue measure. Thus, we do not need different resolution for different areas of the square $[0,1]^d$.

We can extend our approach in several ways. Consider the dyadic partition of $[0,1]^d$ with frequency J_n. Instead of choosing 1 or -1 for each square of this partition (like in our approach), we can do a least square regression in each cell of the partition. Inside a cell $\mathcal{I}_{\mathbf{k}}^{(J_n)}$, where $\mathbf{k} \in I_d(J_n)$, we can compute the line minimizing

$$\sum_{i=1}^{n} (f(X_i) - Y_i)^2 \mathbb{1}_{(X_i \in \mathcal{I}_{\mathbf{k}}^{(J_n)})},$$

where f is taken in the set of all indicators of half spaces of $[0,1]^d$ intersecting $\mathcal{I}_{\mathbf{k}}^{(J_n)}$. Of course, depending on the number of observations inside the cell $\mathcal{I}_{\mathbf{k}}^{(J_n)}$ we can consider bigger classes of indicators than the one made of the indicators of half spaces. Our classifier is close to the histogram estimator in density or regression framework, which has been extended to smoother procedures.

5. Proofs

In all the proofs, we use the analytical representation of the predictions rules to underly the similarity with the techniques used in the wavelet literature. Nevertheless, these proofs can be obtained by using the dyadic decision tree representation.

Proof of Theorem 6.1: Since $\{\eta \geq 1/2\}$ is almost everywhere open there exists an open subset $\mathcal{O}$ of $[0,1]^d$ such that $\lambda_d(\{\eta \geq 1/2\} \Delta \mathcal{O}) = 0$. If $\mathcal{O}$ is the empty set then take $g = -1$, otherwise, for all $x \in \mathcal{O}$ denote by $\mathcal{I}_x$ the biggest subset $\mathcal{I}_{\mathbf{k}}^{(j)}$ for $j \in \mathbb{N}$ and $\mathbf{k} \in I_d(j)$ such that $x \in \mathcal{I}_{\mathbf{k}}^{(j)}$ and $\mathcal{I}_{\mathbf{k}}^{(j)} \subseteq \mathcal{O}$. Remark that $\mathcal{I}_x$ exists because $\mathcal{O}$ is open. We can see that for any $y \in \mathcal{I}_x$ we have $\mathcal{I}_y = \mathcal{I}_x$, thus, $(\mathcal{I}_x : x \in \mathcal{O})$ is a partition of $\mathcal{O}$. We denote by $I_{\mathcal{O}}$ a subset of index $(j, \mathbf{k})$, where $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$ such that $\{\mathcal{O}_x : x \in \mathcal{O}\} = \{\mathcal{I}_k^{(j)} : (j, \mathbf{k}) \in I_{\mathcal{O}}\}$. For any $(j, \mathbf{k}) \in I_{\mathcal{O}}$ we take $a_{\mathbf{k}}^{(j)} = 1$.

Take $\mathcal{O}_1$ an open subset λ_d-almost everywhere equal to $\{\eta < 1/2\}$. If $\mathcal{O}_1$ is the empty set then take $g = 1$. Otherwise, consider the set of index $I_{\mathcal{O}_1}$ built in the same way as previously. For any $(j, \mathbf{k}) \in I_{\mathcal{O}_1}$ we take $a_{\mathbf{k}}^{(j)} = -1$.

For any $(j, \mathbf{k}) \notin I_{\mathcal{O}} \cup I_{\mathcal{O}_1}$, we take $a_{\mathbf{k}}^{(j)} = 0$. Consider

$$g = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

It is easy to check that the function g belongs to $\mathcal{F}^{(d)}$, satisfies the writing convention (W) and, for λ_d-almost $x \in [0,1]^d$, $g(x) = f_\eta(x)$.

Proof of Proposition 6.1: Assume that $\mathcal{F}_w^{(d)} \neq \{\mathbb{1}_{[0,1]^d}\}$. Take $f \in \mathcal{F}_w^{(d)} - \{\mathbb{1}_{[0,1]^d}\}$. Consider the writing of f in the system $\mathcal{S}$ using the convention (W),

$$f = \sum_{j \in \mathbb{N}} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)},$$

where $a_{\mathbf{k}}^{(j)} \in \{-1, 0, 1\}$ for any $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$. Consider $b_{\mathbf{k}}^{(j)} = |a_{\mathbf{k}}^{(j)}|$ for any $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$. Consider $f_2 = \sum_{j \in \mathbb{N}} \sum_{\mathbf{k} \in I_d(j)} b_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}$. Remark that the function $f_2 \in \mathcal{F}^{(d)}$ but does not satisfy the writing convention (W). We have $f_2 = \mathbb{1}_{[0,1]^d}$ a.s.. For any $j \in \mathbb{N}$ we have

$$\text{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\} = \text{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{6.12}$$

Moreover, one coefficient $b_{\mathbf{k}}^{(j)} \neq 0$ contributes to fill a cell of Lebesgue measure 2^{-dj} among the hypercube $[0,1]^d$. Since the mass total of $[0,1]^d$ is 1, we have

$$1 = \sum_{j \in \mathbb{N}} 2^{-dj} \text{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{6.13}$$

Moreover, $f \in \mathcal{F}^{(d)}$ thus, for any $j \in \mathbb{N}$,

$$w(j) \geq \text{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}.$$

We obtain the second assertion of Proposition 6.1 by using the last inequality and both of the assertions (6.12) and (6.13).

Assume that $\sum_{j=1}^{+\infty} 2^{-dj} w(j) \geq 1$. For any integer $j \neq 0$, denote by $\text{Ind}(j)$ the set of indexes $\{(j, \mathbf{k}) : \mathbf{k} \in I_d(j)\}$. We use the lexicographic order of $\mathbb{N}^{d+1}$ to order sets of indexes. Take $\text{Ind}_w(1)$ the family of the first $w(1)$ elements of $\text{Ind}(1)$. Denote by $\text{Ind}_w(2)$ the family made of the first $w(1)$ elements of $\text{Ind}(1)$ and add, at the end of this family in the correct order, the first $w(2)$ elements $(2, \mathbf{k})$ of $\text{Ind}(2)$ such that $\phi_{\mathbf{k}'}^{(1)} \phi_{\mathbf{k}}^{(2)} = 0$ for any $(1, \mathbf{k}') \in \text{Ind}_w(1)$,..., for the step j, construct the family $\text{Ind}_w(j)$ made of all the elements of $\text{Ind}_w(j-1)$ in the same order and add at the end of this family the indexes $(j, \mathbf{k})$ of $\text{Ind}(j)$ among the first $w(j)$ elements of $\text{Ind}(j)$ such that $\phi_{\mathbf{k}'}^{(J)} \phi_{\mathbf{k}}^{(j)} = 0$ for any $(J, \mathbf{k}') \in \text{Ind}_w(j-1)$. If there is no more indexes satisfying this condition then, we stop the construction, otherwise, we go on. Denote by Ind the final family obtained by this construction (Ind can be finite or infinite). Then, we enumerate the indexes of Ind by $(j_1, \mathbf{k}_1) \prec (j_2, \mathbf{k}_2) \prec \cdots$. For the first $(j_1, \mathbf{k}_1) \in \text{Ind}$ take $a_{\mathbf{k}_1}^{(j_1)} = 1$, for the second element $(j_2, \mathbf{k}_2) \in \mathcal{I}$ take $a_{\mathbf{k}_2}^{(j_2)} = -1$,etc. . Consider the function

$$f = \sum_{j \in \mathbb{N}} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

If the construction stops at a given iteration N then f takes its values in $\{-1,1\}$ and the writing convention (W) is fulfilled since every cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ has a neighboring cell associated to a coefficient non equals to 0 with an opposite value. Otherwise, for any integer $j \neq 0$, the number of coefficient $a_{\mathbf{k}}^{(j)}$, for $\mathbf{k} \in I_d(j)$, non equals to 0 is $w(j)$ and the total mass of cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ is $\sum_{j\in\mathbb{N}} 2^{-dj}\text{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}$ which is greater or equal to 1 by assumption. Thus, all the hypercube is filled by cells associated with coefficients non equal to 0. So f takes its values in $\{-1,1\}$ and the writing convention (W) is fulfilled since every cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ has a neighboring cell associated with a coefficient non equals to 0 with an opposite value. Moreover f is not $\mathbb{1}_{[0,1]^d}$.

Proof of Theorem 6.2. Let $\pi = (P^X, \eta)$ be a probability measure on $[0,1]^d \times \{-1,1\}$ in $\mathcal{P}_{w,A}$. Denote by f^* a Bayes rule associated with π (for example $f^* = \text{sign}(2\eta - 1)$) . We have

$$d_\pi(f, f^*) = (1/2)\mathbb{E}[|2\eta(X) - 1||f(X) - f^*(X)|] \leq (A/2)||f - f^*||_{L^1(\lambda_d)}.$$

Let $\epsilon > 0$. Define by J_ϵ the smallest integer satisfying

$$\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w(j) < \frac{\epsilon}{A}.$$

We write f^* in the fundamental system $(\phi_{\mathbf{k}}^{(j)}, j \geq J_\epsilon)$ using the convention of writing of section 3.1. Remark that, we start the expansion of f^* at the level of frequency J_ϵ and then, we use the writing convention (W) on the coefficients of this expansion. Namely, we consider

$$f^* = \sum_{\mathbf{k}\in I_d(J_\epsilon)} A_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)} + \sum_{j=J_\epsilon+1}^{+\infty} \sum_{\mathbf{k}\in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

Next, we define the best approximation of f^* at the frequency level J_ϵ by

$$f_\epsilon = \sum_{\mathbf{k}\in I_d(J_\epsilon)} B_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)}, \tag{6.14}$$

where

$$B_{\mathbf{k}}^{(J_\epsilon)} = \begin{cases} 1 & \text{if } p_{\mathbf{k}}^{(J_\epsilon)} > 1/2 \\ -1 & \text{otherwise} \end{cases} \tag{6.15}$$

and

$$p_{\mathbf{k}}^{(J_\epsilon)} = \mathbb{P}(Y = 1 | X \in \mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}) = \int_{\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}} \eta(x) \frac{dP^X(x)}{P^X(\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)})}, \tag{6.16}$$

for all $\mathbf{k} \in I_d(J_\epsilon)$. Note that, if $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $A_{\mathbf{k}}^{(J_\epsilon)} = B_{\mathbf{k}}^{(J_\epsilon)}$, moreover f^* takes its values in $\{-1,1\}$, thus ,we have

$$\begin{aligned} ||f_\epsilon - f^*||_{L^1(\lambda_d)} &= \sum_{\substack{\mathbf{k}\in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)}\neq 0}} \int_{\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}} |f^*(x) - f_\epsilon(x)|dx + \sum_{\substack{\mathbf{k}\in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)}=0}} \int_{\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}} |f^*(x) - f_\epsilon(x)|dx \\ &\leq 2^{-dJ_\epsilon+1}\text{card}\left\{\mathbf{k} \in I_d(J_\epsilon) : A_{\mathbf{k}}^{(J_\epsilon)} = 0\right\} \\ &\leq 2\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w(j) < 2\epsilon/A. \end{aligned}$$

■

Proof of Theorem 6.3. Let $\pi = (P^X, \eta)$ be a probability measure on $[0,1]^d \times \{-1,1\}$ satisfying (A1), (SMA) and such that $f^* = \text{sign}(2\eta - 1)$, a Bayes classifier associated with π, belongs to $\mathcal{F}_w^{(d)}$ (an L^1-ellipsoid of Bayes rules).

Let $\epsilon > 0$ and J_ϵ the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w(j) < \epsilon/A$. We decompose the risk in the bias term and variance term:

$$\mathcal{E}(\hat{f}_n^{(J_\epsilon)}) = \mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f^*)\right] \leq \mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f_\epsilon)\right] + d_\pi(f_\epsilon, f^*),$$

where $\hat{f}_n^{(J_\epsilon)}$ is introduced in (6.6) and f_ϵ in (6.14).

Using the definition of J_ϵ and according to the approximation Theorem (Theorem 6.2), the bias term satisfies:

$$d_\pi(f_\epsilon, f^*) \leq \epsilon.$$

For the variance term we have (using the notations introduced in (6.6) and (6.15)):

$$\begin{aligned}
\mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f_\epsilon)\right] &= \frac{1}{2}\left|\mathbb{E}\left[Y(f_\epsilon(X) - \hat{f}_n^{(J_\epsilon)}(X))\right]\right| \leq \frac{1}{2}\mathbb{E}\left[\int_{[0,1]^d} |f_\epsilon(x) - \hat{f}_n^{(J_\epsilon)}(x)| dP^X(x)\right] \\
&= \frac{1}{2}\sum_{\mathbf{k}\in I_d(J_\epsilon)} \mathbb{E}\left[\int_{\mathcal{I}_\mathbf{k}^{(J_\epsilon)}} |B_\mathbf{k}^{(J_\epsilon)} - \hat{A}_\mathbf{k}^{(J_\epsilon)}| dP^X\right] \\
&\leq \frac{A}{2^{dJ_\epsilon+1}} \sum_{\mathbf{k}\in I_d(J_\epsilon)} \mathbb{E}[|B_\mathbf{k}^{(J_\epsilon)} - \hat{A}_\mathbf{k}^{(J_\epsilon)}|] \\
&\leq \frac{A}{2^{dJ_\epsilon}} \sum_{\mathbf{k}\in I_d(J_\epsilon)} \mathbb{P}\left(|B_\mathbf{k}^{(J_\epsilon)} - \hat{A}_\mathbf{k}^{(J_\epsilon)}| = 2\right).
\end{aligned}$$

Now, we apply a concentration inequality in each cell of the dyadic partition $(\mathcal{I}_\mathbf{k}^{(J_\epsilon)} : \mathbf{k} \in I_d(J_\epsilon))$. Let $\mathbf{k} \in I_d(J_\epsilon)$. We introduce the following events:

$$\Omega_\mathbf{k}^{(m)} = \left\{\text{Card}\{i \in \{1,\dots,n\} : X_i \in \mathcal{I}_\mathbf{k}^{(J_\epsilon)}\} = m\right\}, \forall m \in \{0,\dots,n\}$$

and

$$\Omega_\mathbf{k} = \{N_\mathbf{k}^{(J_\epsilon)+} \leq N_\mathbf{k}^{(J_\epsilon)-}\},$$

where $N_\mathbf{k}^{(J_\epsilon)+}$ and $N_\mathbf{k}^{(J_\epsilon)-}$ have been defined in subsection 3.2. We have

$$\mathbb{P}(\hat{A}_\mathbf{k}^{(J_\epsilon)} = -1) = \mathbb{P}(\Omega_\mathbf{k}^{(0)c} \cap \Omega_\mathbf{k}) + \mathbb{P}(\Omega_\mathbf{k}^{(0)})$$

and

$$\begin{aligned}
\mathbb{P}(\Omega_\mathbf{k}^{(0)c} \cap \Omega_\mathbf{k}) &= \sum_{m=1}^{n} \mathbb{P}(\Omega_\mathbf{k}^{(m)} \cap \Omega_\mathbf{k}) \\
&= \sum_{m=1}^{n} \mathbb{P}(\Omega_\mathbf{k}|\Omega_\mathbf{k}^{(m)})\mathbb{P}(\Omega_\mathbf{k}^{(m)}).
\end{aligned}$$

Moreover, if we denote by $Z_1, \dots, Z_n$ n i.i.d. random variables with a Bernoulli with parameter $p_\mathbf{k}^{(J_\epsilon)}$ for common probability distribution (we recall that $p_\mathbf{k}^{(J_\epsilon)}$ is introduced in (6.16) and is equal to $\mathbb{P}(Y = 1 | X \in I_\mathbf{k}^{(J_\epsilon)})$), we have for any $m = 1, \dots, n$,

$$\mathbb{P}(\Omega_\mathbf{k}|\Omega_\mathbf{k}^{(m)}) = \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \leq \frac{1}{2}\right).$$

The concentration inequality of Hoeffding leads to

(6.17)

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \geq p_{\mathbf{k}}^{(J_\epsilon)} + t\right) \leq \exp(-2mt^2) \text{ and } \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \leq p_{\mathbf{k}}^{(J_\epsilon)} - t\right) \leq \exp(-2mt^2),$$

for all $t > 0$ and $m = 1, \ldots, n$.

Denote by $b_{\mathbf{k}}^{(J_\epsilon)}$ the probability $\mathbb{P}(X \in \mathcal{I}_{\mathbf{k}}^{(J_\epsilon)})$. If $p_{\mathbf{k}}^{(J_\epsilon)} > 1/2$, applying second inequality of (6.17) leads to

$$\begin{aligned}
\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) &= \mathbb{P}(\hat{A}_{\mathbf{k}}^{(J_\epsilon)} = -1) \\
&\leq \sum_{m=1}^{n} \mathbb{P}\left[\frac{1}{m}\sum_{j=1}^{m} Z_j \leq p_{\mathbf{k}}^{(J_\epsilon)} - (p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)\right] \left(\begin{array}{c} n \\ m \end{array}\right) (b_{\mathbf{k}}^{(J_\epsilon)})^m (1 - b_{\mathbf{k}}^{(J_\epsilon)})^{n-m} \\
&+ \mathbb{P}(\Omega_{\mathbf{k}}^{(0)}) \\
&\leq \sum_{m=0}^{n} \exp\left(-2m(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2\right) \left(\begin{array}{c} n \\ m \end{array}\right) (b_{\mathbf{k}}^{(J_\epsilon)})^m (1 - b_{\mathbf{k}}^{(J_\epsilon)})^{n-m} \\
&= \left(1 - b_{\mathbf{k}}^{(J_\epsilon)}(1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))\right)^n \\
&\leq \exp\left(-na(1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))2^{-dJ_\epsilon}\right).
\end{aligned}$$

If $p_{\mathbf{k}}^{(J_\epsilon)} < 1/2$ then, similar arguments used in the previous case and first inequality of (6.17) lead to

$$\begin{aligned}
\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) &= \mathbb{P}(\hat{A}_{\mathbf{k}}^{(J_\epsilon)} = 1) \\
&\leq \exp\left(-na(1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))2^{-dJ_\epsilon}\right).
\end{aligned}$$

If $p_{\mathbf{k}}^{(J_\epsilon)} = 1/2$, we use $\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) \leq 1$. Like in the proof of Theorem 6.2, we use the writing

$$f^* = \sum_{\mathbf{k} \in I_d(J_\epsilon)} A_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)} + \sum_{j=J_\epsilon+1}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

Since $P^X(\eta = 1/2) = 0$, if $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $p_{\mathbf{k}}^{(J_\epsilon)} \neq 1/2$. Thus, the variance term satisfies:

$$\begin{aligned}
&\mathbb{E}\left[d_\pi(\hat{f}_n, f_\epsilon^*)\right] \\
&\leq \frac{A}{2^{dJ_\epsilon}}\Big(\sum_{\substack{\mathbf{k} \in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} \neq 0}} \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) + \sum_{\substack{\mathbf{k} \in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} = 0}} \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right)\Big) \\
&\leq \frac{A}{2^{dJ_\epsilon}} \sum_{\substack{\mathbf{k} \in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} \neq 0}} \exp\left(-na(1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))2^{-dJ_\epsilon}\right) + A\epsilon.
\end{aligned}$$

If $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $\eta > 1/2$ or $\eta < 1/2$ over the whole set $\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}$, so

$$\left|\frac{1}{2} - p_{\mathbf{k}}^{(J_\epsilon)}\right| = \int_{\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}} \left|\eta(x) - \frac{1}{2}\right| \frac{dP^X(x)}{P^X(\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)})}.$$

Moreover π satisfies $\mathbb{P}(|2\eta(X)-1| \geq h) = 1$, so

$$\left| \frac{1}{2} - p_{\mathbf{k}}^{(J_\epsilon)} \right| \geq \frac{h}{2}.$$

We have shown that for all $\epsilon > 0$,

$$\mathcal{E}(\hat{f}_n) = \mathbb{E}[d_\pi(\hat{f}_n, f^*)] \leq (1+A)\epsilon + \exp\left(-na(1-\exp(-2(h/2)^2))2^{-dJ_\epsilon}\right),$$

where J_ϵ is the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w(j) < \epsilon/A$.

■

Proof of Theorem 6.4. For all $q \in \mathbb{N}$ we consider G_q a net of $[0,1]^d$ defined by:

$$G_q = \left\{ \left(\frac{2k_1+1}{2^{q+1}}, \ldots, \frac{2k_d+1}{2^{q+1}} \right) : (k_1, \ldots, k_d) \in \{0, \ldots, 2^q - 1\} \right\}$$

and the function η_q from $[0,1]^d$ to G_q such that $\eta_q(x)$ is the closest point of G_q from x (in the case of ex aequo, we choose the smallest point for the usual order on $\mathbb{R}^d$). Associated to this grid, the partition $\mathcal{X'}_1^{(q)}, \ldots, \mathcal{X'}_{2^{dq}}^{(q)}$ of $[0,1]^d$ is defined by $x, y \in \mathcal{X'}_i^{(q)}$ iff $\eta_q(x) = \eta_q(y)$ and we use a special indexation for this partition. Denote by $x'^{(q)}_{k_1,\ldots,k_d} = \left(\frac{2k_1+1}{2^{q+1}}, \ldots, \frac{2k_d+1}{2^{q+1}} \right)$. We say that $x'^{(q)}_{k_1,\ldots,k_d} \prec x'^{(q)}_{k'_1,\ldots,k'_d}$ if

$$\eta_{q-1}(x'^{(q)}_{k_1,\ldots,k_d}) \prec \eta_{q-1}(x'^{(q)}_{k'_1,\ldots,k'_d})$$

or

$$\eta_{q-1}(x'^{(q)}_{k_1,\ldots,k_d}) = \eta_{q-1}(x'^{(q)}_{k'_1,\ldots,k'_d}) \text{ and } (k_1, \ldots, k_d) < (k'_1, \ldots, k'_d),$$

for the lexicographical order on $\mathbb{N}^d$. Thus, the partition $(\mathcal{X'}_j^{(q)} : j = 1, \ldots, 2^{dq})$ has an increasing indexation according to the order of $(x'^{(q)}_{k_1,\ldots,k_d})$ for the order defined above. This order take care of the previous partition by splitting blocks in the given right order and, inside a block of a partition, we take the lexicographic order of $\mathbb{N}^d$. We introduce an other parameter $m \in \{1, \ldots, 2^{qd}\}$ and we define for all $i = 1, \ldots, m$, $\mathcal{X}_i^{(q)} = \mathcal{X'}_i^{(q)}$ and $\mathcal{X}_0^{(q)} = [0,1]^d - \cup_{i=1}^m \mathcal{X}_i^{(q)}$. Parameters q and m will be chosen later. We consider $W \in [0, m^{-1}]$, chosen later, and define the function f_X from $[0,1]^d$ to $\mathbb{R}$ by $f_X = W/\lambda_d(\mathcal{X}_1)$ (where λ_d is the Lebesgue measure on $[0,1]^d$) on $\mathcal{X}_1, \ldots, \mathcal{X}_m$ and $(1-mW)/\lambda_d(\mathcal{X}_0)$ on $\mathcal{X}_0$. We denote by P^X the probability distribution on $[0,1]^d$ with the density f_X w.r.t. the Lebesgue measure. For all $\sigma = (\sigma_1, \ldots, \sigma_m) \in \Omega = \{-1, 1\}^m$ we consider η_σ defined, for any $x \in [0,1]^d$, by

$$\eta_\sigma(x) = \begin{cases} \frac{1+\sigma_j h}{2} & \text{if } x \in \mathcal{X}_j, j = 1, \ldots, m, \\ 1 & \text{if } x \in \mathcal{X}_0. \end{cases}$$

We have a set of probability measures $\{\pi_\sigma : \sigma \in \Omega\}$ on $[0,1]^d \times \{-1, 1\}$ indexed by the hypercube Ω where P^X is the marginal on $[0,1]^d$ of π_σ and η_σ its conditional probability function of $Y = 1$ given X. We denote by f_σ^* the Bayes rule associated to π_σ, we have $f_\sigma^*(x) = \sigma_j$ if $x \in \mathcal{X}_j$ for $j = 1, \ldots, m$ and 1 if $x \in \mathcal{X}_0$, for any $\sigma \in \Omega$.

Now we give conditions on q, m and W such that for all σ in Ω, π_σ belongs to $\mathcal{P}_{w,h,a,A}$. If we choose

$$W = 2^{-dq}, \tag{6.18}$$

then, $f_X = \mathbb{1}_{[0,1]^d}$ (so $P^X << \lambda$ and $\forall x \in [0,1]^d, a \leq dP^X/d\lambda(x) \leq A$). We have clearly $|2\eta(x) - 1| \geq h$ for any $x \in [0,1]^d$. We can see that $f_\sigma^* \in \mathcal{F}_w^{(d)}$ for all $\sigma \in \{-1,1\}^m$ iff

$$\begin{array}{l} w(q+1) \geq \inf(x \in 2^d\mathbb{N} : x \geq m) \\ w(q) \geq \left\{ \begin{array}{ll} 2^d - 1 & \text{if} \quad m < 2^d \\ \inf(x \in 2^d\mathbb{N} : x \geq 2^{-d}m) & \text{otherwise} \end{array} \right. \\ \ldots \\ w(1) \geq \left\{ \begin{array}{ll} 2^d - 1 & \text{if} \quad m < 2^{dq} \\ \inf(x \in 2^d\mathbb{N} : x \geq 2^{-dq}m) & \text{otherwise} \end{array} \right. \\ w(0) \geq 1 \end{array} .$$

Since we have $w(0) = 1$, $w(j) \geq 2^d - 1$ and $w(j-1) \geq w(j)/2^d$ for all $j \geq 1$ then, $f_\sigma^* \in \mathcal{F}_w^{(d)}$ for all $\sigma \in \Omega$ iff

$$w(q+1) \geq \inf(x \in 2^d\mathbb{N} : x \geq m). \tag{6.19}$$

Take q, m and W such that (6.18) and (6.19) are fulfilled then, $\{\pi_\sigma : \sigma \in \Omega\}$ is a subset of $\mathcal{P}_{w,h,a,A}$. Let $\sigma \in \Omega$ and $\hat{f}_n$ be a classifier, we have

$$\begin{aligned} \mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^*\right] &= (1/2)\mathbb{E}_{\pi_\sigma}\left[|2\eta_\sigma(X) - 1||\hat{f}_n(X) - f_\sigma^*(X)|\right] \\ &\geq (h/2)\mathbb{E}_{\pi_\sigma}\left[|\hat{f}_n(X) - f_\sigma^*(X)|\right] \\ &\geq (h/2)\mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m \int_{\mathcal{X}_i} |\hat{f}_n(x) - f_\sigma^*(x)| dP^X(x) + \int_{\mathcal{X}_0} |\hat{f}_n(x) - f_\sigma^*(x)| dP^X(x)\right] \\ &\geq (Wh/2)\sum_{i=1}^m \mathbb{E}_{\pi_\sigma}\left[\int_{\mathcal{X}_i} |\hat{f}_n(x) - \sigma_i| \frac{dx}{\lambda(\mathcal{X}_1)}\right] \\ &\geq (Wh/2)\mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m \left|\sigma_i - \int_{\mathcal{X}_i} \hat{f}_n(x) \frac{dx}{\lambda(\mathcal{X}_1)}\right|\right]. \end{aligned}$$

We deduce that

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq (Wh/2) \inf_{\hat{\sigma}_n \in [-1,1]^m} \sup_{\sigma \in \{-1,1\}^m} \mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m |\sigma_i - \hat{\sigma}_i|\right].$$

Now, we control the Hellinger distance between two neighboring probability measures. Let ρ be the Hamming distance on Ω. Let σ, σ' in Ω such that $\rho(\sigma, \sigma') = 1$. We have

$$H^2(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}) = 2\left(1 - \left(1 - \frac{H^2(\pi_\sigma, \pi_{\sigma'})}{2}\right)^n\right),$$

and a straightforward calculus leads to $H^2(\pi_\sigma, \pi_{\sigma'}) = 2W\left(1 - \sqrt{1-h^2}\right)$. If we have $W \leq 1/n$ then, $H^2(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}) \leq \beta < 2$ where $\beta = 2\left(1 - \exp(1 - \sqrt{1-h^2})\right)$. One version of the Assouad Lemma (cf. [6] or Chapter 3) yields

$$\inf_{\hat{\sigma}_n \in [-1,1]^m} \sup_{\sigma \in \{-1,1\}^m} \mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m |\sigma_i - \hat{\sigma}_i|\right] \geq (m/4)\left(1 - (\beta/2)\right)^2.$$

We conclude that

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq Wh\frac{m}{8}\left(1 - \frac{\beta}{2}\right)^2. \tag{6.20}$$

Finally, we take $q = \lfloor \log n/(d \log 2) \rfloor$, $W = 2^{-dq} \leq 1/n$ and $m = w\left(\lfloor \log n/(d \log 2) \rfloor + 1\right) - (2^d - 1)$. Next, replacing these values in (6.20), we obtain

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1} \left(w\left(\lfloor \log n/(d \log 2) \rfloor + 1\right) - (2^d - 1) \right).$$

where $C_0 = (h/8) \exp\left(-(1 - \sqrt{1 - h^2})\right)$.

■

Proof of Corollary 6.1: It suffices to apply Theorem 6.4 to the function w defined by $w(j) = 2^{dj}$ for any integer j and $a = A = 1$ for $P^X = \lambda_d$.

■

Proof of Theorem 6.5:

(1) If we assume that $J_\epsilon \geq K$ then $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w_K^{(d)}(j) = (2^{dK})/(2^{dJ_\epsilon}(2^d - 1))$. We take

$$J_\epsilon = \left\lceil \frac{\log\left((A 2^{dK})/(\epsilon(2^d - 1))\right)}{d \log 2} \right\rceil$$

and ϵ_n the unique solution of $(1 + A)\epsilon_n = \exp(-nC\epsilon_n)$, where $C = a(1 - e^{-h^2/2})(2^d - 1)[A 2^{d(K+1)}]^{-1}$. Thus, $\epsilon_n \leq (\log n)/(Cn)$. For $J_n(K) = J_{\epsilon_n}$, we have

$$\mathcal{E}\left(\hat{f}_n^{(J_n(K))}\right) \leq C_{K,d,h,a,A} \frac{\log n}{n},$$

for any integer n such that $\log n \geq 2^{d(K+1)}(2^d - 1)^{-1}$ and $J_n(K) \geq K$, where $C_{K,d,h,a,A} = 2(1 + A)/C$.

If we have $\lfloor \log n/(d \log 2) \rfloor \geq 2$ then $w\left(\lfloor \log n/(d \log 2) \rfloor + 1\right) - (2^d - 1) \geq 2^d$, so we obtain the lower bound with the constant $C_{0,K} = 2^d C_0$ and if $\lfloor \log n/(d \log 2) \rfloor \geq K$ the constant can be $C_{0,K} = C_0(2^{dK} - (2^d - 1))$.

(2) If we have $J_\epsilon \geq N^{(d)}(\alpha)$, then $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w_\alpha^{(d)}(j) \leq (2^{d(1-\alpha)J_\epsilon}(2^{d(1-\alpha)} - 1))^{-1}$. We take

$$J_\epsilon = \left\lceil \frac{\log(A/(\epsilon(2^{d(1-\alpha)} - 1)))}{d(1 - \alpha) \log 2} \right\rceil.$$

Denote by ϵ_n the unique solution of $(1 + A)\epsilon_n = \exp(-nC\epsilon_n^{1/(1-\alpha)})$ where $C = a(1 - e^{-h^2/2})2^{-d}(A^{-1}(2^{d(1-\alpha)} - 1))^{1/(1-\alpha)}$. We have $\epsilon_n \leq (\log n/(nC))^{1-\alpha}$. For $J_n(\alpha) = J_{\epsilon_n}$, we have

$$\mathcal{E}\left(\hat{f}_n^{(J_n(\alpha))}\right) \leq \frac{2(1 + A)A}{2^{d(1-\alpha)} - 1} \left[\frac{2^d}{a(1 - e^{-h^2/2})}\right]^{1-\alpha} \left(\frac{\log n}{n}\right)^{1-\alpha}.$$

For the lower bound we have for any integer n,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 \max\left(1, n^{-1}\left(2^d n^\alpha - (2^d - 1)\right)\right).$$

■

Proof of Theorem 6.6: Let $\alpha \in (0, 1)$. For n large enough, we have $J^{(n)} \geq J_m(\alpha)$. Since the (SMA) assumption is equivalent to the margin assumption introduced by [91] and [116] with margin parameter equal to 1 (cf. proof of Proposition 3.1 of Chapter 3) we have, according to Corollary 8.1 of Chapter 8,

$$\mathbb{E}[R(\hat{f}_n) - R^*] \leq 3 \min_{J=0,\ldots,J^{(n)}} \mathbb{E}[R(\hat{f}_m^{(J)}) - R^*] + C \frac{(\log n) \log(J^{(n)} + 1)}{n}. \tag{6.21}$$

According to Theorem 6.5, we have

$$\mathbb{E}[R(\hat{f}_m^{(J)}) - R^*] \leq C'_{\alpha,h,a,A}\Big(\frac{\log m}{m}\Big)^{1-\alpha}.$$

Then, combining the last inequality, the fact that $m \leq n/2$ and (6.21), we complete the proof.

■

Proof of Theorem 6.7: Let $\epsilon > 0$. Denote by ϵ_0 the greatest positive number satisfying $\delta(\epsilon_0)\epsilon_0^2 \leq \epsilon$. Consider $N(\epsilon_0) = \mathcal{N}(\partial A, \epsilon_0, ||.||_\infty)$ and $x_1, \ldots, x_{N(\epsilon_0)} \in \mathbb{R}^2$ such that $\partial A \subset \cup_{j=1}^{N(\epsilon_0)} B_\infty(x_j, \epsilon_0)$. Since $2^{-J_{\epsilon_0}} \geq \epsilon_0$, only nine dyadic sets of frequency J_{ϵ_0} can be used to cover a ball of radius ϵ_0 for the infinity norm of $\mathbb{R}^2$. Thus, we only need $9N(\epsilon_0)$ dyadic sets of frequency J_{ϵ_0} to cover ∂A. Consider the partition of $[0,1]^2$ by dyadic sets of frequency J_{ϵ_0}. Except on the $9N(\epsilon_0)$ dyadic sets used to cover the boundary ∂A, the prediction rule f_A is constant, equal to 1 or -1, on the other dyadic sets. Thus, by taking $f_{\epsilon_0} = \sum_{k_1,k_2=0}^{2^{J_{\epsilon_0}}-1} a_{k_1,k_2}^{(J_{\epsilon_0})} \phi_{k_1,k_2}^{(J_{\epsilon_0})}$, where $a_{k_1,k_2}^{(J_{\epsilon_0})}$ is equal to one value of f_A in the dyadic set $\mathcal{I}_{k_1,k_2}^{(J_{\epsilon_0})}$, we have

$$||f_{\epsilon_0} - f_A||_{L^1(\lambda_2)} \leq 9N(\epsilon_0)2^{-2J_{\epsilon_0}} \leq 36\delta(\epsilon_0)\epsilon_0^2 \leq 36\epsilon.$$

■

Part 3

Applications for Concrete Models

CHAPTER 7

Simultaneous Adaptation to the Margin and to Complexity in Classification

We consider the problem of adaptation to the margin and to complexity in binary classification. We suggest an exponential weighting aggregation scheme. We use this aggregation procedure to construct classifiers which adapt automatically to margin and complexity. Two main examples are worked out in which adaptivity is achieved in frameworks proposed by Scovel and Steinwart (2004, 2005) and Tsybakov (2004). Adaptive schemes, like ERM or penalized ERM, usually involve a minimization step. It is not the case of our procedure.

Contents

The material of this chapter is an article accepted for publication in the *Annals of Statistics* (cf. [82]).

1. Introduction

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Denote by D_n a sample $((X_i, Y_i))_{i=1,\dots,n}$ of i.i.d. random pairs of observations where $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$. Denote by π the joint distribution of (X_i, Y_i) on $\mathcal{X} \times \{-1, 1\}$, and P^X the marginal distribution of X_i. Let (X, Y) be a random pair distributed according to π and independent of the data, and let the component X of the pair be observed. The problem of statistical learning in classification (pattern recognition) consists in predicting the corresponding value $Y \in \{-1, 1\}$.

A *prediction rule* is a measurable function $f : \mathcal{X} \longmapsto \{-1, 1\}$. The *misclassification error* associated to f is

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

It is well known (see, e.g., Devroye, Györfi and Lugosi (1996)) that

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

and η is the *a posteriori probability* defined by

$$\eta(x) = \mathbb{P}(Y = 1 | X = x),$$

for all $x \in \mathcal{X}$ (where sign(y) denotes the sign of $y \in \mathbb{R}$ with the convention sign(0) = 1). The prediction rule f^* is called the *Bayes rule* and R^* is called the *Bayes risk*. A *classifier* is a function, $\hat{f}_n = \hat{f}_n(X, D_n)$, measurable with respect to D_n and X with values in $\{-1, 1\}$, that assigns to every sample D_n a prediction rule $\hat{f}_n(., D_n) : \mathcal{X} \longmapsto \{-1, 1\}$. A key characteristic of $\hat{f}_n$ is the *generalization error* $\mathbb{E}[R(\hat{f}_n)]$, where

$$R(\hat{f}_n) = \mathbb{P}(Y \neq \hat{f}_n(X)|D_n).$$

The aim of statistical learning is to construct a classifier $\hat{f}_n$ such that $\mathbb{E}[R(\hat{f}_n)]$ is as close to R^* as possible. Accuracy of a classifier $\hat{f}_n$ is measured by the value $\mathbb{E}[R(\hat{f}_n)] - R^*$ called *excess risk* of $\hat{f}_n$.

Classical approach due to Vapnik and Chervonenkis (see, e.g. Devroye, Györfi and Lugosi (1996)) consists in searching for a classifier that minimizes the *empirical risk*

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(Y_i f(X_i) \leq 0)}, \tag{7.1}$$

over all prediction rules f in a source class $\mathcal{F}$, where $\mathbb{I}_A$ denotes the indicator of the set A. Minimizing the empirical risk (7.1) is computationally intractable for many sets $\mathcal{F}$ of classifiers, because this functional is neither convex nor continuous. Nevertheless, we might base a tractable estimation procedure on minimization of a convex surrogate ϕ for the loss (Cortes and Vapnik (1995), Freund and Schapire (1997), Lugosi and Vayatis (2004), Friedman, Hastie and Tibshirani (2000), Bühlmann and Yu (2002)). It has been recently shown that these classification methods often give classifiers with small Bayes risk (Blanchard, Lugosi and Vayatis (2004), Scovel and Steinwart (2004, 2005)). The main idea is that the sign of the minimizer of $A^{(\phi)}(f) = \mathbb{E}[\phi(Yf(X))]$ the *ϕ-risk*, where ϕ is a convex loss function and f a real valued function, is in many cases equal to the Bayes classifier f^*. Therefore minimizing $A_n^{(\phi)}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$ the *empirical ϕ-risk* and taking $\hat{f}_n = \text{sign}(\hat{F}_n)$ where $\hat{F}_n \in \text{Arg}\min_{f \in \mathcal{F}} A_n^{(\phi)}(f)$ leads to an approximation for f^*. Here, $\text{Arg}\min_{f \in \mathcal{F}} P(f)$, for a functional P, denotes the set of all $f \in \mathcal{F}$ such that $P(f) = \min_{f \in \mathcal{F}} P(f)$. Lugosi and Vayatis (2004), Blanchard, Lugosi and Vayatis (2004), Zhang (2004), Scovel and Steinwart (2004, 2005) and Bartlett, Jordan and McAuliffe (2003) give results on statistical properties of classifiers obtained by minimization of such a convex risk. A wide variety of classification methods in machine learning are based on this idea, in particular, on using the convex loss associated to support vector machines (Cortes and Vapnik (1995), Schölkopf and Smola (2002)),

$$\phi(x) = (1 - x)_+,$$

called the *hinge-loss*, where $z_+ = \max(0, z)$ denotes the positive part of $z \in \mathbb{R}$. Denote by

$$A(f) = \mathbb{E}[(1 - Yf(X))_+]$$

the *hinge risk* of $f : \mathcal{X} \longmapsto \mathbb{R}$ and set

$$A^* = \inf_f A(f), \tag{7.2}$$

where the infimum is taken over all measurable functions f. We will call A^* the *optimal hinge risk*. One may verify that the Bayes rule f^* attains the infimum in (7.2) and Lin (1999) and Zhang (2004) have shown that,

$$R(f) - R^* \leq A(f) - A^*, \tag{7.3}$$

for all measurable functions f with values in $\mathbb{R}$. Thus minimization of $A(f) - A^*$, the *excess hinge risk*, provides a reasonable alternative for minimization of excess risk.

The difficulty of classification is closely related to the behavior of the a posteriori probability η. Mammen and Tsybakov (1999), for the problem of discriminant analysis which is close to our classification problem, and Tsybakov (2004) have introduced an assumption on the closeness of η to $1/2$, called *margin assumption* (or *low noise assumption*). Under this assumption, the risk of a minimizer of the empirical risk over some fixed class $\mathcal{F}$ converges to the minimum risk over the class with *fast rates*, namely faster than $n^{-1/2}$. In fact, with no assumption on the joint distribution π, the convergence rate of the excess risk is not faster than $n^{-1/2}$ (cf. Devroye et al. (1996)). However, under the margin assumption, it can be as fast as n^{-1}. Minimizing penalized empirical hinge risk, under this assumption, also leads to fast convergence rates (Blanchard, Bousquet and Massart (2004), Scovel and Steinwart (2004, 2005)). Massart (2000), Massart and Nédélec (2003) and Massart (2004) also obtain results that can lead to fast rates in classification using penalized empirical risk in a special case of low noise assumption. Audibert and Tsybakov (2005) show that fast rates can be achieved for plug-in classifiers.

In this chapter we consider the problem of adaptive classification. Mammen and Tsybakov (1999) have shown that fast rates depend on both the *margin parameter* κ and complexity ρ of the class of candidate sets for $\{x \in \mathcal{X} : \eta(x) \geq 1/2\}$. Their results were non-adaptive supposing that κ and ρ were known. Tsybakov (2004) suggested an adaptive classifier that attains fast optimal rates, up to a logarithmic factor, without knowing κ and ρ. Tsybakov and van de Geer (2005) suggest a penalized empirical risk minimization classifier that adaptively attain, up to a logarithmic factor, the same fast optimal rates of convergence. Tarigan and van de Geer (2004) extend this result to l_1-penalized empirical hinge risk minimization. Koltchinskii (2005) uses Rademacher averages to get similar result without the logarithmic factor. Related works are those of Koltchinskii (2001), Koltchinskii and Panchenko (2002), Lugosi and Wegkamp (2004).

Note that the existing papers on fast rates either suggest classifiers that can be easily implementable but are non-adaptive, or adaptive schemes that are hard to apply in practice and/or do not achieve the minimax rates (they pay a price for adaptivity). The aim of the present chapter is to suggest and to analyze an exponential weighting aggregation scheme which does not require any minimization step unlike others adaptation schemes like ERM (Empirical Risk Minimization) or penalized ERM, and does not pay a price for adaptivity. This scheme is used a first time to construct minimax adaptive classifiers (cf. Theorem 7.3) and a second time to construct easily implementable classifiers that are adaptive simultaneously to complexity and to the margin parameters and that achieves the fast rates.

The chapter is organized as follows. In Section 2 we prove an oracle inequality which corresponds to the adaptation step of the procedure that we suggest. In Section 3 we apply the oracle inequality to two types of classifiers one of which is constructed by minimization on sieves (as in Tsybakov (2004)), and gives an adaptive classifier which attains fast optimal rates without logarithmic factor, and the other one is based on the support vector machines (SVM), following Scovel and Steinwart (2004, 2005). The later is realized as a computationally feasible procedure and it adaptively attains fast rates of convergence. In particular, we suggest a method of adaptive choice of the parameter of $L1$-SVM classifiers with gaussian RBF kernels. Proofs are given in Section 4.

2. Oracle inequalities

In this section we give an oracle inequality showing that a specifically defined convex combination of classifiers mimics the best classifier in a given finite set.

Suppose that we have $M \geq 2$ different classifiers $\hat{f}_1, \dots, \hat{f}_M$ taking values in $\{-1, 1\}$. The problem of model selection type aggregation, as studied in Nemirovski (2000), Yang (1999), Catoni (1997), Tsybakov (2003), consists in construction of a new classifier $\tilde{f}_n$ (called *aggregate*) which is approximatively at least as good, with respect to the excess risk, as the best among $\hat{f}_1, \dots, \hat{f}_M$. In most of these papers the aggregation is based on splitting of the sample in two independent subsamples D_m^1 and D_l^2 of sizes m and l respectively, where $m \gg l$ and $m + l = n$. The first subsample D_m^1 is used to construct the classifiers $\hat{f}_1, \dots, \hat{f}_M$ and the second subsample D_l^2 is used to aggregate them, i.e., to construct a new classifier that mimics in a certain sense the behavior of the best among the classifiers $\hat{f}_i$.

In this section we will not consider the sample splitting and concentrate only on the construction of aggregates (following Nemirovski (2000), Juditsky and Nemirovski (2000), Tsybakov (2003), Birgé (2004), Bunea, Tsybakov and Wegkamp (2004)). Thus, the first subsample is fixed and instead of classifiers $\hat{f}_1, \dots, \hat{f}_M$, we have fixed prediction rules $f_1, \dots, f_M$. Rather than working with a part of the initial sample we will suppose, for notational simplicity, that the whole sample D_n of size n is used for the aggregation step instead of a subsample D_l^2.

Our procedure is using exponential weights. The idea of exponential weights is well known, see, e.g., Augustin, Buckland and Burnham (1997), Yang (2000), Catoni (2001), Hartigan (2002) and Barron and Leung (2004). This procedure has been widely used in on-line prediction, see, e.g., Vovk (1990) and Lugosi and Cesa-Bianchi (2006). We consider the following aggregate which is a convex combination with exponential weights of M classifiers,

$$\tilde{f}_n = \sum_{j=1}^{M} w_j^{(n)} f_j, \tag{7.4}$$

where

$$w_j^{(n)} = \frac{\exp\left(\sum_{i=1}^n Y_i f_j(X_i)\right)}{\sum_{k=1}^M \exp\left(\sum_{i=1}^n Y_i f_k(X_i)\right)}, \quad \forall j = 1, \dots, M. \tag{7.5}$$

Since $f_1, \dots, f_M$ take their values in $\{-1, 1\}$, we have,

$$w_j^{(n)} = \frac{\exp\left(-nA_n(f_j)\right)}{\sum_{k=1}^M \exp\left(-nA_n(f_k)\right)}, \tag{7.6}$$

for all $j \in \{1, \dots, M\}$, where

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ \tag{7.7}$$

is the empirical analog of the hinge risk. Since $A_n(f_j) = 2R_n(f_j)$ for all $j = 1, \dots, M$, these weights can be written in terms of the empirical risks of f_j's,

$$w_j^{(n)} = \frac{\exp\left(-2nR_n(f_j)\right)}{\sum_{k=1}^M \exp\left(-2nR_n(f_k)\right)}, \ \forall j = 1, \dots, M.$$

The aggregation procedure defined by (7.4) with weights (7.6) does not need any minimization algorithm contrarily to the ERM procedure. Moreover, the following proposition shows that this exponential weighting aggregation scheme has similar theoretical property as the ERM procedure up to the residual $(\log M)/n$. In what follows the aggregation procedure defined by (7.4) with exponential weights (7.6) is called Aggregation procedure with Exponential Weights and is denoted by AEW.

PROPOSITION 7.1. *Let $M \geq 2$ be an integer, $f_1, \ldots, f_M$ be M prediction rules on $\mathcal{X}$. For any integers n, the AEW procedure $\tilde{f}_n$ satisfies*

$$A_n(\tilde{f}_n) \leq \min_{i=1,\ldots,M} A_n(f_i) + \frac{\log(M)}{n}. \tag{7.8}$$

Obviously, inequality (7.8) is satisfied when $\tilde{f}_n$ is the ERM aggregate defined by

$$\tilde{f}_n \in \operatorname{Arg} \min_{f \in \{f_1,\ldots,f_M\}} R_n(f).$$

It is a convex combination of f_j's with weights $w_j = 1$ for one $j \in \operatorname{Arg}\min_j R_n(f_j)$ and 0 otherwise.

We will use the following assumption (cf. Mammen and Tsybakov (1999), Tsybakov (2004)) that will allow us to get fast learning rates for the classifiers that we aggregate.

(MA1) Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the margin assumption (MA1)(κ) with margin parameter $1 \leq \kappa < +\infty$ if there exists $c > 0$ such that,*

$$\mathbb{E}\left\{|f(X) - f^*(X)|\right\} \leq c\left(R(f) - R^*\right)^{1/\kappa}, \tag{7.9}$$

for all measurable functions f with values in $\{-1, 1\}$.

We first give the following proposition which is valid not necessarily for the particular choice of weights given in (7.5).

PROPOSITION 7.2. *Let assumption (MA1)(κ) hold with some $1 \leq \kappa < +\infty$. Assume that there exist two positive numbers $a \geq 1, b$ such that $M \geq an^b$. Let $w_1, \ldots, w_M$ be M statistics measurable w.r.t. the sample D_n, such that $w_j \geq 0$, for all $j = 1, \ldots, M$, and $\sum_{j=1}^M w_j = 1, (\pi^{\otimes n} - a.s.)$. Define $\tilde{f}_n = \sum_{j=1}^M w_j f_j$, where $f_1, \ldots, f_M$ are prediction rules. There exists a constant $C_0 > 0$ (for instance, $C_0 = 10 + ca^{-1/(2b)} + a^{-1/b} \exp\left[\left(b(8c/6)^2\right) \vee \left(((8c/3) \vee 1)/b\right)^2\right]$) such that*

$$(1 - (\log M)^{-1/4})\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \mathbb{E}[A_n(\tilde{f}_n) - A_n(f^*)] + C_0 n^{-\frac{\kappa}{2\kappa-1}} (\log M)^{7/4},$$

where f^* is the Bayes rule.

As a consequence, we obtain the following oracle inequality.

THEOREM 7.1. *Let assumption (MA1)(κ) hold with some $1 \leq \kappa < +\infty$. Assume that there exist two positive numbers $a \geq 1, b$ such that $M \geq an^b$. Let $\tilde{f}_n$ satisfying (7.8), for instance the AEW or the ERM procedure. Then, $\tilde{f}_n$ satisfies*

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq \left(1 + \frac{2}{\log^{1/4}(M)}\right)\left\{2 \min_{j=1,\ldots,M} (R(f_j) - R^*) + C_0 \frac{\log^{7/4}(M)}{n^{\kappa/(2\kappa-1)}}\right\}, \tag{7.10}$$

for all integers $n \geq 1$, where $C_0 > 0$ appears in Proposition 7.2.

REMARK 7.1. *The factor* 2 *multiplying* $\min_{j=1,\dots,M}(R(f_j)-R^*)$ *in (7.10) is due to the relation between the hinge excess risk and the usual excess risk (cf. inequality (7.3)). The hinge-loss is more adapted for our convex aggregate, since we have the same statement without this factor, namely:*

$$\mathbb{E}\left[A(\tilde{f}_n)-A^*\right] \le \left(1+\frac{2}{\log^{1/4}(M)}\right)\left\{\min_{j=1,\dots,M}(A(f_j)-A^*)+C_0\frac{\log^{7/4}(M)}{n^{\kappa/(2\kappa-1)}}\right\}.$$

Moreover, linearity of the hinge-loss on $[-1,1]$ *leads to*

$$\min_{j=1,\dots,M}(A(f_j)-A^*)=\min_{f\in Conv}(A(f)-A^*),$$

where Conv is the convex hull of the set $\{f_j : j=1,\dots,M\}$*. Therefore, the excess hinge risk of* $\tilde{f}_n$ *is approximately the same as the one of the best convex combination of* f_j*'s.*

REMARK 7.2. *For a convex loss function* ϕ*, consider the empirical* ϕ*-risk* $A_n^{(\phi)}(f)$*. Our proof implies that the aggregate*

$$\tilde{f}_n^{(\phi)}(x)=\sum_{j=1}^M w_j^\phi f_j(x) \text{ with } w_j^\phi=\frac{\exp\left(-nA_n^{(\phi)}(f_j)\right)}{\sum_{k=1}^M\exp\left(-nA_n^{(\phi)}(f_k)\right)},\quad \forall j=1,\dots,M,$$

satisfies the inequality (7.8) with $A_n^{(\phi)}$ *in place of* A_n*.*

We consider next a recursive analog of the aggregate (7.4). It is close to the one suggested by Yang (2000) for the density aggregation under Kullback loss and by Catoni (2004) and Bunea and Nobel (2005) for regression model with squared loss. It can be also viewed as a particular instance of the mirror descent algorithm suggested in Juditsky, Nazin, Tsybakov and Vayatis (2005). We consider

$$\bar{f}_n=\frac{1}{n}\sum_{k=1}^n\tilde{f}_k=\sum_{j=1}^M\bar{w}_j f_j \tag{7.11}$$

where

$$\bar{w}_j=\frac{1}{n}\sum_{k=1}^n w_j^{(k)}=\frac{1}{n}\sum_{k=1}^n\frac{\exp(-kA_k(f_j))}{\sum_{l=1}^M\exp(-kA_k(f_l))}, \tag{7.12}$$

for all $j=1,\dots,M$, where $A_k(f)=(1/k)\sum_{i=1}^k(1-Y_if(X_i))_+$ is the empirical hinge risk of f and $w_j^{(k)}$ is the weight defined in (7.5), for the first k observations. This aggregate is especially useful for the on-line framework. The following theorem says that it has the same theoretical properties as the aggregate (7.4).

THEOREM 7.2. *Let assumption (MA1)(*κ*) hold with some* $1\le\kappa<+\infty$*. Assume that there exist two positive numbers* $a\ge 1, b$ *such that* $M\ge an^b$*. Then the convex aggregate* $\bar{f}_n$ *defined by (7.11) satisfies*

$$\mathbb{E}\left[R(\bar{f}_n)-R^*\right]\le\left(1+\frac{2}{\log^{1/4}(M)}\right)\left\{2\min_{j=1,\dots,M}(R(f_j)-R^*)+C_0\gamma(n,\kappa)\log^{7/4}(M)\right\},$$

for all integers $n\ge 1$*, where* $C_0>0$ *appears in Proposition 7.2 and* $\gamma(n,\kappa)$ *is equal to* $((2\kappa-1)/(\kappa-1))n^{-\frac{\kappa}{2\kappa-1}}$ *if* $\kappa>1$ *and to* $(\log n)/n$ *if* $\kappa=1$*.*

REMARK 7.3. *For all* $k\in\{1,\dots,n-1\}$*, less observations are used to construct* $\tilde{f}_k$ *than for the construction of* $\tilde{f}_n$*, thus, intuitively, we expect that* $\tilde{f}_n$ *will learn better than* $\tilde{f}_k$*.*

In view of (7.11), $\bar{f}_n$ is an average of aggregates whose performances are, a priori, worse than those of $\tilde{f}_n$, therefore its expected learning properties would be presumably worse than those of $\tilde{f}_n$. An advantage of the aggregate $\bar{f}_n$ is in its recursive construction, but the risk behavior of $\tilde{f}_n$ seems to be better than that of $\bar{f}_n$. In fact, it is easy to see that Theorem 7.2 is satisfied for any aggregate $\bar{f}_n = \sum_{k=1}^n w_k \tilde{f}_k$ where $w_k \geq 0$ and $\sum_{k=1}^n w_k = 1$ with $\gamma(n,\kappa) = \sum_{k=1}^n w_k k^{-\kappa/(2\kappa-1)}$, and the remainder term is minimized for $w_j = 1$ when $j = n$ and 0 elsewhere, that is for $\bar{f}_n = \tilde{f}_n$.

REMARK 7.4. *In this section, we have only dealt with the aggregation step. But the construction of classifiers has to take place prior to this step. This needs a split of the sample as discussed at the beginning of this section. The main drawback of this method is that only a part of the sample is used for the initial estimation. However, by using different splits of the sample and taking the average of the aggregates associated with each of them, we get a more balanced classifier which does not depend on a particular split. Since the hinge loss is linear on $[-1,1]$, we have the same result as Theorem 7.1 and 7.2 for an average of aggregates of the form (7.4) and (7.11), respectively, for averaging over different splits of the sample.*

3. Adaptation to the margin and to complexity

In Scovel and Steinwart (2004, 2005) and Tsybakov (2004) two concepts of complexity are used. In this section we show that combining classifiers used by Tsybakov (2004) or $L1$-SVM classifiers of Scovel and Steinwart (2004, 2005) with our aggregation method leads to classifiers that are adaptive both to the margin parameter and to the complexity, in the two cases. Results are established for the first method of aggregation defined in (7.4) but they are also valid for the recursive aggregate defined in (7.11).

We use a sample splitting to construct our aggregate. The first subsample $D_m^1 = ((X_1,Y_1),\dots,(X_m,Y_m))$, where $m = n - l$ and $l = \lceil an/\log n \rceil$ for a constant $a > 0$, is implemented to construct classifiers and the second subsample D_l^2, made of the l last observations $((X_{m+1},Y_{m+1}),\dots,(X_n,Y_n))$, is implemented to aggregate them by the procedure (7.4).

3.1. Adaptation in the framework of Tsybakov. Here we take $\mathcal{X} = \mathbb{R}^d$. Introduce the following pseudo-distance, and its empirical analogue, between the sets $G, G' \subseteq \mathcal{X}$:

$$d_\Delta(G,G') = P^X(G \Delta G') \ , \ d_{\Delta,e}(G,G') = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{(X_i \in G \Delta G')},$$

where $G\Delta G'$ is the symmetric difference between sets G and G'. If $\mathcal{Y}$ is a class of subsets of $\mathcal{X}$, denote by $\mathcal{H}_B(\mathcal{Y},\delta,d_\Delta)$ the *δ-entropy with bracketing of $\mathcal{Y}$ for the pseudo-distance* d_Δ (cf. van de Geer (2000) p.16). We say that $\mathcal{Y}$ has a *complexity bound* $\rho > 0$ if there exists a constant $A > 0$ such that

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

Various examples of classes $\mathcal{Y}$ having this property can be found in Dudley (1974), Korostelev and Tsybakov (1993), Mammen and Tsybakov (1995, 1999).

Let $(\mathcal{G}_\rho)_{\rho_{min} \leq \rho \leq \rho_{max}}$ be a collection of classes of subsets of $\mathcal{X}$, where $\mathcal{G}_\rho$ has a complexity bound ρ, for all $\rho_{min} \leq \rho \leq \rho_{max}$. This collection corresponds to an a priori knowledge on π that the set $G^* = \{x \in \mathcal{X} : \eta(x) > 1/2\}$ lies in one of these classes (typically we have $\mathcal{G}_\rho \subset \mathcal{G}_{\rho'}$ if $\rho \leq \rho'$). The aim of adaptation to the margin and complexity

is to propose $\tilde{f}_n$ a classifier free from κ and ρ such that, if π satisfies (MA1)(κ) and $G^* \in \mathcal{G}_\rho$, then $\tilde{f}_n$ learns with the optimal rate $n^{-\frac{\kappa}{2\kappa+\rho-1}}$ (optimality has been established in Mammen and Tsybakov (1999)), and this property holds for all values of $\kappa \geq 1$ and $\rho_{min} \leq \rho \leq \rho_{max}$. Following Tsybakov (2004), we introduce the following assumption on the collection $(\mathcal{G}_\rho)_{\rho_{min}\leq\rho\leq\rho_{max}}$.

(A1)(Complexity Assumption). *Assume that $0 < \rho_{min} < \rho_{max} < 1$ and $\mathcal{G}_\rho$'s are classes of subsets of $\mathcal{X}$ such that $\mathcal{G}_\rho \subseteq \mathcal{G}_{\rho'}$ for $\rho_{min} \leq \rho < \rho' \leq \rho_{max}$ and the class $\mathcal{G}_\rho$ has complexity bound ρ. For any integer n, we define $\rho_{n,j} = \rho_{min} + \frac{j}{N(n)}(\rho_{max} - \rho_{min})$, $j = 0, \ldots, N(n)$, where $N(n)$ satisfies $A_0' n^{b'} \leq N(n) \leq A_0 n^b$, for some finite $b \geq b' > 0$ and $A_0, A_0' > 0$. Assume that for all $n \in \mathbb{N}$,*

(i) for all $j = 0, \ldots, N(n)$ there exists $\mathcal{N}_n^j$ an ϵ-net on $\mathcal{G}_{\rho_{n,j}}$ for the pseudo-distance d_Δ or $d_{\Delta,e}$, where $\epsilon = a_j n^{-\frac{1}{1+\rho_{n,j}}}$, $a_j > 0$ and $\max_j a_j < +\infty$,

(ii) $\mathcal{N}_n^j$ has a complexity bound $\rho_{n,j}$, for $j = 0, \ldots, N(n)$.

The first subsample D_m^1 is used to construct the ERM classifiers $\hat{f}_m^j(x) = 2\mathbb{1}_{\hat{G}_m^j}(x) - 1$, where $\hat{G}_m^j \in \operatorname{Arg\,min}_{G\in\mathcal{N}_m^j} R_m(2\mathbb{1}_G - 1)$ for all $j = 0, \ldots, N(m)$, and the second subsample D_l^2 is used to construct the exponential weights of the aggregation procedure,

$$w_j^{(l)} = \frac{\exp\left(-lA^{[l]}(\hat{f}_m^j)\right)}{\sum_{k=1}^{N(m)} \exp\left(-lA^{[l]}(\hat{f}_m^k)\right)}, \ \forall j = 0, \ldots, N(m),$$

where $A^{[l]}(f) = (1/l)\sum_{i=m+1}^{n} (1 - Y_i f(X_i))_+$ is the empirical hinge risk of $f : \mathcal{X} \longmapsto \mathbb{R}$ based on the subsample D_l^2. We consider

$$\tilde{f}_n(x) = \sum_{j=0}^{N(m)} w_j^{(l)} \hat{f}_m^j(x), \quad \forall x \in \mathcal{X}. \tag{7.13}$$

The construction of $\hat{f}_m^j$'s does not depend on the margin parameter κ.

THEOREM 7.3. *Let $(\mathcal{G}_\rho)_{\rho_{min}\leq\rho\leq\rho_{max}}$ be a collection of classes satisfying Assumption (A1). Then, the aggregate defined in (7.13) satisfies*

$$\sup_{\pi\in\mathcal{P}_{\kappa,\rho}} \mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq Cn^{-\frac{\kappa}{2\kappa+\rho-1}}, \quad \forall n \geq 1,$$

for all $1 \leq \kappa < +\infty$ and all $\rho \in [\rho_{min}, \rho_{max}]$, where $C > 0$ is a constant depending only on $a, b, b', A, A_0, A_0', \rho_{min}, \rho_{max}$ and κ, and $\mathcal{P}_{\kappa,\rho}$ is the set of all probability measures π on $\mathcal{X} \times \{-1, 1\}$ such that Assumption (MA1)(κ) is satisfied and $G^* \in \mathcal{G}_\rho$.

3.2. Adaptation in the framework of Scovel and Steinwart.

3.2.1. *The case of a continuous kernel.* Scovel and Steinwart (2005) have obtained fast learning rates for SVM classifiers depending on three parameters, the *margin parameter* $0 \leq \alpha < +\infty$, the complexity exponent $0 < p \leq 2$ and the approximation exponent $0 \leq \beta \leq 1$. The margin assumption was first introduced in Mammen and Tsybakov (1999) for the problem of discriminant analysis and in Tsybakov (2004) for the classification problem, in the following way:

(MA2) Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X}\times\{-1, 1\}$ satisfies the margin assumption (MA2)(α) with margin parameter $0 \leq \alpha < +\infty$ if there exists $c_0 > 0$ such that*

$$\mathbb{P}\left(|2\eta(X) - 1| \leq t\right) \leq c_0 t^\alpha, \quad \forall t > 0. \tag{7.14}$$

As shown in Boucheron, Bousquet and Lugosi (2006), the margin assumptions (MA1)(κ) and (MA2)(α) are equivalent with $\kappa = \frac{1+\alpha}{\alpha}$ for $\alpha > 0$.

Let $\mathcal{X}$ be a compact metric space. Let H be a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ (see, e.g., Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002)), B_H its closed unit ball. Denote by $\mathcal{N}\left(B_H, \epsilon, L_2(P_n^X)\right)$ the *ϵ-covering number* of B_H w.r.t. the canonical distance of $L_2(P_n^X)$, the L_2-space w.r.t. the empirical measure, P_n^X, on $X_1, \ldots, X_n$. Introduce the following assumptions as in Scovel and Steinwart (2005):

(A2) *There exists $a_0 > 0$ and $0 < p \leq 2$ such that for any integer n,*

$$\sup_{D_n \in (\mathcal{X} \times \{-1,1\})^n} \log \mathcal{N}\left(B_H, \epsilon, L_2(P_n^X)\right) \leq a_0 \epsilon^{-p}, \quad \forall \epsilon > 0,$$

Note that the supremum is taken over all the samples of size n and the bound is assuming for any n. Every RKHS satisfies (A2) with $p = 2$ (cf. Scovel et al. (2005)). We define the *approximation error function* of the $L1$-SVM as $a(\lambda) \stackrel{\text{def}}{=} \inf_{f \in H}\left(\lambda ||f||_H^2 + A(f)\right) - A^*$.

(A3) *The RKHS H, approximates π with exponent $0 \leq \beta \leq 1$, if there exists a constant $C_0 > 0$ such that $a(\lambda) \leq C_0 \lambda^\beta, \quad \forall \lambda > 0$.*

Note that every RKHS approximates every probability measure with exponent $\beta = 0$ and the other extremal case $\beta = 1$ is equivalent to the fact that the Bayes classifier f^* belongs to the RKHS (cf. Scovel et al. (2005)). Furthermore, $\beta > 1$ only for probability measures such that $\mathbb{P}\left(\eta(X) = 1/2\right) = 1$ (cf. Scovel et al. (2005)). If (A2) and (A3) hold, the parameter (p, β) can be considered as a complexity parameter characterizing π and H.

Let H be a RKHS with a continuous kernel on $\mathcal{X}$ satisfying (A2) with a parameter $0 < p < 2$. Define the $L1$-SVM classifier by

$$\hat{f}_n^\lambda = \operatorname{sign}(\hat{F}_n^\lambda) \text{ where } \hat{F}_n^\lambda \in \operatorname{Arg}\min_{f \in H}\left(\lambda ||f||_H^2 + A_n(f)\right) \tag{7.15}$$

and $\lambda > 0$ is called the *regularization parameter*. Assume that the probability measure π belongs to the set $\mathcal{Q}_{\alpha,\beta}$ of all probability measures on $\mathcal{X} \times \{-1, 1\}$ satisfying (MA2)(α) with $\alpha \geq 0$ and (A3) with a complexity parameter (p, β) where $0 < \beta \leq 1$. It has been shown in Scovel et al. (2005) that the $L1$-SVM classifier, $\hat{f}_n^{\lambda_n^{\alpha,\beta}}$, where the regularization parameter is $\lambda_n^{\alpha,\beta} = n^{-\frac{4(\alpha+1)}{(2\alpha+p\alpha+4)(1+\beta)}}$, satisfies the following excess risk bound: for any $\epsilon > 0$, there exists $C > 0$ depending only on α, p, β and ϵ such that

$$\mathbb{E}\left[R(\hat{f}_n^{\lambda_n^{\alpha,\beta}}) - R^*\right] \leq C n^{-\frac{4\beta(\alpha+1)}{(2\alpha+p\alpha+4)(1+\beta)}+\epsilon}, \quad \forall n \geq 1. \tag{7.16}$$

Remark that if $\beta = 1$, that is $f^* \in H$, then the learning rate in (7.16) is (up to an ϵ) $n^{-2(\alpha+1)/(2\alpha+p\alpha+4)}$ which is a fast rate since $2(\alpha+1)/(2\alpha+p\alpha+4) \in [1/2, 1)$.

To construct the classifier $\hat{f}_n^{\lambda_n^{\alpha,\beta}}$ we need to know parameters α and β that are not available in practice. Thus, it is important to construct a classifier, free from these parameters, which has the same behavior as $\hat{f}_n^{\lambda_n^{\alpha,\beta}}$, if the underlying distribution π belongs to $\mathcal{Q}_{\alpha,\beta}$. Below we give such a construction.

Since the RKHS H is given, the implementation of the $L1$-SVM classifier $\hat{f}_n^\lambda$ only requires the knowledge of the regularization parameter λ. Thus, to provide an easily implementable procedure, using our aggregation method, it is natural to combine $L1$-SVM classifiers constructed for different values of λ in a finite grid. We now define such a procedure.

We consider the $L1$-SVM classifiers $\hat{f}_m^\lambda$, defined in (7.15) for the subsample D_m^1, where λ lies in the grid

$$\mathcal{G}(l) = \{\lambda_{l,k} = l^{-\phi_{l,k}} : \phi_{l,k} = 1/2 + k\Delta^{-1}, k = 0, \dots, \lfloor 3\Delta/2 \rfloor\},$$

where we set $\Delta = l^{b_0}$ with some $b_0 > 0$. The subsample D_l^2 is used to aggregate these classifiers by the procedure (7.4), namely

$$\tilde{f}_n = \sum_{\lambda \in \mathcal{G}(l)} w_\lambda^{(l)} \hat{f}_m^\lambda \tag{7.17}$$

where

$$w_\lambda^{(l)} = \frac{\exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^\lambda(X_i)\right)}{\sum_{\lambda' \in \mathcal{G}(l)} \exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{\lambda'}(X_i)\right)} = \frac{\exp\left(-lA^{[l]}(\hat{f}_m^\lambda)\right)}{\sum_{\lambda' \in \mathcal{G}(l)} \exp\left(-lA^{[l]}(\hat{f}_m^{\lambda'})\right)},$$

and $A^{[l]}(f) = (1/l)\sum_{i=m+1}^n (1 - Y_i f(X_i))_+$.

THEOREM 7.4. *Let H be a RKHS with a continuous kernel on a compact metric space $\mathcal{X}$ satisfying (A2) with a parameter $0 < p < 2$. Let K be a compact subset of $(0, +\infty) \times (0, 1]$. The classifier $\tilde{f}_n$, defined in (7.17), satisfies*

$$\sup_{\pi \in \mathcal{Q}_{\alpha,\beta}} \mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \le Cn^{-\frac{4\beta(\alpha+1)}{(2\alpha+p\alpha+4)(1+\beta)}+\epsilon}$$

for all $(\alpha, \beta) \in K$ and $\epsilon > 0$, where $\mathcal{Q}_{\alpha,\beta}$ is the set of all probability measures on $\mathcal{X} \times \{-1, 1\}$ satisfying (MA2)(α) and (A2) with a complexity parameter (p, β) and $C > 0$ is a constant depending only on ϵ, p, K, a and b_0.

3.2.2. *The case of the Gaussian RBF kernel.* In this subsection we apply our aggregation procedure to $L1$-SVM classifiers using *Gaussian RBF kernel*. Let $\mathcal{X}$ be the closed unit ball of the space $\mathbb{R}^{d_0}$ endowed with the Euclidean norm $||x|| = \left(\sum_{i=1}^{d_0} x_i^2\right)^{1/2}, \forall x = (x_1, \dots, x_{d_0}) \in \mathbb{R}^{d_0}$. Gaussian RBF kernel is defined as $K_\sigma(x, x') = \exp\left(-\sigma^2||x - x'||^2\right)$ for $x, x' \in \mathcal{X}$ where σ is a parameter and σ^{-1} is called the *width* of the gaussian kernel. The RKHS associated to K_σ is denoted by H_σ.

Scovel and Steinwart (2004) introduced the following assumption:

(GNA) Geometric noise assumption. *There exist $C_1 > 0$ and $\gamma > 0$ such that*

$$\mathbb{E}\left[|2\eta(X) - 1| \exp\left(-\frac{\tau(X)^2}{t}\right)\right] \le C_1 t^{\frac{\gamma d_0}{2}}, \quad \forall t > 0.$$

Here τ is a function on $\mathcal{X}$ with values in $\mathbb{R}$ which measures the distance between a given point x and the decision boundary, namely,

$$\tau(x) = \begin{cases} d(x, G_0 \cup G_1), & \text{if } x \in G_{-1}, \\ d(x, G_0 \cup G_{-1}), & \text{if } x \in G_1, \\ 0 & \text{otherwise,} \end{cases}$$

for all $x \in \mathcal{X}$, where $G_0 = \{x \in \mathcal{X} : \eta(x) = 1/2\}$, $G_1 = \{x \in \mathcal{X} : \eta(x) > 1/2\}$ and $G_{-1} = \{x \in \mathcal{X} : \eta(x) < 1/2\}$. Here $d(x, A)$ denotes the Euclidean distance from a point x to the set A. If π satisfies Assumption (GNA) for a $\gamma > 0$, we say that π has a *geometric noise exponent* γ.

The $L1$-SVM classifier associated to the gaussian RBF kernel with width σ^{-1} and regularization parameter λ is defined by $\hat{f}_n^{(\sigma,\lambda)} = \text{sign}(\hat{F}_n^{(\sigma,\lambda)})$ where $\hat{F}_n^{(\sigma,\lambda)}$ is given by (7.15) with $H = H_\sigma$. Using the standard development related to SVM (cf. Schölkopf and

Smola (2002)), we may write $\hat{F}_n^{(\sigma,\lambda)}(x) = \sum_{i=1}^n \hat{C}_i K_\sigma(X_i, x), \forall x \in \mathcal{X}$, where $\hat{C}_1, \dots, \hat{C}_n$ are solutions of the following maximization problem

$$\max_{0 \le 2\lambda C_i Y_i \le n^{-1}} \left\{ 2\sum_{i=1}^n C_i Y_i - \sum_{i,j=1}^n C_i C_j K_\sigma(X_i, X_j) \right\},$$

that can be obtained using a standard quadratic programming software. According to Scovel et al. (2004), if the probability measure π on $\mathcal{X} \times \{-1, 1\}$, satisfies the margin assumption (MA2)(α) with margin parameter $0 \le \alpha < +\infty$ and Assumption (GNA) with a geometric noise exponent $\gamma > 0$, the classifier $\hat{f}_n^{(\sigma_n^{\alpha,\gamma}, \lambda_n^{\alpha,\gamma})}$ where regularization parameter and width are defined by

$$\lambda_n^{\alpha,\gamma} = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} & \text{if } \gamma \le \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2(\gamma+1)(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}} & \text{otherwise,} \end{cases} \quad \text{and } \sigma_n^{\alpha,\gamma} = (\lambda_n^{\alpha,\gamma})^{-\frac{1}{(\gamma+1)d_0}},$$

satisfies

$$\mathbb{E}\left[R(\hat{f}_n^{(\sigma_n^{\alpha,\gamma}, \lambda_n^{\alpha,\gamma})}) - R^*\right] \le C \begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\epsilon} & \text{if } \gamma \le \frac{\alpha+2}{2\alpha}, \\ n^{-\frac{2\gamma(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}+\epsilon} & \text{otherwise,} \end{cases} \tag{7.18}$$

for all $\epsilon > 0$, where $C > 0$ is a constant which depends only on α, γ and ϵ. Remark that fast rates are obtained only for $\gamma > (3\alpha + 4)/(2\alpha)$.

To construct the classifier $\hat{f}_n^{(\sigma_n^{\alpha,\gamma}, \lambda_n^{\alpha,\gamma})}$ we need to know parameters α and γ, which are not available in practice. Like in Subsection 3.2.1 we use our procedure to obtain a classifier which is adaptive to the margin and to the geometric noise parameters. Our aim is to provide an easily computable adaptive classifier. We propose the following method based on a grid for (σ, λ). We consider the finite sets

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

where we let $\Delta = l^{b_0}$ for some $b_0 > 0$, and

$$\mathcal{N}(l) = \left\{ (\sigma_{l,\varphi}, \lambda_{l,\psi}) = \left(l^{\varphi/d_0}, l^{-\psi} \right) : (\varphi, \psi) \in \mathcal{M}(l) \right\}.$$

We construct the family of classifiers $\left(\hat{f}_m^{(\sigma,\lambda)} : (\sigma, \lambda) \in \mathcal{N}(l) \right)$ using the observations of the subsample D_m^1 and we aggregate them by the procedure (7.4) using D_l^2, namely

$$\tilde{f}_n = \sum_{(\sigma,\lambda) \in \mathcal{N}(l)} w_{\sigma,\lambda}^{(l)} \hat{f}_m^{(\sigma,\lambda)} \tag{7.19}$$

where

$$w_{\sigma,\lambda}^{(l)} = \frac{\exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(\sigma,\lambda)}(X_i) \right)}{\sum_{(\sigma',\lambda') \in \mathcal{N}(l)} \exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(\sigma',\lambda')}(X_i) \right)}, \quad \forall (\sigma, \lambda) \in \mathcal{N}(l). \tag{7.20}$$

Denote by $\mathcal{R}_{\alpha,\gamma}$ the set of all probability measures on $\mathcal{X} \times \{-1, 1\}$ satisfying both the margin assumption (MA2)(α) with a margin parameter $\alpha > 0$ and Assumption (GNA) with a geometric noise exponent $\gamma > 0$. Define $\mathcal{U} = \{(\alpha, \gamma) \in (0, +\infty)^2 : \gamma > \frac{\alpha+2}{2\alpha}\}$ and $\mathcal{U}' = \{(\alpha, \gamma) \in (0, +\infty)^2 : \gamma \le \frac{\alpha+2}{2\alpha}\}$.

THEOREM 7.5. *Let K be a compact subset of $\mathcal{U}$ and K' a compact subset of $\mathcal{U}'$. The aggregate $\tilde{f}_n$, defined in (7.19), satisfies*

$$\sup_{\pi\in\mathcal{R}_{\alpha,\gamma}} \mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq C \begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\epsilon} & \text{if } (\alpha,\gamma)\in K', \\ n^{-\frac{2\gamma(\alpha+1)}{2\gamma(\alpha+2)+3\alpha+4}+\epsilon} & \text{if } (\alpha,\gamma)\in K, \end{cases}$$

for all $(\alpha,\gamma)\in K\cup K'$ and $\epsilon>0$, where $C>0$ depends only on ϵ, K, K', a and b_0.

4. Proofs

LEMMA 7.1. *For all positive v,t and all $\kappa\geq 1$: $\quad t+v\geq v^{\frac{2\kappa-1}{2\kappa}} t^{\frac{1}{2\kappa}}$.*

Proof. Since log is concave, we have $\log(ab) = (1/x)\log(a^x) + (1/y)\log(b^y) \leq \log\left(a^x/x + b^y/y\right)$ for all positive numbers a,b and x,y such that $1/x+1/y=1$, thus $ab\leq a^x/x + b^y/y$. Lemma 7.1 follows by applying this relation with $a = t^{1/(2\kappa)}, x = 2\kappa$ and $b = v^{(2\kappa-1)/(2\kappa)}$.

Proof of Proposition 7.1. Observe that $(1-x)_+ = 1-x$ for $x\leq 1$. Since $Y_i\tilde{f}_n(X_i)\leq 1$ and $Y_i f_j(X_i)\leq 1$ for all $i=1,\dots,n$ and $j=1,\dots,M$, we have $A_n(\tilde{f}_n) = \sum_{j=1}^M w_j^{(n)} A_n(f_j)$. We have $A_n(f_j) = A_n(f_{j_0}) + \frac{1}{n}\left(\log(w_{j_0}^{(n)}) - \log(w_i^{(n)})\right)$, for any $j, j_0 = 1,\dots,M$, where weights $w_j^{(n)}$ are defined in (7.6) by

$$w_j^{(n)} = \frac{\exp\left(-nA_n(f_j)\right)}{\sum_{k=1}^M \exp\left(-nA_n(f_k)\right)},$$

and by multiplying the last equation by $w_j^{(n)}$ and summing up over j, we get

$$A_n(\tilde{f}_n) \leq \min_{j=1\dots,M} A_n(f_j) + \frac{\log M}{n}. \tag{7.21}$$

Since $\log(w_{j_0}^{(n)})\leq 0, \forall j_0 = 1,\dots,M$ and $\sum_{j=1}^M w_j^{(n)} \log\left(\frac{w_j^{(n)}}{1/M}\right) = K(w|u)\geq 0$ where $K(w|u)$ denotes the Kullback-leibler divergence between the weights $w = (w_j^{(n)})_{j=1,\dots,M}$ and uniform weights $u = (1/M)_{j=1,\dots,M}$.

Proof of Proposition 7.2. Denote by $\gamma = (\log M)^{-1/4}$, $u = 2\gamma n^{-\frac{\kappa}{2\kappa-1}}\log^2 M$ and $W_n = (1-\gamma)(A(\tilde{f}_n) - A^*) - (A_n(\tilde{f}_n) - A_n(f^*))$. We have:

$$\begin{aligned} \mathbb{E}[W_n] &= \mathbb{E}\left[W_n(\mathbb{1}_{(W_n\leq u)} + \mathbb{1}_{(W_n>u)})\right] \leq u + \mathbb{E}\left[W_n\mathbb{1}_{(W_n>u)}\right] \\ &= u + u\mathbb{P}(W_n>u) + \int_u^{+\infty}\mathbb{P}(W_n>t)\,dt \leq 2u + \int_u^{+\infty}\mathbb{P}(W_n>t)\,dt. \end{aligned}$$

On the other hand $(f_j)_{j=1,\dots,M}$ are prediction rules, so we have $A(f_j) = 2R(f_j)$ and $A_n(f_j) = 2R_n(f_j)$, (recall that $A^* = 2R^*$). Moreover we work in the linear part of the hinge-loss, thus

$$\begin{aligned} \mathbb{P}(W_n>t) &= \mathbb{P}\left(\sum_{j=1}^M w_j\left((A(f_j)-A^*)(1-\gamma) - (A_n(f_j) - A_n(f^*))\right) > t\right) \\ &\leq \mathbb{P}\left(\max_{j=1,\dots,M}\left((A(f_j)-A^*)(1-\gamma) - (A_n(f_j) - A_n(f^*))\right) > t\right) \\ &\leq \sum_{j=1}^M \mathbb{P}\left(Z_j > \gamma\left(R(f_j)-R^*\right) + t/2\right), \end{aligned}$$

for all $t > u$, where $Z_j = R(f_j) - R^* - (R_n(f_j) - R_n(f^*))$ for all $j = 1, \ldots, M$(recall that $R_n(f)$ is the empirical risk defined in (7.1)).

Let $j \in \{1, \ldots, M\}$. We can write $Z_j = (1/n)\sum_{i=1}^n (\mathbb{E}[\zeta_{i,j}] - \zeta_{i,j})$ where $\zeta_{i,j} = \mathbb{1}_{(Y_i f_j(X_i) \leq 0)} - \mathbb{1}_{(Y_i f^*(X_i) \leq 0)}$. We have $|\zeta_{i,j}| \leq 1$ and, under the margin assumption, we have $\mathbb{V}(\zeta_{i,j}) \leq \mathbb{E}(\zeta_{i,j}^2) = \mathbb{E}\left[|f_j(X) - f^*(X)|\right] \leq c\left(R(f_j) - R^*\right)^{1/\kappa}$ where $\mathbb{V}$ is the symbol of the variance. By applying Bernstein's inequality and Lemma 1 respectively, we get

$$\begin{aligned}\mathbb{P}\left[Z_j > \epsilon\right] &\leq \exp\left(-\frac{n\epsilon^2}{2c\left(R(f_j) - R^*\right)^{1/\kappa} + 2\epsilon/3}\right) \\ &\leq \exp\left(-\frac{n\epsilon^2}{4c\left(R(f_j) - R^*\right)^{1/\kappa}}\right) + \exp\left(-\frac{3n\epsilon}{4}\right),\end{aligned}$$

for all $\epsilon > 0$. Denote by $u_j = u/2 + \gamma(R(f_j) - R^*)$. After a standard calculation we get

$$\int_u^{+\infty} \mathbb{P}\left(Z_j > \gamma\left(R(f_j) - R^*\right) + t/2\right) dt = 2\int_{u_j}^{+\infty} \mathbb{P}(Z_j > \epsilon) d\epsilon \leq B_1 + B_2,$$

where

$$B_1 = \frac{4c(R(f_j) - R^*)^{1/\kappa}}{nu_j} \exp\left(-\frac{nu_j^2}{4c(R(f_j) - R^*)^{1/\kappa}}\right)$$

and

$$B_2 = \frac{8}{3n} \exp\left(-\frac{3nu_j}{4}\right).$$

Since $R(f_j) \geq R^*$, Lemma 7.1 yields $u_j \geq \gamma\left(R(f_j) - R^*\right)^{\frac{1}{2\kappa}} (\log M)^{\frac{2\kappa-1}{\kappa}} n^{-1/2}$. For any $a > 0$, the mapping $x \mapsto (ax)^{-1}\exp(-ax^2)$ is decreasing on $(0, +\infty)$ thus, we have,

$$B_1 \leq \frac{4c}{\gamma\sqrt{n}} (\log M)^{-\frac{2\kappa-1}{\kappa}} \exp\left(-\frac{\gamma^2}{4c}(\log(M))^{\frac{4\kappa-2}{\kappa}}\right).$$

The mapping $x \longmapsto (2/a)\exp(-ax)$ is decreasing on $(0, +\infty)$, for any $a > 0$ and $u_j \geq \gamma(\log M)^2 n^{-\frac{\kappa}{2\kappa-1}}$ thus,

$$B_2 \leq \frac{8}{3n} \exp\left(-\frac{3\gamma}{4} n^{\frac{\kappa-1}{2\kappa-1}} (\log M)^2\right).$$

Since $\gamma = (\log M)^{-1/4}$, we have $\mathbb{E}(W_n) \leq 4n^{-\frac{\kappa}{2\kappa-1}}(\log M)^{7/4} + T_1 + T_2$, where

$$T_1 = \frac{4Mc}{\sqrt{n}} (\log M)^{-\frac{7\kappa-4}{4\kappa}} \exp\left(-\frac{3}{4c}(\log M)^{\frac{7\kappa-4}{2\kappa}}\right)$$

and

$$T_2 = \frac{8M}{3n} \exp\left(-(3/4) n^{\frac{\kappa-1}{2\kappa-1}} (\log M)^{7/4}\right).$$

We have $T_2 \leq 6(\log M)^{7/4}/n$ for any integer $M \geq 1$. Moreover $\kappa/(2\kappa - 1) \leq 1$ for all $1 \leq \kappa < +\infty$, so we get $T_2 \leq 6n^{-\frac{\kappa}{2\kappa-1}}(\log M)^{7/4}$ for any integers $n \geq 1$ and $M \geq 2$.
Let B be a positive number. The inequality $T_1 \leq Bn^{-\frac{\kappa}{2\kappa-1}}(\log M)^{7/4}$ is equivalent to

$$2(2\kappa - 1)\left[\frac{3}{4c}(\log M)^{\frac{7\kappa-4}{2\kappa}} - \log M + \frac{7\kappa - 2}{2\kappa}\log(\log M)\right] \geq \log\left((4c/B)^{2(2\kappa-1)} n\right).$$

Since we have $\frac{7\kappa-4}{2\kappa} \geq \frac{3}{2} > 1$ for all $1 \leq \kappa < +\infty$ and $M \geq an^b$ for some positive numbers a and b, there exists a constant B which depends only on a, b and c (for instance $B = 4ca^{-1/(2b)}$ when n satisfies $\log(an^b) \geq (b^2(8c/6)^2) \vee ((8c/3) \vee 1)^2)$ such that $T_1 \leq Bn^{-\frac{\kappa}{2\kappa-1}}(\log M)^{7/4}$.

Proof of Theorem 7.1. Let $\gamma = (\log M)^{-1/4}$. Using (7.21), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(A(\tilde{f}_n) - A^*\right)(1-\gamma)\right] - (A(f_{j_0}) - A^*) \\
&\quad = \mathbb{E}\left[\left(A(\tilde{f}_n) - A^*\right)(1-\gamma) - \left(A_n(\tilde{f}_n) - A_n(f^*)\right)\right] + \mathbb{E}\left[A_n(\tilde{f}_n) - A_n(f_{j_0})\right] \\
&\quad \leq \mathbb{E}\left[\left(A(\tilde{f}_n) - A^*\right)(1-\gamma) - \left(A_n(\tilde{f}_n) - A_n(f^*)\right)\right] + \frac{\log M}{n}.
\end{aligned}$$

For W_n defined in the beginning of the proof of Proposition 7.2 and f^* the Bayes rule, we have

$$(1-\gamma)\left(\mathbb{E}\left[A(\tilde{f}_n)\right] - A^*\right) \leq \min_{j=1,\dots,M}(A(f_j) - A^*) + \mathbb{E}\left[W_n\right] + \frac{\log M}{n}. \tag{7.22}$$

According to Proposition 7.2, $\mathbb{E}\left[W_n\right] \leq C_0 n^{-\frac{\kappa}{2\kappa-1}} (\log M)^{7/4}$ where $C_0 > 0$ is given in Proposition 7.2. Using (7.22) and $(1-\gamma)^{-1} \leq 1 + 2\gamma$ for any $0 < \gamma < 1/2$, we get

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \left(1 + \frac{2}{\log^{1/4}(M)}\right)\left\{\min_{j=1,\dots,M}(A(f_j) - A^*) + C\frac{\log^{7/4}(M)}{n^{\kappa/(2\kappa-1)}}\right\}.$$

We complete the proof by using inequality (7.3) and equality $2(R(f) - R^*) = A(f) - A^*$, which holds for any prediction rule f.

Proof of Theorem 7.2. Since $\tilde{f}_k$'s take there values in $[-1,1]$ and $x \mapsto (1-x)_+$ is linear on $[-1,1]$, we obtain $A(\bar{f}_n) - A^* = \frac{1}{n}\sum_{k=1}^n \left(A(\tilde{f}_k) - A^*\right)$. Applying Theorem 7.1 to every $\tilde{f}_k$ for $k = 1, \dots, n$, then taking the average of the n oracle inequalities satisfied by the $\tilde{f}_k$ for $k = 1, \dots, n$ and seeing that $(1/n)\sum_{k=1}^n k^{-\kappa/(2\kappa-1)} \leq \gamma(n,\kappa)$ we obtain

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \leq \left(1 + \frac{2}{\log^{1/4}(M)}\right)\left\{\min_{j=1,\dots,M}(A(f_j) - A^*) + C\gamma(n,\kappa)\log^{7/4}(M)\right\}.$$

We complete the proof by the same argument as at the end of the previous proof.

Proof of Theorem 7.3. Let $\rho_{min} \leq \rho \leq \rho_{max}$ and $\kappa \geq 1$. Let $\rho_{m,j_0} = \min(\rho_{m,j} : \rho_{m,j} \geq \rho)$. Since $N(m) \geq A_0' m^{b'} \geq C l^{b'}$, where $C > 0$, using the oracle inequality, stated in Theorem 7.1, we have, for π satisfying (MA1)(κ),

$$\begin{aligned}
&\mathbb{E}\left[R(\tilde{f}_n) - R^* | D_m^1\right] \\
&\quad \leq \left(1 + \frac{2}{\log^{1/4} N(m)}\right)\left\{2\min_{j=1,\dots,N(m)}\left(R(\hat{f}_m^j) - R^*\right) + C\frac{\log^{7/4} N(m)}{l^{\kappa/(2\kappa-1)}}\right\},
\end{aligned}$$

where C is a positive number depending only on b', a, A_0' and c. Taking the expectation with respect to the subsample D_m^1 we have

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq \left(1 + \frac{2}{\log^{-1/4} N(m)}\right)\left\{2\mathbb{E}\left[R(\hat{f}_m^{j_0}) - R^*\right] + C\frac{\log^{7/4} N(m)}{l^{\kappa/(2\kappa-1)}}\right\}.$$

It follows from Tsybakov (2004) that, the excess risk of $\hat{f}_m^{j_0}$ satisfies

$$\sup_{\pi \in \mathcal{P}_{\kappa,\rho_{j_0}}} \mathbb{E}\left[R(\hat{f}_m^{j_0}) - R^*\right] \leq C m^{-\frac{\kappa}{2\kappa+\rho_{j_0}-1}},$$

where C is a positive number depending only on A, c, κ, ρ_{min} and ρ_{max} (note that C does not depend on ρ_{j_0}).

Moreover we have $m \geq n\,(1 - a/\log 3 - 1/3)$, $N(m) \leq A_0 m^b \leq A_0 n^b$ and $l \geq an/\log n$, so that there exists a constant C depending only on $a, A_0, A_0', b, b', \kappa, \rho_{min}$ and ρ_{max} such

that

$$\sup_{\pi\in\mathcal{P}_{\kappa,\rho_{j_0}}} \mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq C\left\{n^{-\frac{\kappa}{2\kappa+\rho_{j_0}-1}} + n^{-\frac{\kappa}{2\kappa-1}}(\log n)^{11/4}\right\}. \tag{7.23}$$

Since $\rho_{j_0} \leq \rho + N(m)^{-1} \leq \rho + (A_0')^{-1}\left[n\left(1 - a/\log 3 - 1/3\right)\right]^{-b'}$ there exists a constant C depending only on $a, A_0', b', \kappa, \rho_{min}$ and ρ_{max} such that for all integers n, $n^{-\frac{\kappa}{2\kappa+\rho_{j_0}-1}} \leq Cn^{-\frac{\kappa}{2\kappa+\rho-1}}$. Theorem 7.2 follows directly from (7.23) seeing that $\rho \geq \rho_{min} > 0$ and $\mathcal{P}_{\kappa,\rho} \subseteq \mathcal{P}_{\kappa,\rho_{j_0}}$ since $\rho_{j_0} \geq \rho$.

Proof of Theorem 7.4. Define $0 < \alpha_{min} < \alpha_{max} < +\infty$ and $0 < \beta_{min} < 1$ such that $K \subset [\alpha_{min}, \alpha_{max}] \times [\beta_{min}, 1]$. Let $(\alpha_0, \beta_0) \in K$. We consider the function on $(0, +\infty) \times (0, 1]$ with values in $(1/2, 2)$, $\phi(\alpha, \beta) = 4(\alpha+1)/((2\alpha + p\alpha + 4)(1+\beta))$. We take $k_0 \in \{0, \ldots, \lfloor 3\Delta/2 \rfloor - 1\}$ such that

$$\phi_{l,k_0} = 1/2 + k_0\Delta^{-1} \leq \phi(\alpha_0, \beta_0) < 1/2 + (k_0+1)\Delta^{-1}.$$

For n greater than a constant depending only on K, p, b_0 and a there exists $\bar{\alpha}_0 \in [\alpha_{min}/2, \alpha_{max}]$ such that $\phi(\bar{\alpha}_0, \beta_0) = \phi_{l,k_0}$. Since $\alpha \mapsto \phi(\alpha, \beta_0)$ increases on $\mathbb{R}^+$, we have $\bar{\alpha}_0 \leq \alpha_0$. Moreover, we have $|\phi(\alpha_1, \beta_0) - \phi(\alpha_2, \beta_0)| \geq A|\alpha_1 - \alpha_2|$, $\forall \alpha_1, \alpha_2 \in [\alpha_{min}/2, \alpha_{max}]$, where $A > 0$ depends only on p and α_{max}. Thus $|\bar{\alpha}_0 - \alpha_0| \leq (A\Delta)^{-1}$. Since $\bar{\alpha}_0 \leq \alpha_0$ we have $\mathcal{Q}_{\alpha_0,\beta_0} \subseteq \mathcal{Q}_{\bar{\alpha}_0,\beta_0}$, so

$$\sup_{\pi\in\mathcal{Q}_{\alpha_0,\beta_0}} \mathbb{E}[R(\tilde{f}_n) - R^*] \leq \sup_{\pi\in\mathcal{Q}_{\bar{\alpha}_0,\beta_0}} \mathbb{E}[R(\tilde{f}_n) - R^*].$$

Since $\lceil 3\Delta/2 \rceil \geq (3/2)l^{b_0}$, for π satisfying the margin assumption (MA2)($\bar{\alpha}_0$), Theorem 7.1 leads to

$$\begin{aligned}
&\mathbb{E}\left[R(\tilde{f}_n) - R^* | D_m^1\right] \\
&\quad\leq \left(1 + \frac{2}{\log^{1/4}(\lceil 3\Delta/2 \rceil)}\right)\left\{2\min_{\lambda\in\mathcal{G}(l)}\left(R(\hat{f}_m^\lambda) - R^*\right) + C_0\frac{\log^{7/4}(\lceil 3\Delta/2 \rceil)}{l^{(\bar{\alpha}_0+1)/(\bar{\alpha}_0+2)}}\right\},
\end{aligned}$$

for all integers $n \geq 1$, where $C_0 > 0$ depends only on K, a and b_0. Therefore, taking the expectation w.r.t. the subsample D_m^1 we get

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq C_1\left(\mathbb{E}\left[R(\hat{f}_m^{\lambda_{l,k_0}}) - R^*\right] + l^{-\frac{\bar{\alpha}_0+1}{\bar{\alpha}_0+2}}\log^{7/4}(n)\right),$$

where $\lambda_{l,k_0} = l^{-\phi_{l,k_0}}$ and $C_1 > 0$ depends only on K, a and b_0.

Set $\Gamma : (0, +\infty) \times (0, 1] \longmapsto \mathbb{R}^+$ defined by $\Gamma(\alpha, \beta) = \beta\phi(\alpha, \beta), \forall (\alpha, \beta) \in (0, +\infty) \times (0, 1]$. According to Scovel et al. (2005), if $\pi \in \mathcal{Q}_{\bar{\alpha}_0,\beta_0}$ then for all $\epsilon > 0$, there exists $C > 0$ a constant depending only on K, p and ϵ such that,

$$\mathbb{E}\left[R(\hat{f}_m^{\lambda_{l,k_0}}) - R^*\right] \leq Cm^{-\Gamma(\bar{\alpha}_0,\beta_0)+\epsilon}.$$

Remark that C does not depend on $\bar{\alpha}_0$ and β_0 since $(\bar{\alpha}_0, \beta_0) \in [\alpha_{min}/2, \alpha_{max}] \times [\beta_{min}, 1]$ and that the constant multiplying the rate of convergence, stated in Scovel et al. (2005), is uniformly bounded over (α, β) belonging to a compact subset of $(0, +\infty) \times (0, 1]$.

Let $\epsilon > 0$. Assume that $\pi \in \mathcal{Q}_{\alpha_0,\beta_0}$. We have $n\left(1 - a/\log 3 - 1/3\right) \leq m \leq n$, $l \geq an/\log n$ and $\Gamma(\bar{\alpha}_0, \beta_0) \leq (\bar{\alpha}_0 + 1)/(\bar{\alpha}_0 + 2) \leq 1$, therefore, there exist $C_2, C_2' > 0$ depending only on a, b_0, K, p and ϵ such that for any n greater than a constant depending

only on β_{min}, a and b_0

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq C_2\left(n^{-\Gamma(\bar{\alpha}_0,\beta_0)+\epsilon} + n^{-\frac{\bar{\alpha}_0+1}{\bar{\alpha}_0+2}}(\log n)^{11/4}\right) \leq C_2' n^{-\Gamma(\bar{\alpha}_0,\beta_0)+\epsilon}.$$

Moreover, Γ satisfies $|\Gamma(\bar{\alpha}_0,\beta_0) - \Gamma(\alpha_0,\beta_0)| \leq B\Delta^{-1}$, where B depends only on p and α_{min}, and $\left(n^{B\Delta^{-1}}\right)_{n\in\mathbb{N}}$ is upper bounded. This completes the proof.

Proof of Theorem 7.5. Let $(\alpha_0,\gamma_0) \in K \cup K'$. First assume that (α_0,γ_0) belongs to $K \subset \mathcal{U}$. We consider the set

$$\mathcal{S} = \{(\varphi,\psi) \in (0,1/2)\times(1/2,1) : 2 - 2\psi - \varphi > 0\}.$$

Each point of $\mathcal{S}$ is associated to a margin parameter (7.14) and to a geometric noise exponent by the following functions on $\mathcal{S}$ with values in $(0,+\infty)$,

$$\bar{\alpha}(\varphi,\psi) = \frac{4\psi - 2}{2 - 2\psi - \varphi} \text{ and } \bar{\gamma}(\varphi,\psi) = \frac{\psi}{\varphi} - 1.$$

We take $(\varphi,\psi) \in \mathcal{S} \cap \mathcal{M}(l)$ such that $\bar{\alpha}(\varphi,\psi) \leq \alpha_0$, $\bar{\gamma}(\varphi,\psi) \leq \gamma_0$, $\bar{\alpha}(\varphi,\psi)$ is close enough to α_0, $\bar{\gamma}(\varphi,\psi)$ is close enough to γ_0 and $\bar{\gamma}(\varphi,\psi) > (\bar{\alpha}(\varphi,\psi)+2)/(2\bar{\alpha}(\varphi,\psi))$. Since $\gamma_0 > (\alpha_0+2)/(2\alpha_0)$ there exists a solution $(\varphi_0,\psi_0) \in \mathcal{S}$ of the system of equations

$$\left\{\begin{array}{rcl} \bar{\alpha}(\varphi,\psi) & = & \alpha_0 \\ \bar{\gamma}(\varphi,\psi) & = & \gamma_0. \end{array}\right. \tag{7.24}$$

For all integers n greater than a constant depending only on K, a and b_0, there exists $(p_{1,0}, p_{2,0}) \in \{1,\dots,2\lfloor\Delta\rfloor\} \times \{2,\dots,\lfloor\Delta/2\rfloor\}$ defined by

$$\varphi_{l,p_{1,0}} = \min(\varphi_{l,p} : \varphi_{l,p} \geq \varphi_0) \text{ and } \psi_{l,p_{2,0}} = \max(\psi_{l,p_2} : \psi_{l,p_2} \leq \psi_0) - \Delta^{-1}.$$

We have $2 - 2\psi_{l,p_{2,0}} - \varphi_{l,p_{1,0}} > 0$. Therefore $(\varphi_{l,p_{1,0}}, \psi_{l,p_{2,0}}) \in \mathcal{S} \cap \mathcal{M}(l)$. Define $\bar{\alpha}_0 = \bar{\alpha}(\varphi_{l,p_{1,0}}, \psi_{l,p_{2,0}})$ and $\bar{\gamma}_0 = \bar{\gamma}(\varphi_{l,p_{1,0}}, \psi_{l,p_{2,0}})$. Since (φ_0,ψ_0) satisfies (7.24), we have

$$\psi_{l,p_{2,0}} + \frac{1}{\Delta} \leq \psi_0 = \frac{-\alpha_0}{2\alpha_0+4}\varphi_0 + \frac{1+\alpha_0}{2+\alpha_0} \leq \frac{-\alpha_0}{2\alpha_0+4}\left(\varphi_{l,p_{1,0}} - \frac{1}{2\Delta}\right) + \frac{1+\alpha_0}{2+\alpha_0}$$

and $(\alpha_0/(2\alpha_0+4))(2\Delta)^{-1} \leq \Delta^{-1}$, thus

$$\psi_{l,p_{2,0}} \leq -\frac{\alpha_0}{2\alpha_0+4}\varphi_{l,p_{1,0}} + \frac{1+\alpha_0}{2+\alpha_0} \text{ so } \bar{\alpha}_0 \leq \alpha_0.$$

With a similar argument, we have $\psi_{l,p_{2,0}} \leq (\alpha_0+1)\varphi_{l,p_{1,0}}$, that is $\bar{\gamma}_0 \leq \gamma_0$. Now we show that $\bar{\gamma}_0 > (\bar{\alpha}_0+2)/(2\bar{\alpha}_0)$. Since (α_0,γ_0) belongs to a compact, (φ_0,ψ_0) and $(\varphi_{l,p_{1,0}}, \psi_{l,p_{2,0}})$ belong to a compact subset of $(0,1/2)\times(1/2,1)$ for n greater than a constant depending only on K, a, b_0. Thus, there exists $A > 0$, depending only on K, such that for n large enough, we have

$$|\alpha_0 - \bar{\alpha}_0| \leq A\Delta^{-1} \text{ and } |\gamma_0 - \bar{\gamma}_0| \leq A\Delta^{-1}.$$

Denote by $d_K = d(\partial\mathcal{U}, K)$, where $\partial\mathcal{U}$ is the boundary of $\mathcal{U}$ and $d(A,B)$ denotes the Euclidean distance between sets A and B. We have $d_K > 0$ since K is a compact, $\partial\mathcal{U}$ is closed and $K \cap \partial\mathcal{U} = \emptyset$. Set $0 < \alpha_{min} < \alpha_{max} < +\infty$ and $0 < \gamma_{min} < \gamma_{max} < +\infty$ such that $K \subset [\alpha_{min},\alpha_{max}] \times [\gamma_{min},\gamma_{max}]$. Define

$$\mathcal{U}_\mu = \left\{(\alpha,\gamma) \in (0,+\infty)^2 : \alpha \geq 2\mu \text{ and } \gamma > (\alpha - \mu + 2)/(2(\alpha-\mu))\right\}$$

for $\mu = \min(\alpha_{min}/2, d_K)$. We have $K \subset \mathcal{U}_\mu$ so $\gamma_0 > (\alpha_0 - \mu + 2)/(2(\alpha_0 - \mu))$. Since $\alpha \mapsto (\alpha+2)/(2\alpha)$ is decreasing, $\bar{\gamma}_0 > \gamma_0 - A\Delta^{-1}$ and $\alpha_0 \leq \bar{\alpha}_0 + A\Delta^{-1}$, we have $\bar{\gamma}_0 > \bar{\beta}(\bar{\alpha}_0) - A\Delta^{-1}$ where $\bar{\beta}$ is a positive function on $(0, 2\alpha_{max}]$ defined by $\bar{\beta}(\alpha) = (\alpha - (\mu -$

$A\Delta^{-1})+2)/(2(\alpha-(\mu-A\Delta^{-1})))$. We have $|\bar{\beta}(\alpha_1)-\bar{\beta}(\alpha_2)|\geq(2\alpha_{max})^{-2}|\alpha_1-\alpha_2|$ for all $\alpha_1,\alpha_2\in(0,2\alpha_{max}]$. Therefore $\bar{\beta}(\bar{\alpha}_0)-A\Delta^{-1}\geq\bar{\beta}\left(\bar{\alpha}_0+4A\alpha_{max}^2\Delta^{-1}\right)$. Thus, for n greater than a constant depending only on K,a and b_0 we have $\bar{\gamma}_0>(\bar{\alpha}_0+2)/(2\bar{\alpha}_0)$.

Since $\bar{\alpha}_0\leq\alpha_0$ and $\bar{\gamma}_0\leq\gamma_0$, we have $\mathcal{R}_{\alpha_0,\gamma_0}\subset\mathcal{R}_{\bar{\alpha}_0,\bar{\gamma}_0}$ and

$$\sup_{\pi\in\mathcal{R}_{\alpha_0,\gamma_0}}\mathbb{E}\left[R(\tilde{f}_n)-R^*\right]\leq\sup_{\pi\in\mathcal{R}_{\bar{\alpha}_0,\bar{\gamma}_0}}\mathbb{E}\left[R(\tilde{f}_n)-R^*\right].$$

If π satisfies (MA2)$(\bar{\alpha}_0)$ then we get from Theorem 7.1

$$\mathbb{E}\left[R(\tilde{f}_n)-R^*|D_m^1\right]\leq \tag{7.25}$$

$$\left(1+\frac{2}{\log^{1/4}M(l)}\right)\left\{2\min_{(\sigma,\lambda)\in\mathcal{N}(l)}\left(R(\hat{f}_m^{(\sigma,\lambda)})-R^*\right)+C_2\frac{\log^{7/4}(M(l))}{l^{(\bar{\alpha}_0+1)/(\bar{\alpha}_0+2)}}\right\},$$

for all integers $n\geq1$, where $C_2>0$ depends only on K,a and b_0 and $M(l)$ is the cardinality of $\mathcal{N}(m)$. Remark that $M(l)\geq l^{2b_0}/2$, so we can apply Theorem 7.1.

Let $\epsilon>0$. Since $M(l)\leq n^{2b_0}$ and $\bar{\gamma}_0>(\bar{\alpha}_0+2)/(2\bar{\alpha}_0)$, taking expectations in (7.25) and using the result (7.18) of Scovel et al. (2004), for $\sigma=\sigma_{l,\varphi_{l,p_{1,0}}}$ and $\lambda=\lambda_{l,\psi_{l,p_{2,0}}}$, we obtain

$$\sup_{\pi\in\mathcal{R}_{\bar{\alpha}_0,\bar{\gamma}_0}}\mathbb{E}\left[R(\tilde{f}_n)-R^*\right]\leq C\left(m^{-\Theta(\bar{\alpha}_0,\bar{\gamma}_0)+\epsilon}+l^{-\frac{\bar{\alpha}_0+1}{\bar{\alpha}_0+2}}\log^{7/4}(n)\right),$$

where $\Theta:\mathcal{U}\mapsto\mathbb{R}$ is defined, for all $(\alpha,\gamma)\in\mathcal{U}$, by $\Theta(\alpha,\gamma)=(2\gamma(\alpha+1))/(2\gamma(\alpha+2)+3\alpha+4)$ and $C>0$ depends only on a,b_0,K and ϵ. Remark that the constant before the rate of convergence in (7.18) is uniformly bounded on every compact of $\mathcal{U}$. We have $\Theta(\bar{\alpha}_0,\bar{\gamma}_0)\leq\Theta(\alpha_0,\gamma_0)\leq\Theta(\bar{\alpha}_0,\bar{\gamma}_0)+2A\Delta^{-1}$, $m\geq n\,(1-a/\log 3-1/3)$ and $\left(m^{2A\Delta^{-1}}\right)_{n\in\mathbb{N}}$ is upper bounded, so there exists $C_1>0$ depending only on K,a,b_0 such that $m^{-\Theta(\bar{\alpha}_0,\bar{\gamma}_0)}\leq C_1n^{-\Theta(\alpha_0,\gamma_0)},\quad\forall n\geq1$.

Similar argument as at the end of the proof of Theorem 7.4 and the fact that $\Theta(\alpha,\gamma)<(\alpha+1)/(\alpha+2)$ for all $(\alpha,\gamma)\in\mathcal{U}$, leads to the result of the first part of Theorem 7.5.

Let now $(\alpha_0,\gamma_0)\in K'$. Let $\alpha'_{max}>0$ be such that $\forall(\alpha,\gamma)\in K',\alpha\leq\alpha'_{max}$. Take $p_{1,0}\in\{1,\ldots,2\lfloor\Delta\rfloor\}$ such that $\varphi_{l,p_{1,0}}=\min(\varphi_{l,p}:\varphi_{l,p}\geq(2\gamma_0+1)^{-1}$ and $p\in4\mathbb{N})$, where $4\mathbb{N}$ is the set of all integers multiple of 4. For large values of n, $p_{1,0}$ exists and $p_{1,0}\in4\mathbb{N}$. We denote by $\bar{\gamma}_0\in(0,+\infty)$ such that $\varphi_{l,p_{1,0}}=(2\bar{\gamma}_0+1)^{-1}$, we have $\bar{\gamma}_0\leq\gamma_0$ thus $\mathcal{R}_{\alpha_0,\gamma_0}\subseteq\mathcal{R}_{\alpha_0,\bar{\gamma}_0}$ and

$$\sup_{\pi\in\mathcal{R}_{\alpha_0,\gamma_0}}\mathbb{E}\left[R(\tilde{f}_n)-R^*\right]\leq\sup_{\pi\in\mathcal{R}_{\alpha_0,\bar{\gamma}_0}}\mathbb{E}\left[R(\tilde{f}_n)-R^*\right].$$

If π satisfies the margin assumption (7.14) with the margin parameter α_0 then, using Theorem 7.1, we obtain, for any integer $n\geq1$,

$$\mathbb{E}\left[R(\tilde{f}_n)-R^*|D_m^1\right]\leq \tag{7.26}$$

$$\left(1+\frac{2}{\log^{1/4}(M(l))}\right)\left\{2\min_{(\sigma,\lambda)\in\mathcal{N}(l)}\left(R(\hat{f}_m^{(\sigma,\lambda)})-R^*\right)+C_0\frac{\log^{7/4}M(l)}{l^{(\alpha_0+1)/(\alpha_0+2)}}\right\}$$

where $C>0$ appears in Proposition 7.2 and $M(l)$ is the cardinality of $\mathcal{N}(l)$.

Let $\epsilon>0$ and $p_{2,0}\in\{1,\ldots,\lfloor\Delta/2\rfloor\}$ defined by $p_{2,0}=p_{1,0}/4$ (note that $p_{1,0}\in4\mathbb{N}$). We have

$$\sigma_{l,\varphi_{l,p_{1,0}}}=\left(\lambda_{l,\psi_{l,p_{2,0}}}\right)^{-\frac{1}{d_0(\bar{\gamma}_0+1)}}.$$

Since $\bar{\gamma}_0 \leq (\alpha_0 + 2)/(2\alpha_0)$, using the result (7.18) of Scovel et al. (2004) we have, for $\sigma = \sigma_{l,\varphi_{l,p_{1,0}}}$ and $\lambda = \lambda_{l,\psi_{l,p_{2,0}}}$,

$$\mathbb{E}\left[R(\hat{f}_m^{(\sigma_0,\lambda_0)}) - R^*\right] \leq Cm^{-\bar{\Gamma}(\bar{\gamma}_0)+\epsilon},$$

where $\bar{\Gamma} : (0, +\infty) \longmapsto \mathbb{R}$ is the function defined by $\bar{\Gamma}(\gamma) = \gamma/(2\gamma + 1)$ for all $\gamma \in (0, +\infty)$ and $C > 0$ depends only on a, b_0, K' and ϵ. Remark that, as in the first part of the proof, we can uniformly bound the constant before the rate of convergence in (7.18) on every compact subset of $\mathcal{U}'$. Since $M(l) \leq n^{2b_0}$, taking the expectation, in (7.26), we find

$$\sup_{\pi \in \mathcal{R}_{\alpha_0,\bar{\gamma}_0}} \mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq C\left(m^{-\Gamma(\bar{\gamma}_0)+\epsilon} + l^{-\frac{\alpha_0+1}{\alpha_0+2}} \log^{7/4}(n)\right),$$

where $C > 0$ depends only on a, b_0, K' and ϵ. Moreover $|\gamma_0 - \bar{\gamma}_0| \leq 2(2\alpha'_{max} + 1)^2\Delta^{-1}$ so $|\bar{\Gamma}(\bar{\gamma}_0) - \bar{\Gamma}(\gamma_0)| \leq 2(2\alpha_{max} + 1)\Delta^{-1}$. To achieve the proof we use same argument as for the first part of the proof.

CHAPTER 8

Optimal Oracle Inequality for Aggregation of Classifiers under Low Noise Condition

We consider the problem of optimality, in a minimax sense, and adaptivity to the margin and to regularity in binary classification. We prove an oracle inequality, under the margin assumption (low noise condition), satisfied by an aggregation procedure which uses exponential weights. This oracle inequality has an optimal residual: $(\log M/n)^{\kappa/(2\kappa-1)}$ where κ is the margin parameter, M the number of classifiers to aggregate and n the number of observations. We use this inequality first to construct minimax classifiers under margin and regularity assumptions and second to aggregate them to obtain a classifier which is adaptive both to the margin and regularity. Moreover, by aggregating plug-in classifiers (only $\log n$), we provide an easily implementable classifier adaptive both to the margin and to regularity.

Contents

The material of this chapter has been published in *COLT06* (cf. [83]).

1. Introduction

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We consider a random variable (X, Y) with values in $\mathcal{X} \times \{-1, 1\}$ and denote by π the distribution of (X, Y). We denote by P^X the marginal of π on $\mathcal{X}$ and $\eta(x) = \mathbb{P}(Y = 1|X = x)$ the conditional probability function of $Y = 1$ given that $X = x$. We denote by $D_n = (X_i, Y_i)_{i=1,\dots,n}$, n i.i.d. observations of the couple (X, Y).

We recall some usual notions introduced for the classification framework. A *prediction rule* is a measurable function $f : \mathcal{X} \longmapsto \{-1, 1\}$. The *misclassification error* associated to f is

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

It is well known (see, e.g., [47]) that $\min_f R(f) = R(f^*) \stackrel{\text{def}}{=} R^*$, where the prediction rule f^* is called *Bayes rule* and is defined by

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

The minimal risk R^* is called the *Bayes risk*. A *classifier* is a function, $\hat{f}_n = \hat{f}_n(X, D_n)$, measurable with respect to D_n and X with values in $\{-1, 1\}$, that assigns to the sample D_n a prediction rule $\hat{f}_n(., D_n) : \mathcal{X} \longmapsto \{-1, 1\}$. A key characteristic of $\hat{f}_n$ is the value of

generalization error $\mathbb{E}[R(\hat{f}_n)]$. Here

$$R(\hat{f}_n) = \mathbb{P}(Y \neq \hat{f}_n(X)|D_n).$$

The performance of a classifier $\hat{f}_n$ is measured by the value $\mathbb{E}[R(\hat{f}_n) - R^*]$ called the *excess risk* of $\hat{f}_n$. We say that the classifier $\hat{f}_n$ learns with the convergence rate $\phi(n)$, where $(\phi(n))_{n\in\mathbb{N}}$ is a decreasing sequence, if there exists an absolute constant $C > 0$ such that for any integer n, $\mathbb{E}[R(\hat{f}_n) - R^*] \leq C\phi(n)$. Theorem 7.2 of [47] shows that no classifier can learn with a given convergence rate for arbitrary underlying probability distribution π.

In this chapter, we focus on entropy assumptions which allow us to work with finite sieves. Hence, we first work with a finite model for f^*: it means that we take a finite class of prediction rules $\mathcal{F} = \{f_1, \ldots, f_M\}$. Our aim is to construct a classifier $\hat{f}_n$ which mimics the best one of them w.r.t. to the excess risk and with an optimal residual. Namely, we want to state an oracle inequality

$$\mathbb{E}\left[R(\hat{f}_n) - R^*\right] \leq a_0 \min_{f\in\mathcal{F}}(R(f) - R^*) + C\gamma(M, n), \tag{8.1}$$

where $a_0 \geq 1$ and $C > 0$ are some absolute constants and $\gamma(M, n)$ is the residual. The classical procedure, due to Vapnik and Chervonenkis (see, e.g. [47]), is to look for an ERM classifier,i.e., the one which minimizes the *empirical risk*

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i f(X_i)\leq 0\}}, \tag{8.2}$$

over all prediction rules f in $\mathcal{F}$, where $\mathbb{1}_E$ denotes the indicator of the set E. This procedure leads to optimal theoretical results (see, e.g. Chapter 12 of [47]), but minimizing the empirical risk (8.2) is computationally intractable for sets $\mathcal{F}$ of classifiers with large cardinality (often depending on the sample size n), because this risk is neither convex nor continuous. Nevertheless, we might base a tractable estimation procedure on minimization of a convex surrogate ϕ for the loss ([89], [25], [22], [20], [109] and [108]). A wide variety of classification methods in machine learning are based on this idea, in particular, on using the convex loss associated to support vector machines ([41], [104]),

$$\phi(x) = \max(0, 1 - x),$$

called the *hinge-loss*. The risk associated to this loss is called the *hinge risk* and is defined by

$$A(f) = \mathbb{E}[\max(0, 1 - Yf(X))],$$

for all $f : \mathcal{X} \longmapsto \mathbb{R}$. The *optimal hinge risk* is defined by

$$A^* = \inf_f A(f), \tag{8.3}$$

where the infimum is taken over all measurable functions f. The Bayes rule f^* attains the infimum in (8.3) and, moreover, denoting by $R(f)$ the misclassification error of $\text{sign}(f)$ for all measurable functions f with values in $\mathbb{R}$, Zhang, cf. [130], has shown that,

$$R(f) - R^* \leq A(f) - A^*, \tag{8.4}$$

for any real valued measurable function f. Thus, minimization of the *excess hinge risk* $A(f) - A^*$ provides a reasonable alternative for minimization of the excess risk. In this chapter, we provide a procedure which does not need any minimization step. We use a convex combination of the given prediction rules, as explained in section 2.

The difficulty of classification is closely related to the behavior of the conditional probability function η near $1/2$ (the random variable $|\eta(X) - 1/2|$ is sometimes called the theoretical margin). Tsybakov has introduced, in [116], an assumption on the the margin, called *margin (or low noise) assumption*,

(MA) Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the margin assumption MA(κ) with margin parameter $1 \leq \kappa < +\infty$ if there exists $c_0 > 0$ such that,*

$$\mathbb{E}\left\{|f(X) - f^*(X)|\right\} \leq c_0 \left(R(f) - R^*\right)^{1/\kappa}, \tag{8.5}$$

for all measurable functions f with values in $\{-1, 1\}$.

Under this assumption, the risk of an ERM classifier over some fixed class $\mathcal{F}$ can converge to the minimum risk over the class with *fast rates*, namely faster than $n^{-1/2}$ (cf. [116]). On the other hand, with no margin assumption on the joint distribution π (but combinatorial or complexity assumption on the class $\mathcal{F}$), the convergence rate of the excess risk is not faster than $n^{-1/2}$ (cf. [47]).

In this chapter, we suggest an easily implementable procedure of aggregation of classifiers and prove the following results:

(1) We obtain an oracle inequality for our procedure and we use it to show that our classifiers are adaptive both to the margin parameter (low noise exponent) and to a complexity parameter.
(2) We generalize the lower bound inequality stated in Chapter 14 of [47], by introducing the margin assumption and deduce optimal rates of aggregation under low noise assumption in the spirit of Tsybakov [114].
(3) We obtain classifiers with minimax fast rates of convergence on a Hölder class of conditional probability functions η and under the margin assumption.

The chapter is organized as follows. In Section 2 we prove an oracle inequality for our convex aggregate, with an optimal residual, which will be used in Section 3 to construct minimax classifiers and to obtain adaptive classifiers by aggregation of them. Proofs are given in Section 4.

2. Oracle Inequality

We have M prediction rules $f_1, \ldots, f_M$. We want to mimic the best of them according to the excess risk under the margin assumption. Our procedure is using exponential weights. Similar constructions in other context can be found, e.g., in [10], [125], [35], [87], [119] and Chapter 7. Consider the following aggregate which is a convex combination with exponential weights of M classifiers,

$$\tilde{f}_n = \sum_{j=1}^{M} w_j^{(n)} f_j, \tag{8.6}$$

where

$$w_j^{(n)} = \frac{\exp\left(\sum_{i=1}^n Y_i f_j(X_i)\right)}{\sum_{k=1}^M \exp\left(\sum_{i=1}^n Y_i f_k(X_i)\right)}, \quad \forall j = 1, \ldots, M. \tag{8.7}$$

Since $f_1, \ldots, f_M$ take their values in $\{-1, 1\}$, we have,

$$w_j^{(n)} = \frac{\exp\left(-nA_n(f_j)\right)}{\sum_{k=1}^M \exp\left(-nA_n(f_k)\right)}, \tag{8.8}$$

for all $j \in \{1, \dots, M\}$, where

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - Y_i f(X_i)) \tag{8.9}$$

is the empirical analog of the hinge risk. Since $A_n(f_j) = 2R_n(f_j)$ for all $j = 1, \dots, M$, these weights can be written in terms of the empirical risks of f_j's,

$$w_j^{(n)} = \frac{\exp\left(-2nR_n(f_j)\right)}{\sum_{k=1}^{M} \exp\left(-2nR_n(f_k)\right)}, \ \forall j = 1, \dots, M.$$

Remark that, using the definition (8.8) for the weights, we can aggregate functions with values in $\mathbb{R}$ (like in theorem 8.1) and not only functions with values in $\{-1, 1\}$.

The aggregation procedure defined by (8.6) with weights (8.8), that we can called aggregation with exponential weights (AEW), can be compared to the ERM one. First, our AEW method does not need any minimization algorithm contrarily to the ERM procedure. Second, the AEW is less sensitive to the over fitting problem. Intuitively, if the classifier with smallest empirical risk is over fitted (it means that the classifier fits too much to the observations) then the ERM procedure will be over fitted. But, if other classifiers in $\mathcal{F}$ are good classifiers, our procedure will consider their "opinions" in the final decision procedure and these opinions can balance with the opinion of the over fitted classifier in $\mathcal{F}$ which can be false because of its over fitting property. The ERM only considers the "opinion" of the classifier with the smallest risk, whereas the AEW takes into account all the opinions of the classifiers in the set $\mathcal{F}$. The AEW is more temperate contrarily to the ERM. Finally, the following proposition shows that the AEW has similar theoretical property as the ERM procedure up to the residual $(\log M)/n$.

PROPOSITION 8.1. *Let $M \geq 2$ be an integer, $f_1, \dots, f_M$ be M real valued functions on $\mathcal{X}$. For any integers n, the aggregate defined in (8.6) with weights (8.8) $\tilde{f}_n$ satisfies*

$$A_n(\tilde{f}_n) \leq \min_{i=1,\dots,M} A_n(f_i) + \frac{\log(M)}{n}.$$

The following theorem provides first an exact oracle inequality w.r.t. the hinge risk satisfied by the AEW procedure and second shows its optimality among all aggregation procedures. We deduce from it that, for a margin parameter $\kappa \geq 1$ and a set of M functions with values in $[-1, 1]$, $\mathcal{F} = \{f_1, \dots, f_M\}$,

$$\gamma(\mathcal{F}, \pi, n, \kappa) = \sqrt{\frac{\min_{f \in \mathcal{F}} (A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}$$

is an optimal rate of convex aggregation of M functions with values in $[-1, 1]$ w.r.t. the hinge risk, in the sense of Chapter 3.

THEOREM 8.1 (**Oracle inequality and Lower bound**). *Let $\kappa \geq 1$. We assume that π satisfies MA(κ). We denote by $\mathcal{C}$ the convex hull of a finite set of functions with values in $[-1, 1]$, $\mathcal{F} = \{f_1, \dots, f_M\}$. The AEW procedure, introduced in (8.6) with weights (8.8) (remark that the form of the weights in (8.8) allows to take real valued functions for the f_j's), satisfies for any integer $n \geq 1$ the following inequality*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{C}} (A(f) - A^*) + C_0 \gamma(\mathcal{F}, \pi, n, \kappa),$$

where $C_0 > 0$ depends only on the constants κ and c_0 appearing in MA(κ).

Moreover, there exists a set of prediction rules $\mathcal{F} = \{f_1, \dots, f_M\}$ such that for any procedure $\bar{f}_n$ with values in $\mathbb{R}$, there exists a probability measure π satisfying MA(κ) such that for any integers M, n with $\log M \leq n$ we have

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f \in \mathcal{C}}(A(f) - A^*) + C_0' \gamma(\mathcal{F}, \pi, n, \kappa),$$

where $C_0' > 0$ depends only on the constants κ and c_0 appearing in MA(κ).

The hinge loss is linear on $[-1, 1]$, thus, model selection aggregation or convex aggregation are identical problems if we use the hinge risk and if we aggregate function with values in $[-1, 1]$. Namely, $\min_{f \in \mathcal{F}} A(f) = \min_{f \in \mathcal{C}} A(f)$. Moreover, the result of Theorem 8.1 is obtained for the aggregation of functions with values in $[-1, 1]$ and not only for prediction rules. In fact, only functions with values in $[-1, 1]$ have to be considered when we use the hinge loss since, for any real valued function f, we have $\max(0, 1 - y\psi(f(x))) \leq \max(0, 1 - yf(x))$ for all $x \in \mathcal{X}, y \in \{-1, 1\}$ where ψ is the projection on $[-1, 1]$, thus, $A(\psi(f)) - A^* \leq A(f) - A^*$. Remark that, under MA($\kappa$), there exists $c > 0$ such that,$\mathbb{E}\left[|f(X) - f^*(X)|\right] \leq c\left(A(f) - A^*\right)^{1/\kappa}$for all functions f on $\mathcal{X}$ with values in $[-1, 1]$ (cf. Chapter 3) . The proof of Theorem 8.1 is not given here by the lack of space. It can be found in Chapter 3. Instead, we prove here the following slightly less general result that we will be further used to construct adaptive minimax classifiers.

THEOREM 8.2. *Let $\kappa \geq 1$ and let $\mathcal{F} = \{f_1, \dots, f_M\}$ be a finite set of prediction rules with $M \geq 3$. We denote by $\mathcal{C}$ the convex hull of $\mathcal{F}$. We assume that π satisfies MA(κ). The aggregate defined in (8.6) with the exponential weights (8.7) (or (8.8)) satisfies for any integers n, M and any $a > 0$ the following inequality*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq (1 + a) \min_{f \in \mathcal{C}}(A(f) - A^*) + C\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}},$$

where $C > 0$ is a constant depending only on a.

COROLLARY 8.1. *Let $\kappa \geq 1$, $M \geq 3$ and $\{f_1, \dots, f_M\}$ be a finite set of prediction rules. We assume that π satisfies MA(κ). The AEW procedure satisfies for any number $a > 0$ and any integers n, M the following inequality, with $C > 0$ a constant depending only on a,*

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq 2(1 + a) \min_{j = 1, \dots, M}(R(f_j) - R^*) + C\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}.$$

We denote by $\mathcal{P}_\kappa$ the set of all probability measures on $\mathcal{X} \times \{-1, 1\}$ satisfying the margin assumption MA(κ). Combining Corollary 8.1 and the following theorem, we get that the residual

$$\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}$$

is a near optimal rate of model selection aggregation in the sense of Chapter 3 when the underlying probability measure π belongs to $\mathcal{P}_\kappa$.

THEOREM 8.3. *For any integers M and n satisfying $M \leq \exp(n)$, there exists M prediction rules $f_1, \dots, f_M$ such that for any classifier $\hat{f}_n$ and any $a > 0$, we have*

$$\sup_{\pi \in \mathcal{P}_\kappa}\left[\mathbb{E}\left[R(\hat{f}_n) - R^*\right] - 2(1 + a) \min_{j = 1, \dots, M}(R(f_j) - R^*)\right] \geq C_1\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}},$$

where $C_1 = c_0^\kappa / (4e 2^{2\kappa(\kappa - 1)/(2\kappa - 1)} (\log 2)^{\kappa/(2\kappa - 1)})$.

3. Adaptivity Both to the Margin and to Regularity.

In this section we give two applications of the oracle inequality stated in Corollary 8.1. First, we construct classifiers with minimax rates of convergence and second, we obtain adaptive classifiers by aggregating the minimax ones. Following [9], we focus on the regularity model where η belongs to the Hölder class.

For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we define $|s| = \sum_{j=1}^d s_i$, $s! = s_1! \ldots s_d!$, $x^s = x_1^{s_1} \ldots x_d^{s_d}$ and $||x|| = (x_1^2 + \ldots + x_d^2)^{1/2}$. We denote by D^s the differential operator $\frac{\partial^{s_1+\ldots+s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$.

Let $\beta > 0$. We denote by $\lfloor \beta \rfloor$ the maximal integer that is strictly less than β. For any $x \in (0,1)^d$ and any $\lfloor \beta \rfloor$-times continuously differentiable real valued function g on $(0,1)^d$, we denote by g_x its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x, namely,

$$g_x(y) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(y-x)^s}{s!} D^s g(x).$$

For all $L > 0$ and $\beta > 0$. The $(\beta, L, [0,1]^d)-$*Hölder class* of functions, denoted by $\Sigma(\beta, L, [0,1]^d)$, is the set of all real valued functions g on $[0,1]^d$ that are $\lfloor \beta \rfloor$-times continuously differentiable on $(0,1)^d$ and satisfy, for any $x, y \in (0,1)^d$, the inequality

$$|g(y) - g_x(y)| \leq L||x-y||^\beta.$$

A control of the complexity of Hölder classes is given by Kolmogorov and Tikhomorov (1961):

$$\mathcal{N}\left(\Sigma(\beta, L, [0,1]^d), \epsilon, L^\infty([0,1]^d)\right) \leq A(\beta, d)\epsilon^{-\frac{d}{\beta}}, \forall \epsilon > 0, \tag{8.10}$$

where the LHS is the $\epsilon-$entropy of the $(\beta, L, [0,1]^d)-$Hölder class w.r.t. to the norm in $L^\infty([0,1]^d)-$and $A(\beta, d)$ is a constant depending only on β and d.

If we want to use entropy assumptions on the set which η belongs to, we need to make a link between P^X and the Lebesgue measure, since the distance in (8.10) is the $L^\infty-$norm w.r.t. the Lebesgue measure. Therefore, introduce the following assumption:
(A1) *The marginal distribution P^X on $\mathcal{X}$ of π is absolutely continuous w.r.t. the Lebesgue measure λ_d on $[0,1]^d$, and there exists a version of its density which is upper bounded by $\mu_{max} < \infty$.*

We consider the following class of models. For all $\kappa \geq 1$ and $\beta > 0$, we denote by $\mathcal{P}_{\kappa,\beta}$, the set of all probability measures π on $\mathcal{X} \times \{-1, 1\}$, such that

(1) MA(κ) is satisfied.
(2) The marginal P^X satisfies (A1).
(3) The conditional probability function η belongs to $\Sigma(\beta, L, \mathbb{R}^d)$.

Now, we define the class of classifiers which attain the optimal rate of convergence, in a minimax sense, over the models $\mathcal{P}_{\kappa,\beta}$. Let $\kappa \geq 1$ and $\beta > 0$. For any $\epsilon > 0$, we denote by $\Sigma_\epsilon(\beta)$ an ϵ-net on $\Sigma(\beta, L, [0,1]^d)$ for the $L^\infty-$norm, such that, its cardinal satisfies $\log \mathrm{Card}\left(\Sigma_\epsilon(\beta)\right) \leq A(\beta, d)\epsilon^{-d/\beta}$. We consider the AEW procedure defined in (8.6), over the net $\Sigma_\epsilon(\beta)$:

$$\tilde{f}_n^{(\epsilon,\beta)} = \sum_{\eta \in \Sigma_\epsilon(\beta)} w^{(n)}(f_\eta) f_\eta, \text{ where } f_\eta(x) = 2\mathbb{1}_{(\eta(x) \geq 1/2)} - 1. \tag{8.11}$$

THEOREM 8.4. *Let $\kappa > 1$ and $\beta > 0$. Let $a_1 > 0$ be an absolute constant and consider $\epsilon_n = a_1 n^{-\frac{\beta(\kappa-1)}{\beta(2\kappa-1)+d(\kappa-1)}}$. The aggregate (8.11) with $\epsilon = \epsilon_n$, satisfies, for any $\pi \in \mathcal{P}_{\kappa,\beta}$ and*

any integer $n \geq 1$, the following inequality

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^{(\epsilon_n,\beta)}) - R^* \right] \leq C_2(\kappa,\beta,d) n^{-\frac{\beta\kappa}{\beta(2\kappa-1)+d(\kappa-1)}},$$

where $C_2(\kappa,\beta,d) = 2\max\left(4(2c_0\mu_{max})^{\kappa/(\kappa-1)}, CA(\beta,d)^{\frac{\kappa}{2\kappa-1}}\right)(a_1)^{\frac{\kappa}{\kappa-1}} \vee (a_1)^{-\frac{d\kappa}{\beta(\kappa-1)}}$ *and C is the constant appearing in Corollary 8.1.*

Audibert and Tsybakov (cf. [9]) have shown the optimality, in a minimax sense, of the rate obtained in theorem 8.4. Note that this rate is a fast rate because it can approach $1/n$ when κ is close to 1 and β is large.

The construction of the classifier $\tilde{f}_n^{(\epsilon_n,\beta)}$ needs the knowledge of κ and β which are not available in practice. Thus, we need to construct classifiers independent of these parameters and which learn with the optimal rate $n^{-\beta\kappa/(\beta(2\kappa-1)+d(\kappa-1))}$ if the underlying probability measure π belongs to $\mathcal{P}_{\kappa,\beta}$, for different values of κ and β. We now show that using the procedure (8.6) to aggregate the classifiers $\tilde{f}_n^{(\epsilon,\beta)}$, for different values of (ϵ,β) in a grid, the oracle inequality of Corollary 8.1 provides the result.

We use a split of the sample for the adaptation step. Denote by $D_m^{(1)}$ the subsample containing the first m observations and $D_l^{(2)}$ the one containing the $l(=n-m)$ last ones. Subsample $D_m^{(1)}$ is used to construct classifiers $\tilde{f}_m^{(\epsilon,\beta)}$ for different values of (ϵ,β) in a finite grid. Subsample $D_l^{(2)}$ is used to aggregate these classifiers by the procedure (8.6). We take

$$l = \left\lceil \frac{n}{\log n} \right\rceil \quad \text{and} \quad m = n - l.$$

Set $\Delta = \log n$. We consider a grid of values for (ϵ,β):

$$\mathcal{G}(n) = \left\{ (\epsilon_k,\beta_p) = (m^{-\phi_k}, \frac{p}{\Delta}) : \phi_k = \frac{k}{\Delta} : k \in \{1,\dots,\lfloor \Delta/2 \rfloor\}, p \in \{1,\dots,\lceil \Delta \rceil^2\} \right\}.$$

The classifier that we propose is the sign of

$$\tilde{f}_n^{adp} = \sum_{(\epsilon,\beta)\in\mathcal{G}(n)} w^{[l]}(\tilde{F}_m^{(\epsilon,\beta)}) \tilde{F}_m^{(\epsilon,\beta)},$$

where $\tilde{F}_m^{(\epsilon,\beta)} = \text{sign}(\tilde{f}_m^{(\epsilon,\beta)})$ is the classifier associated to the aggregate $\tilde{f}_m^{(\epsilon,\beta)}$ for any $\epsilon,\beta > 0$ and weights $w^{[l]}(F)$ are the ones introduced in (8.7) constructed with the observations $D_l^{(2)}$ for any $F \in \mathcal{F}(n) = \{\text{sign}(\tilde{f}_m^{(\epsilon,\beta)}) : (\epsilon,\beta) \in \mathcal{G}(n)\}$:

$$w^{[l]}(F) = \frac{\exp\left(\sum_{i=m+1}^n Y_i F(X_i)\right)}{\sum_{G\in\mathcal{F}(n)} \exp\left(\sum_{i=m+1}^n Y_i G(X_i)\right)}.$$

The following Theorem shows that $\tilde{f}_n^{adp}$ is adaptive both to the low noise exponent κ and to the complexity (or regularity) parameter β, provided that (κ,β) belongs to a compact subset of $(1,+\infty)\times(0,+\infty)$.

THEOREM 8.5. *Let K be a compact subset of $(1,+\infty)\times(0,+\infty)$. There exists a constant $C_3 > 0$ that depends only on K and d such that for any integer $n \geq 1$, any $(\kappa,\beta) \in K$ and any $\pi \in \mathcal{P}_{\kappa,\beta}$, we have,*

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^{adp}) - R^* \right] \leq C_3 n^{-\frac{\kappa\beta}{\beta(2\kappa-1)+d(\kappa-1)}}.$$

Classifiers $\tilde{f}_n^{(\epsilon_n,\beta)}$, for ϵ_n given in Theorem 8.4 and $\beta > 0$, are not easily implementable since the cardinality of $\Sigma_{\epsilon_n}(\beta)$ is an exponential of n. An alternative procedure which

is easily implementable is to aggregate plug-in classifiers constructed in Audibert and Tsybakov (cf. [9]).

We introduce the class of models $\mathcal{P}'_{\kappa,\beta}$ composed of all the underlying probability measures π such that:

(1) π satisfies the margin assumption MA(κ).
(2) The conditional probability function $\eta \in \Sigma(\beta, L, [0,1]^d)$.
(3) The marginal distribution of X is supported on $[0,1]^d$ and has a Lebesgue density lower bounded and upper bounded by two constants.

THEOREM 8.6 (**Audibert and Tsybakov (2005)**). *Let $\kappa > 1, \beta > 0$. The excess risk of the plug-in classifier $\hat{f}_n^{(\beta)} = 2\mathbb{1}_{\{\hat{\eta}_n^{(\beta)} \geq 1/2\}} - 1$ satisfies*

$$\sup_{\pi \in \mathcal{P}'_{\kappa,\beta}} \mathbb{E}\left[R(\hat{f}_n^{(\beta)}) - R^*\right] \leq C_4 n^{-\frac{\beta\kappa}{(\kappa-1)(2\beta+d)}},$$

where $\hat{\eta}_n^{(\beta)}(\cdot)$ is the locally polynomial estimator of $\eta(\cdot)$ of order $\lfloor\beta\rfloor$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ and C_4 a positive constant.

In [9], it is shown that the rate $n^{-\frac{\beta\kappa}{(\kappa-1)(2\beta+d)}}$ is minimax over $\mathcal{P}'_{\kappa,\beta}$, if $\beta \leq d(\kappa-1)$. Remark that the fast rate n^{-1} can be achieved.

We aggregate classifiers $\hat{f}_n^{(\beta)}$ for different values of β lying in a finite grid. Contrarily to the previous example of adaptation, we only need to consider a grid for β since $\hat{f}_n^{(\beta)}$ is already adaptive to κ the margin parameter. We use a split of the sample to construct our adaptive classifier: $l = \lceil n/\log n \rceil$ and $m = n - l$. The training sample $D_m^1 = ((X_1, Y_1), \ldots, (X_m, Y_m))$ is used for the construction of the class of plug-in classifiers

$$\mathcal{F} = \left\{ \hat{f}_m^{(\beta_k)} : \beta_k = \frac{kd}{\Delta - 2k}, k \in \{1, \ldots, \lfloor \Delta/2 \rfloor\} \right\}, \text{ where } \Delta = \log n.$$

The validation sample $D_l^2 = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n))$ is used for the construction of weights

$$w^{[l]}(f) = \frac{\exp\left(\sum_{i=m+1}^n Y_i f(X_i)\right)}{\sum_{\bar{f} \in \mathcal{F}} \exp\left(\sum_{i=m+1}^n Y_i \bar{f}(X_i)\right)}, \quad \forall f \in \mathcal{F}.$$

The classifier that we propose is $\tilde{F}_n^{adp} = \text{sign}(\tilde{f}_n^{adp})$, where $\tilde{f}_n^{adp} = \sum_{f \in \mathcal{F}} w^{[l]}(f) f$.

THEOREM 8.7. *Let K be a compact subset of $(1, +\infty) \times (0, +\infty)$. There exists a constant $C_5 > 0$ depending only on K and d such that for any integer $n \geq 1$, any $(\kappa, \beta) \in K$, such that $\beta < d(\kappa - 1)$, and any $\pi \in \mathcal{P}'_{\kappa,\beta}$, we have,*

$$\mathbb{E}_\pi\left[R(\tilde{F}_n^{adp}) - R^*\right] \leq C_5 n^{-\frac{\beta\kappa}{(\kappa-1)(2\beta+d)}}.$$

Adaptive classifiers are obtained in Theorem (8.5) and (8.7) by aggregation of only $\log n$ classifiers. Other construction of adaptive classifiers can be found in Chapter 7. In particular, adaptive SVM classifiers.

4. Proofs

Proof of Proposition 8.1. Using the convexity of the hinge loss, we have $A_n(\tilde{f}_n) \leq \sum_{j=1}^M w_j A_n(f_j)$. Denote by $\hat{i} = \arg\min_{i=1,\ldots,M} A_n(f_i)$, we have

$$A_n(f_i) = A_n(f_{\hat{i}}) + \frac{1}{n}\left(\log(w_{\hat{i}}) - \log(w_i)\right)$$

for all $i = 1, \dots, M$ and by averaging over the w_i we get :

$$A_n(\tilde{f}_n) \leq \min_{i=1,\dots,M} A_n(f_i) + \frac{\log(M)}{n}, \tag{8.12}$$

where we used that $\sum_{j=1}^{M} w_j \log\left(\frac{w_j}{1/M}\right) \geq 0$ since it is the Kullback-leibler divergence between the weights $w = (w_j)_{j=1,\dots,M}$ and uniform weights $u = (M^{-1})_{j=1,\dots,M}$.

Proof of Theorem 8.2. Let $a > 0$. Using Proposition 8.1, we have for any $f \in \mathcal{F}$ and for the Bayes rule f^*:

$$A(\tilde{f}_n) - A^* = (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) + A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))$$
$$\leq (1+a)(A_n(f) - A_n(f^*)) + (1+a)\frac{\log M}{n} + A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)).$$

Taking the expectations, we get

$$\begin{aligned}\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] &\leq (1+a)\min_{f\in\mathcal{F}}(A(f) - A^*) + (1+a)(\log M)/n \\ &\quad + \mathbb{E}\left[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))\right].\end{aligned}$$

The following inequality follows from the linearity of the hinge loss on $[-1, 1]$:

$$A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \leq \max_{f\in\mathcal{F}} \left[A(f) - A^* - (1+a)(A_n(f) - A_n(f^*))\right].$$

Thus, using Bernstein's inequality, we have for all $0 < \delta < 4 + 2a$:

$$\begin{aligned}&\mathbb{P}\left[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \geq \delta\right] \\ &\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left[A(f) - A^* - (A_n(f) - A_n(f^*)) \geq \frac{\delta + a(A(f) - A^*)}{1+a}\right] \\ &\leq \sum_{f\in\mathcal{F}} \exp\left(-\frac{n(\delta + a(A(f) - A^*))^2}{2(1+a)^2(A(f) - A^*)^{1/\kappa} + 2/3(1+a)(\delta + a(A(f) - A^*))}\right).\end{aligned}$$

There exists a constant $c_1 > 0$ depending only on a such that for all $0 < \delta < 4 + 2a$ and all $f \in \mathcal{F}$, we have

$$\frac{(\delta + a(A(f) - A^*))^2}{2(1+a)^2(A(f) - A^*)^{1/\kappa} + 2/3(1+a)(\delta + a(A(f) - A^*))} \geq c_1\delta^{2-1/\kappa}.$$

Thus, $\mathbb{P}\left[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \geq \delta\right] \leq M\exp(-nc_1\delta^{2-1/\kappa})$.

Observe that an integration by parts leads to $\int_a^{+\infty} \exp\left(-bt^\alpha\right) dt \leq \frac{\exp(-ba^\alpha)}{\alpha b a^{\alpha-1}}$, for any $\alpha \geq 1$ and $a, b > 0$, so for all $u > 0$, we get

$$\mathbb{E}\left[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))\right] \leq 2u + M\frac{\exp(-nc_1u^{2-1/\kappa})}{nc_1u^{1-1/\kappa}}.$$

If we denote by $\mu(M)$ the unique solution of $X = M\exp(-X)$, we have $\log M/2 \leq \mu(M) \leq \log M$. For u such that $nc_1u^{2-1/\kappa} = \mu(M)$, we obtain the result.

Proof of Corollary 8.1. We deduce Corollary 8.1 from Theorem 8.2, using that for any prediction rule f we have $A(f) - A^* = 2(R(f) - R^*)$ and applying Zhang's inequality $A(g) - A^* \geq (R(g) - R^*)$ fulfilled by all g from $\mathcal{X}$ to $\mathbb{R}$.

Proof of Theorem 8.3. For all prediction rules $f_1, \dots, f_M$, we have

$$\sup_{f_1,\dots,f_M} \inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}_\kappa} \left(\mathbb{E}\left[R(\hat{f}_n) - R^*\right] - 2(1+a)\min_{j=1,\dots,M}(R(f_j) - R^*)\right)$$

$$\geq \inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\kappa : f^* \in \{f_1, \ldots, f_M\}} \left(\mathbb{E}\left[R(\hat{f}_n) - R^* \right] \right).$$

Thus, we look for a set of cardinality not greater than M, of the worst probability measures $\pi \in \mathcal{P}_\kappa$ from our classification problem point of view and choose $f_1, \ldots, f_M$ as the corresponding Bayes rules.

Let N be an integer such that $2^{N-1} \leq M$. Let $x_1, \ldots, x_N$ be N distinct points of $\mathcal{X}$. Let $0 < w < 1/N$. Denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\}) = w$ for $j = 1, \ldots, N-1$ and $P^X(\{x_N\}) = 1-(N-1)w$. We consider the set of binary sequences $\Omega = \{-1, 1\}^{N-1}$. Let $0 < h < 1$. For all $\sigma \in \Omega$ we consider

$$\eta_\sigma(x) = \begin{cases} (1 + \sigma_j h)/2 & \text{if } x = x_1, \ldots, x_{N-1}, \\ 1 & \text{if } x = x_N. \end{cases}$$

For all $\sigma \in \Omega$ we denote by π_σ the probability measure on $\mathcal{X} \times \{-1, 1\}$ with the marginal P^X on $\mathcal{X}$ and with the conditional probability function η_σ of $Y = 1$ knowing X.

Assume that $\kappa > 1$. We have $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) = (N-1)w \mathbb{1}_{\{h \leq t\}}, \forall 0 \leq t < 1$. Thus, if we assume that $(N-1)w \leq h^{1/(\kappa-1)}$ then $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) \leq t^{1/(\kappa-1)}$, for all $t \geq 0$, and according to [116], π_σ belongs to MA(κ).

We denote by ρ the Hamming distance on Ω (cf. [115] p.88). Let σ, σ' be such that $\rho(\sigma, \sigma') = 1$. We have

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\left(1 - (1 - w(1 - \sqrt{1 - h^2}))^n\right).$$

We take w and h such that $w(1-\sqrt{1-h^2}) \leq 1/n$, thus, $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) \leq \beta = 2(1-e^{-1}) < 2$ for any integer n.

Let $\hat{f}_n$ be a classifier and $\sigma \in \Omega$. Using MA(κ), we have

$$\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^* \right] \geq (c_0 w)^\kappa \mathbb{E}_{\pi_\sigma}\left[\left(\sum_{i=1}^{N-1} |\hat{f}_n(x_i) - \sigma_i| \right)^\kappa \right].$$

By Jensen's Lemma and Assouad's Lemma (cf. [115]) we obtain:

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\kappa : f^* \in \{f_\sigma : \sigma \in \Omega\}} \left(\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^* \right] \right) \geq (c_0 w)^\kappa \left(\frac{N-1}{4} (1 - \beta/2)^2 \right)^\kappa.$$

We obtain the result by taking $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$ and

$$h = \left(\frac{1}{n} \left\lceil \frac{\log M}{\log 2} \right\rceil \right)^{\frac{\kappa-1}{2\kappa-1}}.$$

For $\kappa = 1$, we take $h = 1/2$, thus $|2\eta_\sigma(X) - 1| \geq 1/2$ a.s. so $\pi_\sigma \in$MA(1) (cf.[116]). Putting $w = 4/n$ and $N = \lceil \log M / \log 2 \rceil$, we obtain the result.

Proof of Theorem 8.4. According to Theorem 8.1, where we set $a = 1$, we have, for any $\epsilon > 0$:

$$\mathbb{E}_\pi\left[R(\tilde{f}_n^\epsilon) - R^* \right] \leq 4 \min_{\bar{\eta} \in \Sigma_\epsilon(\beta)} \left(R(f_{\bar{\eta}}) - R^* \right) + C \left(\frac{\log \mathrm{Card}\Sigma_\epsilon(\beta)}{n} \right)^{\frac{\kappa}{2\kappa-1}}.$$

Let $\bar{\eta}$ be a function with values in $[0, 1]$ and denote by $\bar{f} = \mathbb{1}_{\bar{\eta} \geq 1/2}$ the plug-in classifier associated. We have $|2\eta - 1| \mathbb{1}_{\bar{f} \neq f^*} \leq 2|\bar{\eta} - \eta|$, thus:

$$R(\bar{f}) - R^* = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{\bar{f} \neq f^*} \right] = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{\bar{f} \neq f^*} \mathbb{1}_{\bar{f} \neq f^*} \right]$$

$$\leq \left|\left| |2\eta - 1| \mathbb{1}_{\bar{f} \neq f^*} \right|\right|_{L^\infty(P^X)} \mathbb{E}\left[\mathbb{1}_{\bar{f} \neq f^*} \right] \leq \left|\left| |2\eta - 1| \mathbb{1}_{\bar{f} \neq f^*} \right|\right|_{L^\infty(P^X)} c_0 \left(R(\bar{f}) - R^* \right)^{\frac{1}{\kappa}},$$

and assumption (A1) lead to

$$R(f_{\bar{\eta}}) - R^* \leq (2c_0\mu_{max})^{\frac{\kappa}{\kappa-1}} ||\bar{\eta} - \eta||_{L^\infty([0,1]^d)}^{\frac{\kappa}{\kappa-1}}.$$

Hence, for any $\epsilon > 0$, we have

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^\epsilon) - R^* \right] \leq D \left(\epsilon^{\frac{\kappa}{\kappa-1}} + \left(\frac{\epsilon^{-d/\beta}}{n} \right)^{\frac{\kappa}{2\kappa-1}} \right),$$

where $D = \max \left(4(2c_0\mu_{max})^{\kappa/(\kappa-1)}, CA(\beta,d)^{\frac{\kappa}{2\kappa-1}} \right)$. For the value

$$\epsilon_n = a_1 n^{-\frac{\beta(\kappa-1)}{\beta(2\kappa-1)+d(\kappa-1)}},$$

we have

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^{\epsilon_n}) - R^* \right] \leq C_1 n^{-\frac{\beta\kappa}{\beta(2\kappa-1)+d(\kappa-1)}},$$

where $C_1 = 2D(a_1)^{\frac{\kappa}{\kappa-1}} \vee (a_1)^{-\frac{d\kappa}{\beta(\kappa-1)}}$

Proof of Theorem 8.5. Consider the following function on $(1,+\infty) \times (0,+\infty)$ with values in $(0,1/2)$:

$$\phi(\kappa,\beta) = \frac{\beta(\kappa-1)}{\beta(2\kappa-1)+d(\kappa-1)}.$$

For any n greater than $n_1 = n_1(K)$, we have $\Delta^{-1} \leq \phi(\kappa,\beta) \leq \lfloor \Delta/2 \rfloor \Delta^{-1}$ and $\Delta^{-1} \leq \beta \leq \Delta$ for all $(\kappa,\beta) \in K$.

Let $(\kappa_0,\beta_0) \in K$. For any $n \geq n_1$, there exists $k_0 \in \{1,\dots,\lfloor \Delta/2 \rfloor - 1\}$ such that $\phi_{k_0} = k_0\Delta^{-1} \leq \phi(\kappa_0,\beta_0) < (k_0+1)\Delta^{-1}$ and $p_0 \in \{1,\dots,\lceil \Delta \rceil^2 - 1\}$ such that $\beta_{p_0} = p_0\Delta^{-1} \leq \beta_0 < (p_0+1)\Delta^{-1}$. Denote by $f_{\beta_{p_0}}(\cdot)$ the increasing function $\phi(\cdot,\beta_{p_0})$ from $(1,+\infty)$ to $(0,1/2)$ and set

$$\kappa_{0,n} = \left(f_{\beta_{p_0}} \right)^{-1} (\phi_{k_0}).$$

There exists $m = m(K)$ such that $m|\kappa_0 - \kappa_{0,n}| \leq |f_{\beta_{p_0}}(\kappa_0) - f_{\beta_{p_0}}(\kappa_{0,n})| \leq \Delta^{-1}$.

Let $\pi \in \mathcal{P}_{\kappa_0,\beta_0}$. According to the oracle inequality of Corollary 8.1, we have, conditionally to the first subsample D_m^1:

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^{adp}) - R^* | D_m^1 \right] \leq 4 \min_{(\epsilon,\beta)\in\mathcal{G}(n)} \left(R(\tilde{f}_m^{(\epsilon,\beta)}) - R^* \right) + C \left(\frac{\log \mathrm{Card}(\mathcal{G}(n))}{l} \right)^{\frac{\kappa_0}{2\kappa_0-1}}.$$

Using the definition of l and $\mathrm{Card}(\mathcal{G}(n)) \leq (\log n)^3$, there exists $\tilde{C} > 0$ independent of n such that for $\epsilon_m^0 = \epsilon_m^{-\phi_{k_0}}$

$$\mathbb{E}_\pi \left[R(\tilde{f}_n^{adp}) - R^* \right] \leq \tilde{C} \left(\mathbb{E}_\pi \left[R(\tilde{f}_m^{(\epsilon_m^0,\beta_{p_0})}) - R^* \right] + \left(\frac{\log^2 n}{n} \right)^{\frac{\kappa_0}{2\kappa_0-1}} \right).$$

Moreover $\beta_{p_0} \leq \beta_0$ and there exists a constant A, depending only on K, such that $\kappa_0 \leq \kappa_{0,n} + A\Delta^{-1} = \kappa'_{0,n}$, hence, $\mathcal{P}_{\kappa_0,\beta_0} \subseteq \mathcal{P}_{\kappa'_{0,n},\beta_{p_0}}$ and ϵ_m^0 is equal to $m^{-\Theta(\kappa'_{0,n},\beta_0)}$ up to a multiplying constant. Thus $\pi \in \mathcal{P}_{\kappa'_{0,n},\beta_{p_0}}$ and, according to Theorem 8.4, we have

$$\mathbb{E}_\pi \left[R(\tilde{f}_m^{(\epsilon_m^0,\beta_0)}) - R^* \right] \leq C_1(K,d) m^{-\psi(\kappa'_{0,n},\beta_{p_0})},$$

where $C_1(K,d) = \max \left(C_1(\kappa,\beta,d) : (\kappa,\beta) \in K \right)$ and $\psi(\kappa,\beta) = \frac{\beta\kappa}{\beta(2\kappa-1)+d(\kappa-1)}$. By construction, there exists $A_2 = A_2(K,d) > 0$ such that $|\psi(\kappa'_{0,n},\beta_{p_0}) - \psi(\kappa_0,\beta_0)| \leq A_2\Delta^{-1}$. Moreover for any integer n we have $n^{A_2/\log n} = \exp(A_2)$, which is a constant. We conclude

that

$$\mathbb{E}_\pi\left[R(\tilde{f}_n^{adp}) - R^*\right] \leq C_2(K,d)\left(n^{-\psi(\kappa_0,\beta_0)} + \left(\frac{\log^4 n}{n}\right)^{\frac{\kappa_0}{2\kappa_0-1}}\right),$$

where $C_2(K,d) > 0$ is independent of n. We achieve the proof by observing that $\psi(\kappa_0,\beta_0) < \frac{\kappa_0}{2\kappa_0-1}$.

Proof of Theorem 8.7. We consider the following function on $(1,+\infty)\times(0,+\infty)$ with values in $(0,1/2)$:

$$\Theta(\kappa,\beta) = \frac{\beta\kappa}{(\kappa-1)(2\beta+d)}.$$

For any n greater than $n_1 = n_1(K)$, we have $\min(\kappa/(\kappa-1) : (\kappa,\beta)\in K)\Delta^{-1} \leq \Theta(\kappa,\beta) \leq \lfloor\Delta/2\rfloor\,\Delta^{-1}\max(\kappa/(\kappa-1) : (\kappa,\beta)\in K)$, for all $(\kappa,\beta)\in K$.

Let $(\kappa_0,\beta_0)\in K$ be such that $\beta_0 < (\kappa_0-1)d$. For any $n \geq n_1$, there exists $k_0 \in \{1,\ldots,\lfloor\Delta/2\rfloor - 1\}$ such that

$$\frac{\kappa_0}{\kappa_0-1}k_0\Delta^{-1} \leq \Theta(\kappa_0,\beta_0) < \frac{\kappa_0}{\kappa_0-1}(k_0+1)\Delta^{-1}.$$

Let $\pi\in\mathcal{P}_{\kappa_0,\beta_0}$. According to the oracle inequality of Corollary 8.1, we have, conditionally to the first subsample D_m^1:

$$\mathbb{E}_\pi\left[R(\tilde{F}_n^{adp}) - R^*|D_m^1\right] \leq 4\min_{f\in\mathcal{F}}(R(f)-R^*) + C\left(\frac{\log \mathrm{Card}(\mathcal{F})}{l}\right)^{\frac{\kappa_0}{2\kappa_0-1}}.$$

Using the proof of Theorem 8.5 we get that there exists $\tilde{C} > 0$ independent of n such that

$$\mathbb{E}_\pi\left[R(\tilde{f}_n^{adp}) - R^*\right] \leq \tilde{C}\left(\mathbb{E}_\pi\left[R(\hat{f}_m^{(\beta_{k_0})}) - R^*\right] + \left(\frac{\log^2 n}{n}\right)^{\frac{\kappa_0}{2\kappa_0-1}}\right)$$

Moreover $\beta_{k_0} \leq \beta_0$, hence, $\mathcal{P}_{\kappa_0,\beta_0} \subseteq \mathcal{P}_{\kappa_0,\beta_{k_0}}$. Thus, according to Theorem 8.6, we have

$$\mathbb{E}_\pi\left[R(\hat{F}_m^{(\beta_{k_0})}) - R^*\right] \leq C_4(K,d)m^{-\Theta(\kappa_0,\beta_{k_0})},$$

where $C_4(K,d) = \max\left(C_4(\kappa,\beta,d) : (\kappa,\beta)\in K\right)$. We have $|\Theta(\kappa_0,\beta_{k_0}) - \Theta(\kappa_0,\beta_0)| \leq \Delta^{-1}$ by construction. Moreover $n^{1/\log n} = e$ for any integer n. We conclude that

$$\mathbb{E}_\pi\left[R(\tilde{F}_n^{adp}) - R^*\right] \leq \tilde{C}_4(K,d)\left(n^{-\Theta(\kappa_0,\beta_0)} + \left(\frac{\log^2 n}{n}\right)^{\frac{\kappa_0}{2\kappa_0-1}}\right),$$

where $\tilde{C}_4(K,d) > 0$ is independent of n. We achieve the proof by observing that $\Theta(\kappa_0,\beta_0) < \frac{\kappa_0}{2\kappa_0-1}$, if $\beta_0 < (\kappa_0-1)d$.

CHAPTER 9

Adapting to unknown smoothness by aggregation of thresholded Wavelet Estimators

We study the performances of an adaptive procedure based on a convex combination, with data-driven weights, of term-by-term thresholded wavelet estimators. For the bounded regression model, with random uniform design, and the nonparametric density model, we show that the resulting estimator is optimal in the minimax sense over all Besov balls $B^s_{p,q}$ for $s > 1/p$ under the L^2 risk, without any logarithm factor.

Contents

The material of this chapter is a joint work with Christophe Chesneau submitted for publication (cf. [38]).

1. Introduction

Wavelet shrinkage methods have been very successful in nonparametric function estimation. They provide estimators that are spatially adaptive and (near) optimal over a wide range of function classes. Standard approaches are based on the term-by-term thresholds. A well-known example is the hard thresholded estimator introduced by [50]. If we observe n statistical data and if the unknown function f has an expansion of the form $f = \sum_j \sum_k \beta_{j,k}\psi_{j,k}$ where $\{\psi_{j,k},\ j,k\}$ is a wavelet basis and $(\beta_{j,k})_{j,k}$ is the associated wavelet coefficients, then the term-by-term wavelet thresholded method consists in three steps. First, a linear step corresponding to the estimation of the coefficients $\beta_{j,k}$ by some estimators $\hat{\beta}_{j,k}$ constructed from the data. Second, a non-linear step consisting in

a thresholded procedure $T_\lambda(\hat\beta_{j,k})\mathbb{1}_{\{|\hat\beta_{j,k}|\geq\lambda_j\}}$ where $\lambda=(\lambda_j)_j$ is a positive sequence and $T_\lambda(\hat\beta_{j,k})$ denotes a certain transformation of the $\hat\beta_{j,k}$ which may depend on λ. Third, a reconstruction step of the form $\hat f_\lambda=\sum_{j\in\Omega_n}\sum_k T_\lambda(\hat\beta_{j,k})\mathbb{1}_{\{|\hat\beta_{j,k}|\geq\lambda_j\}}\psi_{j,k}$ where Ω_n is a finite set of integers depending on the number n of data. Naturally, the performances of $\hat f_\lambda$ strongly depend on the choice of the threshold λ. For the standard statistical models (regression, density,...), the most common choice is the universal threshold introduced by [50]. It can be expressed in the form: $\lambda^*=(\lambda_j^*)_j$ where $\lambda_j^*=c\sqrt{(\log n)/n}$ where $c>0$ denotes a large enough constant. In the literature, several techniques have been proposed to determine the 'best' adaptive threshold. There are, for instance, the RiskShrink and SureShrink methods (see [49, 50]), the cross-validation methods (see [97], [121] and [69]), the methods based on hypothesis tests (see [1] and [2]), the Lepski methods (see [71]) and the Bayesian methods (see [40] and [3]). Most of them are described in detail in [97] and [4].

In the present chapter, we propose to study the performances of an adaptive wavelet estimator based on a convex combination of $\hat f_\lambda$'s. In the framework of nonparametric density estimation and bounded regression estimation with random uniform design, we prove that, in some sense, it is at least as good as the term-by-term thresholded estimator $\hat f_\lambda$ defined with the 'best' threshold λ. In particular, we show that this estimator is optimal, in the minimax sense, over all Besov balls under the L^2 risk. The proof is based on a non-adaptive minimax result proved by [46] and some powerful oracle inequality satisfied by aggregation methods.

The exact oracle inequality of Section 2 is given in a general framework. Two aggregation procedures satisfy this oracle inequality. The well known ERM (for Empirical Risk Minimization) procedure (cf. [117], [77] and references therein) and an exponential weighting aggregation scheme, which has been studied, among others, by [87], [26] and in the others chapters of this part. There is a recursive version of this scheme studied by [35], [125], [72] and [75]. In the sequential prediction problem, weighted average predictions with exponential weights have been widely studied (cf. e.g. [119] and [37]). A result in Chapter 4 shows that the ERM procedure is suboptimal for strictly convex losses (which is the case for density and regression estimation when the integrated squared risk is used). Thus, in our case it is better to combine the $\hat f_\lambda$'s, for λ lying in a grid, using the aggregation procedure with exponential weights than using the ERM procedure. Moreover, from a computation point of view the aggregation scheme with exponential weights does not require any minimization step contrarily to the ERM procedure.

The chapter is organized as follows. Section 2 presents general oracle inequalities satisfied by two aggregation methods. Section 3 describes the main procedure of the study and investigates its minimax performances over Besov balls for the L^2 risk. All the proofs are postponed in the last section.

2. Oracle Inequalities

2.1. Framework. Let $(\mathcal{Z},\mathcal{T})$ a measurable space. Denote by $\mathcal{P}$ the set of all probability measures on $(\mathcal{Z},\mathcal{T})$. Let F be a function from $\mathcal{P}$ with values in an algebra $\mathcal{F}$. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability measure. Let D_n be a family of n i.i.d. observations $Z_1,\dots,Z_n$ having the common probability measure π. The probability measure π is unknown. Our aim is to estimate $F(\pi)$ from the observations D_n.

In our estimation problem, we assume that we have access to an "empirical risk". It means that there exists $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$ such that the risk of an estimate $f \in \mathcal{F}$ of $F(\pi)$ is of the form

$$A(f) \stackrel{\text{def}}{=} \mathbb{E}\left[Q(Z, f)\right].$$

In what follows, we present several statistical problems which can be written in this way. If the minimum over all f in $\mathcal{F}$

$$A^* \stackrel{\text{def}}{=} \min_{f \in \mathcal{F}} A(f)$$

is achieved by at least one function, we denote by f^* a minimizer in $\mathcal{F}$. In this chapter we will assume that $\min_{f\in\mathcal{F}} A(f)$ is achievable, otherwise we replace f^* by f_n^*, an element in $\mathcal{F}$ satisfying $A(f_n^*) \leq \inf_{f\in\mathcal{F}} A(f) + n^{-1}$.

In most of the cases f^* will be equal to our aim $F(\pi)$ up to some known additive terms. We don't know the risk A, since π is not available from the statistician, thus, instead of minimizing A over $\mathcal{F}$ we consider an empirical version of A constructed from the observations D_n. The main interest of such a framework is that we have access to an empirical version of $A(f)$ for any $f \in \mathcal{F}$. It is denoted by

$$A_n(f) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} Q(Z_i, f). \tag{9.1}$$

We exhibit three statistical models having the previous form of estimation.

Bounded Regression: Take $\mathcal{Z} = \mathcal{X} \times [0, 1]$, where $(\mathcal{X}, \mathcal{A})$ is a measurable space, $Z = (X, Y)$ a couple of random variables on $\mathcal{Z}$, with probability distribution π, such that X takes its values in $\mathcal{X}$ and Y takes its values in $[0, 1]$. We assume that the conditional expectation $\mathbb{E}[Y|X]$ exists. In the regression framework, we want to estimate the regression function

$$f^*(x) = \mathbb{E}\left[Y|X = x\right], \ \forall x \in \mathcal{X}.$$

Usually, the variable Y is not an exact function of X. Given an input $X \in \mathcal{X}$, we are not able to predict the exact value of the output $Y \in [0, 1]$. This issue can be seen in the regression framework as a noised estimation. It means that at each spot X of the input set, the predicted label Y is concentrated around $\mathbb{E}\left[Y|X\right]$ up to an additional noise with null mean denoted by ζ. The regression model can then be written as

$$Y = \mathbb{E}\left[Y|X\right] + \zeta.$$

Take $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $[0, 1]$. Define $||f||^2_{L^2(P^X)} = \int_{\mathcal{X}} f^2 dP^X$ for all functions f in $L^2(\mathcal{X}, \mathcal{A}, P^X)$ where P^X is the probability measure of X. Consider

$$Q((x, y), f) = (y - f(x))^2, \tag{9.2}$$

for any $(x, y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$. Pythagore's Theorem yields

$$A(f) = \mathbb{E}\left[Q((X, Y), f)\right] = ||f^* - f||^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right].$$

Thus f^* is a minimizer of $A(f)$ and $A^* = \mathbb{E}[\zeta^2]$.

Density estimation: Let $(\mathcal{Z}, \mathcal{T}, \mu)$ be a measured space where μ is a finite measure. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability distribution. We assume that π is absolutely continuous w.r.t. to μ and denote by f^* one version of the density. Consider $\mathcal{F}$ the set of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$. We consider

$$Q(z, f) = -\log f(z),$$

for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have

$$A(f) = \mathbb{E}\left[Q(Z,f)\right] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z))d\pi(z).$$

Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} \log(f^*(z))d\pi(z)$.

Instead of using the Kullback-Leibler loss, one can use the quadratic loss. For this setup, consider $\mathcal{F}$ the set $L^2(\mathcal{Z}, \mathcal{T}, \mu)$ of all measurable functions with an integrated square. Define

$$Q(z,f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z), \tag{9.3}$$

for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have, for any $f \in \mathcal{F}$,

$$A(f) = \mathbb{E}\left[Q(Z,f)\right] = ||f^* - f||^2_{L^2(\mu)} - \int_{\mathcal{Z}} (f^*(z))^2 d\mu(z).$$

Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

Classification framework: Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We assume that the space $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ is endowed with an unknown probability measure π. We consider a random variable $Z = (X, Y)$ with values in $\mathcal{Z}$ with probability distribution π. Denote by $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Let ϕ be a function from $\mathbb{R}$ to $\mathbb{R}$. For any $f \in \mathcal{F}$ consider the $\phi-$risk, $A(f) = \mathbb{E}[Q((X,Y),f)]$, where the loss is given by $Q((x,y),f) = \phi(yf(x))$ for any $(x,y) \in \mathcal{X} \times \{-1,1\}$. Most of the time a minimizer f^* of the $\phi-$risk A over $\mathcal{F}$ or its sign is equal to the Bayes rule $f^*(x) = \text{Sign}(2\eta(x) - 1), \forall x \in \mathcal{X}$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$ (cf. [130]).

In this chapter, we obtain an oracle inequality in the general framework described at the beginning of this Subsection. Then, we use it in the density estimation and the bounded regression frameworks. For applications of this oracle inequality in the classification setup, we refer to Chapters 7 and 8.

Now, we introduce an assumption which improve the quality of estimation in our framework. This assumption has been first introduced by [91], for the problem of discriminant analysis, and [116], for the classification problem. With this assumption, parametric rates of convergence can be achieved, for instance, in the classification problem (cf. [116], [109]).

Margin Assumption(MA): *The probability measure π satisfies the margin assumption MA($\kappa, c, \mathcal{F}_0$), where $\kappa \geq 1, c > 0$ and $\mathcal{F}_0$ is a subset of $\mathcal{F}$ if $\mathbb{E}[(Q(Z,f) - Q(Z,f^*))^2] \leq c(A(f) - A^*)^{1/\kappa}, \forall f \in \mathcal{F}_0$.*

In the bounded regression setup, it is easy to see that any probability distribution π on $\mathcal{X} \times [0,1]$ naturally satisfies the margin assumption $\text{MA}(1, 16, \mathcal{F}_1)$, where $\mathcal{F}_1$ is the set of all measurable functions from $\mathcal{X}$ to $[0,1]$. In density estimation with the integrated squared risk, all probability measures π on $(\mathcal{Z}, \mathcal{T})$ absolutely continuous w.r.t. the measure μ with one version of its density a.s. bounded by a constant $B \geq 1$, satisfies the margin assumption $\text{MA}(1, 16B^2, \mathcal{F}_B)$ where $\mathcal{F}_B$ is the set of all non-negative function $f \in L^2(\mathcal{Z}, \mathcal{T}, \mu)$ bounded by B.

The margin assumption is linked to the convexity of the underlying loss. In density and regression estimation it is naturally satisfied with the better margin parameter $\kappa = 1$, but, for non-convex loss (for instance in classification) this assumption does not hold naturally (cf. Chapter 4 for a discussion on the margin assumption and for examples of such losses).

2.2. Aggregation Procedures. Let's work with the notations introduced in the beginning of the previous Subsection. The aggregation framework considered, among

others, by [74], [125], [35],[98], [114], [87], [17] is the following: take $\mathcal{F}_0$ a finite subset of $\mathcal{F}$, our aim is to mimic (up to an additive residual) the best function in $\mathcal{F}_0$ w.r.t. the risk A. For this, we consider two aggregation procedures.

The Aggregation with Exponential Weights aggregate (**AEW**) over $\mathcal{F}_0$ is defined by

$$\tilde{f}_n^{(AEW)} \stackrel{\text{def}}{=} \sum_{f\in\mathcal{F}_0} w^{(n)}(f) f, \tag{9.4}$$

where the exponential weights $w^{(n)}(f)$ are defined by

$$w^{(n)}(f) = \frac{\exp\left(-nA_n(f)\right)}{\sum_{g\in\mathcal{F}_0}\exp\left(-nA_n(g)\right)}, \quad \forall f \in \mathcal{F}_0. \tag{9.5}$$

We consider the Empirical Risk Minimization procedure **(ERM)** over $\mathcal{F}_0$ defined by

$$\tilde{f}_n^{(ERM)} \in \operatorname{Arg}\min_{f\in\mathcal{F}_0} A_n(f). \tag{9.6}$$

2.3. Oracle Inequalities. In this Subsection we state an exact oracle inequality satisfied by the ERM procedure and the AEW procedure (in the convex case) in the general framework of the beginning of Subsection 2.1. From this exact oracle inequality we deduce two other oracle inequalities in the density estimation and the bounded regression framework. We introduce a quantity which is going to be our residual term in the exact oracle inequality. We consider

$$\gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q) = \begin{cases} \left(\frac{\mathcal{B}(\mathcal{F}_0,\pi,Q)^{\frac{1}{\kappa}}\log M}{\beta_1 n}\right)^{1/2} & \text{if } \mathcal{B}(\mathcal{F}_0,\pi,Q) \geq \left(\frac{\log M}{\beta_1 n}\right)^{\frac{\kappa}{2\kappa-1}} \\ \left(\frac{\log M}{\beta_2 n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\mathcal{F}_0, \pi, Q)$ denotes $\min_{f\in\mathcal{F}_0}(A(f) - A^*)$, $\kappa \geq 1$ is the margin parameter, π is the underlying probability measure, Q is the loss function,

$$\beta_1 = \min\Big(\frac{\log 2}{96cK}, \frac{3\sqrt{\log 2}}{16K\sqrt{2}}, \frac{1}{8(4c + K/3)}, \frac{1}{576c}\Big). \tag{9.7}$$

and

$$\beta_2 = \min\Big(\frac{1}{8}, \frac{3\log 2}{32K}, \frac{1}{2(16c + K/3)}, \frac{\beta_1}{2}\Big), \tag{9.8}$$

where the constant $c > 0$ appears in $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$.

THEOREM 9.1. *Consider the general framework introduced in the beginning of Subsection 2.1. Let $\mathcal{F}_0$ denote a finite subset of M elements $f_1, \ldots, f_M$ in $\mathcal{F}$, where $M \geq 2$ is an integer. Assume that the underlying probability measure π satisfies the margin assumption $MA(\kappa, c, \mathcal{F}_0)$ for some $\kappa \geq 1, c > 0$ and $|Q(Z, f) - Q(Z, f^*)| \leq K$ a.s., for any $f \in \mathcal{F}_0$, where $K \geq 1$ is a constant. The Empirical Risk Minimization procedure (9.6) satisfies*

$$\mathbb{E}[A(\tilde{f}_n^{(ERM)}) - A^*] \leq \min_{j=1,\ldots,M}(A(f_j) - A^*) + 4\gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q).$$

Moreover, if $f \longmapsto Q(z, f)$ is convex for π-almost $z \in \mathcal{Z}$, then the AEW procedure satisfies the same oracle inequality as the ERM procedure.

Now, we give two corollaries of Theorem 9.1 in the density estimation and bounded regression framework.

COROLLARY 9.1. *Consider the bounded regression setup. Let $f_1, \ldots, f_M$ be M functions on $\mathcal{X}$ with values in $[0, 1]$. Let $\tilde{f}_n$ denote either the ERM or the AEW procedure. We have,*

for any $\epsilon > 0$,

$$\mathbb{E}[||f^* - \tilde{f}_n||^2_{L^2(P^X)}] \leq (1+\epsilon) \min_{j=1,\dots,M}(||f^* - f_j||^2_{L^2(P^X)}) + \frac{4\log M}{\epsilon\beta_2 n},$$

where β_2 *is defined in (9.8) where we take* K *equals to* 4.

COROLLARY 9.2. *Consider the density estimation framework. Assume that the underlying density function* f^* *to estimate is bounded by* $B \geq 1$. *Let* $f_1, \dots, f_M$ *be* M *functions bounded from above and below by* B. *Let* $\tilde{f}_n$ *denote either the ERM or the AEW procedure. We have, for any* $\epsilon > 0$,

$$\mathbb{E}[||f^* - \tilde{f}_n||^2_{L^2(\mu)}] \leq (1+\epsilon) \min_{j=1,\dots,M}(||f^* - f_j||^2_{L^2(\mu)}) + \frac{4\log M}{\epsilon\beta_2 n}, \tag{9.9}$$

where β_2 *is defined in (9.8) where we replace* K *by* $2B^2\mu(\mathcal{Z}) + 4B$.

In both of the last Corollaries, the ERM and the AEW procedures can both be used to mimic the best f_j among the f_j's. Nevertheless, from a computational point of view the AEW procedure does not require any minimization step contrarily to the ERM procedure. Moreover, from a theoretical point of view the ERM procedure can not mimic the best f_j among the f_j's as fast as the cumulative aggregate with exponential weights (it is an average of AEW procedures). For a comparison between these procedures we refer to Chapter 4.

REMARK 9.1. *The constants of aggregation multiplying the residual term in Theorem 9.1 and in both of the following Corollaries are very large and are certainly not optimal. Nevertheless, this is a constant of aggregation and not a constant of estimation. It means that when we use, for instance, the oracle inequality (9.9), to construct adaptive estimators, the term* $(1+\epsilon)\min_{j=1,\dots,M}(||f^* - f_j||^2_{L^2(\mu)})$ *is equal to* $(1+\epsilon)Cn^{-(2s)/(2s+1)}$, *where* s *is a regularity parameter. In that case, the constant of aggregation is divided by* n, *whereas the constant of estimation* C *is divided by* $n^{-(2s)/(2s+1)} >> n^{-1}$. *Moreover, They come from the proof and does not appear in the simulations (cf. Section 5).*

3. Multi-thresholding wavelet estimator

In the present section, we propose an adaptive estimator constructed from aggregation techniques and wavelet thresholding methods. For the density model and the regression model with uniform random design, we show that it is optimal in the minimax sense over a wide range of function spaces.

3.1. Wavelets and Besov balls. We consider an orthonormal wavelet basis generated by dilation and translation of a compactly supported "father" wavelet ϕ and a compactly supported "mother" wavelet ψ. For the purposes of this chapter, we use the periodized wavelets bases on the unit interval. Let $\phi_{j,k} = 2^{j/2}\phi(2^j x - k), \psi_{j,k} = 2^{j/2}\psi(2^j x - k)$ be the elements of the wavelet basis and $\phi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x-l), \psi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x-l)$, there periodized versions, defined for any $x \in [0,1]$, $j \in \mathbb{N}$ and $k \in \{0, \dots, 2^j - 1\}$. There exists an integer τ such that the collection ζ defined by $\zeta = \{\phi^{per}_{j,k}, k = 0, ..., 2^\tau - 1;\ \psi^{per}_{j,k}, \quad j = \tau, ..., \infty,\ k = 0, ..., 2^j - 1\}$ constitutes an orthonormal basis of $L^2([0,1])$. In what follows, the superscript "*per*" will be suppressed from the notations for convenience. For any integer

$l \geq \tau$, a square-integrable function f^* on $[0,1]$ can be expanded into a wavelet series

$$f^*(x) = \sum_{k=0}^{2^l-1} \alpha_{l,k}\phi_{l,k}(x) + \sum_{j=l}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x),$$

where $\alpha_{j,k} = \int_0^1 f^*(x)\phi_{j,k}(x)dx$ and $\beta_{j,k} = \int_0^1 f^*(x)\psi_{j,k}(x)dx$. Further details on wavelet theory can be found in [95] and [43].

Now, let us define the main function spaces of the study. Let $M \in (0,\infty)$, $s \in (0,N)$, $p \in [1,\infty)$ and $q \in [1,\infty)$. Let us set $\beta_{\tau-1,k} = \alpha_{\tau,k}$. We say that a function f^* belongs to the Besov balls $B^s_{p,q}(M)$ if and only if the associated wavelet coefficients satisfy

$$\Big[\sum_{j=\tau-1}^{\infty}\Big[2^{j(s+1/2-1/p)}\Big(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p\Big)^{1/p}\Big]^q\Big]^{1/q} \leq M, \qquad \text{if } \ q \in [1,\infty),$$

with the usual modification if $q = \infty$. We work with the Besov balls because of their exceptional expressive power. For a particular choice of parameters s, p and q, they contain the Hölder and Sobolev balls (see [95]).

3.2. Term-by-term thresholded estimator. In this Subsection, we consider the estimation of an unknown function f^* in $L^2([0,1])$ from a general situation. We only assume to have n observations gathered in the data set D_n from which we are able to estimate the wavelet coefficients $\alpha_{j,k}$ and $\beta_{j,k}$ of f^* in the basis ζ. We denote by $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ such estimates.

A **term-by-term thresholded wavelet estimator** is given by

$$\hat{f}_\lambda(D_n,x) = \sum_{k=0}^{2^\tau-1}\hat{\alpha}_{\tau,k}\phi_{\tau,k}(x) + \sum_{j=\tau}^{j_1}\sum_{k=0}^{2^j-1}\Upsilon_{\lambda_j}(\hat{\beta}_{j,k})\psi_{j,k}(x), \tag{9.10}$$

where j_1 is an integer satisfying $(n/\log n) \leq 2^{j_1} < 2(n/\log n)$, $\lambda = (\lambda_\tau, ...\lambda_{j_1})$ is a vector of positive integers and, for any $u > 0$, the operator Υ_u is such that there exist two constants $C_1, C_2 > 0$ satisfying, for any $x, y \in \mathbb{R}$,

$$|\Upsilon_u(x) - y|^2 \leq C_1(\min(y, C_2u)^2 + |x-y|^2\mathbb{1}_{\{|x-y|\geq 2^{-1}u\}}). \tag{9.11}$$

The inequality (9.11) holds for the hard thresholding rule $\Upsilon_u^{hard}(x) = x\mathbb{1}_{\{|x|\geqslant u\}}$, the soft thresholding rule $\Upsilon_u^{soft}(x) = sign(x)(|x|-u)\mathbb{1}_{\{|x|\geqslant u\}}$ (see [50], [51] and [46]) and the non-negative garrote thresholding rule $\Upsilon_u^{NG}(x) = \big(x - u^2/x\big)\,\mathbb{1}_{\{|x|\geqslant u\}}$ (see [57]).

If we consider the minimax point of view over Besov balls under the integrated squared risk, then [46] makes the conditions on $\hat{\alpha}_{j,k}$, $\hat{\beta}_{j,k}$ and the threshold λ such that the estimator $\hat{f}_\lambda(D_n,.)$ defined by (9.10) is optimal for numerous statistical models. This result is recalled in Theorem 9.2 below.

THEOREM 9.2 (Delyon and Juditsky (1996)). *Let us consider the general statistical framework described in the beginning of the present section. Assume that there exists a constant $C > 0$ such that, for any $j \in \{\tau-1, ..., j_1\}$, $k \in \{0, ..., 2^j-1\}$ and n large enough, we have*

$$\mathbb{E}(|\hat{\beta}_{j,k} - \beta_{j,k}|^4) \leq Cn^{-2}, \ \textit{where we take } \hat{\beta}_{\tau-1,k} = \hat{\alpha}_{\tau,k}, \tag{9.12}$$

and that there exist two constants $C > 0$ and $\rho_* > 0$ such that, for any $a, j \in \{\tau, ..., j_1\}$, $k \in \{0, ..., 2^j - 1\}$ and n large enough, we have

$$\mathbb{P}\left(2\sqrt{n}|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \rho_*\sqrt{a}\right) \leqslant C2^{-4a}. \tag{9.13}$$

Let us consider the term-by-term thresholded estimator $\hat{f}_{v_{j_s}}(D_n, .)$ defined by (9.10) with the threshold

$$v_{j_s} = (\rho_*(j - j_s)_+)_{j=\tau,...,j_1},$$

where j_s is an integer such that $n^{1/(1+2s)} \leq 2^{j_s} < 2n^{1/(1+2s)}$. Then, there exists a constant $C > 0$ such that, for any $p \in [1, \infty]$, $s \in (1/p, N]$, $q \in [1, \infty]$ and n large enough, we have:

$$\sup_{f \in B^s_{p,q}(L)} \mathbb{E}[\|\hat{f}_{v_{j_s}}(D_n, .) - f^*\|^2_{L^2([0,1])}] \leqslant Cn^{-2s/(2s+1)}.$$

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax for numerous statistical models, where s is a regularity parameter. For the density model and the regression model with uniform design, we refer the reader to [46] for further details about the choice of the estimator $\hat{\beta}_{j,k}$ and the value of the thresholding constant ρ_*. Starting from this non-adaptive result, we use aggregation methods to construct an adaptive estimator at least as good in the minimax sense as $\hat{f}_{v_{j_s}}(D_n, .)$.

3.3. Multi-thresholding estimator. Let us divide our observations D_n into two disjoint subsamples D_m, of size m, made of the first m observations and $D^{(l)}$, of size l, made of the last remaining observations, where we take

$$l = \lceil n/\log n \rceil \text{ and } m = n - l.$$

The first subsample D_m, sometimes called "training sample", is used to construct a family of estimators (in our case this is thresholded estimators) and the second subsample $D^{(l)}$, called the "training sample", is used to construct the weights of the aggregation procedure. For a discussion on the sample splitting we refer to Chapter 7.

Definition 9.1. *Let us consider the term-by-term thresholded estimator described in (9.10). Assume that we want to estimate a function f^* from $[0, 1]$ with values in $[a, b]$. Consider the projection function*

$$h_{a,b}(y) = \max(a, \min(y, b)), \forall y \in \mathbb{R}. \tag{9.14}$$

We define the **multi-thresholding estimator** *$\tilde{f}_n : [0, 1] \to [a, b]$ at a point $x \in [0, 1]$ by the following aggregate*

$$\tilde{f}_n(x) = \sum_{u \in \Lambda_n} w^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .)))h_{a,b}(\hat{f}_{v_u}(D_m, x)), \tag{9.15}$$

where $\Lambda_n = \{0, ..., \log n\}$, $v_u = (\rho(j - u)_+)_{j=\tau,...,j_1}, \forall u \in \Lambda_n$ and ρ is a positive constant depending on the model worked out and

$$w^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .))) = \frac{\exp\left(-lA^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .)))\right)}{\sum_{\gamma \in \Lambda_n} \exp\left(-lA^{(l)}(h_{a,b}(\hat{f}_{v_\gamma}(D_m, .)))\right)}, \quad \forall u \in \Lambda_n,$$

where $A^{(l)}(f) = \frac{1}{l}\sum_{i=m+1}^{n} Q(Z_i, f)$ is the empirical risk constructed from the l last observations, for any function f and for the choice of a loss function Q depending on the model considered (cf. (9.2) and (9.3) for examples).

The multi-thresholding estimator $\tilde{f}_n$ realizes a kind of 'adaptation to the threshold' by selecting the best threshold v_u for u describing the set Λ_n. Since we know that there exists an element in Λ_n depending on the regularity of f^* such that the non-adaptive estimator $\hat{f}_{v_u}(D_m, .)$ is optimal in the minimax sense (see Theorem 9.2), the multi-thresholding estimator is optimal independently of the regularity of f^*. Moreover, the cardinality of Λ_n is only $\log n$, thus the construction of $\tilde{f}_n$ does not require the construction of too many estimators.

4. Performances of the multi-thresholding estimator

In this section we explore the minimax performances of the multi-thresholding estimator defined in (9.15) under the $L^2([0,1])$ risk over Besov balls in the density estimation and the bounded regression with uniform random design models.

4.1. Density model. In the density estimation model, Theorem 9.3 below investigates rates of convergence achieved by the multi-thresholding estimator (defined by (9.15)) under the $L^2([0,1])$ risk over Besov balls.

THEOREM 9.3. *Let us consider the problem of estimating f^* from the density model. Assume that there exists $B \geq 1$ such that the underlying density function f^* to estimate is bounded by B. Let us consider the multi-thresholding estimator defined in (9.15) where we take $a = 0, b = B$, ρ such that $\rho^2 \geq 4(\log 2)(8B + (8\rho/(3\sqrt{2}))(\|\psi\|_\infty + B))$ and*

$$\hat{\alpha}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j,k}(X_i), \qquad \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(X_i). \tag{9.16}$$

Then, there exists a constant $C > 0$ such that

$$\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\|\tilde{f}_n - f^*\|^2_{L^2([0,1])}] \leqslant C n^{-2s/(2s+1)},$$

for any $p \in [1, \infty]$, $s \in (p^{-1}, N]$, $q \in [1, \infty]$ and integer n.

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax over $B^s_{p,q}(L)$. Further details about the minimax rate of convergence over Besov balls under the $L^2([0,1])$ risk for the density model can be found in [46] and [62]. For further details about the density estimation via adaptive wavelet thresholded estimators, see [52], [46] and [99]. See also [65] for a practical study.

4.2. Bounded regression. In the framework of the bounded regression model with uniform random design, Theorem 9.4 below investigates the rate of convergence achieved by the multi-thresholding estimator defined by (9.15) under the $L^2([0,1])$ risk over Besov balls.

THEOREM 9.4. *Let us consider the problem of estimating the regression function f^* in the bounded regression model with random uniform design. Let us consider the multi-thresholding estimator (9.15) with ρ such that $\rho^2 \geq 4(\log 2)(8 + (8\rho/(3\sqrt{2}))(\|\psi\|_\infty + 1))$ and*

$$\hat{\alpha}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_{j,k}(X_i), \qquad \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{j,k}(X_i). \tag{9.17}$$

Then, there exists a constant $C > 0$ such that, for any $p \in [1, \infty]$, $s \in (p^{-1}, N]$, $q \in [1, \infty]$ and integer n, we have

$$\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\|\tilde{f}_n - f^*\|^2_{L^2([0,1])}] \leqslant C n^{-2s/(2s+1)}.$$

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax over $B^s_{p,q}(L)$. The multi-thresholding estimator has better minimax properties than several other wavelet estimators developed in the literature. To the authors's knowledge, the result obtained, for instance, by the hard thresholded estimator (see [50]), by the global wavelet block thresholded estimator (see [76]), by the localized wavelet block thresholded estimator (see [29, 32, 30], [61, 60], [39] and [31]) and, in particular, the penalized Blockwise Stein method (see [36]) are worse than the one obtained by the multi-thresholding estimator and stated in Theorems 9.3 and 9.4. This is because, on the difference of those works, we obtain the optimal rate of convergence without any extra logarithm factor. In fact, the multi-thresholding estimator has similar minimax performances than the empirical Bayes wavelet methods (see [128] and [70]) and several term-by-term wavelet thresholded estimators defined with a random threshold (see [71] and [18]). Finally, it is important to mention that the multi-thresholding estimator does not need any minimization step and is relatively easy to implement.

TABLE 1. Theoretical performances of some well known adaptive wavelet estimators and of the multi-thresholding estimator.

Estimators \ $B^s_{p,q}(L)$	$1 < \pi < 2$	$2 \leq \pi$
Hard thresholding	near optimal	near optimal
Block thresholding	near optimal	optimal
Multi-thresholding	optimal	optimal

In the table 1, 'near optimal' means that the estimation procedure achieves the minimax rate up to a logarithm factor.

5. Simulated Illustrations

This section illustrates the performances of the multi-thresholding estimator. Let us consider the regression model with random uniform design and with the noise $\zeta = \max(-2^{-1}, min(N, 2^{-1}))\sigma$ where $\sigma = 0,05$ and N is a standard Gaussian variable. For the simulations we take $n = 2^{13}$ observations.

Let us define the multi-thresholding estimator (9.15) with

- the non-negative garrote thresholding operator $\Upsilon_u(x) = \left(x - u^2/x\right) 1_{\{|x| \geqslant u\}}$. The reason why we chose this thresholding rule is that, for the universal threshold, it provides better numerical and graphical result than the hard and soft thresholding rules (cf. [57]).
- the wavelet basis 'sym8' (Symlet 8, see for instance [43])
- the function f is 'Heavisine',

$$f(x) = 3,3662 * [4 \sin(4\pi x) - sgn(x - 0,3) - sgn(0,72 - x)].$$

- the estimators $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ defined by (9.17),
- the thresholding constant $\rho = \sigma\sqrt{2}$.

In the simulation below, the multi-thresholding estimator is called Estimator Multi-NG, (NG is for "nonnegative garotte") and we use all the observations for the construction of the estimators to aggregate and for the construction of the exponential weights. The Estimator NG is the usual nonnegative garotte thresholding estimator taken with the threshold

$$\lambda = (\lambda_\tau, \ldots, \lambda_{j_1}), \text{ where } \lambda_j = \sigma\sqrt{2}\sqrt{(j/n)}, \forall j \in \{\tau, ..., j_1\}$$

proposed by [52]. The resulting estimator is near optimal in the minimax sense over Besov balls under the L^2 risk (cf. [57]). The multi-thresholding estimator is visually better

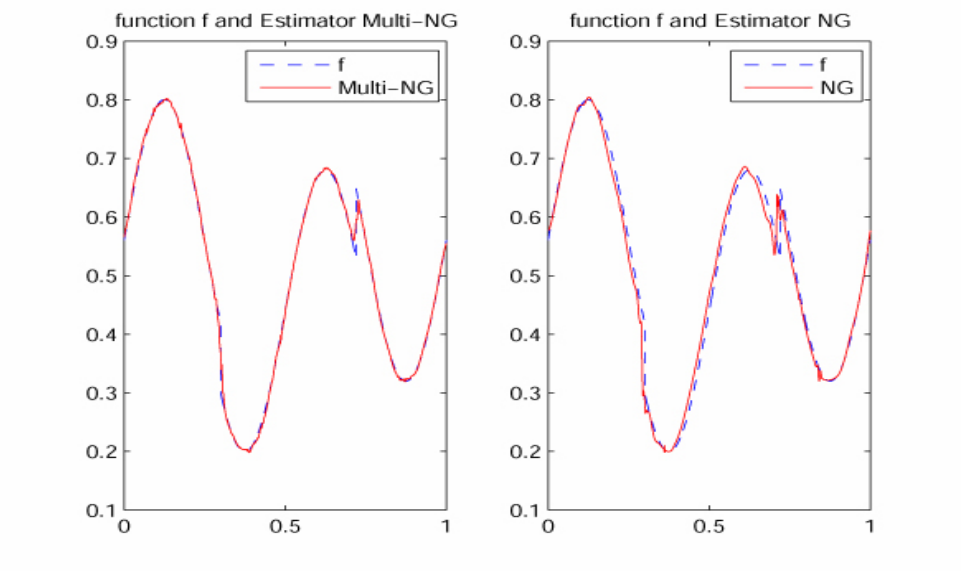

figure 1: Visual comparisons of the reconstructions of the Estimator Multi-NG and the conventional Estimator NG.

than the usual nonnegative garotte thresholding estimator. The next figure shows the repartition of the 'mass' between the aggregated estimators.

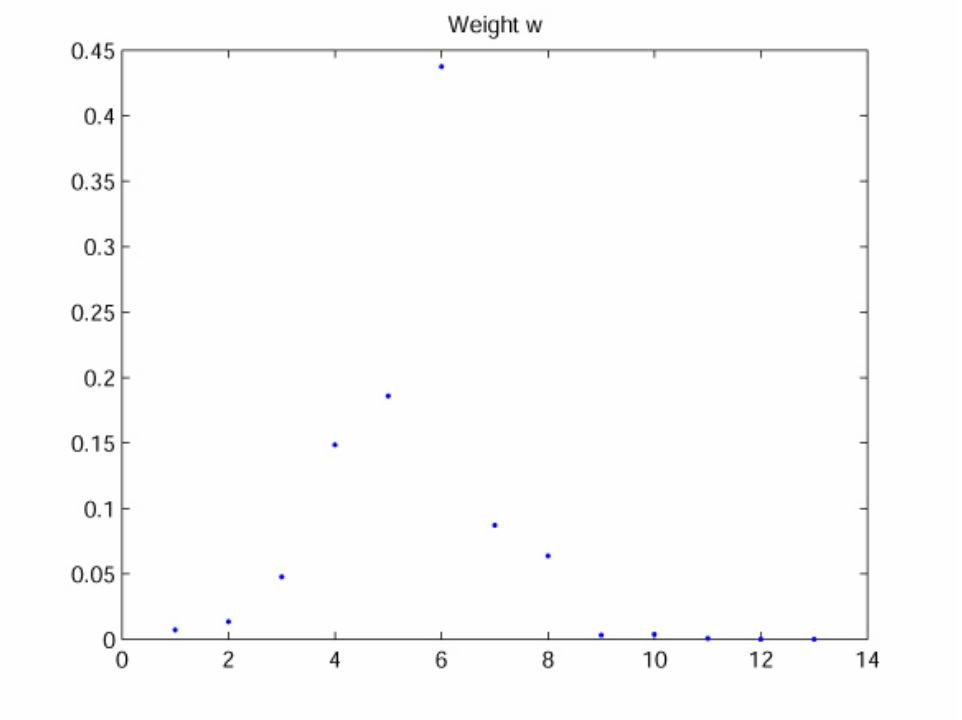

figure 2: Spatial repartition of the weights w.

On figure 2, we can see a concentration of the weights around the estimator with threshold v_u where $u = 6$. Around this estimator there are five others estimators which share most of the remaining mass. This figure shows how the multi-thresholding estimator proceeds by concentrating around the best estimator among the f_{v_u} for u in Λ_n.

6. Proofs

Proof of Theorem 9.1. We recall the notations of the general framework introduced in the beginning of Subsection 2.1. Consider a loss function $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$, the risk $A(f) = \mathbb{E}[Q(Z, f)]$, the minimum risk $A^* = \min_{f \in \mathcal{F}} A(f)$, where we assume, w.l.o.g., that it is achieved by an element f^* in $\mathcal{F}$ and the empirical risk $A_n(f) = (1/n) \sum_{i=1}^n Q(Z_i, f)$, for any $f \in \mathcal{F}$. The following proof is a generalization of the proof of Theorem 3.1 in Chapter 3.

We first start by a 'linearization' of the risk. Consider the convex set

$$\mathcal{C} = \left\{ (\theta_1, \ldots, \theta_M) : \theta_j \geq 0 \text{ and } \sum_{j=1}^M \theta_j = 1 \right\}$$

and define the following functions on $\mathcal{C}$

$$\tilde{A}(\theta) \stackrel{\text{def}}{=} \sum_{j=1}^M \theta_j A(f_j) \text{ and } \tilde{A}_n(\theta) \stackrel{\text{def}}{=} \sum_{j=1}^M \theta_j A_n(f_j)$$

which are linear versions of the risk A and its empirical version A_n.

Using the Lagrange method of optimization we find that the exponential weights $w \stackrel{\text{def}}{=} (w^{(n)}(f_j))_{1\le j\le M}$ are the unique solution of the minimization problem

$$\min\Big(\tilde{A}_n(\theta) + \frac{1}{n}\sum_{j=1}^{M}\theta_j\log\theta_j : (\theta_1,\ldots,\theta_M)\in\mathcal{C}\Big),$$

where we use the convention $0\log 0 = 0$. Take $\hat{\jmath}\in\{1,\ldots,M\}$ such that $A_n(f_{\hat{\jmath}}) = \min_{j=1,\ldots,M} A_n(f_j)$. The vector of exponential weights w satisfies

$$\tilde{A}_n(w) \le \tilde{A}_n(e_{\hat{\jmath}}) + \frac{\log M}{n},$$

where e_j denotes the vector in $\mathcal{C}$ with 1 for j-th coordinate (and 0 elsewhere).

Let $\epsilon > 0$. Denote by $\tilde{A}_{\mathcal{C}}$ the minimum $\min_{\theta\in\mathcal{C}}\tilde{A}(\theta)$. We consider the subset of $\mathcal{C}$

$$\mathcal{D} \stackrel{\text{def}}{=} \Big\{\theta\in\mathcal{C} : \tilde{A}(\theta) > \tilde{A}_{\mathcal{C}} + 2\epsilon\Big\}.$$

Let $x > 0$. If

$$\sup_{\theta\in\mathcal{D}}\frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} \le \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then for any $\theta\in\mathcal{D}$, we have

$$\tilde{A}_n(\theta) - A_n(f^*) \ge \tilde{A}(\theta) - A^* - \frac{\epsilon(\tilde{A}(\theta) - A^* + x)}{(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)} \ge \tilde{A}_{\mathcal{C}} - A^* + \epsilon,$$

because $\tilde{A}(\theta) - A^* \ge \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$. Hence,

$$\begin{aligned}
&\mathbb{P}\left[\inf_{\theta\in\mathcal{D}}\Big(\tilde{A}_n(\theta) - A_n(f^*)\Big) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right]\\
&\le \mathbb{P}\left[\sup_{\theta\in\mathcal{D}}\frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right]. \qquad (9.18)
\end{aligned}$$

Observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. Thus, for $j_0\in\{1,\ldots,M\}$ such that $\tilde{A}(e_{j_0}) = \min_{j=1,\ldots,M}\tilde{A}(e_j)$ $(= \min_{j=1,\ldots,M} A(f_j))$, we have $\tilde{A}(e_{j_0}) = \min_{\theta\in\mathcal{C}}\tilde{A}(\theta)$. We obtain the last inequality by linearity of $\tilde{A}$ and the convexity of $\mathcal{C}$. Let $\hat{w}$ denotes either the exponential weights w or $e_{\hat{\jmath}}$. According to (9.18), we have

$$\tilde{A}_n(\hat{w}) \le \min_{j=1,\ldots,M}\tilde{A}_n(e_j) + \frac{\log M}{n} \le \tilde{A}_n(e_{j_0}) + \frac{\log M}{n}$$

So, if $\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon$ then $\hat{w}\in\mathcal{D}$ and thus, there exists $\theta\in\mathcal{D}$ such that $\tilde{A}_n(\theta) - \tilde{A}_n(f^*) \le \tilde{A}_n(e_{j_0}) - \tilde{A}_n(f^*) + (\log M)/n$. Hence, we have

$$\begin{aligned}
\mathbb{P}\Big[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\Big] &\le \mathbb{P}\left[\inf_{\theta\in\mathcal{D}}\tilde{A}_n(\theta) - A_n(f^*) \le \tilde{A}_n(e_{j_0}) - A_n(f^*) + \frac{\log M}{n}\right]\\
&\le \mathbb{P}\left[\inf_{\theta\in\mathcal{D}}\tilde{A}_n(\theta) - A_n(f^*) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right]\\
&\quad + \mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \ge \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right]\\
&\le \mathbb{P}\left[\sup_{\theta\in\mathcal{C}}\frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(f) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right]
\end{aligned}$$

$$+\mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right].$$

If we assume that

$$\sup_{\theta\in\mathcal{C}} \frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then, there exists $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_M^{(0)}) \in \mathcal{C}$, such that

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - (\tilde{A}_n(\theta^{(0)}) - A_n(f^*))}{\tilde{A}(\theta^{(0)}) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

The linearity of $\tilde{A}$ yields

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - (\tilde{A}_n(\theta^{(0)}) - A_n(f^*))}{\tilde{A}(\theta^{(0)}) - A^* + x} = \frac{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* + x]}$$

and since, for any numbers $a_1, \ldots, a_M$ and positive numbers $b_1, \ldots, b_M$, we have

$$\frac{\sum_{j=1}^M a_j}{\sum_{j=1}^M b_j} \leq \max_{j=1,\ldots,M}\left(\frac{a_j}{b_j}\right),$$

then, we obtain

$$\max_{j=1,\ldots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{A_{\mathcal{F}_0} - A^* + 2\epsilon + x},$$

where $A_{\mathcal{F}_0} \stackrel{\text{def}}{=} \min_{j=1,\ldots,M} A(f_j)$ $(= \tilde{A}_{\mathcal{C}})$.

Now, we use the relative concentration inequality of Lemma 9.1 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\max_{j=1,\ldots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{A_{\mathcal{F}_0} - A^* + 2\epsilon + x}\right] \\
&\quad\leq\quad M\left(1 + \frac{4c(A_{\mathcal{F}_0} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right)\exp\left(-\frac{n(\epsilon x)^2}{4c(A_{\mathcal{F}_0} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
&\qquad +M\left(1 + \frac{4K(A_{\mathcal{F}_0} - A^* + 2\epsilon + x)}{3n\epsilon x}\right)\exp\left(-\frac{3n\epsilon x}{4K(A_{\mathcal{F}_0} - A^* + 2\epsilon + x)}\right).
\end{aligned}$$

Using the margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$ to upper bound the variance term and applying Bernstein's inequality, we get

$$\begin{aligned}
\mathbb{P}\Big[A_n(f_{j_0}) \quad - \quad & A_n(f^*) \geq A_{\mathcal{F}_0} - A^* + \epsilon - \frac{\log M}{n}\Big] \\
&\leq \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{2c(A_{\mathcal{F}_0} - A^*)^{1/\kappa} + (2K/3)(\epsilon - (\log M)/n)}\right),
\end{aligned}$$

for any $\epsilon > (\log M)/n$. From now, we take $x = A_{\mathcal{F}_0} - A^* + 2\epsilon$, then, for any $(\log M)/n < \epsilon < 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\tilde{A}(\hat{w}) > A_{\mathcal{F}_0} + 2\epsilon\right) &\leq \exp\left(-\frac{n(\epsilon - \log M/n)^2}{2c(A_{\mathcal{F}_0} - A^*)^{1/\kappa} + (2K/3)(\epsilon - (\log M)/n)}\right) \\
&+ \quad M\left(1 + \frac{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right)\exp\left(-\frac{n\epsilon^2}{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}\right) \\
&+ \quad M\left(1 + \frac{8K}{3n\epsilon}\right)\exp\left(-\frac{3n\epsilon}{8K}\right).
\end{aligned}$$

If $\hat{w}$ denotes $e_{\hat{\jmath}}$ then, $\tilde{A}(\hat{w}) = \tilde{A}(e_{\hat{\jmath}}) = A(\tilde{f}^{(ERM)})$. If $\hat{w}$ denotes the vector of exponential weights w and if $f \longmapsto Q(z,f)$ is convex for π-almost $z \in \mathcal{Z}$, then, $\tilde{A}(\hat{w}) = \tilde{A}(w) \geq A(\tilde{f}_n^{(AEW)})$. If $f \longmapsto Q(z,f)$ is assumed to be convex for π-almost $z \in \mathcal{Z}$ then, let $\tilde{f}_n$ denote either the ERM procedure or the AEW procedure, otherwise, let $\tilde{f}_n$ denote the ERM procedure $\tilde{f}_n^{(ERM)}$. We have for any $2(\log M)/n < u < 1$,

$$\mathbb{E}[A(\tilde{f}_n) - A_{\mathcal{F}_0}] \leq \mathbb{E}\left[\tilde{A}(\hat{w}) - A_{\mathcal{F}_0}\right] \leq 2u + 2\int_{u/2}^{1} [T_1(\epsilon) + M(T_2(\epsilon) + T_3(\epsilon))]\, d\epsilon, \tag{9.19}$$

where

$$T_1(\epsilon) = \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{2c(A_{\mathcal{F}_0} - A^*)^{1/\kappa} + (2K/3)(\epsilon - (\log M)/n)}\right),$$

$$T_2(\epsilon) = \left(1 + \frac{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right)\exp\left(-\frac{n\epsilon^2}{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}\right)$$

and

$$T_3(\epsilon) = \left(1 + \frac{8K}{3n\epsilon}\right)\exp\left(-\frac{3n\epsilon}{8K}\right).$$

We recall that β_1 is defined in (9.7). Consider separately the following cases $(C1)$ and $(C2)$.

$(C1)$ The case $A_{\mathcal{F}_0} - A^* \geq ((\log M)/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

Denote by $\mu(M)$ the unique solution of $\mu_0 = 3M\exp(-\mu_0)$. Then, clearly $(\log M)/2 \leq \mu(M) \leq \log M$. Take u such that $(n\beta_1 u^2)/(A_{\mathcal{F}_0} - A^*)^{1/\kappa} = \mu(M)$. Using the definition of case (1) and of $\mu(M)$ we get $u \leq A_{\mathcal{F}_0} - A^*$. Moreover, $u \geq 4\log M/n$, then

$$\begin{aligned}\int_{u/2}^{1} T_1(\epsilon)d\epsilon \quad &\leq \quad \int_{u/2}^{(A_{\mathcal{F}_0}-A^*)/2} \exp\left(-\frac{n(\epsilon/2)^2}{(2c + K/6)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right) d\epsilon \\ &\quad + \int_{(A_{\mathcal{F}_0}-A^*)/2}^{1} \exp\left(-\frac{n(\epsilon/2)^2}{(4c + K/3)\epsilon^{1/\kappa}}\right) d\epsilon.\end{aligned}$$

Using Lemma 9.2 and the inequality $u \leq A_{\mathcal{F}_0} - A^*$, we obtain

$$\int_{u/2}^{1} T_1(\epsilon)d\epsilon \leq \frac{8(4c + K/3)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{8(4c + K/3)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \tag{9.20}$$

We have $16c(A_{\mathcal{F}_0} - A^* + 2u) \leq nu^2$ thus, using Lemma 9.2, we get

$$\begin{aligned}\int_{u/2}^{1} T_2(\epsilon)d\epsilon \quad &\leq \quad 2\int_{u/2}^{(A_{\mathcal{F}_0}-A^*)/2} \exp\left(-\frac{n\epsilon^2}{64c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right) d\epsilon \\ &\quad + 2\int_{(A_{\mathcal{F}_0}-A^*)/2}^{1} \exp\left(-\frac{n\epsilon^{2-1/\kappa}}{128c}\right) d\epsilon \\ &\leq \quad \frac{2148c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{2148c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \qquad (9.21)\end{aligned}$$

We have $16(3n)^{-1} \leq u \leq A_{\mathcal{F}_0} - A^*$, thus,

$$\int_{u/2}^{1} T_3(\epsilon)d\epsilon \leq \frac{16K(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{3nu}\exp\left(-\frac{3nu^2}{16K(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \tag{9.22}$$

From (9.20), (9.21), (9.22) and (9.19) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 2u + 6M\frac{(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{n\beta_1 u}\exp\left(-\frac{n\beta_1 u^2}{(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right).$$

The definition of u leads to $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 4\sqrt{\frac{(A_{\mathcal{F}_0}-A^*)^{1/\kappa}\log M}{n\beta_1}}$.

(C2)The case $A_{\mathcal{F}_0} - A^* \leq ((\log M)/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

We now choose u such that $n\beta_2 u^{(2\kappa-1)/\kappa} = \mu(M)$, where $\mu(M)$ denotes the unique solution of $\mu_0 = 3M\exp(-\mu_0)$ and β_2 is defined in (9.8). Using the definition of case (2) and of $\mu(M)$ we get $u \geq A_{\mathcal{F}_0} - A^*$ (since $\beta_1 \geq 2\beta_2$). Using the fact that $u > 4\log M/n$ and Lemma 9.2, we have

$$\int_{u/2}^{1} T_1(\epsilon)d\epsilon \leq \frac{2(16c+K/3)}{nu^{1-1/\kappa}}\exp\left(-\frac{3nu^{2-1/\kappa}}{2(16c+K/3)}\right). \tag{9.23}$$

We have $u \geq (128c/n)^{\kappa/(2\kappa-1)}$ and using Lemma 9.2, we obtain

$$\int_{u/2}^{1} T_2(\epsilon)d\epsilon \leq \frac{256c}{nu^{1-1/\kappa}}\exp\left(-\frac{nu^{2-1/\kappa}}{256c}\right). \tag{9.24}$$

Since $u > 16K/(3n)$ we have

$$\int_{u/2}^{1} T_3(\epsilon)d\epsilon \leq \frac{16K}{3nu^{1-1/\kappa}}\exp\left(-\frac{3nu^{2-1/\kappa}}{16K}\right). \tag{9.25}$$

From (9.23), (9.24), (9.25) and (9.19) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 2u + 6M\frac{\exp\left(-n\beta_2 u^{(2\kappa-1)/\kappa}\right)}{n\beta_2 u^{1-1/\kappa}}.$$

The definition of u yields $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 4\left(\frac{\log M}{n\beta_2}\right)^{\frac{\kappa}{2\kappa-1}}$. This completes the proof.

LEMMA 9.1. *Consider the framework introduced in the beginning of Subsection 2.1. Let $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ be a finite subset of $\mathcal{F}$. We assume that π satisfies MA($\kappa, c, \mathcal{F}_0$), for some $\kappa \geq 1, c > 0$ and $|Q(Z,f) - Q(Z,f^*)| \leq K$ a.s., for any $f \in \mathcal{F}_0$, where $K \geq 1$ is a constant. We have for any positive numbers t, x and any integer n*

$$\mathbb{P}\left[\max_{f\in\mathcal{F}} \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x} > t\right]$$
$$\leq M\left(\left(1 + \frac{4cx^{1/\kappa}}{n(tx)^2}\right)\exp\left(-\frac{n(tx)^2}{4cx^{1/\kappa}}\right) + \left(1 + \frac{4K}{3ntx}\right)\exp\left(-\frac{3ntx}{4K}\right)\right).$$

Proof. We use a "peeling device". Let $x > 0$. For any integer j, we consider $\mathcal{F}_j = \{f \in \mathcal{F} : jx \leq A(f) - A^* < (j+1)x\}$. Define the empirical process

$$Z_x(f) = \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x}.$$

Using Bernstein's inequality and margin assumption MA($\kappa, c, \mathcal{F}_0$) to upper bound the variance term, we have

$$\mathbb{P}\left[\max_{f\in\mathcal{F}} Z_x(f) > t\right] \leq \sum_{j=0}^{+\infty} \mathbb{P}\left[\max_{f\in\mathcal{F}_j} Z_x(f) > t\right]$$
$$\leq \sum_{j=0}^{+\infty} \mathbb{P}\Big[\max_{f\in\mathcal{F}_j} A(f) - A_n(f) - (A(f^*) - A_n(f^*)) > t(j+1)x\Big]$$

$$
\begin{aligned}
&\leq M\sum_{j=0}^{+\infty}\exp\Big(-\frac{n[t(j+1)x]^2}{2c((j+1)x)^{1/\kappa}+(2K/3)t(j+1)x}\Big)\\
&\leq M\Big(\sum_{j=0}^{+\infty}\exp\Big(-\frac{n(tx)^2(j+1)^{2-1/\kappa}}{4cx^{1/\kappa}}\Big)+\exp\Big(-(j+1)\frac{3ntx}{4K}\Big)\Big)\\
&\leq M\Big(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{4c}\right)+\exp\left(-\frac{3ntx}{4K}\right)\Big)\\
&\quad +M\int_1^{+\infty}\Big(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{4c}u^{2-1/\kappa}\right)+\exp\left(-\frac{3ntx}{4K}u\right)\Big)du.
\end{aligned}
$$

Lemma 9.2 completes the proof.

LEMMA 9.2. *Let $\alpha \geq 1$ and $a, b > 0$. An integration by part yields*

$$\int_a^{+\infty}\exp\left(-bt^\alpha\right)dt \leq \frac{\exp(-ba^\alpha)}{\alpha b a^{\alpha-1}}$$

Proof of Corollaries 9.1 and 9.2. In the bounded regression setup, any probability distribution π on $\mathcal{X}\times[0,1]$ satisfies the margin assumption $\mathrm{MA}(1,16,\mathcal{F}_1)$, where $\mathcal{F}_1$ is the set of all measurable functions from $\mathcal{X}$ to $[0,1]$. In density estimation with the integrated squared risk, any probability measure π on $(\mathcal{Z},\mathcal{T})$, absolutely continuous w.r.t. the measure μ with one version of its density a.s. bounded by a constant $B\geq 1$, satisfies the margin assumption $\mathrm{MA}(1,16B^2,\mathcal{F}_B)$ where $\mathcal{F}_B$ is the set of all non-negative function $f\in L^2(\mathcal{Z},\mathcal{T},\mu)$ bounded by B. To complete the proof we use that for any $\epsilon>0$,

$$\Big(\frac{\mathcal{B}(\mathcal{F}_0,\pi,Q)\log M}{\beta_1 n}\Big)^{1/2}\leq \epsilon\mathcal{B}(\mathcal{F}_0,\pi,Q)+\frac{\log M}{\beta_2 n\epsilon}$$

and in both cases $f\longmapsto Q(z,f)$ is convex for any $z\in\mathcal{Z}$.

Proof of Theorem 9.3. We apply Theorem 9.2, with $\epsilon=1$, to the multi-thresholding estimator $\hat{f}_n$ defined in (9.15). Since the density function f^* to estimate takes its values in $[0,B]$, $Card(\Lambda_n)=\log n$ and $m\geq n/2$, we have, conditionally to the first subsample D_m,

$$
\begin{aligned}
&\mathbb{E}[\|f^*-\hat{f}_n\|^2_{L^2([0,1])}\ |D_m]\\
&\quad\leq 2\min_{u\in\Lambda_n}(||f^*-h_{0,B}(\hat{f}_{v_u}(D_m,.))||^2_{L^2([0,1])})+\frac{4(\log n)\log(\log n)}{\beta_2 n}\\
&\quad\leq 2\min_{u\in\Lambda_n}(||f^*-\hat{f}_{v_u}(D_m,.)||^2_{L^2([0,1])})+\frac{4(\log n)\log(\log n)}{\beta_2 n},
\end{aligned}
$$

where $h_{0,B}$ is the projection function introduced in (9.14) and β_2 is given in (9.8). Now, for any $s>0$, let us consider j_s an integer in Λ_n such that $n^{1/(1+2s)}\leq 2^{j_s}<2n^{1/(1+2s)}$. Since the estimators $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ defined by (9.16) satisfy the inequalities (9.12) and (9.13), Theorem 9.2 implies that, for any $p\in[1,\infty]$, $s\in(1/p,N]$, $q\in[1,\infty]$ and n large enough, we have

$$
\begin{aligned}
&\sup_{f^*\in B^s_{p,q}(L)}\mathbb{E}[\|\tilde{f}-f^*\|^2_{L^2([0,1])}]=\sup_{f^*\in B^s_{p,q}(L)}\mathbb{E}[\mathbb{E}[\|\tilde{f}-f^*\|^2_{L^2([0,1])}\ |D_m]]\\
&\quad\leq 2\sup_{f^*\in B^s_{p,q}(L)}\mathbb{E}[\min_{u\in\Lambda_n}(||f^*-\hat{f}_{v_u}(D_m,.)||^2_{L^2([0,1])}]+\frac{4(\log n)\log(\log n)}{\beta_2 n}
\end{aligned}
$$

$$\begin{aligned} &\leq 2 \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[||f^* - \hat{f}_{v_{j_s}}(D_m, .)||^2_{L^2([0,1])}] + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\ &\leq Cn^{-2s/(1+2s)}. \end{aligned}$$

This completes the proof of Theorem 9.3.

Proof of Theorem 9.4. The proof of Theorem 9.4 is similar to the proof of Theorem 9.3. We only need to prove that, for any $j \in \{\tau, ..., j_1\}$ and $k \in \{0, ..., 2^j - 1\}$, the estimators $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ defined by (9.17) satisfy the inequalities (9.12) and (9.13). First of all, let us notice that the random variables $Y_1\psi_{j,k}(X_1), ..., Y_n\psi_{j,k}(X_n)$ are i.i.d and that there $m-$th moment, for $m \geq 2$, satisfies

$$\mathbb{E}(|\psi_{j,k}(X_1)|^m) \leq \|\psi\|^{m-2}_\infty 2^{j(m/2-1)} \mathbb{E}(|\psi_{j,k}(X_1)|^2) = \|\psi\|^{m-2}_\infty 2^{j(m/2-1)}.$$

For the first inequality (cf. inequality (9.12)), Rosenthal's inequality (see [62, p.241]) yields, for any $j \in \{\tau, ..., j_1\}$,

$$\begin{aligned} \mathbb{E}(|\hat{\beta}_{j,k} - \beta_{j,k}|^4) &\leq C(n^{-3}\mathbb{E}(|Y_1\psi_{j,k}(X_1)|^4) + n^{-2}[\mathbb{E}(|Y_1\psi_{j,k}(X_1)|^2)]^2) \\ &\leq C\|Y\|^4_\infty \|\psi\|^4_\infty (n^{-3}2^{j_1} + n^{-2}) \leq Cn^{-2}. \end{aligned}$$

For second inequality (cf. inequality (9.13)), Bernstein's inequality yields

$$\mathbb{P}\Big(2\sqrt{n}|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \rho\sqrt{a}\Big) \leq 2\exp\Big(-\frac{\rho^2 a}{8\sigma^2 + (8/3)M\rho\sqrt{a}/(2\sqrt{n})}\Big),$$

where $a \in \{\tau, ..., j_1\}$, $\rho \in (0, \infty)$,

$$\begin{aligned} M &= \|Y\psi_{j,k}(X) - \beta_{j,k}\|_\infty \leq 2^{j/2}\|Y\|_\infty\|\psi\|_\infty + \|f^*\|^2_{L^2([0,1])} \\ &\leq 2^{j_1/2}(\|\psi\|_\infty + 1) \leq 2^{1/2}(n/\log n)^{1/2}(\|\psi\|_\infty + 1), \end{aligned}$$

and $\sigma^2 = \mathbb{E}(|Y_1\psi_{j,k}(X_1) - \beta_{j,k}|^2) \leq \mathbb{E}(|Y_1\psi_{j,k}(X_1)|^2) \leq \|Y\|^2_\infty \leq 1$. Since $a \leq \log n$, we complete the proof by seeing that for ρ large enough, we have

$$\exp\Big(-\frac{\rho^2 a}{8\sigma^2 + (8/3)M\rho\sqrt{a}/(2\sqrt{n})}\Big) \leq 2^{-4a}.$$

CHAPTER 10

Optimal rates and adaptation in the single-index model using aggregation

We want to recover the regression function in the single-index model. Using an aggregation algorithm with local polynomial estimators, we answer in particular to Question 2 from Stone (1982) [110] on the optimal convergence rate within this model. The procedure constructed here has strong adaptation properties: it adapts both to the smoothness of the link function and to the unknown index. Moreover, the procedure locally adapts to the distribution of the data, which allows to prove the results for a fairly general design. The behavior of this algorithm is studied through numerical simulations. In particular, we show empirically that it improves strongly empirical risk minimization.

Contents

The material of this chapter is a joint work with Stéphane Gaïffas, submitted for publication (cf. [56]).

1. Introduction

The single-index model is standard in statistical literature. It is widely used in several fields, since it provides a simple trade-off between purely nonparametric and purely parametric approaches. Moreover, it is well-known that it allows to deal with the so-called "curse of dimensionality" phenomenon. Within the minimax theory, this phenomenon is explained by the fact that the minimax rate linked to this model (which is multivariate, in the sense that the number of explanatory variables is larger than 1) is the same as in the univariate model. Indeed, if n is the sample size, the minimax rate over an isotropic s-Hölder ball is $n^{-2s/(2s+d)}$ for mean integrated square error (MISE) in the d-dimensional regression model without the single-index constraint, while in the single-index model, this rate is conjectured to be $n^{-2s/(2s+1)}$ by [110]. Hence, even for small values of d (larger than 2), the dimension has a strong impact on the quality of estimation when no prior assumption on the structure of the multivariate regression function is made. In this sense, the single-index model provides a simple way to reduce the dimension of the problem.

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random variable satisfying

$$Y = g(X) + \sigma(X)\varepsilon, \tag{10.1}$$

where ε is independent of X with law $N(0,1)$ and where $\sigma(\cdot)$ is such that $\sigma_0 < \sigma(X) \leq \sigma$ a.s. for some $\sigma_0 > 0$ and a known $\sigma > 0$. We denote by P the probability distribution of (X, Y) and by P_X the margin law in X or *design* law. In the single-index model, the regression function as a particular structure. Indeed, we assume that g can be written has

$$g(x) = f(\vartheta^\top x) \tag{10.2}$$

for all $x \in \mathbb{R}^d$, where $f : \mathbb{R} \to \mathbb{R}$ is the *link* function and where the direction $\vartheta \in \mathbb{R}^d$, or *index*, belongs to the half-unit sphere

$$S_+^{d-1} = \big\{v \in \mathbb{R}^d \mid \|v\|_2 = 1 \text{ and } v_d \geq 0\big\},$$

where $\|\cdot\|_2$ is the Euclidean norm over $\mathbb{R}^d$. The assumption $\vartheta \in S_+^{d-1}$ entails the unicity of (f, ϑ) in (10.2) and thus the identifiability of the model. We assume that the available data

$$D_n := [(X_i, Y_i); 1 \leq i \leq n] \tag{10.3}$$

is a sample of n i.i.d. copies of (X, Y) satisfying (10.1) and (10.2). In this model, we can focus on the estimation of the index ϑ based on D_n when the link function f is unknown, or we can focus on the estimation of the regression g when both f and ϑ are unknown. In this chapter, we consider the latter problem. It is assumed below that f belongs to some family of Hölder balls, that is, we do not suppose its smoothness to be known.

Statistical literature on this model is wide. Among many other references, see [66] for applications in econometrics, an application in medical science can be found in [122], see also [44], [45] and the survey paper by [58]. For the estimation of the index, see for instance [67]; for testing the parametric versus the nonparametric single-index assumption, see [111]. See also a chapter in [59] which is devoted to dimension reduction techniques in the bounded regression model. While the literature on single-index modelling is vast, several problems remain open. For instance, Question 2 from [110] concerning the minimax rate over Hölder balls in model (10.1),(10.2) is still open.

This chapter provides new minimax results about the single-index model, which answer in particular to latter question. Indeed, we prove that in model (10.1),(10.2), we can achieve

the rate $n^{-2s/(2s+1)}$ for a link function in a whole family of Hölder balls with smothness s, see Theorem 10.1. The optimality of this rate is proved in Theorem 10.2. To prove the upper bound, we use an estimator which adapts both to the index parameter and to the smoothness of the link function. This result is stated under fairly general assumptions on the design, which include any "non-pathological" law for P_X. Moreover, this estimator has a nice "design-adaptation" property, since it does not depend within its construction on P_X.

2. Construction of the procedure

The procedure developed here for recovering the regression does not use a plugin estimator by direct estimation of the index. Instead, it *adapts* to it, by aggregating several univariate estimators based on projected samples

$$D_m(v) := [(v^\top X_i, Y_i), 1 \le i \le m], \tag{10.4}$$

where $m < n$, for several v in a lattice of S_+^{d-1}. This "adaptation to the direction" uses a split of the sample, like in cross-validation for instance. We split the whole sample D_n into a *training sample*

$$D_m := [(X_i, Y_i); 1 \le i \le m]$$

and a *learning sample*

$$D_{(m)} := [(X_i, Y_i); m+1 \le i \le n].$$

The choice of the split size can be quite general (see Section 3 for details). In the numerical study (conducted in Section 4 below), we consider simply $m = 3n/4$ (the learning sample size is a quarter of the whole sample), which provides good results, but other splits can be considered as well.

Using the training sample, we compute a family $\{\bar{g}^{(\lambda)}\,;\,\lambda \in \Lambda\}$ of linear (or *weak*) estimators of the regression g. Each of these estimators depend on a parameter $\lambda = (v, s)$ which make them work based on the data "as if" the true underlying index were v and "as if" the smoothness of the link function were s (in the Hölder sense, see Section 3).

Then, using the learning sample, we compute a weight $w(\bar{g}) \in [0, 1]$ for each $\bar{g} \in \{\bar{g}^{(\lambda)}\,;\,\lambda \in \Lambda\}$, satisfying $\sum_{\lambda\in\Lambda} w(\bar{g}^{(\lambda)}) = 1$. These weights give a level of significance to each weak estimator. Finally, the adaptive, or *aggregated* estimator, is simply the convex combination of the weak estimators:

$$\hat{g} := \sum_{\lambda\in\Lambda} w(\bar{g}^{(\lambda)})\bar{g}^{(\lambda)}.$$

The family of weak estimators consists of univariate local polynomial estimators (LPE), with a data-driven bandwidth that fits locally to the amount of data. In the next section the parameter $\lambda = (v, s)$ is fixed and known, thus we contruct a univariate LPE based on the sample $D_m(v) = [(Z_i, Y_i); 1 \le i \le m] = [(v^\top X_i, Y_i); 1 \le i \le m]$.

2.1. Weak estimators: univariate LPE. The LPE is standard in statistical literature, see for instance [115], among many others. The reason why we consider local polynomials instead of some other method (like smoothing splines, for instance) is theoretical. It is linked with the fact that we need rate-optimal weak estimators under the general design Assumption (D), so that the aggregated estimator is also rate-optimal. We construct an estimator $\bar{f}$ of f based on i.i.d. copies $[(Z_i, Y_i); 1 \le i \le m]$ of a couple $(Z, Y) \in \mathbb{R} \times \mathbb{R}$

such that

$$Y = f(Z) + \sigma(Z)\epsilon, \tag{10.5}$$

where ϵ is standard Gaussian noise independent of Z, $\sigma : \mathbb{R} \to [\sigma_0, \sigma_1] \subset (0, +\infty)$ and $f \in H(s, L)$ where $H(s, L)$ is the set of s-Hölderian functions such that

$$|f^{(\lfloor s \rfloor)}(z_1) - f^{(\lfloor s \rfloor)}(z_2)| \leq L|z_1 - z_2|^{s - \lfloor s \rfloor}$$

for any $z_1, z_2 \in \mathbb{R}$, where $L > 0$ and $\lfloor s \rfloor$ stands for the largest integer smaller than s. This Hölder assumption is standard in nonparametric literature.

Let $r \in \mathbb{N}$ and $h > 0$ be fixed. If z is fixed, we consider the polynomial $\bar{P}_{(z,h)} \in \mathrm{Pol}_r$ (the set of real polynomials with degree at most r) which minimizes in P:

$$\sum_{i=1}^{m} \big(Y_i - P(Z_i - z)\big)^2 \mathbf{1}_{Z_i \in I(z,h)}, \tag{10.6}$$

where $I(z, h) := [z - h, z + h]$ and we define the LPE at z by

$$\bar{f}(z, h) := \bar{P}_{(z,h)}(z).$$

The polynomial $\bar{P}_{(z,h)}$ is well-defined and unique when the symmetrical matrix $\bar{\mathbf{Z}}_m(z, h)$ with entries

$$(\bar{\mathbf{Z}}_m(z, h))_{a,b} := \frac{1}{m\bar{P}_Z[I(z, h)]} \sum_{i=1}^{m} \Big(\frac{Z_i - z}{h}\Big)^{a+b} \mathbf{1}_{Z_i \in I(z,h)} \tag{10.7}$$

for $(a, b) \in \{0, \ldots, R\}^2$ is definite positive, where $\bar{P}_Z$ is the empirical distribution of $(Z_i)_{1 \leq i \leq m}$, given by

$$\bar{P}_Z[A] := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{Z_i \in A} \tag{10.8}$$

for any $A \subset \mathbb{R}$. When $\bar{\mathbf{Z}}_m(z, h)$ is degenerate, we simply take $\bar{f}(z, h) := 0$. The tuning parameter $h > 0$, which is called *bandwidth*, localizes the least square problem around the point z in (10.6). Of course, the choice of h is of first importance in this estimation method (as with any linear method). An important remark is then about the design law. Indeed, the law of $Z = v^\top X$ varies with v strongly: even if P_X is very simple (for instance uniform over some subset of $\mathbb{R}^d$ with positive Lebesgue measure), $P_{v^\top X}$ can be "far" from the uniform law, namely with a density that can vanish at the boundaries of its support, or inside the support, see the examples in Figure 1. This remark motivates the following choice for the bandwidth.

If $f \in H(s, L)$ for known s and L, a "natural" bandwidth, which makes the balance between the bias and the variance of the LPE is given by

$$H_m(z) := \operatorname*{argmin}_{h \in (0,1)} \Big\{ Lh^s \geq \frac{\sigma}{(m\bar{P}_Z[I(z, h)])^{1/2}} \Big\}. \tag{10.9}$$

This bandwidth choice stabilizes the LPE, since it fits point-by-point to the local amount of data. We consider then

$$\bar{f}(z) := \bar{f}(z, H_m(z)), \tag{10.10}$$

for any $z \in \mathbb{R}$, which is in view of Theorem 10.3 (see Section 3) a rate-optimal estimator over $H(s, L)$ in model (10.5).

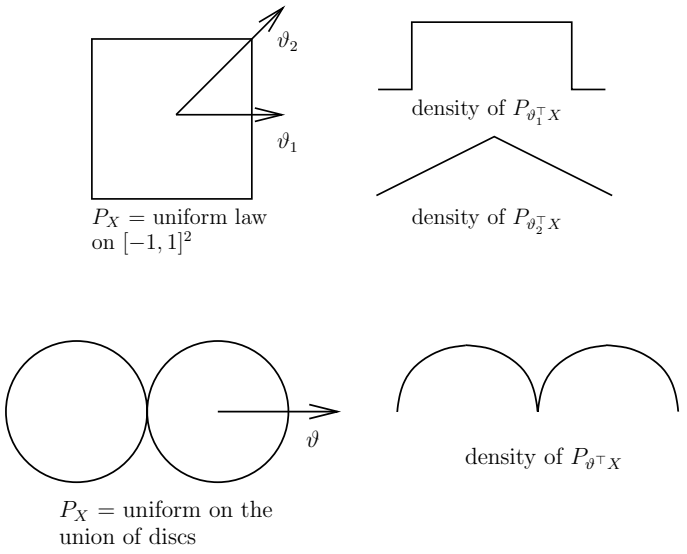

figure 1: Simple design examples

2.2. Adaptation by aggregation. If $\lambda := (v, s)$ is fixed, we consider the LPE $\bar{f}^{(\lambda)}$ given by (10.10), and we take

$$\bar{g}^{(\lambda)}(x) := \tau_Q(\bar{f}^{(\lambda)}(\vartheta^\top x)), \tag{10.11}$$

for any $x \in \mathbb{R}^d$ as an estimator of g, where $\tau_Q(f) := \max(-Q, \min(Q, f))$ is the truncation operator by $Q > 0$. The reason why we need to truncate the weak estimators is related to the theoretical results concerning the aggregation procedure described below, see Theorem 10.4 in Section 3. In order to adapt to the index ϑ and to the smoothness s of the link function, we aggregate the weak estimators from the family $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ with the following algorithm: we take the convex combination

$$\hat{g} := \sum_{\lambda \in \Lambda} w(\bar{g}^{(\lambda)}) \bar{g}^{(\lambda)} \tag{10.12}$$

where for a function $\bar{g} \in \{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$, the weight is given by

$$w(\bar{g}) := \frac{\exp\big(-TR_{(m)}(g)\big)}{\sum_{\lambda \in \Lambda} \exp\big(-TR_{(m)}(\bar{g}^{(\lambda)})\big)}, \tag{10.13}$$

with a *temperature* parameter $T > 0$ and

$$R_{(m)}(\bar{g}) := \sum_{i=m+1}^{n} (Y_i - \bar{g}(X_i))^2, \tag{10.14}$$

which is the empirical least squares of $\bar{g}$ over the training sample (up to a division by the sample size). The set of parameters Λ is given by $\Lambda := \bar{S} \times G$, where G is the grid with step $(\log n)^{-1}$ given by

$$G := \big\{s_{\min}, s_{\min} + (\log n)^{-1}, s_{\min} + 2(\log n)^{-1}, \ldots, s_{\max}\big\} \tag{10.15}$$

The tuning parameters $s_{\min}$ and $s_{\max}$ correspond to the minimum and maximum "allowed" smoothness for the link function: with this choice of G, the aggregated estimator converges with the optimal rate for a link function in $H(s, L)$ for any $s \in [s_{\min}, s_{\max}]$ in view of Theorem 10.1. The set $\bar{S} = \bar{S}_\Delta^{d-1}$ is the regular lattice of the half unit-sphere S_+^{d-1} with step Δ. Namely, $\bar{S}_\Delta^{d-1}$ is such that for any lattitude, any consecutive points in the same

lattitude have distance Δ (if $d \geq 3$, a couple of points in S_+^{d-1} belongs to the same lattitude if they have one common coordinate). The step is taken as

$$\Delta = (n \log n)^{-1/(2s_{\min})}, \tag{10.16}$$

which relies on the minimal allowed smoothness of the link function. For instance, if we want the estimator to be adaptive over Hölder classes of functions at least Lipschitz, we take $\Delta = (n \log n)^{-1/2}$.

We can understand this algorithm in the following way: first, we compute the least squares of each weak estimators. This is the most natural way of assessing the level of significance of some estimator among the other ones. Then, we put a Gibbs law over the set of weak estimators. The mass of each estimator relies on its least squares (over the learning sample). Finally, the aggregate is simply the mean expected estimator according to this law.

REMARK 10.1. *This aggregation algorithm (with Gibbs weights) can be found in [87] in the regression framework, for projection-type weak estimators. Iterative versions of this algorithm can be found in [35], [72], [125]. This aggregation algorithm is also a simplified version of the one from [75]. Indeed, the algorithm proposed therein is a refinement of a stochastic gradient descent, namely a so-called mirror descent in the dual space with averaging, see [75] and [72] for more details. It makes an extra summation of weights relying to the cummulative least squares over the learning sample, that we do not make here.*

If T is small, the weights (10.13) are close to the uniform law over the set of weak estimators, and of course, the resulting aggregate is inaccurate. If T is large, only one weight will equal 1, and the others equal to 0: in this situation, the aggregate is equal to the estimator obtained by empirical risk minimization (ERM). This behavior can be also explained by equation (10.30) in the proof of Theorem 10.4. Indeed, the exponential weights (10.13) realize an optimal tradeoff between the ERM procedure and the uniform weights procedure. The parameter T is somehow a regularization parameter of this tradeoff.

The ERM already gives good results, but if T is chosen carefully, we expect to obtain an estimator which outperforms the ERM. It has been proved theoretically in Chapter 4 that the aggregate outperforms the ERM in the regression framework. This fact is confirmed by the numerical study conducted in Section 4, where the choice of T is done using a simple leave-one-out cross-validation algorithm over the whole sample for aggregates obtained with several T. Namely, we consider the temperature

$$\hat{T} := \operatorname*{argmin}_{T \in \mathcal{T}} \sum_{j=1}^{n} \sum_{i \neq j} \big(Y_i - \hat{g}_{-i}^{(T)}(X_i)\big)^2, \tag{10.17}$$

where $\hat{g}_{-i}^{(T)}$ is the aggregated estimator (10.12) with temperature T, based on the sample $D_n^{-i} = [(X_j, Y_j); j \neq i]$, and where $\mathcal{T}$ is some set of temperatures (in Section 4, we take $\mathcal{T} = \{0.1, 0.2, \ldots, 4.9, 5\}$).

2.3. Reduction of the complexity of the algorithm. The procedure described below requires the computation of the LPE for each parameter $\lambda \in \tilde{\Lambda} := \Lambda \times \mathcal{L}$ (in the simulations, we do also a grid $\mathcal{L}$ over the radius parameter L). Hence, there are $|\bar{S}_\Delta^{d-1}| \times |G| \times |\mathcal{L}|$ LPE to compute. Namely, this is $(\pi/\Delta)^{d-1} \times |G| \times |\mathcal{L}|$, which equals, if $|G| = |\mathcal{L}| = 4$ and $\Delta = (n \log n)^{-1/2}$ (as in the simulation, see Section 4) to 1079 when $d = 2$ and to 72722 when $d = 3$, which is much too large. Hence, the complexity of this

procedure must be reduced: we propose a recursive algorithm which improves strongly the complexity of the estimator. Indeed, most of the coefficients $w(\bar{g}^{(\lambda)})$ are very close to zero (see Figures 6 and 7 in Section 4) when $\lambda = (s, v)$ is such that v is "far" from the true index ϑ. Hence, these coefficients should not be computed at all, since the corresponding weak estimators do not contribute to the aggregated estimator (10.12). Hence, the computation of the lattice should be done iteratively, only around the coefficients which are significative among the other ones. This is done with the following algorithm, which makes a preselection of weak estimators to aggregate ($B^{d-1}(v, \delta)$ stands for the ball in $(\mathbb{R}^d, \|\cdot\|_2)$ centered at v with radius δ and $R_{(m)}(\bar{g})$ is given by (10.14)).

(1) Define $\Delta = (n \log n)^{-1/2}$ and $\Delta_0 = (2d \log n)^{-1/(2(d-1))}$;
(2) compute the lattice $\hat{S} = \bar{S}_{\Delta_0}^{d-1}$;
(3) find the point $\hat{v}$ such that $(\hat{s}, \hat{v}) = \hat{\lambda} = \operatorname{argmin}_{\lambda \in \Lambda} R_{(m)}(\bar{g}^{(\lambda)})$;
(4) divide Δ_0 by 2;
(5) put $\hat{S} = \bar{S}_{\Delta_0}^{d-1} \cap B^{d-1}(\hat{v}, 2^{1+1/(d-1)}\Delta_0)$;
(6) stop if $\Delta_0 \leq \Delta$, otherwise continue with step 3.

When the algorithm exits, $\hat{S}$ is a section of the lattice $\bar{S}_{\Delta}^{d-1}$ centered at $\hat{v}$ with radius $2^{d-1}\Delta$, which contains (with a high probability) the points $v \in \bar{S}_{\Delta}^{d-1}$ corresponding to the largest coefficients $w(\bar{g}^{(\lambda)})$ where $\lambda = (v, s, L) \in \bar{S}_{\Delta}^{d-1} \times G \times \mathcal{L}$. The aggegate is then computed for a set of parameters $\hat{\Lambda} = \hat{S} \times G \times \mathcal{L}$ using (10.12) with weights (10.13). The parameter Δ_0 is chosen so that the surface of $B^{d-1}(v, \Delta_0)$ is $C_d(2d \log n)^{-1/2}$ for any d, which gets larger with the dimension. Moreover, the number of iterations is $O(\log n)$, thus the complexity is much smaller than the full aggregation algorithm. This procedure gives nice empirical results, see Section 4. We give a numerical illustration of the iterative construction of $\hat{S}$ in Figure 2.

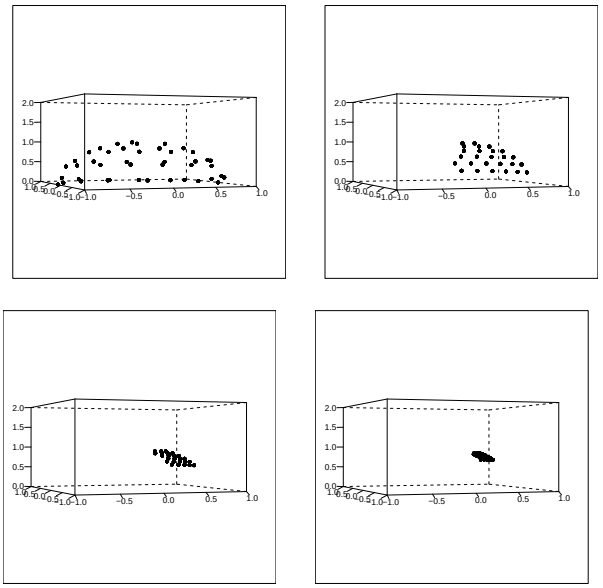

figure 2: Iterative construction of $\hat{S}$

REMARK 10.2. *Most of the weak estimators* $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ *are constructed with a sample* $D_m(v)$ *where* v *is "far" from the true index* ϑ*. Thus, most of these estimators are quite inaccurate, and it is very unlikely to have overfitted estimation in* $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ *(with*

respect to the true sample $D_m(\vartheta)$). This is the reason why we do not add a penalization term in Step 3 of the algorithm.

3. Main results

The error of estimation is measured with the $L^2(P_X)$-norm, defined by

$$\|f\|_{L^2(P_X)} := \Big(\int_{\mathbb{R}^d} f(x)^2 P_X(dx) \Big)^{1/2},$$

where we recall that P_X is the design law. We consider the set $H^Q(s,L) := H(s,L) \cap \{f \mid \|f\|_\infty := \sup_x |f(x)| \le Q\}$. Since we want the adaptive procedure to work whatever $\vartheta \in S_+^{d-1}$ is, we need to work with as general assumptions on the law of $\vartheta^\top X$ as possible. As mentioned in Section 2, even if P_X is simple, $P_{\vartheta^\top X}$ can be quite complicated. The following assumption generalizes the usual assumptions on random designs (when P_X has a density with respect to the Lebesgue measure) that can be met in literature, namely, we do not assume that the design density is bounded away from zero, since even with very simple designs, this assumption is not met (see Figure 1). We say that a real random variable Z satisfies Assumption (D) if:

Assumption (D). *There is a density μ of P_Z with respect to the Lebesgue measure which is continuous. Moreover, we assume that*

- *μ is compactly supported*;
- *There is a finite number of z in the support of μ such that $\mu(z) = 0$;*
- *For any such z, there is an interval $I_z = [z - a_z, z + b_z]$ such that μ is decreasing over $[z - a_z, z]$ and increasing over $[z, z + b_z]$;*
- *There is $\beta \ge 0$ and $\gamma > 0$ such that*

$$P_Z(I) \ge \gamma |I|^{\beta+1}$$

for any I, where $|I|$ stands for the length of I.

This assumption includes any design with continuous density with respect to the Lebesgue measure that can vanish at several points, but not faster than some power function.

3.1. Upper and lower bounds. The next Theorem provides an upper bound for the adaptive estimator constructed in Section 2. For the upper bound to hold, the tuning parameters of the procedure must be as follows: $T > 0$ can be arbitrary (for the proof of the upper bound, but not in practice of course), the choice of the training sample size is quite general: we consider

$$m = [n(1 - \ell_n)], \tag{10.18}$$

where $[x]$ is the integral part of x, and where ℓ_n is a positive sequence such that for all n, $(\log n)^{-\alpha} \le \ell_n < 1$ with $\alpha > 0$. Note that in methods involving data splitting, the optimal choice of the split size is open. The degree r of the LPE and the grid choice G must be such that $s_{\max} \le r + 1$. The upper bound below shows that the estimator converges with the optimal rate for a link function in a whole family of Hölder classes, and for any index. In what follows, E^n stands for the expectation with respect to the joint law P^n of the whole sample D_n.

THEOREM 10.1. *Let $\hat{g}$ be the aggregated estimator given by* (10.12) *with the weights* (10.13). *If for all $\vartheta \in S_+^{d-1}$, $\vartheta^\top X$ satisfies Assumption (D), we have*

$$\sup_{\vartheta \in S_+^{d-1}} \sup_{f \in H^Q(s,L)} E^n \|\hat{g} - g\|^2_{L^2(P_X)} \leq C n^{-2s/(2s+1)},$$

for any $s \in [s_{\min}, s_{\max}]$ when n is large enough, where we recall that $g(\cdot) = f(\vartheta^\top \cdot)$. The constant $C > 0$ depends on $\sigma, L, s_{\min}, s_{\max}$ and P_X only.

Note that $\hat{g}$ does not depend within its construction on the index ϑ, nor the smoothness s of the link function f, nor the design law P_X. In Theorem 10.2 below, we prove in our setting (when Assumption (D) holds on the design) that $n^{-2s/(2s+1)}$ is indeed the minimax rate for a link function in $H(s, L)$ in the single-index model.

THEOREM 10.2. *Let $s, L, Q > 0$ and $\vartheta \in S_+^{d-1}$ be such that $\vartheta^\top X$ satisfies Assumption (D). We have*

$$\inf_{\tilde{g}} \sup_{f \in H^Q(s,L)} E^n \|\tilde{g} - g\|^2_{L^2(P_X)} \geq C' n^{-2s/(2s+1)},$$

where the infimum is taken among all estimators based on data from (10.1),(10.2), *and where $C' > 0$ is a constant depending on σ, s, L and $P_{\vartheta^\top X}$ only.*

Theorem 10.1 and Theorem 10.2 together entail that $n^{-2s/(2s+1)}$ is the minimax rate for the estimation of g in model (10.1) under the constraint (10.2) when the link function belongs to an s-Hölder class. It answers in particular to Question 2 from [110].

3.2. A new result for the LPE. In this section, we give upper bounds for the LPE in the univariate regression model (10.5). Despite the fact that the literature about LPE is wide, the Theorem below is new. It provides a minimax optimal upper bound for the $L^2(P_X)$-integrated risk of the LPE over Hölder balls under Assumption (D), which is a general assumption for random designs (having a density with respect to the Lebesgue measure). This generalization is important in the situation where the univariate explanatory variables Z_i are equal to $\vartheta^\top X_i$ for some $\vartheta \in S_+^{d-1}$, like in the single-index model for instance, see also Figure 1.

In this section, the smoothness s is supposed known and fixed, and we assume that the degree r of the local polynomials satisfies $r + 1 \geq s$. First, we give an upper bound for the pointwise risk conditionally on the design. Then, we derive from it an upper bound for the $L^2(P_Z)$-integrated risk, using standard tools from empirical process theory (see Appendix). Here, E^m stands for the expectation with respect to the joint law P^m of the observations $[(Z_i, Y_i); 1 \leq i \leq m]$. Let us define the matrix

$$\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$$

where $\bar{\mathbf{Z}}_m(z, h)$ is given by (10.7) and $H_m(z)$ is given by (10.9). Let us denote by $\lambda(M)$ the smallest eigenvalue of a matrix M and introduce $Z_1^m := (Z_1, \ldots, Z_m)$.

THEOREM 10.3. *For any $z \in \operatorname{Supp} P_Z$, let $\bar{f}(z)$ be given by* (10.10). *We have on the event $\{\lambda(\bar{\mathbf{Z}}_m(z)) > 0\}$:*

$$\sup_{f \in H(s,L)} E^m \big[(\bar{f}(z) - f(z))^2 | Z_1^m \big] \leq 2\lambda(\bar{\mathbf{Z}}_m(z))^{-2} L^2 H_m(z)^{2s}. \tag{10.19}$$

Moreover, if Z satisfies Assumption (D), we have

$$\sup_{f \in H^Q(s,L)} E^m \big[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \big] \leq C_2 m^{-2s/(2s+1)} \tag{10.20}$$

for m large enough, where we recall that τ_Q is the truncation operator by $Q > 0$ and where $C_2 > 0$ is a constant depending on s, Q, and P_Z only.

REMARK 10.3. *Note that while inequality* (10.19) *in Theorem 10.3 is stated over* $\{\lambda(\bar{\mathbf{Z}}_m(z)) > 0\}$*, which entails existence and unicity of a solution to the linear system* (10.6) *(this inequality is stated conditionally on the design), we only need Assumption (D) for inequality* (10.20) *to hold.*

3.3. Oracle inequality. In this section, we provide an oracle inequality for the aggregation algorithm (10.12) with weights (10.13). This result, which is of independent interest, is stated for a general finite set $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ of deterministic functions such that $\|\bar{g}^{(\lambda)}\|_\infty \leq Q$ for all $\lambda \in \Lambda$. These functions are for instance weak estimators computed with the training sample (or *frozen* sample), which is independent of the learning sample. Let $D := [(X_i, Y_i); 1 \leq i \leq |D|]$ (where $|D|$ stands for the cardinality of D) be an i.i.d. sample of (X, Y) from the multivariate regression model (10.1), where no particular structure like (10.2) is assumed.

The aim of aggregation schemes is to mimic (up to an additive residual) the oracle in $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$. This aggregation framework has been considered, among others, by [17], [35], [74], [87], [98], [114] and [125].

THEOREM 10.4. *The aggregation procedure $\hat{g}$ based on the learning sample D defined by* (10.12) *and* (10.13) *satisfies*

$$E^D \|\hat{g} - g\|^2_{L^2(P_X)} \leq (1+a) \min_{\lambda \in \Lambda} \|\bar{g}^{(\lambda)} - g\|^2_{L^2(P_X)} + \frac{C \log|\Lambda| (\log|D|)^{1/2}}{|D|}$$

for any $a > 0$, where $|\Lambda|$ denotes the cardinality of Λ, where E^D stands for the joint law of D, and where $C := 3[8Q^2(1+a)^2/a + 4(6Q^2 + 2\sigma 2\sqrt{2})(1+a)/3] + 2 + 1/T$.

This theorem is a model-selection type oracle inequality for the aggregation procedure given by (10.12) and (10.13). Sharper oracle inequalities for more general models can be found in [75], where the algorithm used therein requires an extra summation, see Remark 10.1.

REMARK 10.4. *Inspection of the proof shows that the ERM (which is the estimator minimizing the empirical risk $R_{(m)}(g) := \sum_{i=m+1}^n (Y_i - g(X_i))^2$ over all g in $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$) satisfies the same oracle inequality of Theorem 10.4 as the exponential weighted average scheme $\hat{g}$. Nevertheless, it has been proved in Chapter 4 that the ERM is theoretically suboptimal in this framework. The simulation study of Section 4 (especially Figures 3, 4, 5) confirm this suboptimality.*

4. Numerical illustrations

We implemented the procedure described in Section 2 using the `R` software[1]. In order to increase computation speed, we implemented the computation of local polynomials and the bandwidth selection (10.9) in `C` language. We simulate samples from the single index model (10.1),(10.2). We consider Gaussian noise with variance

$$\sigma = \Big[\sum_{1 \leq i \leq n} f(\vartheta^\top X_i)^2 / (n \times \texttt{rsnr}) \Big]^{1/2},$$

where $\texttt{rsnr} = 5$. We consider the following link functions:

[1] see `http://www.r-project.org/`

- `oscsine`$(x) = 4(x+1)\sin(4\pi x^2)$,
- `hardsine`$(x) = 2\sin(1+x)\sin(2\pi x^2+1)$.

The simulations are done with a uniform design on $[-1,1]^d$, with dimensions $d \in \{2,3,4\}$ and we consider several indexes ϑ that makes $P_{\vartheta^\top X}$ not uniform.

In all the computations below, the parameters for the procedure are $\Lambda = \hat{S} \times G \times \mathcal{L}$ where $\hat{S}$ is computed using the algorithm described in Section 2.3 and where $G = \{1,2,3,4\}$ and $\mathcal{L} = \{0.1, 0.5, 1, 1.5\}$. The degree of the local polynomials is $r = 5$. The learning sample has size $[n/4]$, and is chosen randomly in the whole sample. We do not use a jackknife procedure (that is, the average of estimators obtained with several learning subsamples), since the results are stable enough (at least when $n \geq 100$) when we consider only one learning sample.

In Tables 1, 2, 3 and Figures 3, 4, 5, we show the mean MISE for 100 replications and its standard deviation for several Gibbs temperatures, several sample sizes and indexes. These results give the empirical proof that the aggregated estimator outperforms the ERM (which is computed as the aggregated estimator with a large temperature $T = 30$) since in each case, the aggregated estimator with cross-validated temperature (`aggCVT`, given by (10.17), with $\mathcal{T} = \{0.1, 0.2, \ldots, 4.9, 5\}$), has a MISE up to three times smaller than the MISE of the ERM. Moreover, `aggCVT` is more stable than the ERM in view of the standard deviations (in brackets). Note also that as expected, the dimension parameter has no impact on the accuracy of estimation: the mises are barely the same when $d = 2, 3, 4$.

The aim of Figures 6 and 7 is to give an illustration of the aggregation phenomenon. In these figures, we show the weights obtained for a single run, using the aggregation procedure with the parameter set $\Lambda = \bar{S}_\Delta^{d-1} \times \{3\} \times \{1\}$ (that is, $s = 3$ and $L = 1$ are fixed and we do not use the reduction of complexity algorithm). These figures motivates indeed the use the reduction of complexity algorithm, since only weights corresponding to a point of $\bar{S}_\Delta^{d-1}$ which is close to the true index are significant (at least numerically). Finally, we show typical realisations for several index functions, indexes and sample sizes in Figures 8, 9, 10, 11.

TABLE 1. Mise against the Gibbs temperature ($f =$ `hardsine`, $d = 2$, $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$.)

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.026	0.017	0.015	**0.014**	**0.014**	0.015	0.034	0.015
	(.009)	(.006)	(.006)	(.005)	(.005)	(.006)	(.018)	(.005)
n = 200	0.015	0.009	**0.008**	**0.008**	0.009	0.011	0.027	0.009
	(.004)	(.002)	(.003)	(.003)	(.005)	(.007)	(.014)	(.004)
n = 400	0.006	0.005	**0.004**	0.005	0.006	0.007	0.016	0.005
	(.001)	(.001)	(.001)	(.001)	(.002)	(.002)	(.003)	(.002)

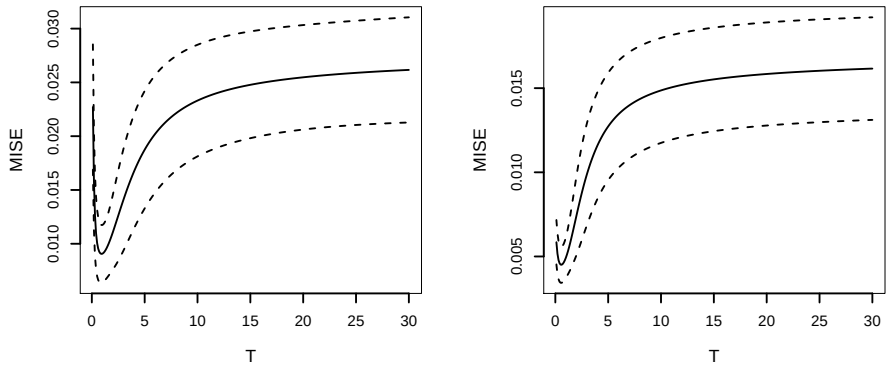

figure 3: MISE against the Gibbs temperature for $f =$ `hardsine`, $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $n = 200, 400$ (solid line = mean of the MISE for 100 replications, dashed line = mean MISE $\pm$ standard deviation.)

TABLE 2. Mise against the Gibbs temperature ($f =$ `hardsine`, $d = 3$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$).

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.029	0.021	0.019	0.018	**0.017**	0.018	0.037	0.020
	(.011)	(.008)	(.008)	(0.007)	(.008)	(.009)	(.022)	(.008)
n = 200	0.016	0.010	0.010	**0.009**	**0.009**	0.010	0.026	0.010
	(.005)	(.003)	(.003)	(.002)	(.002)	(.003)	(0.008)	(.003)
n = 400	0.007	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.003)	(.001)

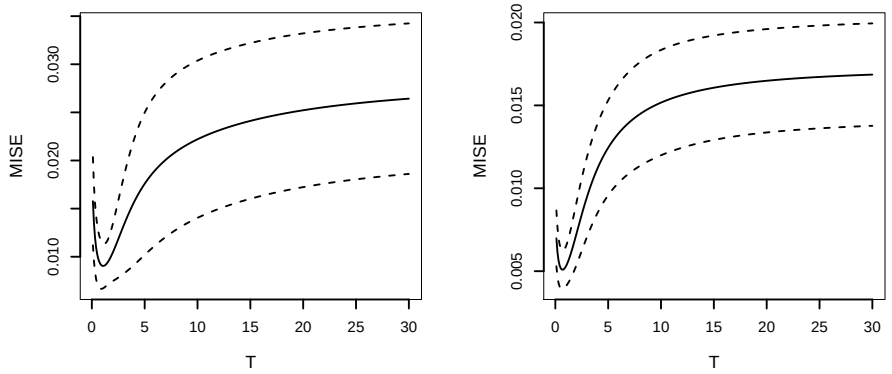

figure 4: MISE against the Gibbs temperature for $f =$ `hardsine`, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $n = 200, 400$ (solid line = mean of the MISE for 100 replications, dashed line = mean MISE $\pm$ standard deviation.)

TABLE 3. Mise against the Gibbs temperature (f = `hardsine`, $d = 4$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$)

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.038	0.027	0.021	0.019	**0.017**	**0.017**	0.038	0.020
	(.016)	(.010)	(.009)	(.008)	(.007)	(.007)	(.025)	(.010)
n = 200	0.019	0.013	**0.012**	**0.012**	0.013	0.014	0.031	0.013
	(.014)	(.009)	(.010)	(.011)	(.012)	(.012)	(.016)	(.010)
n = 400	0.009	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.004)	(.001)

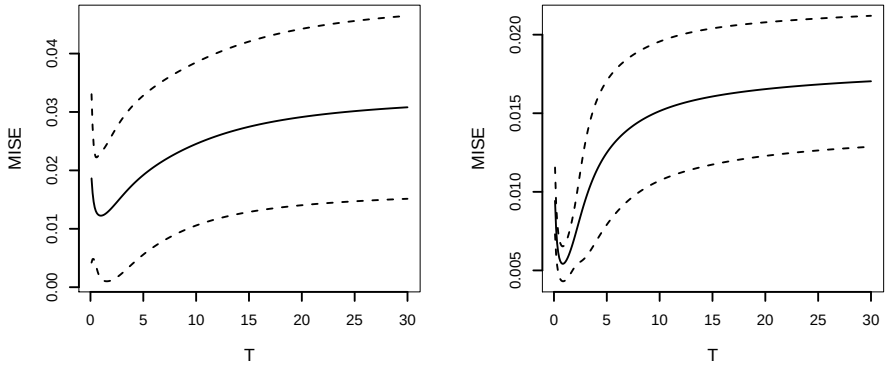

figure 5: MISE against the Gibbs temperature for f = `hardsine`, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $n = 200, 400$.

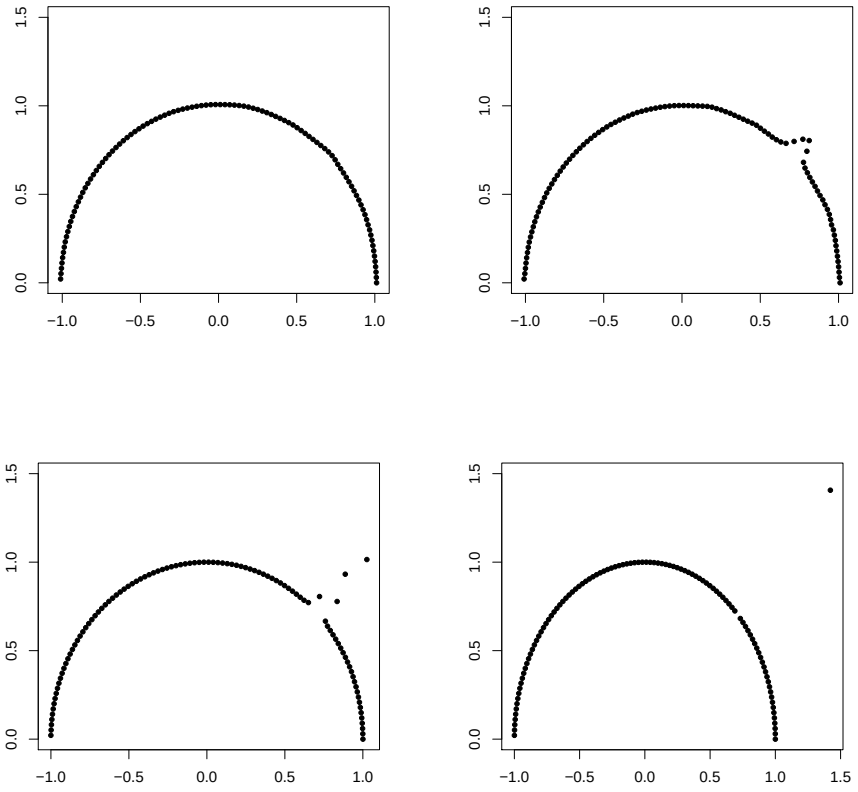

figure 6: Weights (for a single run) at each points of the lattice $\bar{S}^1_\Delta$ for $\Delta = 0.03$, $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ and $T = 0.05, 0.2, 0.5, 10$ (from top to bottom and left to right).

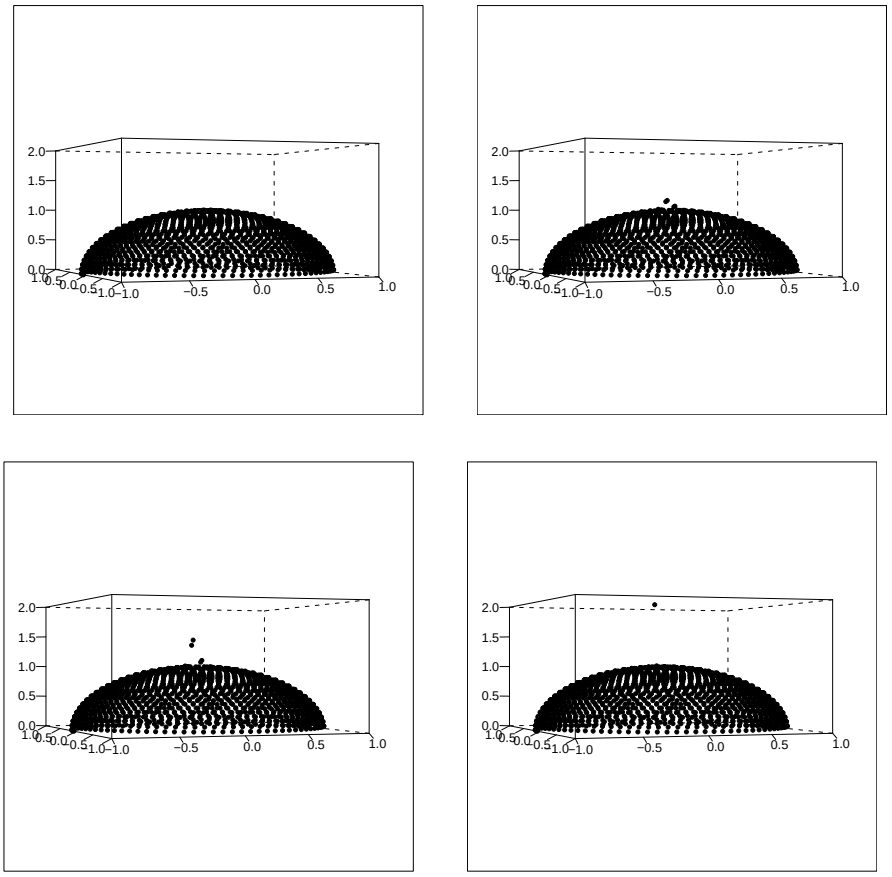

figure 7: Weights (for a single run) at each points of the lattice $\bar{S}^2_\Delta$ for $\Delta = 0.07$, $\vartheta = (0, 0, 1)$, and $T = 0.05, 0.3, 0.5, 10$ (from top to bottom and left to right).

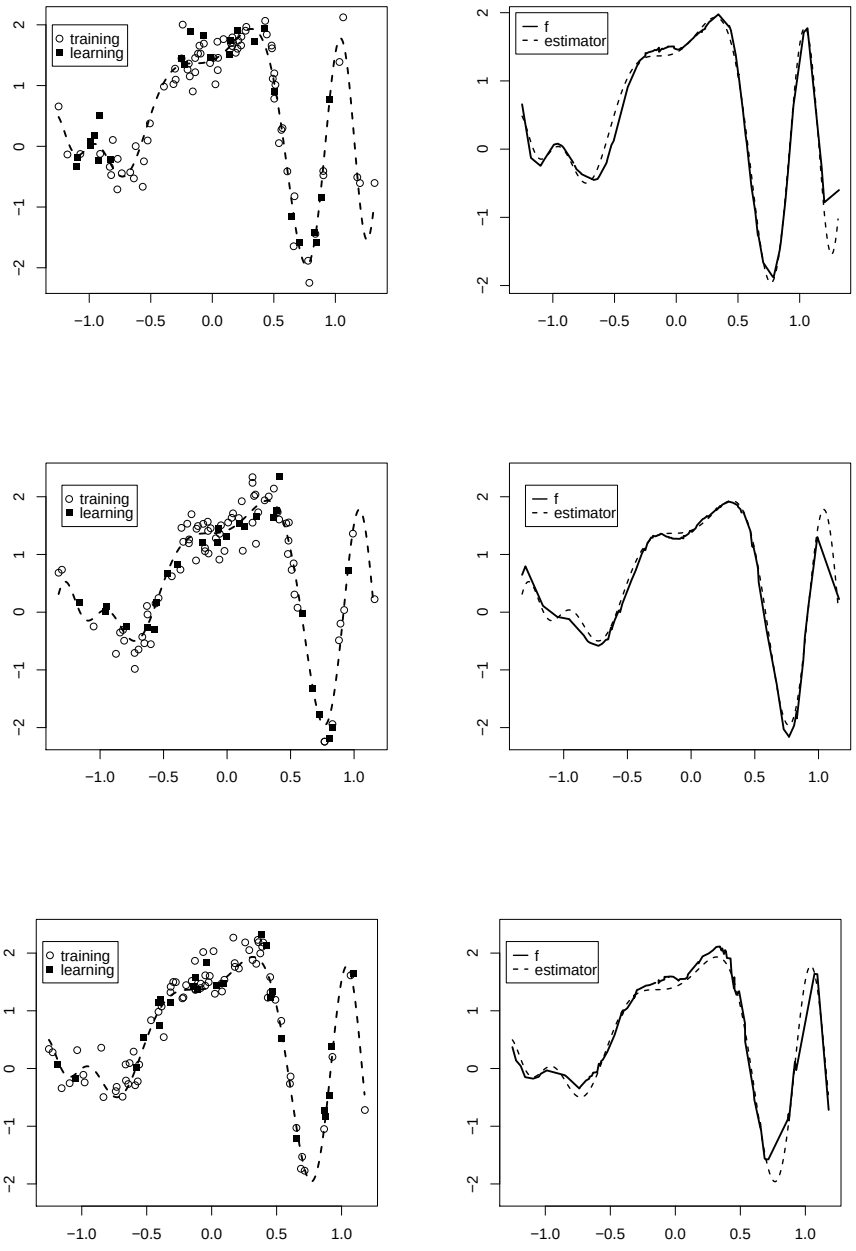

figure 8: Simulated datasets and aggregated estimators with cross-validated temperature for $f = \texttt{hardsine}$, $n = 100$, and indexes $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ from top to bottom.

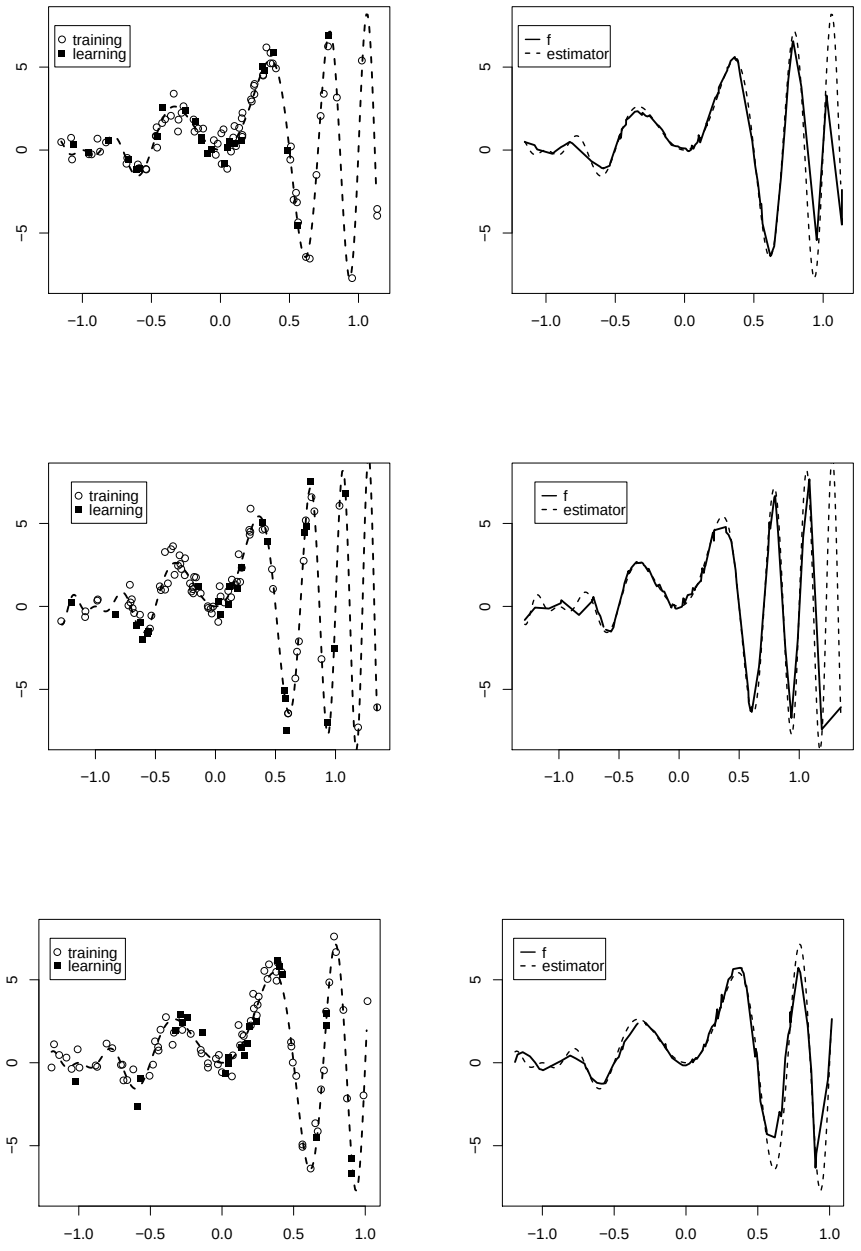

figure 9: Simulated datasets and aggregated estimators with cross-validated temperature for $f =$ `oscsine`, $n = 100$, and indexes $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ from top to bottom.

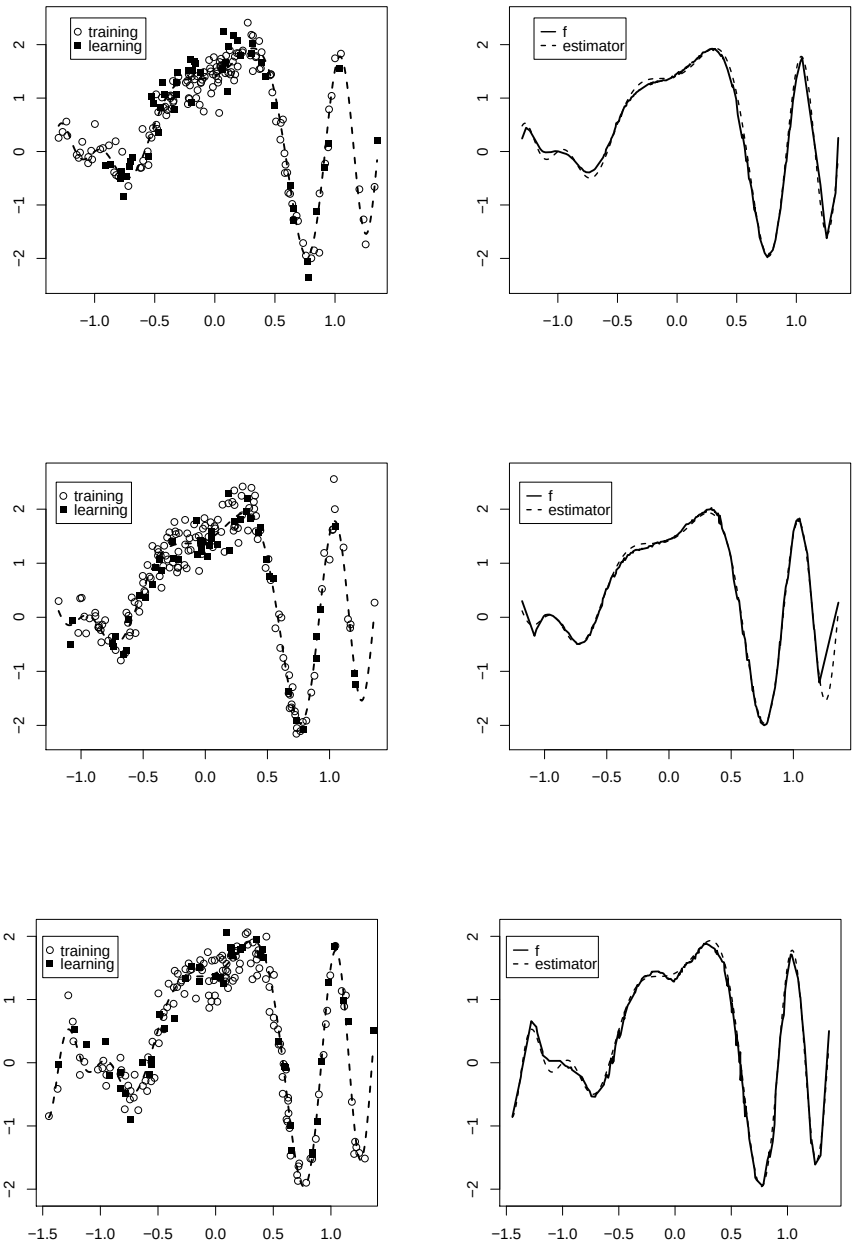

figure 10: Simulated datasets and aggregated estimators with cross-validated temperature for $f = \texttt{hardsine}$, $n = 200$, and indexes $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ from top to bottom.

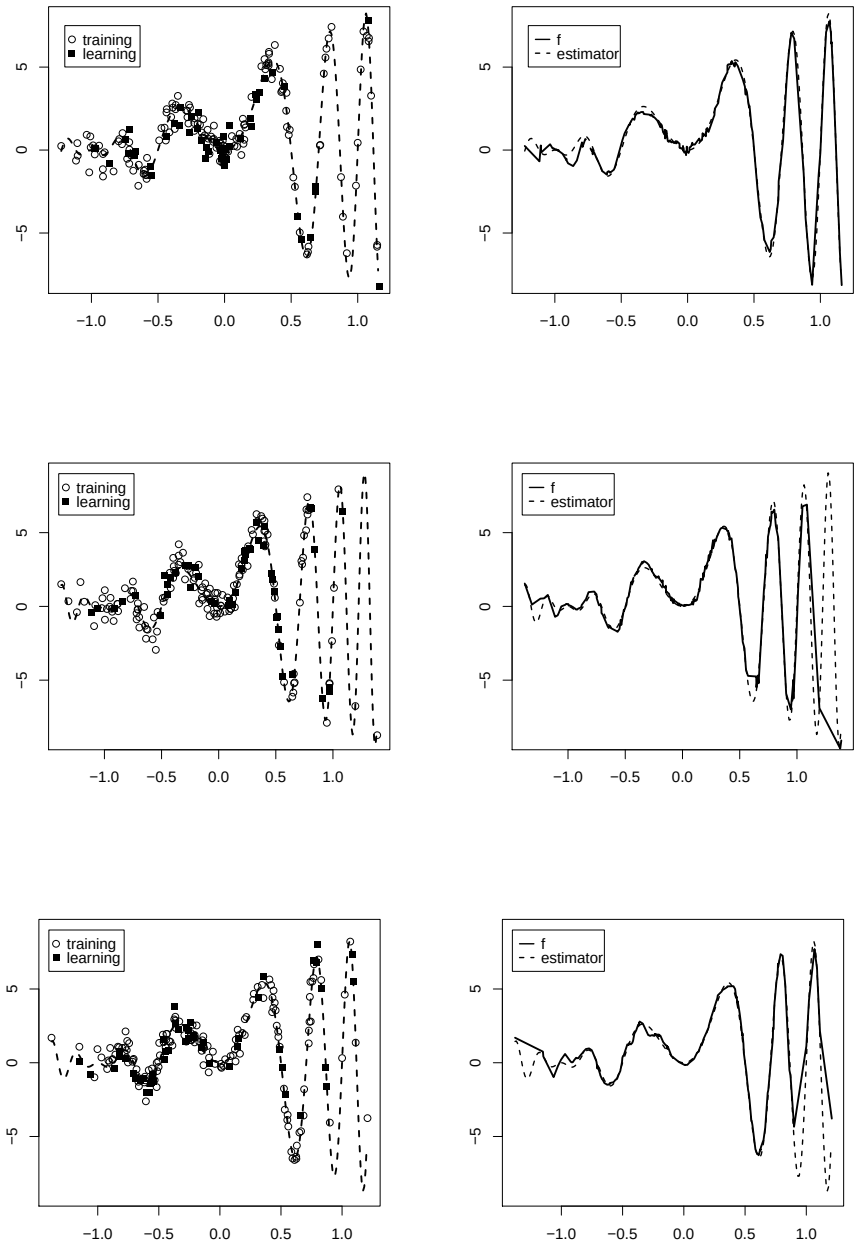

figure 11: Simulated datasets and aggregated estimators with cross-validated temperature for $f = \texttt{oscsine}$, $n = 200$, and indexes $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ from top to bottom.

5. Proofs

Proof of Theorem 10.1. First, we use Theorem 10.4. The functions $\bar{g}^{(\lambda)}$ are given by (10.11). They are computed based on the training (or "frozen") sample D_m, which is independent of the learning sample $D_{(m)}$. If $E^{(m)}$ denotes the integration with respect to the joint law of $D_{(m)}$, we obtain using Theorem 10.4:

$$E^{(m)}\|\hat{g}-g\|^2_{L^2(P_X)} \le (1+a)\min_{\lambda\in\Lambda}\|\bar{g}^{(\lambda)}-g\|^2_{L^2(P_X)} + \frac{C\log|\Lambda|(\log|D_{(m)}|)^{1/2}}{|D_{(m)}|}$$
$$\le (1+a)\|\bar{g}^{(\bar{\lambda})}-g\|^2_{L^2(P_X)} + o(n^{-2s/(2s+1)}),$$

since $\log|\Lambda|(\log|D_{(m)}|)^{1/2}/|D_{(m)}| \le d(\log n)^{3/2+\gamma}/(2\tau n)$ (see (10.18) and (10.16)), and where $\bar{\lambda}=(\bar{\vartheta},\bar{s})\in\Lambda$ is such that $\|\bar{\vartheta}-\vartheta\|_2\le\Delta$ and $\lfloor\bar{s}\rfloor=\lfloor s\rfloor$ with $s\in[\bar{s},\bar{s}+(\log n)^{-1}]$. By integration with respect to P^m, we obtain

$$E^n\|\hat{g}-g\|^2_{L^2(P_X)} \le (1+a)E^m\|\bar{g}^{(\bar{\lambda})}-g\|^2_{L^2(P_X)} + o(n^{-2s/(2s+1)}). \tag{10.21}$$

The choice of $\bar{\lambda}$ entails $H^Q(s,L)\subset H^Q(\bar{s},\bar{L})$ and

$$n^{-2\bar{s}/(2\bar{s}+1)} \le e^{1/2}n^{-2s/(2s+1)}.$$

Thus, together with (10.21), the Theorem follows if we prove that

$$\sup_{f\in H^Q(\bar{s},\bar{L})} E^m\|\bar{g}^{(\bar{\lambda})}-g\|^2_{L^2(P_X)} \le Cm^{-2\bar{s}/(2\bar{s}+1)}. \tag{10.22}$$

for n large enough, where $C>0$. We cannot use directly Theorem 10.3 to prove this, since the weak estimator $\bar{g}^{(\bar{\lambda})}$ works based on data $D_m(\bar{\vartheta})$ (see (10.4)) while the true index is ϑ. In order to simplify notations, we replace the dependence upon $\bar{\lambda}$ by $\bar{\vartheta}$, since in the upper bound (10.22), the estimator uses the "correct" smoothness parameter $\bar{s}$. We have

$$E^m\|\bar{g}^{(\bar{\vartheta})}-g\|^2_{L^2(P_X)} \le 2\big(E^m\|\bar{g}^{(\bar{\vartheta})}(\cdot)-f(\bar{\vartheta}^\top\cdot)\|^2_{L^2(P_X)} + \|f(\bar{\vartheta}^\top\cdot)-f(\vartheta^\top\cdot)\|^2_{L^2(P_X)}\big)$$

and using together (10.16) and fact that $f\in H^Q(s,L)$ for $s\ge\tau$, we obtain

$$\|f(\bar{\vartheta}^\top\cdot)-f(\vartheta^\top\cdot)\|^2_{L^2(P_X)} \le L^2\int\|x\|_2^{2\tau}P_X(dx)\Delta^{2\tau} \le C(n\log n)^{-1}.$$

Let us denote by $Q_\vartheta(\cdot|X_1^m)$ the joint law of $(X_i,Y_i)_{1\le i\le m}$ from model (10.1) (when the index is ϑ) conditional on the $(X_i)_{1\le i\le m}$, which is given by

$$Q_\vartheta(dy_1^m|x_1^m) := \prod_{i=1}^m \frac{1}{(\sigma(x_i)(2\pi)^{1/2})}\exp\Big(-\frac{(y_i-f(\vartheta^\top x_i))^2}{2\sigma(x_i)^2}\Big)dy_i.$$

Under $Q_{\bar{\vartheta}}(\cdot|X_1^m)$, we have

$$L_X(\vartheta,\bar{\vartheta}) := \frac{dQ_\vartheta(\cdot|X_1^m)}{dQ_{\bar{\vartheta}}(\cdot|X_1^m)}$$
$$\overset{\text{(law)}}{=} \exp\Big(-\sum_{i=1}^m\frac{\epsilon_i(f(\bar{\vartheta}^\top X_i)-f(\vartheta^\top X_i))}{\sigma(X_i)} - \frac{1}{2}\sum_{i=1}^m\frac{(f(\bar{\vartheta}^\top X_i)-f(\vartheta^\top X_i))^2}{\sigma(X_i)^2}\Big).$$

Hence, if P_X^m denotes the joint law of $(X_1,\dots,X_m)$,

$$E^m\|\bar{g}^{(\bar{\vartheta})}(\cdot)-f(\bar{\vartheta}^\top\cdot)\|^2_{L^2(P_X)}$$
$$= \int\int\|\bar{g}^{(\bar{\vartheta})}(\cdot)-f(\bar{\vartheta}^\top\cdot)\|^2_{L^2(P_X)}L_X(\vartheta,\bar{\vartheta})dQ_{\bar{\vartheta}}(\cdot|X_1^m)dP_X^m$$

$$\leq C \int\int \|\bar{f}^{(\bar{\vartheta})}(\bar{\vartheta}^\top \cdot) - f(\bar{\vartheta}^\top \cdot)\|^2_{L^2(P_X)} dQ_{\bar{\vartheta}}(\cdot|X_1^m) dP_X^m \tag{10.23}$$
$$+ 4Q^2 \int\int L_X(\vartheta, \bar{\vartheta}) \mathbf{1}_{\{L_X(\vartheta,\bar{\vartheta}) \geq C\}} dQ_{\bar{\vartheta}}(\cdot|X_1^m) dP_X^m,$$

where we decomposed the integrand over $\{L_X(\vartheta, \bar{\vartheta}) \geq C\}$ and $\{L_X(\vartheta, \bar{\vartheta}) \leq C\}$ for some constant $C \geq 3$, and where we used the fact that $\|\bar{g}^{(\bar{\vartheta})}\|_\infty, \|f\|_\infty \leq Q$. Under $Q_{\bar{\vartheta}}(\cdot|X_1^m)$, the (X_i, Y_i) have the same law as (X, Y) from model (10.1) where the index is $\bar{\vartheta}$. Moreover, we assumed that $P_{\bar{\vartheta}^\top X}$ satisfies Assumption (D). Hence, Theorem 10.3 entails that, uniformly for $f \in H^Q(\bar{s}, \bar{L})$,

$$\int\int \|\bar{f}^{(\bar{\vartheta})}(\bar{\vartheta}^\top \cdot) - f(\bar{\vartheta}^\top \cdot)\|^2_{L^2(P_X)} dQ_{\bar{\vartheta}}(\cdot|X_1^m) dP_X^m \leq C' m^{-2\bar{s}/(2\bar{s}+1)}.$$

Moreover, the second term in the right hand side of (10.23) is smaller than

$$4Q^2 \int \Big(\int L_X(\vartheta, \bar{\vartheta})^2 dQ_{\bar{\vartheta}}(\cdot|X_1^m) \Big)^{1/2} Q_{\bar{\vartheta}}\big[L_X(\vartheta, \bar{\vartheta}) \geq C|X_1^m\big]^{1/2} dP_X^m.$$

Using together (10.16), the fact that $f \in H^Q(s, L)$ for $s \geq \tau$, the fact that P_X is compactly supported and the fact that $\sigma(X) > \sigma_0$ a.s., we obtain

$$\int L_X(\vartheta, \bar{\vartheta})^2 dQ_{\bar{\vartheta}}(\cdot|X_1^m) \leq \exp\Big(\frac{1}{2} \sum_{i=1}^m \frac{(f(\bar{\vartheta}^\top X_i) - f(\vartheta^\top X_i))^2}{\sigma(X_i)^2}\Big) \leq 1$$

P_X^m-a.s. when m is large enough. Moreover, we have with the same arguments

$$Q_{\bar{\vartheta}}\big[L_X(\vartheta, \bar{\vartheta}) \geq C|X_1^m\big] \leq m^{-(\log C)^2/2} \leq m^{-4\bar{s}/(2\bar{s}+1)}$$

for C large enough, where we use the standard Gaussian deviation $P[N(0, b^2) \geq a] \leq \exp(-a^2/(2b^2))$. This concludes the proof of Theorem 10.1. □

Proof of Theorem 10.2. We want to bound the minimax risk

$$\inf_{\hat{g}_n} \sup_{f \in H^Q(s,L)} E^n \int \big(\hat{g}_n(x) - f(\vartheta^\top x)\big)^2 P_X(dx) \tag{10.24}$$

from below, where the infimum is taken among all estimators $\mathbb{R}^d \to \mathbb{R}$ based on data from model (10.1),(10.2). We recall that $\vartheta^\top X$ satisfies Assumption (D). We consider $\vartheta^{(2)}, \ldots, \vartheta^{(d)}$ in $\mathbb{R}^d$ such that $(\vartheta, \vartheta^{(2)}, \ldots, \vartheta^{(d)})$ is an orthogonal basis of $\mathbb{R}^d$. We denote by $\mathbf{O}$ the matrix with columns $\vartheta, \vartheta^{(2)}, \ldots, \vartheta^{(d)}$. We define $Y := \mathbf{O}X = (Y^{(1)}, \ldots, Y^{(d)})$ and $Y_2^d := (Y^{(2)}, \ldots, Y^{(d)})$. By a change of variable, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d} \big(\hat{g}_n(x) - f(\vartheta^\top x)\big)^2 P_X(dx) \\
&\quad = \int_{\mathbb{R}^d} \big(\hat{g}_n(\mathbf{O}^{-1}y) - f(y^{(1)})\big)^2 P_Y(dy) \\
&\quad = \int_{\mathbb{R}} \int_{\mathbb{R}^{d-1}} \big(\hat{g}_n(\mathbf{O}^{-1}y) - f(y^{(1)})\big)^2 P_{Y_2^d|Y^{(1)}}(dy_2^d|y^{(1)}) P_{Y^{(1)}}(dy^{(1)}) \\
&\quad \geq \int_{\mathbb{R}} \big(\hat{f}_n(y^{(1)}) - f(y^{(1)})\big)^2 P_{\vartheta^\top X}(dy^{(1)}),
\end{aligned}$$

where $\hat{f}_n(y^{(1)}) := \int \hat{g}_n(\mathbf{O}^{-1}y) P_{Y_2^d|Y^{(1)}}(dy_2^d|y^{(1)})$. Hence, if $Z := \vartheta^\top X$, (10.24) is larger than

$$\inf_{\hat{f}_n} \sup_{f \in H^Q(s,L)} E^n \int \big(\hat{f}_n(z) - f(z)\big)^2 P_Z(dz), \tag{10.25}$$

where the infimum is taken among all estimators $\mathbb{R} \to \mathbb{R}$ based on data from model (10.1) with $d = 1$ (univariate regression). In order to bound (10.25) from below, we use the following Theorem, from [115], which is a standard tool for the proof of such a lower bound. We say that ∂ is a *semi-distance* on some set Θ if it is symmetric, if it satisfies the triangle inequality and if $\partial(\theta, \theta) = 0$ for any $\theta \in \Theta$. We consider $K(P|Q) := \int \log(\frac{dP}{dQ})dP$ the Kullback-Leibler divergence between probability measures P and Q.

THEOREM 10.5. *Let (Θ, ∂) be a set endowed with a semi-distance ∂. We suppose that $\{P_\theta; \theta \in \Theta\}$ is a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and that $(v_n)_{n\in\mathbb{N}}$ is a sequence of positive numbers. If there exist $\{\theta_0, \dots, \theta_M\} \subset \Theta$, with $M \geq 2$, such that*

- $\partial(\theta_j, \theta_k) \geq 2v_n \quad \forall 0 \leq j < k \leq M$
- $P_{\theta_j} \ll P_{\theta_0} \quad \forall 1 \leq j \leq M$,
- $\frac{1}{M}\sum_{j=1}^{M} K(P^n_{\theta_j}|P^n_{\theta_0}) \leq \alpha \log M$ *for some* $\alpha \in (0, 1/8)$,

then

$$\inf_{\hat\theta_n} \sup_{\theta\in\Theta} E^n_\theta[(v_n^{-1}\partial(\hat\theta_n, \theta))^2] \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log M}}\right),$$

where the infimum is taken among all estimators based on a sample of size n.

Let us define $m := \lfloor c_0 n^{1/(2s+1)} \rfloor$, the largest integer smaller than $c_0 n^{1/(2s+1)}$, where $c_0 > 0$. Let $\varphi : \mathbb{R} \to [0, +\infty)$ be a function in $H^Q(s, 1/2; \mathbb{R})$ with support in $[-1/2, 1/2]$. We take $h_n := m^{-1}$ and $z_k := (k - 1/2)/m$ for $k \in \{1, \dots, m\}$. For $\omega \in \Omega := \{0, 1\}^m$, we consider the functions

$$f(\cdot; \omega) := \sum_{k=1}^{m} \omega_k \varphi_k(\cdot) \quad \text{where} \quad \varphi_k(\cdot) := Lh_n^s \varphi\Big(\frac{\cdot - z_k}{h_n}\Big).$$

We have

$$\begin{aligned}\|f(\cdot; \omega) - f(\cdot; \omega')\|_{L^2(P_Z)} &= \Big(\sum_{k=1}^{m} (\omega_k - \omega_{k'})^2 \int \varphi_k(z)^2 P_Z(dz)\Big)^{1/2} \\ &\geq \mu_0^{1/2} \rho(\omega, \omega') L^2 h_n^{2s+1} \int_{S_\mu} \varphi(u)^2 du,\end{aligned}$$

where $S_\mu := \operatorname{Supp} P_Z - \cup_z [a_z, b_z]$ (the union is over the z such that $\mu(z) = 0$, see Assumption (D)), where $\mu_0 := \min_{z\in S_\mu} \mu(z) > 0$ and where

$$\rho(\omega, \omega') := \sum_{k=1}^{m} \mathbf{1}_{\omega_k \neq \omega'_k}$$

is the Hamming distance on Ω. Using a result of Varshamov-Gilbert (see [115]) we can find a subset $\{\omega^{(0)}, \dots, \omega^{(M)}\}$ of Ω such that $\omega^{(0)} = (0, \dots, 0)$, $\rho(\omega^{(j)}, \omega^{(k)}) \geq m/8$ for any $0 \leq j < k \leq M$ and $M \geq 2^{m/8}$. Therefore, we have

$$\|f(\cdot; \omega^{(j)}) - f(\cdot; \omega^{(k)})\|_{L^2(P_Z)} \geq D n^{-s/(2s+1)},$$

where $D = \mu_0^{1/2} \int_{S_\mu} \varphi(u)^2 du/(8c_0^{2s}) \geq 2$ for c_0 small enough. Moreover,

$$\begin{aligned}\frac{1}{M}\sum_{k=1}^{M} K(P^n_{f(\cdot,\omega^{(0)})}|P^n_{f(\cdot,\omega^{(k)})}) &\leq \frac{n}{2M\sigma_0^2}\sum_{k=1}^{M} \|f(\cdot; \omega^{(0)}) - f(\cdot; \omega^{(k)})\|^2_{L^2(P_Z)} \\ &\leq \frac{n}{2\sigma_0^2} L^2 h_n^{2s+1} \|\varphi\|_2^2 m \leq \alpha \log M,\end{aligned}$$

where $\alpha := (L^2\|\varphi\|_2^2)/(\sigma^2 c_0^{2s+1}\log 2) \in (0, 1/8)$ for c_0 small enough. The conclusion follows from Theorem 10.5. □

Proof of Theorem 10.3. We recall that $R = \lfloor s \rfloor$ is the largest integer smaller than s, and that $\lambda(M)$ stands for the smallest eigenvalue of a matrix M.

Proof of (10.19). First, we prove a bias-variance decomposition of the LPE at a fixed point $z \in \operatorname{Supp} P_Z$. This kind of result is commonplace, see for instance [115]. We introduce the following weighted pseudo-inner product, for fixed $z \in \mathbb{R}$ and $h > 0$, as

$$\langle f, g\rangle_h := \frac{1}{m\bar{P}_Z[I(z,h)]}\sum_{i=1}^m f(Z_i)g(Z_i)\mathbf{1}_{Z_i\in I(z,h)},$$

where we recall that $I(z,h) = [z-h, z+h]$, and that $\bar{P}_Z$ is given by (10.8), and we consider the associated pseudo-norm $\|g\|_h^2 := \langle g, g\rangle_h$. We introduce the power functions $\varphi_a(\cdot) := ((\cdot - x)/h)^a$ for $a \in \{0, \dots, R\}$, which satisfy $\|\varphi_a\|_h \le 1$.

Note that the entries of the matrix $\bar{\mathbf{Z}}_m = \bar{\mathbf{Z}}_m(z,h)$ (see (10.7)) satisfy $(\bar{\mathbf{Z}}_m(z,h))_{a,b} := \langle\varphi_a, \varphi_b\rangle_h$ for $(a,b) \in \{0,\dots,R\}^2$. Hence, (10.6) is equivalent to find $\bar{P} \in \mathrm{Pol}_R$ such that

$$\langle \bar{P}, \varphi_a\rangle_h = \langle Y, \varphi_a\rangle_h \tag{10.26}$$

for any $a \in \{0,\dots,R\}$, where $\langle Y, \varphi\rangle_h := (m\bar{P}_Z[I(z,h)])^{-1}\sum_{i=1}^m Y_i\varphi(Z_i)\mathbf{1}_{Z_i\in I(z,h)}$. In other words, $\bar{P}$ is the projection of Y onto Pol_R with respect to the inner product $\langle\cdot,\cdot\rangle_h$. For $e_1 := (1, 0, \dots, 0) \in \mathbb{R}^{R+1}$, we have

$$\bar{f}(z) - f(z) = e_1^\top \bar{\mathbf{Z}}_m^{-1}\bar{\mathbf{Z}}_m(\bar{\theta} - \theta)$$

whenever $\lambda(\bar{\mathbf{Z}}_m) > 0$, where $\bar{\theta}$ is the coefficient vector of $\bar{P}$ and θ is the coefficient vector of the Taylor polynomial P of f at z with degree R. In view of (10.26):

$$(\bar{\mathbf{Z}}_m(\bar{\theta} - \theta))_a = \langle\bar{P} - P, \varphi_a\rangle_h = \langle Y - P, \varphi_a\rangle_h,$$

thus $\bar{\mathbf{Z}}_m(\bar{\theta} - \theta)) = B + V$ where $(B)_a := \langle f - P, \varphi_a\rangle_h$ and $(V)_a := \langle\sigma(\cdot)\xi, \varphi_a\rangle_h$. The bias term satisfies $|e_1^\top\bar{\mathbf{Z}}_m^{-1}B| \le (R+1)^{1/2}\|\bar{\mathbf{Z}}_m^{-1}\|\|B\|_\infty$ where for any $a \in \{0,\dots,R\}$

$$|(B)_a| \le \|f - P\|_h \le Lh^s/R!.$$

Let $\bar{\mathbf{Z}}_m^\sigma$ be the matrix with entries $(\bar{\mathbf{Z}}_m^\sigma)_{a,b} := \langle\sigma(\cdot)\varphi_a, \sigma(\cdot)\varphi_b\rangle_h$. Since V is, conditionally on $Z_1^m = (Z_1, \dots, Z_m)$, centered Gaussian with covariance matrix $(m\bar{P}_Z[I(z,h)])^{-1}\bar{\mathbf{Z}}_m^\sigma$, $e_1^\top\bar{\mathbf{Z}}_m^{-1}V$ is centered Gaussian with variance smaller than

$$(m\bar{P}_Z[I(z,h)])^{-1}e_1^\top\bar{\mathbf{Z}}_m^{-1}\bar{\mathbf{Z}}_m^\sigma\bar{\mathbf{Z}}_m^{-1}e_1 \le \sigma^2(m\bar{P}_Z[I(z,h)])^{-1}\lambda(\bar{\mathbf{Z}}_m)^{-1}$$

where we used $\sigma(\cdot) \le \sigma$. Hence, if $C_R := (R+1)^{1/2}/R!$, we obtain

$$E^m[(\bar{f}(z) - f(z))^2|Z_1^m] \le \lambda(\bar{\mathbf{Z}}_m(z,h))^{-2}\big(C_R L h^s + \sigma(m\bar{P}_Z[I(z,h)])^{-1/2}\big)^2$$

for any z, and the bandwidth choice (10.9) entails (10.19).

Proof of (10.20)**.** Let us consider the sequence of positive curves $h_m(\cdot)$ defined as the point-by-point solution to

$$Lh_m(z)^s = \frac{\sigma}{(mP_Z[I(z, h_m(z))])^{1/2}} \tag{10.27}$$

for all $z \in \operatorname{Supp} P_Z$, where we recall $I(z,h) = [z-h, z+h]$, and let us define

$$r_m(z) := Lh_m(z)^s.$$

The sequence $h_m(\cdot)$ is the deterministic equivalent to the bandwidth $H_m(\cdot)$ given by (10.9). Indeed, with a large probability, $H_m(\cdot)$ and $h_m(\cdot)$ are close to each other in view of Lemma 10.1 below. Under Assumption (D) we have $P_Z[I] \geq \gamma |I|^{\beta+1}$, which entails together with (10.27) that

$$h_m(z) \leq D m^{-1/(1+2s+\beta)} \tag{10.28}$$

uniformly for $z \in \operatorname{Supp} P_Z$, where $D = (\sigma/L)^{2/(1+2s+\beta)} (\gamma 2^{\beta+1})^{-1/(1+2s+\beta)}$. Moreover, since P_Z has a continuous density μ with respect to the Lebesgue measure, we have

$$h_m(z) \geq D m^{-1/(1+2s)} \tag{10.29}$$

uniformly for $z \in \operatorname{Supp} P_Z$, where $D = (\sigma/L)^{2/(1+2s)} (2\mu_\infty)^{-1/(2s+1)}$. We recall that P_Z^m stands for the joint law of $(Z_1, \ldots, Z_m)$.

LEMMA 10.1. *If P_Z satisfies Assumption (D), we have for any $\epsilon \in (0, 1/2)$*

$$P_Z^m \Big[\sup_{z \in \operatorname{Supp}(P_Z)} \Big| \frac{H_m(z)}{h_m(z)} - 1 \Big| > \epsilon \Big] \leq \exp(-D\epsilon^2 m^\alpha)$$

for m large enough, where $\alpha := 2s/(1+2s+\beta)$ and D is a constant depending on σ and L.

The next lemma provides an uniform control on the smallest eigenvalue of $\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$ under Assumption (D).

LEMMA 10.2. *If P_Z satisfies Assumption (D), there exists $\lambda_0 > 0$ depending on β and s only such that*

$$P_Z^m \big[\inf_{z \in \operatorname{Supp} P_Z} \lambda(\bar{\mathbf{Z}}_m(z)) \leq \lambda_0 \big] \leq \exp(-D m^\alpha),$$

for m large enough, where $\alpha = 2s/(1+2s+\beta)$, and D is a constant depending on $\gamma, \beta, s, L, \sigma$.

The proofs of the lemmas are given in Section 6. We consider the event

$$\Omega_m(\epsilon) := \big\{ \inf_{z \in \operatorname{Supp} P_Z} \lambda(\bar{\mathbf{Z}}_m(z)) > \lambda_0 \big\} \cap \big\{ \sup_{z \in \operatorname{Supp} P_Z} |H_m(z)/h_m(z) - 1| \leq \epsilon \big\},$$

where $\epsilon \in (0, 1/2)$. We have for any $f \in H^Q(s, L)$

$$E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)}] \leq \lambda_0^{-2} (1+\epsilon)^{2s} \frac{\sigma^2}{m} \int \frac{P_Z(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)},$$

where we used together the definition of $\Omega_m(\epsilon)$, (10.19) and (10.27). Let us denote $I := \operatorname{Supp} P_Z$ and let I_{z*} be the intervals from Assumption (D). Using together the fact that $\min_{z \in I - \cup_{z*} I_{z*}} \mu(z) > 0$ and (10.29), we obtain

$$\frac{\sigma^2}{m} \int_{I - \cup_{z*} I_{z*}} \frac{P_Z(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)} \leq C m^{-2s/(2s+1)}.$$

Using the monoticity constraints from Assumption (D), we obtain

$$\begin{aligned} \frac{\sigma^2}{m} \int_{I_{z*}} \frac{P(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)} &\leq \frac{\sigma^2}{m} \Big(\int_{z*-a_{z*}}^{z*} \frac{\mu(z)dz}{\int_{z-h_m(z)}^{z} \mu(t)dt} + \int_{z*}^{z*+b_{z*}} \frac{\mu(z)dz}{\int_{z}^{z+h_m(z)} \mu(t)dt} \Big) \\ &\leq \frac{\sigma^2}{m} \int_{I_{z*}} h_m(z)^{-1} dz \leq C m^{-2s/(2s+1)}, \end{aligned}$$

hence $E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)}] \leq C m^{-2s/(2s+1)}$ uniformly for $f \in H^Q(s, L)$. Using together Lemmas 10.1 and 10.2, we obtain $E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)^\complement}] = o(n^{-2s/(2s+1)})$, and (10.20) follows. □

Proof of Theorem 10.4. In model (10.1), when the noise ϵ is centered and such that $E(\epsilon^2) = 1$, the risk of a function $\bar{g} : \mathbb{R}^d \to \mathbb{R}$ is given by

$$A(\bar{g}) := E[(Y - \bar{g}(X))^2] = E[\sigma(X)^2] + \|\bar{g} - g\|^2_{L^2(P_X)},$$

where g is the regression function. Therefore, the excess risk satisfies

$$A(\bar{g}) - A = \|\bar{g} - g\|^2_{L^2(P_X)},$$

where $A := A(g) = E[\sigma(X)^2]$. Let us introduce $n := |D|$ the size of the learning sample, and $M := |\Lambda|$ the size of the dictionary of functions $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$. The empirical risk of $\bar{g}$ over the $D = [(X_i, Y_i); 1 \leq i \leq n]$ is given by

$$A_n(\bar{g}) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{g}(X_i))^2.$$

We begin with a linearization of these risks. We consider the convex set

$$\mathcal{C} := \Big\{ (\theta_\lambda)_{\lambda \in \Lambda} \text{ such that } \theta_\lambda \geq 0 \text{ and } \sum_{\lambda \in \Lambda} \theta_\lambda = 1 \Big\},$$

and define the linearized risks on $\mathcal{C}$

$$\tilde{A}(\theta) := \sum_{\lambda \in \Lambda} \theta_\lambda A(\bar{g}^{(\lambda)}), \quad \tilde{A}_n(\theta) := \sum_{\lambda \in \Lambda} \theta_\lambda A_n(\bar{g}^{(\lambda)}),$$

which are linear versions of the risk A and its empirical version A_n. The exponential weights $w = (w_\lambda)_{\lambda \in \Lambda} := (w(\bar{g}^{(\lambda)}))_{\lambda \in \Lambda}$ are actually the unique solution of the minimization problem

$$\min \Big(\tilde{A}_n(\theta) + \frac{1}{Tn} \sum_{\lambda \in \Lambda} \theta_\lambda \log \theta_\lambda \;\Big|\; (\theta_\lambda) \in \mathcal{C} \Big), \tag{10.30}$$

where $T > 0$ is the temperature parameter in the weights (10.13), and where we use the convention $0 \log 0 = 0$. Let $\hat{\lambda} \in \Lambda$ be such that $A_n(\bar{g}^{(\hat{\lambda})}) = \min_{\lambda \in \Lambda} A_n(\bar{g}^{(\lambda)})$. Since $\sum_{\lambda \in \Lambda} w_\lambda \log\big(\frac{w_\lambda}{1/M}\big) = K(w|u) \geq 0$ where $K(w|u)$ denotes the Kullback-Leibler divergence between the weights w and the uniform weights $u := (1/M)_{\lambda \in \Lambda}$, we have together with (10.30):

$$\begin{aligned}
\tilde{A}_n(w) &\leq \tilde{A}_n(w) + \frac{1}{Tn} K(w|u) \\
&= \tilde{A}_n(w) + \frac{1}{Tn} \sum_{\lambda \in \Lambda} w_\lambda \log w_\lambda + \frac{\log M}{Tn} \\
&\leq \tilde{A}_n(e_{\hat{\lambda}}) + \frac{\log M}{Tn},
\end{aligned}$$

where $e_\lambda \in \mathcal{C}$ is the vector with 1 for the λ-th coordinate and 0 elsewhere. Let $a > 0$ and $A_n := A_n(g)$. For any $\lambda \in \Lambda$, we have

$$\begin{aligned}
\tilde{A}(w) - A &= (1+a)(\tilde{A}_n(w) - A_n) + \tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \\
&\leq (1+a)(\tilde{A}_n(e_\lambda) - A_n) + (1+a)\frac{\log M}{Tn} \\
&+ \tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n).
\end{aligned}$$

Let us denote by E_K the expectation with respect to P_K, the joint law of D for a noise ϵ which is bounded almost surely by $K > 0$. We have

$$E_K\big[\tilde{A}(w) - A\big] \le (1+a)\min_{\lambda\in\Lambda}(\tilde{A}_n(e_\lambda) - A_n) + (1+a)\frac{\log M}{Tn} + E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big].$$

Using the linearity of $\tilde{A}$ on $\mathcal{C}$, we obtain

$$\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \le \max_{g\in\mathcal{G}_\Lambda}\big(A(g) - A - (1+a)(A_n(g) - A_n)\big),$$

where $\mathcal{G}_\Lambda := \{\bar{g}^{(\lambda)}\ ;\ \lambda\in\Lambda\}$. Then, using Bernstein inequality, we obtain for all $\delta > 0$

$$\begin{aligned} &P_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \ge \delta\big] \\ &\le \sum_{g\in\mathcal{G}_\Lambda} P_K\Big[A(g) - A - (A_n(g) - A_n) \ge \frac{\delta + a(A(g) - A)}{1+a}\Big] \\ &\le \sum_{g\in\mathcal{G}_\Lambda} \exp\Big(-\frac{n(\delta + a(A(g) - A))^2(1+a)^{-1}}{8Q^2(1+a)(A(g) - A) + 2(6Q^2 + 2\sigma K)(\delta + a(A(g) - A))/3}\Big). \end{aligned}$$

Moreover, we have for any $\delta > 0$ and $g \in \mathcal{G}_\Lambda$,

$$\frac{(\delta + a(A(g) - A))^2(1+a)^{-1}}{8Q^2(A(g) - A) + 2(6Q^2(1+a) + 2\sigma K)(\delta + a(A(g) - A))/3} \ge C(a,K)\delta,$$

where $C(a,K) := \big(8Q^2(1+a)^2/a + 4(6Q^2 + 2\sigma K)(1+a)/3\big)^{-1}$, thus

$$E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big] \le 2u + M\frac{\exp(-nC(a,K)u)}{nC(a,K)}$$

If we denote by γ_A the unique solution of $\gamma = A\exp(-\gamma)$, where $A > 0$, we have $\log A/2 \le \gamma_A \le \log A$. Thus, if we take $u = \gamma_M/(nC(a,K))$, we obtain

$$E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big] \le \frac{3\log M}{C(a,K)n}.$$

By convexity of the risk, we have

$$\tilde{A}(w) - A \ge A(\hat{g}) - A,$$

thus

$$E_K\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] \le (1+a)\min_{\lambda\in\Lambda}\|\bar{g}^{(\lambda)} - g\|^2_{L^2(P_X)} + C_1\frac{\log M}{n},$$

where $C_1 := (1+a)(T^{-1} + 3C(a,K)^{-1})$. It remains to prove the result when the noise is Gaussian. Let us denote $\epsilon^n_\infty := \max_{1\le i\le n}|\epsilon_i|$. For any $K > 0$, we have

$$\begin{aligned} E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] &= E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\mathbf{1}_{\epsilon^n_\infty\le K}\big] + E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\mathbf{1}_{\epsilon^n_\infty > K}\big] \\ &\le E_K\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] + 2Q^2P[\epsilon^n_\infty > K]. \end{aligned}$$

For $K = K_n := 2(2\log n)^{1/2}$, we obtain using standard results about the maximum of Gaussian vectors that $P[\epsilon^n_\infty > K_n] \le P[\epsilon^n_\infty - E[\epsilon^n_\infty] > (2\log n)^{1/2}] \le 1/n$, which concludes the proof of the Theorem. □

6. Proof of the lemmas

Proof of Lemma 10.1. Using together (10.9) and (10.27), if $I_m^\epsilon(z) := [z - (1+\epsilon)h_m(z), z + (1+\epsilon)h_m(z)]$ and $I_m(z) := I_m^0(z)$, we obtain for any $\epsilon \in (0, 1/2)$:

$$\begin{aligned}\{H_m(z) \le (1+\epsilon)h_m(z)\} &= \big\{(1+\epsilon)^{2s}\bar{P}_Z[I_m^\epsilon(z)] \ge P_Z[I_m(z)]\big\} \\ &\supset \big\{(1+\epsilon)^{2s}\bar{P}_Z[I_m(z)] \ge P_Z[I_m(z)]\big\},\end{aligned}$$

where we used the fact that $\epsilon \mapsto P_Z[I_m^\epsilon(z)]$ is nondecreasing. Similarly, we have on the other side

$$\{H_m(z) > (1-\epsilon)h_m(z)\} \supset \big\{(1-\epsilon)^{2s}\bar{P}_Z[I_m(z)] \le P_Z[I_m(z)]\big\}.$$

Thus, if we consider the set of intervals

$$\mathcal{I}_m := \bigcup_{z \in \operatorname{Supp} P_Z} \big\{I_m(z)\big\},$$

we obtain

$$\Big\{ \sup_{z \in \operatorname{Supp} P_Z} \Big| \frac{H_m(z)}{h_m(z)} - 1 \Big| \ge \epsilon \Big\} \subset \Big\{ \sup_{I \in \mathcal{I}_m} \Big| \frac{\bar{P}_Z[I]}{P_Z[I]} - 1 \Big| \ge \epsilon/2 \Big\}.$$

Using together (10.27) and (10.28), we obtain

$$P_Z[I_m(z)] = \sigma^2/(mL^2 h_m(z)^{2s}) \ge D m^{-(\beta+1)/(1+2s+\beta)} =: \alpha_m. \tag{10.31}$$

Hence, if $\epsilon' := \epsilon(1+\epsilon/2)/(\epsilon+2)$, we have

$$\begin{aligned}\Big\{ \sup_{I \in \mathcal{I}_m} \Big| \frac{\bar{P}_Z[I]}{P_Z[I]} - 1 \Big| \ge \epsilon/2 \Big\} \subset & \Big\{ \sup_{I \in \mathcal{I}_m} \frac{\bar{P}_Z[I] - P_Z[I]}{\sqrt{\bar{P}_Z[I]}} \ge \epsilon' \alpha_m^{1/2} \Big\} \\ & \cup \Big\{ \sup_{I \in \mathcal{I}_m} \frac{P_Z[I] - \bar{P}_Z[I]}{\sqrt{P_Z[I]}} \ge \epsilon \alpha_m^{1/2}/2 \Big\}.\end{aligned}$$

Hence, Theorem 10.6 and the fact that the shatter coefficient satisfies $\mathcal{S}(\mathcal{I}_m, m) \le m(m+1)/2$ (see Appendix) entails the Lemma. □

Proof of Lemma 10.2. Let us denote $\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$ where $\bar{\mathbf{Z}}_m(z,h)$ is given by (10.7) and $H_m(z)$ is given by (10.9). Let us define the matrix $\tilde{\mathbf{Z}}_m(z) := \tilde{\mathbf{Z}}_m(z, h_m(z))$ where

$$(\tilde{\mathbf{Z}}_m(z,h))_{a,b} := \frac{1}{m P_Z[I(z,h)]} \sum_{i=1}^m \Big(\frac{Z_i - z}{h}\Big)^{a+b} \mathbf{1}_{Z_i \in I(z,h)}.$$

Step 1. Let us define for $\epsilon \in (0,1)$ the event

$$\Omega_1(\epsilon) := \Big\{ \sup_{z \in \operatorname{Supp} P_Z} \Big| \frac{H_m(z)}{h_m(z)} - 1 \Big| \le \epsilon \Big\} \cap \Big\{ \sup_{z \in \operatorname{Supp} P_Z} \Big| \frac{\bar{P}_Z[I(z, H_m(z))]}{P_Z[I(z, h_m(z))]} - 1 \Big| \le \epsilon \Big\}.$$

For a matrix A, we denote $\|A\|_\infty := \max_{a,b} |(A)_{a,b}|$. We can prove that on $\Omega_1(\epsilon)$, we have

$$\|\bar{\mathbf{Z}}_m(z) - \tilde{\mathbf{Z}}_m(z)\|_\infty \le \epsilon.$$

Moreover, using Lemma 10.1, we have $P_Z^m[\Omega_1(\epsilon)^\complement] \le C \exp(-D\epsilon^2 m^\alpha)$. Hence, on $\Omega_1(\epsilon)$, we have for any $v \in \mathbb{R}^d$, $\|v\|_2 = 1$

$$v^\top \bar{\mathbf{Z}}_m(z) v \ge v^\top \tilde{\mathbf{Z}}_m(z) v - \epsilon$$

uniformly for $z \in \operatorname{Supp} P_Z$.

Step 2. We define the deterministic matrix $\mathbf{Z}(z) := \mathbf{Z}(z, h_m(z))$ where

$$(\mathbf{Z}(z,h))_{a,b} := \frac{1}{P_Z[I(z,h)]} \int_{I(z,h)} \Big(\frac{t-z}{h}\Big)^{a+b} P_Z(dt),$$

and

$$\lambda_0 := \mathrm{liminf}_m \inf_{z \,\mathrm{Supp}\, P_Z} \lambda\big(\mathbf{Z}(z, h_m(z))\big).$$

We prove that $\lambda_0 > 0$. Two cases can occur: either $\mu(z) = 0$ or $\mu(z) > 0$. We show that in both cases, the liminf is positive. If $\mu(z) > 0$, the entries $(\mathbf{Z}(z, h_m(z)))_{a,b}$ have limit $(1 + (-1)^{a+b})/(2(a+b+1))$, which defines a positive definite matrix. If $\mu(z) = 0$, we know that the density $\mu(\cdot)$ of P_Z behaves as the power function $|\cdot - z|^{\beta(z)}$ around z for $\beta(z) \in (0, \beta)$. In this case, $(\mathbf{Z}(z, h_m(z)))_{a,b}$ has limit $(1 + (-1)^{a+b})(\beta(z)+1)/(2(1+a+b+\beta(z)))$, which defines also a definite positive matrix.

Step 3. We prove that

$$P_Z^m[\sup_{z \in \mathrm{Supp}\, P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon] \le \exp(-D\epsilon^2 m^\alpha).$$

We consider the sets of nonnegative functions (we recall that $I(z,h) = [z-h, z+h]$)

$$F^{(\mathrm{even})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ even and } 0 \le a \le 2R}} \Big\{ \Big(\frac{\cdot - z}{h_m(z)}\Big)^a \mathbf{1}_{I(z, h_m(z))}(\cdot) \Big\},$$

$$F_+^{(\mathrm{odd})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ odd and } 0 \le a \le 2R}} \Big\{ \Big(\frac{\cdot - z}{h_m(z)}\Big)^a \mathbf{1}_{[z, z+h_m(z)]}(\cdot) \Big\},$$

$$F_-^{(\mathrm{odd})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ odd and } 0 \le a \le 2R}} \Big\{ \Big(\frac{z - \cdot}{h_m(z)}\Big)^a \mathbf{1}_{[z-h_m(z), z]}(\cdot) \Big\}.$$

Writing $I(z, h_m(z)) = [z - h_m(z), z) \cup [z, z + h_m(z)]$ when $a+b$ is odd, and since

$$P_Z[I(z, h_m(z))] \ge E f(Z_1)$$

for any $f \in F := F^{(\mathrm{even})} \cup F_+^{(\mathrm{odd})} \cup F_-^{(\mathrm{odd})}$, we obtain

$$\|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty \le \sup_{f \in F} \frac{|\frac{1}{m}\sum_{i=1}^m f(Z_i) - E f(Z_1)|}{E f(Z_1)}.$$

Hence, since $x \mapsto x/(x+\alpha)$ is increasing for any $\alpha > 0$, and since $\alpha := E f(Z_1) \ge D m^{-(\beta+1)/(1+2s+\beta)} =: \alpha_m$ (see (10.31)), we obtain

$$\Big\{ \sup_{z \in \mathrm{Supp}\, P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon \Big\} \subset \Big\{ \sup_{f \in F} \frac{|\frac{1}{m}\sum_{i=1}^m f(Z_i) - E f(Z_1)|}{\alpha_m + \frac{1}{m}\sum_{i=1}^m f(Z_i) + E f(Z_1)} > \epsilon/2 \Big\}.$$

Then, using Theorem 10.7 (note that any $f \in F$ is non-negative), we obtain

$$P_Z^m[\sup_{z \in \mathrm{Supp}\, P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon] \le 4E[\mathcal{N}_1(\alpha_m \epsilon/8, F, Z_1^m)] \exp\big(-D\epsilon^2 m^{2s/(1+2s+\beta)}\big).$$

Together with the inequality

$$E[\mathcal{N}_1(\alpha_m \epsilon/8, F, Z_1^m)] \le D(\alpha_m \epsilon)^{-1} m^{1/(2s+1) + (\beta-1)/(2s+\beta)}, \tag{10.32}$$

(see the proof below), this entails the Lemma. □

Proof of (10.32). It suffices to prove the inequality for $F^{(\mathrm{even})}$ and a fixed $a \in \{0, \dots, 2R\}$, since the proof is the same for $F_+^{(\mathrm{odd})}$ and $F_-^{(\mathrm{odd})}$. We denote $f_z(\cdot) := ((\cdot -$

$z)/h_m(z))^a \mathbf{1}_{I(z,h_m(z))}(\cdot)$. We prove the following statement

$$\mathcal{N}(\epsilon, F, \|\cdot\|_\infty) \le D\epsilon^{-1} m^{1/(2s+1)+(\beta-1)/(2s+\beta)},$$

which is stronger than (10.32), where $\|\cdot\|_\infty$ is the uniform norm over the support of P_Z. Let $z, z_1, z_2 \in \operatorname{Supp} P_Z$. We have

$$|f_{z_1}(z) - f_{z_2}(z)| \le \max(a,1)\Big|\frac{z-z_1}{h_1} - \frac{z-z_2}{h_2}\Big|\mathbf{1}_{I_1\cup I_2},$$

where $h_j := h_m(z_j)$ and $I_j := [z_j - h_j, z_j + h_j]$ for $j = 1, 2$. Hence,

$$|f_{z_1}(z) - f_{z_2}(z)| \le \frac{|h_1 - h_2| + |z_1 - z_2|}{\min(h_1, h_2)}.$$

Using (10.27) together with a differentiation of $z \mapsto h_m(z)^{2s} P_Z[I(z, h_m(z))]$, we obtain that

$$|h_m(z_1) - h_m(z_2)| \le \sup_{z_1 \le z \le z_2} \Big|\frac{h_m(z)^{2s+1}(\mu(z - h_m(z)) - \mu(z + h_m(z)))}{(2s\sigma^2)/(mL) + h_m(z)^{2s+1}(\mu(z - h_m(z)) + \mu(z + h_m(z)))}\Big| |z_1 - z_2|,$$

for any $z_1 < z_2$ in $\operatorname{Supp}\mu$. This entails together with Assumption (D), (10.28) and (10.29):

$$|h_m(z_1) - h_m(z_2)| \le \frac{\mu_\infty}{2s(\gamma L)^{(2s+1)/(2s+\beta+1)}} \Big(\frac{m}{\sigma^2}\Big)^{\frac{\beta}{2s+\beta+1}} |z_1 - z_2|,$$

for any $z_1 < z_2$ in $\operatorname{Supp}\mu$. Hence,

$$|f_{z_1}(z) - f_{z_2}(z)| \le D m^{\frac{1}{2s+1} + \frac{\beta-1}{2s+\beta}} |z_1 - z_2|,$$

which concludes the proof of (10.32). □

7. Some tools form empirical process theory

Let $\mathcal{A}$ be a set of Borelean subsets of $\mathbb{R}$. If $x_1^n := (x_1, \dots, x_n) \in \mathbb{R}^n$, we define

$$N(\mathcal{A}, x_1^n) := \big|\big\{\{x_1, \dots, x_n\} \cap A | A \in \mathcal{A}\big\}\big|$$

and we define the *shatter* coefficient

$$S(\mathcal{A}, n) := \max_{x_1^n \in \mathbb{R}^n} N(\mathcal{A}, (x_1, \dots, x_n)). \tag{10.33}$$

For instance, if $\mathcal{A}$ is the set of all the intervals $[a, b]$ with $-\infty \le a < b \le +\infty$, we have $S(\mathcal{A}, n) = n(n+1)/2$.

Let $X_1, \dots, X_n$ be i.i.d. random variables with values in $\mathbb{R}$, and let us define $\mu[A] := P(X_1 \in A)$ and $\bar{\mu}_n[A] := n^{-1} \sum_{i=1}^n \mathbf{1}_{X_i \in A}$. The following inequalities for relative deviations are due to Vapnik and Chervonenkis (1974), see for instance in [117].

THEOREM 10.6 (Vapnik and Chervonenkis (1974)). *We have*

$$P\Big[\sup_{A\in\mathcal{A}} \frac{\mu(A) - \bar{\mu}_n(A)}{\sqrt{\mu(A)}} > \epsilon\Big] \le 4\mathcal{S}(\mathcal{A}, 2n) \exp(-n\epsilon^2/4)$$

and

$$P\Big[\sup_{A\in\mathcal{A}} \frac{\bar{\mu}_n(A) - \mu(A)}{\sqrt{\bar{\mu}_n(A)}} > \epsilon\Big] \le 4\mathcal{S}(\mathcal{A}, 2n) \exp(-n\epsilon^2/4)$$

where $\mathcal{S}_\mathcal{A}(2n)$ is the shatter coefficient of $\mathcal{A}$ defined by (10.33).

Let $(\mathcal{X}, \tau)$ be a measured space and $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [-K, K]$. Let us fix $p \geq 1$ and $z_1^n \in \mathcal{X}^n$. Define the semi-distance $d_p(f, g)$ between f and g by

$$d_p(f, g) := \Big(\frac{1}{n} \sum_{i=1}^{n} |f(z_i) - g(z_i)|^p \Big)^{1/p}$$

and denote by $\mathcal{B}_p(f, \epsilon)$ the d_p-ball with center f and radius ϵ. The $\epsilon-$covering number of $\mathcal{F}$ w.r.t d_p is defined as

$$\mathcal{N}_p(\epsilon, \mathcal{F}, z_1^n) := \min \big(N \,|\, \exists f_1, \ldots, f_N \text{ s.t. } \mathcal{F} \subseteq \cup_{j=1}^{M} \mathcal{B}_p(f_j, \epsilon) \big).$$

THEOREM 10.7 (Haussler (1992)). *If $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [0, K]$, we have*

$$P\Big[\sup_{f \in \mathcal{F}} \frac{\big| E[f(X_1)] - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \big|}{\alpha + E[f(X_1)] + \frac{1}{n} \sum_{i=1}^{n} f(X_i)} \geq \epsilon \Big] \leq 4E[\mathcal{N}_p(\alpha\epsilon/8, \mathcal{F}, X_1^n)] \exp\Big(- \frac{n\alpha\epsilon^2}{16K^2} \Big).$$

Bibliography

[1] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computat. Stat. Data Anal.*, 22:351–361, 1996.

[2] F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[3] F. Abramovich, T. Sapatinas, and B.W. Silverman. Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B*, 60:725–749, 1998.

[4] A. Antoniadis and J. Bigot. Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. *J. Statist. Software*, 6(3):1–83, 2001.

[5] A. Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:643–645, 1999.

[6] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, 296(23):1021–1024, 1983. French.

[7] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincare (B), Probability and Statistics*, 40(6):685–736, 2004.

[8] J.-Y. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint PMA-908, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, 2004.

[9] J.-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers under margin condition. *Ann. Statist.*, 35(2), April 2007.

[10] N.H. Augustin, S.T. Buckland, and K.P. Burnham. Model selection: An integral part of inference. *Boimetrics*, 53:603–618, 1997.

[11] A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields*, 113:301–413, 1999.

[12] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

[13] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[14] P.L. Bartlett, Y. Freund, W.S. Lee, and R.E. Schapire. Boosting the margin: a new explanantion for the effectiveness of voting methods. *Ann. Statist.*, 26:1651–1686, 1998.

[15] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, Classification and Risk Bounds. *J. Am. Statist. Assoc.*, 101:138–156, 2006.

[16] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics Volume 1.* Prentice Hall, 2001.

[17] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-862.pdf, 2005.

[18] L Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[19] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. Available at http//mahery.math.u-psud.fr/∼blanchard/publi/, 2004.

[20] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *JMLR*, 4:861–894, 2003.

[21] G. Blanchard, C. Schäfer, Y. Rozenholc, and K-R. Müller. Optimal dyadic decision trees. To appear, Machine Learning, 2006.

[22] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323 – 375, 2005.

[23] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16(3):277–292, 2000.

[24] L. Breiman, J. Freidman, J. Olshen, and C. Stone. Classification and regression trees. Wadsworth, 1984.

[25] P. Bühlmann and B. Yu. Analyzing bagging. *Ann. Statist.*, 30(4):927–961, 2002.

[26] F. Bunea and A. Nobel. Online prediction algorithms for aggregation of arbitrary estimators of a conditional mean. Submitted to IEEE Transactions in Information Theory, 2005.

[27] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. to appear in Ann. Statist.. Available at http://www.stat.fsu.edu/ wegkamp, 2005.

[28] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via l1 penalized least squares. volume 32, pages 379–391, 2006. COLT 2006.

[29] T. Cai. On adaptivity of Blockshrink wavelet estimator over Besov spaces. *Technical Report, 97-05, Department of Statistics, Purdue University*, 1997.

[30] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat.*, 27:898–924, 1999.

[31] T. Cai and E. Chicken. Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis*, 95:76–106, 2005.

[32] T. Cai and B.W. Silverman. Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya*, (63):127–148, 2001.

[33] O. Catoni. A mixture approach to universal model selection. preprint LMENS-97-30, available at http://www.dma.ens.fr/EDITION/preprints/, 1997.

[34] O. Catoni. "universal" aggregation rules with exact bias bounds. Preprint n.510, LPMA, available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html, 1999.

[35] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, N.Y., 2001.

[36] L. Cavalier and A. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Meth. Statist.*, 10(3):247–282, 2001.

[37] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, 2006.

[38] C. Chesneau and G. Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. Submitted, 2006.

[39] E. Chicken. Nonparametric regression on random processes and design. *Florida State University Department of Statistics, Technical Report*, 2003.

[40] H. A. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, 92:1413–1421, 1997.

[41] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[42] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 1991. Second edition, 2006.

[43] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Reg. Conf. Series in Applied Math. SIAM, Philadelphia, 1992.

[44] M. Delecroix, W. Härdle, and M. Hristache. Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, 86(2):213–226, 2003.

[45] M. Delecroix, M. Hristache, and V. Patilea. On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, 136(3):730–769, 2006.

[46] B. Delyon and A. Juditsky. On minimax wavelet estimators. *Applied Computational Harmonic Analysis*, 3:215–228, 1996.

[47] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, Berlin, Heidelberg, 1996.

[48] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, New-York, 2001.

[49] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[50] D.L. Donoho and I.M. Johnstone. Adaptating to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, 90(432):1200–1224, 1995.

[51] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptotia ? *J. Royal Statist. Soc. Ser. B.*, 57:301–369, 1995.

[52] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[53] U. Einmahl and D. Mason. Some universal results on the behavior of increments of partial sums. *Ann. Probab.*, 24:2626–2635, 1996.

[54] Y. Freund and R. Schapire. A decision-theoric generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[55] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, 28:337–374, 2000.

[56] S. Gaïffas and G. Lecué. Optimal rates and adaptation in the single-index model using aggregation. Submitted.

[57] H. Gao. Wavelet shrinkage denoising using the nonnegative garrote. *J. Comput. Graph. Statist.*, 7:469–488, 1998.

[58] G. Geenens and M. Delecroix. A survey about single-index models theory, 2005.

[59] 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.

[60] P. Hall, G. Kerkyacharian, and D. Picard. Block thresholding rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26:942–962, 1998.

[61] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, 9(1):33–49, 1999.

[62] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelet, Approximation and Statistical Applications*, volume 129 of *Lectures Notes in Statistics*. Springer

Verlag, New York, 1998.

[63] J.A. Hartigan. Bayesian regression using akaike priors. Yale University, New Haven, Preprint, 2002.

[64] R. Herbei and H. Wegkamp. Classification with reject option. 2005.

[65] D.R.M. Herrick, G.P. Nason, and B.W. Silverman. Some new methods for wavelet density estimation. *Sankhya Series A*, 63:394–411, 2001.

[66] J.L. Horowitz. *Semiparametric methods in econometrics*, volume 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[67] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001.

[68] I.A. Ibragimov and R.Z. Hasminskii. An estimate of density of a distribution. In *Studies in mathematical stat. IV.*, volume 98, pages 61–85. Zap. Nauchn. Semin., LOMI, 1980.

[69] M. Jansen. *Noise reduction by wavelet thresholding*, volume 161. Springer–Verlag, New York, lecture notes in statistics edition, 2001.

[70] I.M Johnstone and B.W. Silverman. Empirical bayes selection of wavelet thresholds. *Ann. Statist.*, 33(4):1700–1752, 1998.

[71] A. Juditsky. Wavelet estimators: adapting to unknown smoothness. *Math. Methods of Statistics*, (1):1–20, 1997.

[72] A. Juditsky, A. Nazin, A.B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41:368 – 384, 2005.

[73] A. Juditsky, A. Nazin, A.B. Tsybakov, and N. Vayatis. Generalization error bounds for aggregation by mirror descent. In J.Platt Y.Weiss, B.Schölkopf, editor, *Advances in Neural Information Processing 18. Proceedings of NIPS-2005*. MIT Press, Cambridge, MA(2006), 2006.

[74] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Statist.*, 28(3):681–712, 2000.

[75] A.B. Juditsky, Ph. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. Preprint n.1034, Laboratoire de Probabilités et Modèle aléatoires, Universités Paris 6 and Paris 7 (available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005), 2006.

[76] G. Kerkyacharian, D. Picard, and K. Tribouley. Lp adaptive density estimation. *Bernoulli*, 2:229–247, 1996.

[77] V. Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6):1–50, December 2006. 2004 IMS Medallion Lecture.

[78] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30:1–50, 2002.

[79] G. Lecué. Classification with fast rates for sparsity class of Bayes rules. To appear in *Electronic Journal of Statistics*, 2005.

[80] G. Lecué. Lower bounds and aggregation in density estimation. *Journal of Machine Learning research*, 7(Jun):971–981, 2005.

[81] G. Lecué. Optimal rates of aggregation in classification. To appear in *Bernoulli*, 2005.

[82] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. To appear in *Ann. Statist.*, 2005.

[83] G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. *In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006*, 32(4):364–378, 2006.

[84] G. Lecué. Suboptimality of Penalized Empirical Risk Minimization. 2006.

[85] G. Lecué. Suboptimality of Penalized Empirical Risk Minimization in Classification. Accepted in COLT07, 2006.

[86] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[87] G. Leung and A. Barron. Information theory and mixing least-square regressions. *IEEE Transactions on Information Theory*, 52 (8):3396–3410, 2006.

[88] Y. Lin. A note on margin-based loss functions in classification. Technical report, Technical Report 1029r, Departement of Statistics, University of Wisconsin, Madison., 1999.

[89] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1):30–55, 2004.

[90] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.

[91] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

[92] P. Massart. Some applications of concentration inequalities to statistics. *Probability Theory. Annales de la Faculté des Sciences de Toulouse*, (2):245–303, 2000. volume spécial dédié à Michel Talagrand.

[93] P. Massart. *Concentration inequalities and model selection.* Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer, 2006.

[94] P. Massart and E. Nédélec. Risk Bound for Statistical Learning. *Ann. Statist.*, 34(5), October 2006.

[95] Y. Meyer. *Ondelettes et Opérateurs.* Hermann, Paris, 1990.

[96] S. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and knowledge Discovery*, 2(4):345–389, 1998.

[97] G.P. Nason. *Choice of the Threshold Parameter in Wavelet Function Estimation*, volume 103. 1995.

[98] A. Nemirovski. *Topics in Non-parametric Statistics*, volume 1738 of *Ecole d'été de Probabilités de Saint-Flour 1998, Lecture Notes in Mathematics.* Springer, N.Y., 2000.

[99] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335, 2000.

[100] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, 1993.

[101] P. Rigollet. *Inégalités d'oracle, agrégation et adaptation.* PhD thesis, Université Paris 6, 2006.

[102] P. Rigollet and A.B.Tsybakov. Linear and convex aggregation of density estimators. available at http://hal.archives-ouvertes.fr/ccsd-00068216/en/, 2006.

[103] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, (June 1970).

[104] B. Schölkopf and A. Smola. *Learning with kernels.* MIT press, Cambridge University, 2002.

[105] C. Scott and R. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, April 2006.

[106] H. Simon. General lower bounds on the number of examples needed for learning probabilistics concepts. *Proceedings of the sixth Annual ACM conference on Computational Learning Theory*, pages 402–412, 1993. Association for Computing Machinery, New-York.

[107] I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the bayes risk. *In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006*, 32(4):79–93, 2006.

[108] I. Steinwart and C. Scovel. Fast Rates for Support Vector Machines. *In Proceeding of the 18th Annual Conference on Learning Theory, COLT 2005. Lecture Notes in Computer Science 3559 Springer 2005*, 2005.

[109] I. Steinwart and C. Scovel. Fast Rates for Support Vector Machines using Gaussian Kernels. *Ann. Statist.*, 35(2), April 2007.

[110] C.J. Stone. Optimal global rates of convergnece for nonparametric estimators. *Ann. Statist.*, 10(4):1040–1053, 1982.

[111] Winfried Stute and Li-Xing Zhu. Nonparametric checks for single-index models. *Ann. Statist.*, 33(3):1048–1083, 2005.

[112] B. Tarigan and S.A. van de Geer. Adaptivity of Support Vector Machines with $l1$ Penalty. *PASCAL. Technical Report*, MI 2004-14, 2004.

[113] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Series BB 58*, pages 267–288, 1996.

[114] A. B. Tsybakov. Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003. Springer, Heidelberg.

[115] A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, 2004.

[116] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[117] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[118] V.N. Vapnik and A.Ya. Chervonenkis. Theory of pattern recognition. *Nauka, Moscow*, 1974. In Russian.

[119] V.G. Vovk. Aggregating Strategies. *In: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT1990, CA: Morgan Kaufmann*, pages 371–386, 1990.

[120] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.

[121] N. Weyrich and G.T. Warhola. Wavelet shrinkage and generalized cross-validation for image denoising. *IEEE Trans. Im. Proc.*, 7:82–90, 1998.

[122] Yingcun Xia and Wolfgang Härdle. Semi-parametric estimation of partially linear single-index models. (97):1162–1184, 2006.

[123] Y. Yang. Minimax nonparametric classification—part I: Rates of convergence. *IEEE Transaction on Information Theory*, 45:2271–2284, 1999.

[124] Y. Yang. Minimax nonparametric classification—partII: Model selection for adaptation. *IEEE Transaction on Information Theory*, 45:2285–2292, 1999.

[125] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.

[126] Y. Yang. Adaptive regression by mixing. *J. Am. Statist. Ass.*, 96:574–588, 2001.

[127] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10:25–47, 2004.

[128] C.H. Zhang. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.*, 33:54–100, 2005.

[129] T. Zhang. On the convergence of mdl density estimation. In *COLT*, 2004.

[130] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

Premier trimestre 2007

Méthodes d'agrégation : optimalité et vitesses rapides.

Résumé : Le principal travail de cette thèse porte sur l'étude des méthodes d'agrégation sous l'hypothèse de marge. Nous avons mis en avant que l'hypothèse de marge améliore les vitesses d'agrégation. Un autre résultat de cette thèse montre que certaines méthodes de minimisation du risque empirique pénalisé sont sous-optimales quand le risque est convexe, même sous l'hypothèse de marge. Contrairement aux procédures d'agrégation à poids exponentiels, ces méthodes n'arrivent pas à profiter de la marge du modèle. Nous avons ensuite appliqué les méthodes d'agrégation à la résolution de quelques problèmes d'adaptation. Une dernière contribution apportée dans cette thèse a été de proposer une approche du contrôle du biais en classification par l'introduction d'espaces de règles de prédiction parcimonieuses. Des vitesses minimax ont été obtenues pour ces modèles et une méthode d'agrégation a donné une version adaptative de ces procédures d'estimation.

Mots-clés : Estimation non-paramétrique, classification, régression, estimation de densité, adaptation, optimalité, réduction de dimension, vitesses minimax, inégalités d'oracle.

Aggregation procedures: optimality and fast rates.

Abstract: In this thesis we deal with aggregation procedures under the margin assumption. We prove that the margin assumption improves the rate of aggregation. Another contribution of this thesis is to show that some empirical risk minimization procedures are suboptimal when the loss function is convex, even under the margin assumption. Contrarily to some aggregation procedures with exponential weights, these model selection methods cannot benefit from the large margin. Then, we apply aggregation methods to construct adaptive estimators in several different problems. The final contribution of this thesis is to purpose a new approach to the control of the bias term in classification by introducing some spaces of sparse prediction rules. Minimax rates of convergence have been obtained for these classes of functions and, by using an aggregation method, we provide an adaptive version of these estimators.

Keywords: Non-parametric estimation, classification, regression, density estimation, adaptation, optimality, dimension reduction, minimax rates of convergence, oracle inequalities.

AMS Classification: Primary: 62G05. Secondary: 62H30, 68T10, 62G07, 62G08, 68T05, 68Q32.