Oracle inequalities, aggregation and adaptation

Philippe Rigollet

THESE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS–VI

Spécialité : **Mathématiques**

présentée par

Philippe RIGOLLET

Inégalités d'oracle, agrégation et adaptation

Rapporteurs :	M.	Peter	**BICKEL**	UC Berkeley
	Mme	Sara	**van de GEER**	ETH Zürich

Soutenue publiquement le **20 novembre 2006** devant le jury composé de

M.	Lucien	**BIRGÉ**	Université Paris–VI	Président
M.	Stéphane	**BOUCHERON**	Université Paris–VII	Examinateur
Mme.	Sara	**van de GEER**	ETH Zürich	Rapporteur
M.	Anatoli	**IOUDITSKI**	Université Grenoble–I	Examinateur
M.	Pascal	**MASSART**	Université Paris–XI, Orsay	Examinateur
M.	Alexandre	**TSYBAKOV**	Université Paris–VI	Directeur

Remerciements

Mes premiers remerciements vont en toute logique à Alexandre "Sacha" Tsybakov qui, pendant ces trois années de thèse a été mon directeur de thèse, mon collaborateur et mon ami. J'ai pu à ses côtés faire l'expérience d'un chercheur brillant auprès de qui j'ai appris énormément, mathématiquement et humainement parlant. Je ne pense pas être un jour en mesure de lui rendre tout ce qu'il m'a apporté en si peu de temps. Merci Sacha!

Cette thèse a été jalonnée de rencontres, de discussions et de personnes à qui je voudrais rendre hommage, à commencer par Lucien Birgé qui fut à l'origine de mon intérêt pour la statistique mathématique . Je garde un excellent souvenir de son cours d'introduction, de nos discussions estivales et de ses conseils éclairés. Je tiens aussi à exprimer l'admiration que j'ai pour sa capacité à sans cesse renouveler ses travaux et à développer de nouvelles théories. Il fait partie de ceux qui font avancer les choses. C'est pour moi un plaisir et un honneur de le compter parmi les membres de mon jury.

La présence de Pascal Massart dans mon jury est aussi un grand honneur et je le remercie d'avoir accepté d'en faire. Ses réponses ou simplement ses remarques m'ont permis d'avancer dans mes recherches à plusieurs reprises au cours de cette thèse.

Lors de mes premiers pas dans l'apprentissage statistique, Stéphane Boucheron a répondu gentillement à mes questions, parfois naïves, me faisant ainsi profiter de sa grande expérience. C'est avec la même gentillesse et toujours autant de spontanéité qu'il a accepté de faire partie de mon jury.

Ma rencontre avec Anatoli Iouditski a aussi été un moment fort de cette thèse. Il m'a introduit au domaine de l'optimisation pour lequel mon intérêt ne fait que croître. Travailler à ses côtés est une expérience très stimulante et instructive.

J'adresse un remerciement tout particulier à Sara van de Geer et Peter Bickel pour avoir accepté d'être mes rapporteurs.

Je souhaite également remercier:

Régis Vert et Peng Zhao, mes coauteurs, avec qui travailler fut un réel plaisir et une formidable source d'inspiration.

Nicolas Vayatis, Gérard Biau, Jean-Philippe Vert, Arnak Dalalyan, Jean-Yves Audibert, Sacha Nazine et Cristina Butucea. Ils ont fait preuve d'une grande disponibilité et j'ai appris énormément au cours de discussions plus ou moins formelles que nous avons eues.

Vladimir Koltchinskii, Vladimir Spokoiny, Mark Low, Peter Bartlett et Bin Yu. Je garde un excellent souvenir des moments passés à partager les idées de ces brillants chercheurs, en particulier lors de mon sejour à Berkeley. Je les remercie du temps qu'ils ont bien voulu m'accorder si simplement.

Florence Merlevède, Jérôme Dedecker, Dominique Picard, Gérard Kerkyacharian et Fabienne Comte pour m'avoir délivré quelques secrets de la statistique lors de leurs cours. Leur pédagogie a suscité chez moi l'intérêt actuel que j'ai pour les statistiques en général.

L'équipe administrative du Laboratoire de Probabilités et Modèles aléatoire: Nelly pour son incroyable efficacité, Josette et Salima pour leur sympathie et leur disponibilité, Jacques Portes qui résout tous les problèmes - même le dimanche . Bien qu' il ne fasse pas partie

du laboratoire, je voudrais aussi remercier ici Hubert Andrique chez qui j'ai pris l'habitude de m'évader le temps d'une reliure.

Quelque part entre le soutien amical, scientifique et logistique, se trouve la contribution de Gilles Stoltz à cette thèse. Un grand merci pour toutes tes astuces et pour m'avoir conseillé à maintes reprises.

Que seraient ces lignes sans remercier mes plus proches collègues, mes amis: Vivian, Oliv, Juan-Carlos, Bertrand, Guillaume, Arvind, Fabien, Elie et Anne-Laure? Merci à vous pour tous les bons moments au bureau ou ailleurs, ...

Je voudrais remercier mes amis, Antoine, Rodolphe, Thomas et Joëlle, mon oncle Jean-Pierre et ma tante Juliette. Leurs remarques bienveillantes et dénuées d'un quelconque caractère mathématique m'ont aidé à prendre le recul nécessaire dans les moments difficiles.

Encore quelques lignes, sans doute trop courtes, pour exprimer toute la gratitude que j'ai pour ma sœur Lucie, mes parents et bien sûr Claire, qui m'ont toujours encouragé et soutenu de manière inconditionnelle dans l'accomplissement de ce travail.

Contents

CHAPITRE 1

Introduction et présentation des résultats

Historiquement, les inégalités d'oracle ont été développées comme des outils particulièrement efficaces pour l'adaptation à un paramètre inconnu en statistique mathématique. Initialement dédiées à la démonstration de propriétés statistiques de certains estimateurs, elles peuvent s'inscrire dans le cadre plus général du problème d'agrégation où elles sont au centre de la définition d'une *vitesse optimale d'agrégation*. Elles constituent alors d'une part des outils mathématiques et d'autre part des résultats précis et non asymptotiques. Les travaux faisant l'objet de cette thèse présentent différentes utilisations des inégalités d'oracle, d'abord dans un cadre général d'agrégation puis dans des modèles statistiques plus particuliers, comme l'estimation de densité et la classification. Les résultats obtenus sont une palette non exhaustive mais représentative de l'utilisation des inégalités d'oracle en statistique mathématique.

Les sections de ce chapitre décrivent chacune un problème statistique traité dans cette thèse, dans lequel les inégalités d'oracle interviennent. Après une brève introduction au problème en question, on y explique le rôle des inégalités d'oracle.

Contents

1. Inégalités d'oracle et modèle de suite gaussienne

Le terme *oracle* remonte à la Grèce antique et désignait alors la réponse d'une divinité faite à une personne qui la consultait. Il désigne aujourd'hui un avis considéré comme infaillible. En statistique, ce terme réfère à une quantité qui n'est calculable que si l'on connaît la distribution sous-jacente des données. Etant donnée une collection de procédures statistiques $\mathcal{M}$ et une fonction de risque R à valeurs positives, on appelle *oracle* (ou procédure oracle) une procédure $m^\star \in \mathcal{M}$ telle que

$$R(m^\star) = \min_{m \in \mathcal{M}} R(m)\,.$$

En général, la fonction de risque R dépend d'une distribution de probabilité inconnue, c'est pourquoi $m^\star$ est aussi inconnue et appelée *oracle*. Bien qu'elle soit inconnue, il est parfois possible de construire une procédure $\hat{m}$ qui imite la procédure $m^\star$ en termes de risque R sans pour autant estimer $m^\star$ directement. Cette propriété se traduit par une *inégalité d'oracle* :

$$R(\hat{m}) \leq CR(m^\star) + \varepsilon\,, \tag{1.1}$$

où $C \geq 1$ est une quantité déterministe bornée et $\varepsilon > 0$ est un terme résiduel en général négligeable devant $R(m^\star)$. On constate que l'inégalité d'oracle (1.1) garantit que lorsque ε est négligeable devant $R(m^\star)$, le risque de $\hat{m}$ est du même ordre que celui de l'oracle $m^\star$. Pourtant les deux procédures $\hat{m}$ et $m^\star$ peuvent être très différentes l'une de l'autre.

Le terme d'*oracle* et plus précisément celui d'*inégalité d'oracle* a été introduit par Donoho and Johnstone (1994). Les premièrs exemples d'inégalités d'oracle ont ensuite été développés dans des articles proches (Donoho *et al.*, 1995; Donoho and Johnstone, 1995) dans lesquels le risque de l'oracle $R(m^\star)$ est appelé *risque idéal*.

Pour fixer les idées, considérons le modèle de suite gaussienne. Il présente l'intérêt d'être simple et d'avoir été un cadre propice au développement des inégalités d'oracle.

1.1. Le modèle de suite gaussienne. Soit le modèle

$$y_k = \theta_k + \varepsilon\xi_k, \quad k = 1, 2, \ldots, \tag{1.2}$$

où les y_k sont des observations, les ξ_k des variables aléatoires indépendantes et identiquement distribuées (i.i.d) de loi $\mathcal{N}(0,1)$, $0 < \varepsilon < 1$, et $\theta = (\theta_1, \theta_2, \ldots) \in \ell_2$ est le paramètre à estimer.

Considérons alors la classe des estimateurs linéaires de θ :

$$\hat{\theta} = \hat{\theta}(h) = (\hat{\theta}_1, \hat{\theta}_2, \ldots),\ \hat{\theta}_k = h_k y_k, \quad k = 1, 2, \ldots,$$

où $h = (h_1, h_2, \ldots) \in \ell_2$ est une suite de poids. Le risque quadratique de l'estimateur linéaire associé à la suite h s'écrit alors :

$$R_\varepsilon(h, \theta) = \mathbb{E}_\theta \left\| \hat{\theta}(h) - \theta \right\|^2 = \sum_{k=1}^{\infty} \left((1 - h_k)^2 \theta_k^2 + \varepsilon^2 h_k^2 \right)\,.$$

Ici, $\|\,.\,\|$ désigne la norme ℓ_2 et $\mathbb{E}_\theta$ désigne l'espérance par rapport à la loi de $y = (y_1, y_2, \ldots)$ dans le modèle (1.2).

Des exemples d'estimateurs linéaires classiques sont :

Les estimateurs par projection: $h_k = \mathbb{1}_{\{k \le N\}}$ où $\mathbb{1}_{\{.\}}$ désigne la fonction indicatrice et $N \ge 1$ est un entier.

Les estimateurs spline: $h_k = 1/(1 + \lambda k^{2m})$ où $m \ge 2$ est un entier et $\lambda > 0$ est un paramètre de lissage.

Les estimateurs de Pinsker: $h_k = (1 - \lambda k^{2\beta})_+$ où $\beta > 0, \lambda > 0$ sont des paramètres et $x_+ = \max(x, 0)$.

Plus généralement, considérons une classe de poids $\mathcal{H}$.

DÉFINITION 1.1. *Soit $\mathcal{H}$ une classe de poids, $\Theta \subseteq \ell_2$ un ensemble, tels que pour tout $\theta \in \Theta$, il existe $h^{\mathcal{H}} = h^{\mathcal{H}}(\theta)$ vérifiant*

$$h^{\mathcal{H}}(\theta) = \operatorname*{argmin}_{h \in \mathcal{H}} R_\varepsilon(h, \theta)\,.$$

Alors la fonction sur Θ à valeurs dans $\mathcal{H} : \theta \mapsto h^{\mathcal{H}}(\theta)$ est appelée **oracle linéaire sur $\mathcal{H}$.** *Par extension, le pseudo-estimateur $\hat{\theta}(h^{\mathcal{H}})$ est aussi appelé oracle.*

L'oracle $\hat{\theta}(h^{\mathcal{H}})$ n'est pas un estimateur puisqu'il dépend du paramètre inconnu θ et n'est pas calculable à partir des données. C'est pourquoi on l'appelle pseudo-estimateur. Par exemple, quand $\mathcal{H}$ est la classe de tous les estimateurs linéaires, l'oracle correspondant

$$h^{\mathrm{lin}}(\theta) = \operatorname*{argmin}_{h \in \ell_2} R_\varepsilon(h, \theta)$$

dépend du paramètre inconnu θ de manière explicite. En effet, on a $h^{\mathrm{lin}}(\theta) = (h_1^{\mathrm{lin}}, h_2^{\mathrm{lin}}, \ldots)$ avec

$$h_k^{\mathrm{lin}} = \frac{\theta_k^2}{\varepsilon^2 + \theta_k^2}\,, \quad k = 1, 2, \ldots$$

1.2. Inégalités d'oracle. Pour une classe de poids $\mathcal{H}$ donnée, le but est alors de trouver une suite de poids $\hat{h} = \hat{h}(y) = (\hat{h}_1(y), \hat{h}_2(y), \ldots)$, à valeurs dans $\mathcal{H}$, telle que pour une quantité déterministe $C_\varepsilon \ge 1$, on ait

$$R_\varepsilon(\hat{h}, \theta) \le C_\varepsilon \inf_{h \in \mathcal{H}} R_\varepsilon(h, \theta)\,, \tag{1.3}$$

pour tout $\theta \in \Theta \subseteq \ell_2$.

Si $C_\varepsilon \to 1$, quand $\varepsilon \to 0$, l'inégalité (1.3) signifie que l'estimateur $\hat{\theta} = \hat{\theta}(\hat{h})$ imite l'oracle sur $\mathcal{H}$ en termes de risque R_ε. Dans ce cas, l'inégalité est appelée *inégalité d'oracle asymptotiquement exacte*. En outre, si $C_\varepsilon \equiv 1$ l'inégalité d'oracle est qualifiée d'*exacte*. Lorsque $C_\varepsilon \ge C$ pour une certaine constante $C > 1$, l'estimateur $\hat{\theta}$ imite seulement la vitesse de convergence (quand ε tend vers 0) de l'oracle. L'inégalité (1.3) est alors appelée simplement *inégalité d'oracle* ou *inégalité d'oracle approximative*.

Vraisemblablement, les premières inégalités d'oracle asymptotiquement exactes furent obtenues pour la classe des « poids linéaires de lissage ordonnés » (Kneip, 1994). Les premiers articles de Shibata (1981); Li (1987); Golubev (1990); Polyak and Tsybakov (1990, 1992); Golubev and Nussbaum (1992) contiennent aussi, implicitement, des inégalités d'oracle asymptotiquement exactes sur certaines classes $\mathcal{H}$. Tous ces articles utilisent le C_p de Mallows ou une de ses variations pour définir $\hat{h}$. Birgé (2001) et Birgé and Massart (2001) obtiennent des inégalités d'oracle asymptotiquement exactes sur la classe des estimateurs par projection en utilisant respectivement la méthode de Lepskiĭ et une version modifiée du C_p de Mallows. Cavalier *et al.* (2002) utilisent aussi le C_p pour obtenir des inégalités d'oracle asymptotiquement exactes pour une grande variété de classes $\mathcal{H}$ qui ne satisfont

pas nécessairement aux hypothèses restrictives de Kneip. Ils étendent aussi la théorie au cas où les ξ_k sont hétéroscédastiques pour couvrir le cas des problèmes inverses. Cavalier and Tsybakov (2001) utilisent la méthode par blocs de Stein avec pénalisation pour obtenir un estimateur non linéaire qui vérifie des inégalités d'oracle asymptotiquement exactes sur toute classe d'estimateurs linéaires à poids monotones décroissants. Dans le contexte des ondelettes, nous ne citons que les articles fondateurs de Donoho and Johnstone (1994); Donoho *et al.* (1995) où des inégalités d'oracle sont utilisées pour l'adaptation à un paramètre de régularité inconnu. En outre, Donoho and Johnstone (1995) obtiennent des inégalités d'oracle asymptotiquement exactes sur la classe des estimateurs de James-Stein.

Les inégalités d'oracle ci-dessus correspondent à la définition initiale qui en a été faite. Nous retiendrons cependant une définition légèrement plus générale dans laquelle $\hat{h}$ n'est pas nécessairement dans la classe $\mathcal{H}$. Ainsi l'interprétation d'une inégalité d'oracle reste la même, c'est-à-dire qu'étant donnée une collection de méthodes (ici les estimateurs $\{\hat{\theta}(h) : h \in \mathcal{H}\}$) le but est d'imiter le meilleur d'entre eux au sens d'un certain risque (ici R_ε) sans nécessairement utiliser une méthode issue de la collection. Nous verrons dans la suite que cette modification sera utilisée de deux manières : soit en imitant l'oracle directement par un élément qui n'est pas dans la collection (cf. partie 2), soit en remarquant que l'oracle sur $\mathcal{H}$ imite lui-même l'oracle sur une classe plus grande $\mathcal{H}' \supset \mathcal{H}$ (cf. chapitre 6).

2. Agrégation et optimisation stochastique

Soit $(\mathcal{Z}, \mathfrak{F})$ un espace mesurable et Z une variable aléatoire à valeurs dans $\mathcal{Z}$ de loi P inconnue. Pour un espace fonctionnel $\mathcal{F}$ donné, considérons la fonction de perte $Q : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ et la fonction de risque correspondante $A : \mathcal{F} \to \mathbb{R}$ définie par

$$A(f) = \mathbb{E}Q(Z, f), \quad f \in \mathcal{F},$$

où $\mathbb{E}$ est le symbole de l'espérance. Le but de l'agrégation est d'imiter (quand il existe) l'oracle

$$f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} A(f).$$

Supposons dans la suite que l'oracle $f^\star$ existe. Etant donné un échantillon $(Z_1, \dots, Z_n)$ tiré selon la loi P, il s'agit de construire un estimateur $\hat{f}_n$ mesurable par rapport à l'échantillon et vérifiant l'inégalité d'oracle exacte suivante

$$\mathbb{E}A(\hat{f}_n) \le \min_{f \in \mathcal{F}} A(f) + \Delta_n^{\mathcal{F}}, \tag{1.4}$$

où $\Delta_n^{\mathcal{F}} > 0$ est un terme résiduel qui ne dépend pas de la loi P.

2.1. Optimisation stochastique. L'oracle $f^\star$ minimise le risque A sur $\mathcal{F}$ et le but de l'agrégation est d'imiter cet optimum. Comme la fonction A n'est pas connue, une optimisation directe n'est pas possible. Construire $\hat{f}_n$ revient donc à résoudre le problème d'*optimisation stochastique* énoncé comme suit.

Etant donnée une suite de fonctions sur $\mathcal{F}$ et à valeurs dans $\mathbb{R}$, $Q(Z_1, \cdot), \dots, Q(Z_n, \cdot)$ indépendantes et de même loi que $Q(Z, \cdot)$, construire un estimateur f_n^* tel que

$$\mathbb{E}A(f_n^*) = \min_{f \in \mathcal{F}} A(f).$$

Il n'est pas possible en général de résoudre ce problème pour un nombre fini d'observations. Cependant, lorsque le terme résiduel $\Delta_n^{\mathcal{F}}$ tend vers 0 avec n grand, l'inégalité d'oracle exacte (1.4) garantit que $\hat{f}_n$ résout le problème d'optimisation stochastique asymptotiquement. En effet, si $\Delta_n^{\mathcal{F}}$ tend vers 0, l'inégalité (1.4) implique

$$\mathbb{E}A(\hat{f}_n) \to \min_{f\in\mathcal{F}} A(f), \quad n \to \infty\,.$$

Remarquons qu'une inégalité d'oracle *exacte* est nécessaire pour assurer la résolution asymptotique du problème d'optimisation stochastique. En effet, supposons qu'à la place de (1.4), on ait

$$\mathbb{E}A(\hat{f}_n) \le C_n \min_{f\in\mathcal{F}} A(f) + \Delta_n^{\mathcal{F}}\,,$$

avec $C_n \to C > 1,\ n \to \infty$. Dans ce cas,

$$\mathbb{E}A(\hat{f}_n) - \min_{f\in\mathcal{F}} A(f) \le (C_n - 1)\min_{f\in\mathcal{F}} A(f) + \Delta_n^{\mathcal{F}} \to (C-1)\min_{f\in\mathcal{F}} A(f), \quad n \to \infty\,.$$

On voit alors que quand $\min_{f\in\mathcal{F}} A(f) \neq 0$, l'inégalité d'oracle approximative ne permet pas de garantir la résolution asymptotique du problème d'optimisation stochastique.

2.2. Interprétation statistique. Soit $\mathcal{X}$ un borelien de $\mathbb{R}^d$ et $F = (f_1, \dots, f_M)^\top$ un vecteur de fonctions tel que pour tout $j = 1, \dots, M,\ f_j : \mathcal{X} \to \mathbb{R}$ est un estimateur ou un classifieur. On se place pour l'instant dans le cadre d'agrégation pure où les f_j sont construits à partir d'un échantillon indépendant de $(Z_1, \dots, Z_n)$ et gelé, de sorte que le vecteur F peut être considéré comme déterministe (cf. chapitre 5 pour plus de details).

Pour tout $\theta \in \mathbb{R}^M$, définissons la combinaison linéaire des f_j

$$\mathsf{f}_\theta = \theta^\top F\,,$$

et considérons la classe de fonctions

$$\mathcal{F}_\Theta = \{\mathsf{f}_\theta : \theta \in \Theta\}\,,$$

pour un certain sous-ensemble Θ de $\mathbb{R}^M$. Le choix de l'ensemble Θ caractérise le type du problème d'agrégation. Trois choix particuliers sont habituellement considérés (Nemirovski, 2000; Tsybakov, 2003). Dans la suite, ils sont appelés les *trois problèmes standard de l'agrégation*.

(1) **Agrégation linéaire.** Le but est de construire $\hat{\theta}_n^{\mathbf{L}}$ qui imite la meilleure *combinaison linéaire* des f_j, c'est-à-dire qui vérifie l'inégalité d'oracle

$$\mathbb{E}A(\mathsf{f}_{\hat{\theta}_n^{\mathbf{L}}}) \le \inf_{\theta\in\mathbb{R}^M} A(\mathsf{f}_\theta) + \Delta_{n,M}^{\mathbf{L}}\,,$$

pour toute distribution P dans une grande classe $\mathcal{P}$, où $\Delta_{n,M}^{\mathbf{L}}$ est un terme résiduel qui ne dépend pas de la distribution P.

(2) **Agrégation convexe.** Le but est de construire $\hat{\theta}_n^{\mathbf{C}}$ qui imite la meilleure *combinaison convexe* des f_j, c'est-à-dire qui vérifie l'inégalité d'oracle

$$\mathbb{E}A(\mathsf{f}_{\hat{\theta}_n^{\mathbf{C}}}) \le \inf_{\theta\in H} A(\mathsf{f}_\theta) + \Delta_{n,M}^{\mathbf{C}}\,,$$

pour toute distribution P dans une grande classe $\mathcal{P}$, où $\Delta_{n,M}^{\mathbf{C}}$ est un terme résiduel qui ne dépend pas de la distribution P et H est un sous ensemble convexe non trivial de $\mathbb{R}^M$.

(3) **Sélection de modèle.** Le but est de construire $\hat{\theta}_n^{\mathbf{MS}}$ qui imite le meilleur des f_j, c'est-à-dire qui vérifie l'inégalité d'oracle

$$\mathbb{E}A(\mathsf{f}_{\hat{\theta}_n^{\mathbf{MS}}}) \leq \inf_{\theta\in\{e_1,\dots,e_M\}} A(\mathsf{f}_\theta) + \Delta_{n,M}^{\mathbf{MS}}\,,$$

pour toute distribution P dans une grande classe $\mathcal{P}$, où $\Delta_{n,M}^{\mathbf{MS}}$ est un terme résiduel qui ne dépend pas de la distribution P et les e_j sont les vecteurs de la base canonique de $\mathbb{R}^M$ définis par

$$e_j = (0,\dots,0,1,0,\dots,0) \in \mathbb{R}^M,\ j=1,\dots,M\,,$$

où le 1 est en j^{e} position. On a alors, $\mathsf{f}_{e_j} = f_j$ pour tout $j=1,\dots,M$ et

$$\inf_{\theta\in\{e_1,\dots,e_M\}} A(\mathsf{f}_\theta) = \min_{1\leq j\leq M} A(f_j)\,.$$

Dans le cadre de l'agrégation convexe, une attention particulière sera accordée au simplexe Λ^M défini par

$$\Lambda^M = \Big\{\theta = (\theta^{(1)},\dots,\theta^{(M)}) \in \mathbb{R}^M :\ \theta^{(j)} \geq 0,\ \sum_{j=1}^M \theta^{(j)} = 1\Big\}.$$

Le term *simplexe* n'est pas défini de manière précise dans la littérature et nous utiliserons parfois la définition alternative suivante (cf. chapitre 3) :

$$\Lambda^M = \Big\{\theta = (\theta^{(1)},\dots,\theta^{(M)}) \in \mathbb{R}^M :\ \theta^{(j)} \geq 0,\ \sum_{j=1}^M \theta^{(j)} \leq 1\Big\}.$$

Clairement,

$$\inf_{\theta\in\mathbb{R}^M} A(\mathsf{f}_\theta) \leq \inf_{\theta\in\Lambda^M} A(\mathsf{f}_\theta) \leq \inf_{\theta\in\{e_1,\dots,e_M\}} A(\mathsf{f}_\theta)\,,$$

alors que les plus petits termes résiduels possibles (étant donnés une fonction de perte Q et une classe de distribution $\mathcal{P}$) vérifient en général

$$\Delta_{n,M}^{\mathbf{MS}} \leq c_1\Delta_{n,M}^{\mathbf{C}} \leq c_2\Delta_{n,M}^{\mathbf{L}}\,,$$

où c_1 et c_2 sont des constantes positives. Ces termes résiduels caractérisent les *vitesses d'agrégation* pour chacun des trois problèmes standard de l'agrégation.

Le modèle général ci-dessus permet d'inclure de nombreux modèles statistiques comme cas particuliers. Ainsi, on verra qu'il permet de traiter les problèmes d'estimation non-paramétrique d'une fonction de régression et d'une densité de probabilité, la classification avec perte convexe ou encore l'estimation paramétrique d'une densité de probabilité avec la perte de Kullback-Leibler. Pour certains choix de pertes Q, il est possible de calculer les termes résiduels optimaux au sens minimax. Ils sont alors appelés *vitesses optimales d'agrégation.*

2.3. Vitesses optimales d'agrégation. Etant donnés une classe de distributions $\mathcal{P}$ sur $(\mathcal{Z},\mathfrak{F})$ et une classe de fonctions $\mathcal{F}$, le plus petit terme résiduel $\psi_n = \psi_n(\mathcal{F})$ possible caractérise la vitesse optimale d'agrégation des fonctions de $\mathcal{F}$. En voici une définition plus précise.

Définition 1.2. *Soit $\mathcal{P}$ une classe de distributions sur $(\mathcal{Z},\mathfrak{F})$ et soit $\mathcal{F}$ une classe de fonctions sur $\mathcal{Z}$ à valeurs réelles.*

Une suite positive (ψ_n) est appelée **vitesse optimale d'agrégation pour** $(\mathcal{P},\mathcal{F})$ *si elle vérifie les deux propriétés suivantes.*

– *Il existe un estimateur* $\hat{f}_n$ *(*agrégat*) tel que*

$$\sup_{P\in\mathcal{P}}\Big[\mathbb{E}A(\hat{f}_n) - \inf_{f\in\mathcal{F}} A(f)\Big] \le C\psi_n\,, \tag{1.5}$$

pour une certaine constante $C>0$ *et pour tout entier* $n\ge 1$ *.*

– *Il existe une sous classe* $\mathcal{F}'\subset\mathcal{F}$ *telle que pour tout estimateur* T_n *construit à partir d'un échantillon de taille* n *tiré selon la loi* $P\in\mathcal{P}$*, on ait*

$$\sup_{P\in\mathcal{P}}\Big[\mathbb{E}A(T_n) - \inf_{f\in\mathcal{F}'} A(f)\Big] \ge c\psi_n\,, \tag{1.6}$$

pour une certaine constante $c>0$ *et pour tout entier* $n\ge 1$ *.*

Afin d'obtenir des bornes inférieures du type (1.6), nous spécifions le modèle, c'est-à-dire la forme de la fonction Q ainsi que les classes $\mathcal{F}$ et $\mathcal{P}$.

3. Estimation adaptative d'une densité de probabilité

3.1. Agrégation d'estimateurs de densité. Une grande partie de cette thèse est consacrée à l'étude du problème de l'agrégation optimale dans le cadre particulier du modèle d'estimation d'une densité de probabilité. Il s'agit d'un modèle bien connu et étudié de manière considérable. Un estimateur de densité est un outil particulièrement informatif pour l'étude d'une loi inconnue à partir d'un ensemble de n observations $(X_1,\dots,X_n)$ issues de cette loi. C'est aussi un outil déterminant pour la communication d'informations à des non mathématiciens puisqu'il révèle certaines caractéristiques essentielles sur la loi des observations telles que la symétrie ou l'unimodalité d'une manière assez simple. Pour l'estimation d'une densité, deux approches peuvent être considérées. L'approche *paramétrique* suppose que les données sont issues d'une loi dont seulement quelques paramètres finidimensionnels sont inconnus : un exemple est donné par la distribution gaussienne de moyenne μ et de variance σ^2 inconnues. Dans ce cas, l'estimation de densité se résume à l'estimation des paramètres inconnus.

Les hypothèses paramétriques sont relativement fortes et exigent une connaissance profonde du phénomène que l'on étudie. Elles peuvent être abandonnées en utilisant des méthodes *non paramétriques* dans lesquelles les données « parlent d'elles même » lors de l'estimation (Silverman, 1986). Les estimateurs à noyau de la densité sont particulièrement étudiés et utilisés en pratique. Dans le cas particulier où les X_i sont des variables aléatoires réelles, ils sont de la forme

$$\hat{p}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\,, \tag{1.7}$$

où $h>0$ est un paramètre de lissage appelé *fenêtre* et $K:\mathbb{R}\to\mathbb{R}$ est un noyau qui vérifie certaines hypothèses techniques. Il est bien connu que le choix de h conditionne la qualité de l'estimateur de manière critique. De nombreuses méthodes utilisant les observations $(X_1,\dots,X_n)$ pour le choix de h ont été proposées. Certaines sont appréciées pour leur simplicité telle que la *règle du pouce* (Silverman, 1986; Scott, 1992). D'autres méthodes appelées *méthodes de validation croisée* minimisent un estimateur du risque. Elles remontent à Rudemo (1982); Bowman (1984) lorsque le critère de performance est la perte quadratique intégrée. Un aperçu d'un grand nombre de ces méthodes est disponible par exemple, dans Wand and Jones (1995). Toutes ces méthodes utilisent les observations pour choisir un unique paramètre h_n. Il s'agit d'un problème de sélection de modèle similaire à

celui qui a été défini dans la section précédente. Pour s'en rendre compte, introduisons le modèle d'estimation de densité de manière plus formelle.

Soient $X_1, \dots, X_n$, des variables aléatoires réelles i.i.d de densité $p \in L_2(\mathbb{R})$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$. La performance d'un estimateur $\hat{p}$ construit à partir de l'échantillon $\mathbb{X}^n = (X_1, \dots, X_n)$ est mesurée par le risque L_2 :

$$R_n(\hat{p}, p) = \mathbb{E}\|\hat{p} - p\|^2 ,$$

où $\mathbb{E}$ est le symbole générique de l'espérance et pour toute fonction $g \in L_2(\mathbb{R})$,

$$\|g\| = \Big(\int_{\mathbb{R}} g^2(x)\mathrm{d}x \Big)^{1/2} . \tag{1.8}$$

Supposons que l'on dispose d'un ensemble $\mathcal{H} = \{h_1, \dots, h_M\}$ de paramètres de lissage. Le but de la sélection du paramètre de lissage est de construire un estimateur $\tilde{p}_n$ tel que

$$R_n(\tilde{p}_n, p) \le \min_{h \in \mathcal{H}} R_n(\hat{p}_{n,h}, p) + \Delta_n , \tag{1.9}$$

où Δ_n est un terme résiduel qui ne dépend pas de p. Les méthodes de type plug-in choisissent $\tilde{p}_n$ de la forme d'un estimateur à noyau mais nous verrons que ce n'est pas nécessaire pour obtenir une inégalité du type (1.9). Ce modèle s'inscrit dans le cadre plus général de la section 2 (cf. exemple 5 p. 37).

Le choix du paramètre de lissage pour les estimateurs à noyau n'est pas la seule motivation pour agréger des estimateurs $\hat{p}_1, \dots, \hat{p}_M$ de la densité. L'agrégation est en effet un outil général qui permet de combiner toutes sortes d'estimateurs, pas nécessairement de même nature. Considérons l'exemple suivant : $M = 2$ et $\hat{p}_1$ est un bon estimateur paramétrique appartenant à une certaine famille paramétrique et $\hat{p}_2$ est un estimateur non paramétrique. Si la densité p appartient à la famille paramétrique, $\hat{p}_1$ est un excellent estimateur : son risque L_2 converge vers 0 à la vitesse paramétrique $O(1/n)$. Mais si p n'appartient pas à cette famille paramétrique, $\hat{p}_1$ peut ne même pas converger vers p. Dans les deux cas, $\hat{p}_2$ converge à une vitesse non paramétrique relativement lente. Dans cet exemple, l'agrégation combine les avantages des approches paramétrique et non paramétrique en produisant un estimateur $\tilde{p}_n$ qui est une moyenne pondérée de $\hat{p}_1$ et $\hat{p}_2$. Le poids accordé à chacun des estimateurs est d'autant plus grand qu'il est performant au vu des observations $X_1, \dots, X_n$.

3.2. Estimation adaptative. Les inégalités d'oracle constituent un résultat non asymptotique d'adaptation. Il s'agit d'*adaptation à l'oracle*. Dans le cadre de l'estimation de densité décrit ci-dessus, une inégalité d'oracle est de la forme

$$R_n(\tilde{p}_n, p) \le C_n \inf_{T_n \in \mathcal{M}_n} R_n(T_n, p) + r_n(\mathcal{M}_n) , \tag{1.10}$$

où $C_n \ge 1$, $\mathcal{M}_n$ est une classe d'estimateurs construits à partir de l'échantillon $\mathbb{X}^n$ et $r_n(\mathcal{M}_n)$ est un terme résiduel qui ne dépend pas de p. Par exemple, $\mathcal{M}_n$ peut être la classe des estimateurs à noyau K fixé pour différents paramètres de lissage :

$$\mathcal{M}_n = \{\hat{p}_{n,h} : h \in \mathcal{H}\} ,$$

où $\mathcal{H}$ est un sous ensemble de $]0, +\infty[$.

Lorsque l'oracle, possède des propriétés statistiques intéressantes, l'inégalité d'oracle (1.10) permet de *transmettre* ces propriétés à l'estimateur $\tilde{p}_n$. En effet, l'oracle est par définition proche de la densité au sens du risque R_n pourvu que la classe $\mathcal{M}_n$ soit assez grande. Une conséquence directe de l'inégalité (1.10) est que $\tilde{p}_n$ est lui aussi proche de p

quand le terme résiduel $r_n(\mathcal{M}_n)$ est négligeable devant le risque de l'oracle. Ainsi pour dériver des propriétés d'adaptation à certaines caractéristiques de la densité inconnue p, il faut choisir la classe $\mathcal{M}_n$ de manière convenable : plus $\mathcal{M}_n$ est grand (au sens de l'inclusion), plus l'oracle sur $\mathcal{M}_n$ a de bonnes propriétés statistiques mais plus $r_n(\mathcal{M}_n)$ est grand.

Pour choisir $\mathcal{M}_n$ de manière optimale, il faut spécifier la classe de densités $\mathcal{P}$ à laquelle p appartient. Elle détermine en effet la vitesse optimale de convergence au sens minimax.

DÉFINITION 1.3. *Une suite positive (ψ_n) est appelée* **vitesse de convergence minimax** *sur la classe de densités $\mathcal{P}$ si*

– *Il existe une constante $c > 0$ telle que*

$$\liminf_{n\to\infty} \psi_n^{-1}\big[\inf_{T_n}\sup_{p\in\mathcal{P}} R_n(T_n,p)\big] \geq c\,, \tag{1.11}$$

où $\inf_{T_n}$ désigne l'infimum sur l'ensemble de tous les estimateurs.

– *Il existe une constante $C > 0$ et un estimateur $\hat{p}_n$ tels que*

$$\limsup_{n\to\infty} \psi_n^{-1}\big[\sup_{p\in\mathcal{P}} R_n(\hat{p}_n,p)\big] \leq C\,. \tag{1.12}$$

Un estimateur vérifiant (1.12) *lorsque* (1.11) *est satisfaite est appelé* **estimateur optimal en vitesse de convergence**.

Dans la définition précédente, la vitesse de convergence minimax ψ_n est définie à une constante multiplicative près. On choisit donc de la représenter sous la forme conventionnelle n^α ou $n^\alpha(\log n)^\beta$, $\alpha < 0, \beta \in \mathbb{R}$, c'est-à-dire avec la constante 1. Les inégalités d'oracle exactes sont plus intéressantes à obtenir que les inégalités d'oracle approximatives quand on s'intéresse aux valeurs des constantes c et C.

DÉFINITION 1.4. *Soit (ψ_n) la vitesse de convergence minimax sur la classe de densités $\mathcal{P}$. Un estimateur p_n^* est dit* **asymptotiquement exact** *pour $\mathcal{P}$, s'il existe une constante $C^* > 0$ telle que*

$$\lim_{n\to\infty} \psi_n^{-1}\big[\inf_{T_n}\sup_{p\in\mathcal{P}} R_n(T_n,p)\big] = \lim_{n\to\infty} \psi_n^{-1}\big[\sup_{p\in\mathcal{P}} R_n(p_n^*,p)\big] = C^*\,.$$

La constante C^* est alors appelée **constante exacte** *associée à la classe $\mathcal{P}$.*

Les vitesses de convergence minimax ont été calculées pour de nombreux choix de $\mathcal{P}$, en particulier quand $\mathcal{P}$ est une classe de densités ayant une régularité fixée β. Considérons l'exemple suivant.

DÉFINITION 1.5. *Pour $\beta = 1, 2\ldots$, et $Q > 0$, on définit la classe de densités de Sobolev*

$$\mathcal{P}(\beta, Q) = \big\{p : \mathbb{R} \to \mathbb{R},\ p \geq 0, \int p = 1,\ \|p^{(\beta)}\|^2 \leq Q\big\}\,,$$

où $p^{(\beta)}$ désigne la dérivée d'ordre β de p et $\|\cdot\|$ est la norme L_2 définie en (1.8).

Dans la continuité des travaux de Efroĭmovich and Pinsker (1982) et Golubev (1992), Schipper (1996) a montré que la vitesse de convergence minimax sur $\mathcal{P}(\beta, Q)$ est $n^{-\frac{2\beta}{2\beta+1}}$ et la constante exacte est la *constante de Pinsker*

$$C^*(\beta, Q) = (2\beta+1)\Big[\frac{\pi(2\beta+1)(\beta+1)}{\beta}\Big]^{-2\beta/(2\beta+1)} Q^{1/(2\beta+1)}\,.$$

Un estimateur asymptotiquement exact pour $\mathcal{P}(\beta, Q)$ est donné par l'estimateur de densité de type Pinsker $\hat{p}_\ell^*$ défini à l'aide de la transformée de Fourier inverse par

$$\hat{p}_\ell^*(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega x} \hat{\varphi}_\ell(\omega) \mathrm{d}\omega \,.$$

La fonction $\hat{\varphi}_\ell$ est définie par $\hat{\varphi}_\ell(\omega) = (1 - \kappa^*|\omega|^\beta)_+ \varphi_n(\omega)$, où φ_n est la fonction caractéristique empirique, $(x)_+$ désigne la partie positive de x et

$$\kappa^* = \left(\frac{2\beta}{(2\beta+1)(\beta+1)Q} \right)^{\frac{\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \,.$$

On peut montrer que l'estimateur $\hat{p}_\ell^*$ est un estimateur à noyau de la forme (1.7) avec le noyau $K = K_\beta$ donné par

$$K_\beta(x) = \frac{\beta!}{\pi} \sum_{j=1}^{\beta} \frac{\sin^{(j)}(x)}{(\beta - j)! x^{j+1}} \,,$$

et le paramètre de lissage

$$h = D^* n^{-\frac{1}{2\beta+1}} \quad \text{où} \quad D^* = \left(\frac{2\beta}{Q(\beta+1)(2\beta+1)} \right)^{\frac{1}{2\beta+1}} .$$

Ce résultat est généralisé à des β non entiers, $\beta > 1/2$, dans le corollaire 4.1 ainsi que dans Dalelane (2005a). Il est à noter que l'estimateur $\hat{p}_\ell^*$ dépend des quantités β et Q inconnues et n'est donc en pratique pas calculable. On cherche alors à s'*adapter* aux paramètres inconnus.

Supposons plus généralement que la classe de densités $\mathcal{P}$ inconnue appartient à une famille de classes $\{\mathcal{P}_\gamma\}_{\gamma \in G}$, indexée par un paramètre inconnu γ à valeurs dans un ensemble G donné. Dans l'exemple des classes de Sobolev, on a

$$\gamma = (\beta, Q), \quad G = \mathbb{N}^* \times \mathbb{R}_+^*, \quad \mathcal{P}_\gamma = \mathcal{P}(\beta, Q) \,.$$

Définition 1.6. *Un estimateur $\hat{p}_n$ est dit* **adaptatif en vitesse de convergence** *sur la famille de classes $\{\mathcal{P}_\gamma\}_{\gamma \in G}$, s'il existe une constante $C > 0$ telle que*

$$\limsup_{n \to \infty} \sup_{\gamma \in G} \left[\psi_n^{-1}(\gamma) \sup_{p \in \mathcal{P}_\gamma} R_n(\hat{p}_n, p) \right] \leq C \,,$$

où $\psi_n(\gamma)$ est la vitesse de convergence minimax sur la classe $\mathcal{P}_\gamma$.

Un estimateur p_n^* est dit **asymptotiquement exact adaptatif** *sur la famille $\{\mathcal{P}_\gamma\}_{\gamma \in G}$, s'il existe une famille de constantes $C^*(\gamma) > 0$ telles que*

$$\lim_{n \to \infty} \sup_{\gamma \in G} \left[\big(C^*(\gamma)\psi_n(\gamma)\big)^{-1} \inf_{T_n} \sup_{p \in \mathcal{P}} R_n(T_n, p) \right] = \lim_{n \to \infty} \sup_{\gamma \in G} \left[\big(C^*(\gamma)\psi_n(\gamma)\big)^{-1} \sup_{p \in \mathcal{P}} R_n(p_n^*, p) \right] = 1 \,.$$

Il existe différentes méthodes de construction d'estimateurs adaptatifs de la densité. La plus générale est la méthode de Lepskiĭ (Lepskiĭ, 1990, 1991, 1992a,b) qui est appliquée au problème d'estimation de la densité dans Butucea (2001). Les méthodes de validation croisée ou plus généralement de minimisation d'un estimateur sans biais du risque sont plus connues et sont utilisées de manières très variées (cf., e.g., Efroĭmovich, 1985; Golubev, 1992; Dalelane, 2005b; Rigollet, 2006a). Citons enfin les méthodes d'adaptation par seuillage des coefficients d'ondelettes dont un aperçu est donné, par exemple, dans Härdle *et al.* (1998). Plus récemment, des méthodes basées sur l'agrégation ont été explorées. Certaines sont présentées dans cette these au chapitre 5. Une autre idée d'utilisation de

l'agrégation pour l'adaptation est liée à l'algorithme de BOOSTING (Bühlmann and Yu, 2003; Bickel and Ritov, 2004; Bickel *et al.*, 2006).

3.3. Inégalités d'oracle et adaptation. Les inégalités d'oracle constituent un outil particulièrement commode pour prouver le caractère adaptatif d'un estimateur. Aussi, l'agrégation d'estimateurs non adaptatifs permet d'obtenir des estimateurs adaptatifs. Une méthode générale est décrite ci-dessous.

Soit une famille de classes de densités $\{\mathcal{P}_\gamma\}_{\gamma\in G}$ telle que la densité p appartienne à l'une des classes $\mathcal{P}_\gamma, \gamma \in G$. Supposons que l'on dispose d'un estimateur $\tilde{p}_n$ vérifiant une inégalité d'oracle du type (1.10) pour toute densité $p \in \mathcal{P}_\gamma$, $\gamma \in G$. Supposons de plus que $\mathcal{M}_n$ est une classe d'estimateurs qui contient l'ensemble $\{p^*_{n,\gamma}, \gamma \in G\}$ où, pour chaque $\gamma \in G$, l'estimateur $p^*_{n,\gamma}$ est asymptotiquement exact pour $\mathcal{P}_\gamma$ pour un γ fixé dans G. En prenant le supremum sur $\mathcal{P}_\gamma$ dans chaque membre de (1.10), on obtient

$$\begin{aligned}\sup_{p\in\mathcal{P}_\gamma} R_n(\tilde{p}_n, p) &\le C_n \sup_{p\in\mathcal{P}_\gamma} \inf_{T_n\in\mathcal{M}_n} R_n(T_n, p) + r_n(\mathcal{M}_n)\\ &\le C_n \sup_{p\in\mathcal{P}_\gamma} R_n(p^*_{n,\gamma}, p) + r_n(\mathcal{M}_n)\\ &\le C_n C^*(\gamma)\psi_n(\gamma)(1+o(1)) + r_n(\mathcal{M}_n)\,.\end{aligned}$$

Donc si le terme résiduel $r_n(\mathcal{M}_n)$ est négligeable par rapport à $\psi_n(\gamma)$, l'estimateur $\tilde{p}_n$ est adaptatif en vitesse de convergence. Si de plus, la quantité C_n tend vers 1, c'est-à-dire que l'inégalité d'oracle (1.10) est asymptotiquement exacte, alors l'estimateur $\tilde{p}_n$ est asymptotiquement exact adaptatif sur la famille $\{\mathcal{P}_\gamma\}_{\gamma\in G}$.

4. Excès de risque en classification

Le problème de classification binaire donne lieu à certains résultats qui peuvent être interprétés comme des inégalités d'oracle.

4.1. Le modèle de classification binaire. Soit $(\mathcal{X}, \mathfrak{X})$ un sous espace mesurable de $\mathbb{R}^d$. Soit (X, Y) un couple aléatoire de loi P où $X \in \mathcal{X} \subset \mathbb{R}^d$ est un vecteur de $d \ge 1$ caractéristiques et $Y \in \{0, 1\}$ est une étiquette indiquant à quelle classe X appartient. La loi jointe P du couple (X, Y) est entièrement déterminée par le couple (P_X, η) où P_X est la loi marginale de X et η est la fonction de régression de Y sur X, c'est-à-dire que $\eta(x) = P(Y = 1|X = x)$, $x \in \mathcal{X}$. Le but de la classification est de prédire Y étant donnée la valeur de X, c'est-à-dire de construire une fonction mesurable $g : \mathcal{X} \to \{0, 1\}$ appelée *classifieur* ou *fonction de décision*. La performance de g est mesurée par son *erreur moyenne de classification*

$$R(g) = P(g(X) \ne Y)\,.$$

Soit la fonction de décision de Bayes

$$g^\star(x) = \mathbb{1}_{\{\eta(x)\ge 1/2\}},\ x \in \mathcal{X}\,.$$

Il s'agit de la fonction de décision qui attribue à $x \in \mathcal{X}$, l'étiquette la plus probable. La proposition suivante fait de $g^\star$ un oracle.

PROPOSITION 1.1. *Pour toute fonction de décision $g : \mathcal{X} \to \{0, 1\}$,*

$$R(g^\star) \le R(g)\,.$$

Une preuve de cette proposition est disponible par exemple dans Devroye *et al.* (1996, Théorème 2.1).

4.2. Excès de risque et inégalités d'oracle. Supposons maintenant que l'on dispose de n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, i.i.d. de loi P et indépendantes de (X, Y). Une *fonction de décision empirique* est une fonction aléatoire $\hat{g}_n : \mathcal{X} \to \{0, 1\}$ construite à partir de l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$. Puisque $g^\star$ possède le plus petit risque R possible, il est naturel de mesurer la performance d'une fonction de décision empirique par son *excès de risque*

$$\mathcal{E}(\hat{g}_n) = \mathbb{E}_n R(\hat{g}_n) - R(g^\star)\,,$$

où $\mathbb{E}_n$ désigne l'espérance par rapport à la loi jointe de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Supposons alors qu'il existe une fonction de décision $\hat{g}_n$ et une suite positive (ψ_n) tendant vers 0 telles que

$$\mathcal{E}(\hat{g}_n) \leq c\psi_n, \quad \forall\ n \geq 1\,, \tag{1.13}$$

où c est une constante positive. La borne supérieure sur l'excès de risque (1.13) peut être interprétée comme une inégalité d'oracle exacte où l'oracle est la fonction de décision de Bayes $g^\star$. On a en effet

$$\mathbb{E}_n R(\hat{g}_n) \leq \inf_g R(g) + c\psi_n\,.$$

4.3. Classification semi-supervisée. Dans ce modèle, en plus de l'échantillon initial $(X_1, Y_1), \ldots, (X_n, Y_n)$, on dispose d'un échantillon indépendant $(X_{n+1}, \ldots, X_{n+m})$ où les $X_i, i = n+1, \ldots, n+m$, sont i.i.d. de loi P_X, la loi marginale de X. Ce type de modèle apparaît notamment quand la récolte des données non étiquetées est une tâche relativement aisée alors que l'étiquetage nécessite l'intervention d'experts entraînant des coûts élevés. On peut citer comme exemples la classification de contenu de pages web ou d'images. Il est donc naturel de supposer que $m \gg n$. Ces observations apportent une information supplémentaire sur la loi marginale P_X. Il existe essentiellement deux façons d'incorporer cette information pour améliorer les bornes sur l'excès de risque.

La première ne demande pas d'hypothèse supplémentaire mais s'inscrit dans le cadre d'une méthodologie particulière. Supposons que l'on dispose d'un ensemble de fonctions de décision à agréger. On peut dans ce cas utiliser les données non étiquetées pour mesurer la compatibilité des fonctions de décision à agréger afin d'en réduire la complexité.

La méthode retenue au Chapitre 7 consiste à faire une hypothèse qui relie le comportement de la fonction de régression η à la loi marginale P_X. Sous ce type d'hypothèse, une meilleure connaissance de P_X à travers les données non étiquetées implique une meilleure connaissance de η et permet donc d'améliorer la performance d'une fonction de décision.

5. Estimation des ensembles de niveau de densité par la méthode plug-in

5.1. Présentation du problème. Soit $(\mathcal{X}, \mathfrak{X})$ un sous espace mesurable de $\mathbb{R}^d$ et soit Q une mesure positive σ-finie sur $\mathcal{X}$. Soit P une mesure de probabilité admettant une densité p par rapport à Q. Pour une valeur fixée $\lambda > 0$, on définit l'ensemble de niveau λ de la densité p par

$$\Gamma_p(\lambda) = \left\{x \in \mathcal{X} \ :\ p(x) \geq \lambda\right\}.$$

Soit $(X_1, \ldots, X_n)$ un échantillon de n vecteurs de $\mathcal{X}$ ayant comme densité commune p inconnue. Un *estimateur plug-in* de $\Gamma_p(\lambda)$ est de la forme

$$\widehat{\Gamma}(\lambda) = \left\{x \in \mathcal{X} \ :\ \hat{p}_n(x) \geq \lambda\right\},$$

où $\hat{p}_n$ est un estimateur de la densité p construit à partir des données.

5.2. Mesures de performance. Soit $\hat{G}_n$ un estimateur de l'ensemble $\Gamma_p(\lambda)$, par exemple de type plug-in mais pas nécessairement. Etant donnée une pseudo-distance d entre deux ensembles $G_1 \subset \mathcal{X}$ et $G_2 \subset \mathcal{X}$, on mesure la performance de l'estimateur $\hat{G}_n$ par

$$\mathbb{E}d\big(\hat{G}_n, \Gamma_p(\lambda)\big)\,,$$

où $\mathbb{E}$ est le symbole de l'espérance. Définissons la mesure $\tilde{Q}_\lambda$ sur $(\mathcal{X}, \mathfrak{X})$ par

$$\tilde{Q}_\lambda(\mathrm{d}x) = |p(x) - \lambda| Q(\mathrm{d}x)\,.$$

Deux pseudo-distances s'imposent naturellement dans le cadre de l'estimation des ensembles de niveau de la densité.

(i) La mesure Q de la différence symétrique entre les ensembles :

$$d_\triangle(G_1, G_2) = Q(G_1 \triangle G_2)\,.$$

(*ii*) La mesure $\tilde{Q}_\lambda$ de la différence symétrique entre les ensembles :

$$d_H(G_1, G_2) = \tilde{Q}_\lambda(G_1 \triangle G_2) = \int_{G_1 \triangle G_2} |p(x) - \lambda| \mathrm{d}Q(x)\,.$$

La pseudo-distance $d_\triangle$ est naturelle lorsqu'on estime des ensembles sans connaissance a priori sur le problème. Cependant la qualité d'un estimateur $\hat{G}$ de l'ensemble de niveau λ de la densité peut aussi être mesurée par son *excès de masse* $H(\hat{G})$ défini comme suit (voir Hartigan, 1987; Müller and Sawitzki, 1987) :

$$H(\hat{G}) = P(\hat{G}) - \lambda Q(\hat{G})\,.$$

L'excès de masse mesure la concentration de la mesure de probabilité P dans l'ensemble $\hat{G}$ et présente la particularité d'être maximisé par $\Gamma_p(\lambda)$. Il est donc naturel de mesurer la performance d'un estimateur $\hat{G}_n$ par son *déficit moyen d'excès de masse*

$$\mathcal{D}(\hat{G}_n) = H\big(\Gamma_p(\lambda)\big) - \mathbb{E}H\big(\hat{G}_n\big)\,.$$

On peut montrer que, pour tout ensemble $G \subset \mathcal{X}$,

$$H\big(\Gamma_p(\lambda)\big) - H\big(G\big) = d_H\big(\Gamma_p(\lambda), G\big)\,,$$

et par conséquent

$$\mathcal{D}(\hat{G}_n) = \mathbb{E}d_H(\Gamma_p(\lambda), \hat{G}_n)\,.$$

Supposons alors qu'il existe un estimateur $\hat{G}_n$ de l'ensemble de niveau λ de la densité, $\Gamma_p(\lambda)$ et une suite positive (ψ_n) tendant vers 0 telles que

$$\mathcal{D}(\hat{G}_n) \leq c\psi_n, \quad \forall\ n \geq 1\,, \tag{1.14}$$

où c est une constante positive. La borne supérieure sur le déficit d'excès de masse (1.14) peut être interprétée comme une inégalité d'oracle exacte où l'oracle est l'ensemble de niveau $\Gamma_p(\lambda)$ et la performance est mesurée par $-H$, où H est l'excès de masse. En effet l'inégalité (1.14) est équivalente à

$$\mathbb{E}\big[-H\big(\hat{G}_n\big)\big] \leq -H\big(\Gamma_p(\lambda)\big) + c\psi_n\,. \tag{1.15}$$

6. Plan de la thèse

Les contributions de cette thèse s'articulent en trois parties.

La première partie intitulée *Aggregation* traite des trois problèmes d'agrégation pure décrits précédemment. Les résultats du chapitre 2 concernent la sélection de modèle pour une perte convexe dans le cadre général de la section 2. On prouve deux inégalités d'oracle

du type (1.4). Les résultats obtenus s'appliquent directement à différents modèles d'estimation non paramétrique tels que la régression, la classification ou l'estimation de densité. Le cadre de travail est proche de celui de Iouditski *et al.* (2005) où des résultats similaires pour l'agrégation convexe sont énoncés. L'algorithme utilisé est aussi inspiré de celui de Iouditski *et al.* (2005), à savoir la méthode de descente miroir avec moyennisation. Le chapitre 3 examine l'agrégation dans le modèle d'estimation de la densité avec une perte quadratique. Pour chacun des trois problèmes standard de l'agrégation on exhibe les vitesses optimales en construisant des inégalités du type (1.5) et (1.6).

La seconde partie contient différents résultats sur l'estimation de densité en perte quadratique. Après avoir énoncé un théorème de type Pinsker (1980) transposé au modèle d'estimation de densité, on s'intéresse à l'utilisation des inégalités d'oracle pour dériver des propriétés d'adaptation au sens minimax exact. Une des méthodes consiste à agréger des estimateurs à noyau de la densité tandis que l'autre repose sur la méthode par blocs de Stein. En outre, on montre que les estimateurs proposés vérifient des inégalités d'oracle dites « à noyau », c'est-à-dire qu'ils imitent l'oracle sur de grandes classes d'estimateurs à noyau.

La troisième et dernière partie est la compilation de deux résultats qui illustrent comment les inégalités d'oracle peuvent être perçues comme des résultats précis et non plus comme de simples outils mathématiques. Dans le problème de classification semi-supervisée traité au chapitre 7, on donne des bornes sur une partie de l'excès de risque permettant d'observer distinctement les effets des données étiquetées et des données non étiquetées sur les vitesses de convergence. La méthodologie utilisée fait appel à l'estimation des ensembles de niveau d'une densité par méthodes plug-in dont une étude détaillée est réalisée au chapitre 8. On y introduit notamment la notion de γ-exposant d'une densité, $\gamma > 0$ qui permet de prouver des inégalités d'oracle du type (1.15) avec un terme résiduel qui peut être proche de $O(1/n)$ lorsque γ tend vers l'infini. Des bornes inférieures attestant de l'optimalité du terme résiduel sont également démontrées.

Part 1

Aggregation

CHAPTER 2

Learning by mirror averaging

This chapter describes a general method for the model selection problem in aggregation. Given a collection of M different estimators or classifiers, we study the problem of model selection type aggregation, i.e., we construct a new estimator or classifier, called aggregate, which is nearly as good as the best among them with respect to a given risk criterion. We define our aggregate by a simple recursive procedure which solves an auxiliary stochastic linear programming problem related to the original non-linear one and constitutes a special case of the mirror averaging algorithm. We show that the aggregate satisfies sharp oracle inequalities under some general assumptions. The results allow one to construct in an easy way sharp adaptive nonparametric estimators for several problems including regression, classification and density estimation.

Contents

The material of this chapter is a joint work with Anatoli Juditsky and Alexandre Tsybakov (Juditsky *et al.*, 2006).

1. Introduction

Several problems in statistics and machine learning can be stated as follows: given a collection of M different estimators, construct a new estimator which is nearly as good as the best among them with respect to a given risk criterion. This target is called model selection (MS) type aggregation, and it can be described in terms of the following stochastic optimization problem.

Let $(\mathcal{Z}, \mathfrak{F})$ be a measurable space and let Λ^M be the simplex

$$\Lambda^M = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^{M} \theta^{(j)} = 1,\ \theta^{(j)} \geq 0,\ j = 1, \dots, M \right\}.$$

Here and throughout the chapter we suppose that $M \geq 2$ and we denote by $z^{(j)}$ the jth component of a vector $z \in \mathbb{R}^M$. We denote by $\left[z^{(j)}\right]_{j=1}^{M}$ the vector $z = (z^{(1)}, \dots, z^{(M)})^\top \in \mathbb{R}^M$.

Let Z be a random variable with values in $\mathcal{Z}$. The distribution of Z is denoted by P and the corresponding expectation by E. Suppose that P is unknown and that we observe n i.i.d. random variables $Z_1, \dots, Z_n$ with values in $\mathcal{Z}$ having the same distribution as Z. The distribution (respectively, expectation) w.r.t. the sample $Z_1, \dots, Z_n$ is denoted by P_n (respectively, by E_n).

Consider a measurable function $Q : \mathcal{Z} \times \Lambda^M \to \mathbb{R}$ and the corresponding average risk function

$$A(\theta) = EQ(Z, \theta)\,,$$

assuming that this expectation exists for all $\theta \in \Lambda^M$. Stochastic optimization problems that are usually studied in this context consist in minimization of A on some subsets of Λ^M, given the sample $Z_1, \dots, Z_n$. Note that since the distribution of Z is unknown, direct (deterministic) minimization of A is not possible.

For $j \in \{1, \dots, M\}$, denote by e_j the jth coordinate unit vector in $\mathbb{R}^M$: $e_j = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^M$, where 1 appears in jth position.

The stochastic optimization problem associated to MS aggregation is

$$\min_{\theta \in \{e_1, \dots, e_M\}} A(\theta)\,.$$

The aim of MS aggregation is to "mimic the oracle" $\min_{1 \le j \le M} A(e_j)$, i.e., to construct an estimator $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \dots, Z_n$ and called aggregate, such that

$$E_n A(\tilde{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \Delta_{n,M}\,, \tag{2.1}$$

where $\Delta_{n,M} > 0$ is a remainder term that should be as small as possible.

As an example, one may consider the loss function of the form $Q(z, \theta) = \ell(z, \theta^\top H)$ where $\ell : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ and $H = (h_1, \dots, h_M)^\top$ is a vector of preliminary estimators (classifiers) constructed from a training sample which is supposed to be frozen in our considerations (thus, h_j can be viewed as fixed functions). The value $A(e_j) = E\ell(Z, h_j)$ is the risk corresponding to h_j. We now make a slight abuse of language and also call the estimator $\tilde{\theta}_n^\top H$ an aggregate. Inequality (2.1) can then be interpreted as follows: the aggregate $\tilde{\theta}_n^\top H$, i.e. the convex combination of initial estimators (classifiers) h_j, with the vector of mixture coefficients $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \dots, Z_n$, is nearly as good as the best among $h_1, \dots, h_M$. The word "nearly" here means that the value $\min_{1 \le j \le M} A(e_j)$ is reproduced up to a reasonably small remainder term $\Delta_{n,M}$. Lower bounds can be established showing that, under some assumptions, the smallest possible value of $\Delta_{n,M}$ in a minimax sense has the form

$$\Delta_{n,M} = \frac{C \log M}{n}\,, \tag{2.2}$$

with some constant $C > 0$ (cf. Chapter 3).

The aim of this chapter is to obtain bounds of the form (2.1) – (2.2) under some general conditions on the loss function Q. For two special cases (density estimation with the Kullback-Leibler (KL) loss, and regression model with squared loss) such bounds have been proved earlier in the benchmark works of Catoni (1997, 1999, 2004) and Yang (2000). They independently obtained the bound for density estimation with the KL loss, and Catoni (1999, 2004) solved the problem for the regression model with squared loss. Bunea and Nobel (2005) suggested another proof of the regression result of Catoni (1999, 2004) improving it in the case of bounded response, and obtained some inequalities with suboptimal remainder terms under weaker conditions. For a problem which is different

but close to ours (MS aggregation in the Gaussian white noise model with squared loss) Nemirovski (2000, p. 226) established an inequality similar to (2.1), with a suboptimal remainder term. Leung and Barron (2004) improved upon this result to achieve optimality.

Several other works provided less precise bounds than (2.1) – (2.2), with the risk of the oracle $\min_{1\le j\le M} A(e_j)$ replaced by $K \min_{1\le j\le M} A(e_j)$ in (2.1) and with a remainder term which is sometimes larger than the optimal one (2.2). A detailed account can be found in the survey of Boucheron *et al.* (2005) or in the lecture notes of Massart (2006). We mention here only some recent work where aggregation of arbitrary estimators is considered: Wegkamp (2003); Bartlett *et al.* (2002); Lugosi and Wegkamp (2004); Yang (2004); Zhang (2006); Bunea *et al.* (2004); Samarov and Tsybakov (2005). These results are useful for statistical applications, especially if K is close to 1. However, the inequalities with $K > 1$ do not provide valid bounds for the excess risk $\mathcal{E}(\tilde\theta_n)$ defined by

$$\mathcal{E}(\tilde\theta_n) = E_n A(\tilde\theta_n) - \min_{1\le j\le M} A(e_j)\,.$$

As a consequence, they do not show that $\tilde\theta_n$ approximately solves the stochastic optimization problem.

Here we study the aggregate $\hat\theta_n$ which is defined by a simple recursive procedure. The procedure solves an auxiliary stochastic linear programming problem related to the original non-linear one $\min_{1\le j\le M} A(e_j)$, and it constitutes a special case of the mirror averaging algorithm of Iouditski *et al.* (2005). In particular cases, it yields the methods described by Catoni (2004) and Yang (2000). We prove that the mirror averaging aggregate $\hat\theta_n$ satisfies oracle inequalities (2.1) – (2.2) under some general assumptions on Q such as, for example, exponential concavity. We show that these assumptions are fulfilled for several statistical models including regression, classification and density estimation.

Our results have a connection to the theory of on-line prediction of individual deterministic sequences (cf. Cesa-Bianchi and Lugosi, 2006). A general problem considered there is for an agent to compete against the observed predictions of a group of experts, so that the agent's error is close to that of the best expert. The results and the methods are similar to ours, in particular, oracle inequalities can be obtained for deterministic bounded or binary sequences of observations (Vovk, 1990; Kivinen and Warmuth, 1999). However, they cannot be meaningfully applied to our stochastic setup because they deal with the cumulative loss rather than with the expected loss, and the algorithms of deterministic prediction do not have the averaging step [cf. (2.5), (2.8) below]. For more references and discussion we refer to Cesa-Bianchi and Lugosi (2006).

2. The algorithm

We first recall the definition of a particular version of the mirror averaging algorithm. For $\beta > 0$ define the function $W_\beta : \mathbb{R}^M \to \mathbb{R}$ by

$$W_\beta(z) \triangleq \beta \log\left(\frac{1}{M}\sum_{j=1}^{M} \mathrm{e}^{-z^{(j)}/\beta}\right), \qquad z = (z^{(1)},\dots,z^{(M)}). \tag{2.3}$$

The gradient of W_β is given by

$$\nabla W_\beta(z) = \left[-\frac{\mathrm{e}^{-z^{(j)}/\beta}}{\sum_{k=1}^{M}\mathrm{e}^{-z^{(k)}/\beta}}\right]_{j=1}^{M}.$$

Consider now $Q(z,\theta)$ which is convex and differentiable in θ for all $z \in \mathcal{Z}$, with gradient $\nabla_\theta Q(z,\theta)$. Consider a non decreasing sequence $(\beta_i)_{i\geq 0}$. The algorithm described in Figure 1 is a particular case of the general mirror averaging algorithm of Iouditski *et al.* (2005). The name mirror averaging is due to the fact that (2.4) does a stochastic gradient

MIRROR AVERAGING

- Fix the initial values $\bar\theta_0 \in \Lambda^M$, $\zeta_0 = 0 \in \mathbb{R}^M$ and $\beta_0 > 0$.
- For $i = 1, \dots, n-1$, do the recursive update

$$\begin{aligned} \beta_i &= \beta_0\sqrt{i+1}\,, \\ \zeta_i &= \zeta_{i-1} + \nabla_\theta Q(Z_i, \bar\theta_{i-1})\,, \\ \bar\theta_i &= -\nabla W_{\beta_i}(\zeta_i)\ . \end{aligned} \tag{2.4}$$

- Output at iteration n the average:

$$\tilde\theta_n = \frac{1}{n}\sum_{i=1}^{n} \bar\theta_{i-1}\,. \tag{2.5}$$

FIGURE 1. A particular instance of the Mirror Averaging algorithm (MA) introduced in Iouditski *et al.* (2005).

descent in the dual space with further "mirroring" to the primal space (so-called mirror descent, cf. Ben-Tal and Nemirovski, 1999) and averaging. For more details and discussion see Iouditski *et al.* (2005). They show that when $\beta_0 = C_\beta(\log M)^{-1/2}$, for an appropriate constant C_β and under some additional assumptions the following oracle inequality holds (cf. Iouditski *et al.*, 2005, Theorem 2)

$$E_n A(\tilde\theta_n) \leq \min_{\theta\in\Lambda^M} A(\theta) + C_0\sqrt{\frac{\log M}{n}}\,, \tag{2.6}$$

where $C_0 > 0$ is a constant depending only on C_β and on the supremum norm of the gradient $\nabla_\theta Q(\cdot,\cdot)$.

Note that in (2.6) the minimum is taken over the whole simplex Λ^M, so an inequality of the type (2.1) holds as well, but for large n the remainder term in (2.6) is of larger order than the optimal one given in (2.2).

To improve upon this, consider another version of the mirror averaging method. If A is a convex function, we can bound it from above by a linear function:

$$A(\theta) \leq \sum_{j=1}^{M} \theta^{(j)} A(e_j) \triangleq \tilde A(\theta), \quad \forall\ \theta \in \Lambda^M,$$

where

$$\tilde A(\theta) = E\,\tilde Q(Z,\theta) \quad \text{with} \quad \tilde Q(Z,\theta) \triangleq \theta^\top u(Z), \quad u(Z) \triangleq \Big(Q(Z,e_1), \dots, Q(Z,e_M)\Big)^\top .$$

Note that

$$\tilde A(e_j) = A(e_j), \quad j = 1, \dots, M.$$

Since Λ^M is a simplex, the minimum of the linear function $\tilde A$ is attained at one of its vertices. Therefore,

$$\min_{\theta\in\Lambda^M} \tilde A(\theta) = \min_{1\leq j\leq M} A(e_j).$$

Thus, the problem of MS aggregation can be solved using a mirror averaging algorithm for the linear stochastic programming problem of minimization of $\tilde{A}$ on Λ^M. It is defined as follows.

Consider the vector

$$u_i \triangleq \Big(Q(Z_i, e_1), \dots, Q(Z_i, e_M)\Big)^\top = u(Z_i) = \nabla_\theta \tilde{Q}(Z_i, \theta)\,,$$

and the iterations:

LINEAR MIRROR AVERAGING

- Fix the initial values $\theta_0 \in \Lambda^M$ and $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1, \dots, n-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + u_i\,, \\ \theta_i &= -\nabla W_\beta(\zeta_i)\,. \end{aligned} \tag{2.7}$$

- Output at iteration n the average

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \theta_{i-1}\,. \tag{2.8}$$

FIGURE 2. Linear Mirror Averaging algorithm (LMA).

Remark that we define the algorithm LMA for a general function Q, not necessarily for a convex one. Note also that $\hat{\theta}_n$ is measurable w.r.t. the subsample $(Z_1, \dots, Z_{n-1})$. The components $\theta_i^{(j)}$ of the vector θ_i from (2.7) can be written in the form

$$\theta_i^{(j)} = \frac{\exp\left(-\beta^{-1}\sum_{m=1}^{i} Q(Z_m, e_j)\right)}{\sum_{k=1}^{M}\exp\left(-\beta^{-1}\sum_{m=1}^{i} Q(Z_m, e_k)\right)}, \quad j = 1, \dots, M.$$

Since LMA is a particular case of the mirror averaging algorithm MA corresponding to a linear function $\tilde{A}$, inequality (2.6) remains valid with A replaced by $\tilde{A}$. But we show below that in fact $\hat{\theta}_n$ satisfies a stronger inequality, i.e., one with the optimal remainder term (2.2).

Finally, note that W_β defined in (2.3) is not the only possible choice: other functions W_β satisfying the conditions described in Iouditski *et al.* (2005) can be used to construct the updates (2.7).

3. Main results

In this section we give two theorems. They establish results in the form of oracle inequalities satisfied by $\hat{\theta}_n$. Theorem 2.1 requires a more conservative assumption on the loss functions Q than Theorem 2.2. This assumption is easier to check, and it often leads to a sharper bound but not for such models as nonparametric density estimation with the L_2 loss which will be treated using Theorem 2.2. In some cases (for example, in regression with Gaussian noise) Theorem 2.1 yields a suboptimal remainder term, while Theorem 2.2

does the correct job. In both theorems it is supposed that the values $A(e_1), \dots, A(e_M)$ are finite. We will also need the following definition.

DEFINITION 2.1. *A function $T : \mathbb{R}^M \to \mathbb{R}$ is* exponentially concave *if the composite function $\exp \circ T$ is concave.*

It is straightforward to see that exponential concavity of a function $-T$ implies that T is convex. Furthermore, if $-T/\beta$ is exponentially concave for some $\beta > 0$, then $-T/\beta'$ is exponentially concave for all $\beta' > \beta$. Let Q_1 be the function on $\mathcal{Z} \times \Lambda^M \times \Lambda^M$ defined by $Q_1(z, \theta, \theta') = Q(z, \theta) - Q(z, \theta')$ for all $z \in \mathcal{Z}$ and all $\theta, \theta' \in \Lambda^M$.

THEOREM 2.1. *Assume that Q_1 can be decomposed into the sum of two functions $Q_1 = Q_2 + Q_3$ such that:*

- *The mapping $\theta \mapsto -Q_2(z, \theta, \theta')/\beta$ is exponentially concave on the simplex Λ^M, for all $z \in \mathcal{Z}$, $\theta' \in \Lambda^M$, and $Q_2(z, \theta, \theta) = 0$ for all $z \in \mathcal{Z}, \theta \in \Lambda^M$.*
- *There exists a function R on $\mathcal{Z}$ integrable w.r.t. P and such that $-Q_3(z, \theta, \theta') \le R(z)$, for all $z \in \mathcal{Z}$, $\theta, \theta' \in \Lambda^M$.*

Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \ge 2, n \ge 1$, the following oracle inequality

$$E_{n-1} A(\hat{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \frac{\beta \log M}{n} + E\big[R(Z)\big] .$$

THEOREM 2.2. *Assume that for some $\beta > 0$ there exists a Borel function $\Psi_\beta : \Lambda^M \times \Lambda^M \to \mathbb{R}_+$ such that the mapping $\theta \mapsto \Psi_\beta(\theta, \theta')$ is concave on the simplex Λ^M for any fixed $\theta' \in \Lambda^M$, $\Psi_\beta(\theta, \theta) = 1$ and $E \exp(-Q_1(Z, \theta, \theta')/\beta) \le \Psi_\beta(\theta, \theta')$ for all $\theta, \theta' \in \Lambda^M$. Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \ge 2, n \ge 1$, the following oracle inequality*

$$E_{n-1} A(\hat{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \frac{\beta \log M}{n} .$$

Proofs of both theorems are based on the following lemma. Introduce the discrete random variable ω with values in the set $\{e_1, \dots, e_M\}$ and with the distribution $\mathbb{P}$ defined conditionally on $(Z_1, \dots, Z_{n-1})$ by $\mathbb{P}[\omega = e_j] = \hat{\theta}_n^{(j)}$, where $\hat{\theta}_n^{(j)}$ is the jth component of $\hat{\theta}_n$. The expectation corresponding to $\mathbb{P}$ is denoted by $\mathbb{E}$.

LEMMA 2.1. *For any measurable function Q and any $\beta > 0$ we have*

$$E_{n-1} A(\hat{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \frac{\beta \log M}{n} + S_1. \tag{2.9}$$

where

$$S_1 \triangleq \beta E_n \log \left(\mathbb{E} \exp \Big[- \frac{Q_1(Z_n, \omega, \mathbb{E}[\omega])}{\beta} \Big] \right).$$

PROOF. By definition of $W_\beta(\cdot)$, for $i = 1, \dots, n$,

$$W_\beta(\zeta_i) - W_\beta(\zeta_{i-1}) = \beta \log \left(\frac{\sum_{j=1}^M \mathrm{e}^{-\zeta_i^{(j)}/\beta}}{\sum_{j=1}^M \mathrm{e}^{-\zeta_{i-1}^{(j)}/\beta}} \right) = \beta \log \big(- v_i^\top \nabla W_\beta(\zeta_{i-1}) \big) = \beta \log \big(v_i^\top \theta_{i-1} \big) , \tag{2.10}$$

where

$$v_i = \left[\exp \left(- \frac{u_i^{(j)}}{\beta} \right) \right]_{j=1}^M .$$

Taking expectations on both sides of (2.10), summing up over i, using the fact that (θ_{i-1}, Z_i) has the same distribution as (θ_{i-1}, Z_n) for $i = 1, \dots, n$, and applying the Jensen

inequality, we get

$$\begin{aligned}\frac{E_n\big[W_\beta(\zeta_n)-W_\beta(\zeta_0)\big]}{n} &= \frac{\beta}{n}\sum_{i=1}^n E_n\log\Big(\sum_{j=1}^M \theta_{i-1}^{(j)}\exp\Big[-\frac{Q(Z_i,e_j)}{\beta}\Big]\Big)\\ &= \frac{\beta}{n}\sum_{i=1}^n E_n\log\Big(\sum_{j=1}^M \theta_{i-1}^{(j)}\exp\Big[-\frac{Q(Z_n,e_j)}{\beta}\Big]\Big)\\ &\le \beta E_n\log\Big(\sum_{j=1}^M \hat\theta_n^{(j)}\exp\Big[-\frac{Q(Z_n,e_j)}{\beta}\Big]\Big)\triangleq S\,.\end{aligned}\tag{2.11}$$

Since $Q_1(z,\omega,\mathbb{E}[\omega]) = Q(z,\omega)-Q(z,\mathbb{E}[\omega])$, and $\mathbb{E}[\omega]=\hat\theta_n$, the RHS of (2.11) can be written in the form

$$\begin{aligned}S &= \beta E_n\log\Big(\mathbb{E}\exp\Big[-\frac{Q(Z_n,\omega)}{\beta}\Big]\Big)\\ &= \beta E_n\log\Big(\exp\Big[-\frac{Q(Z_n,\mathbb{E}[\omega])}{\beta}\Big]\Big)+\beta E_n\log\Big(\mathbb{E}\exp\Big[-\frac{Q_1(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Big]\Big)\\ &= -E_{n-1}A(\hat\theta_n)+S_1\end{aligned}\tag{2.12}$$

We now bound from below the LHS of equation (2.11). As in Iouditski *et al.* (2005), denote by βV the Fenchel-Legendre dual of $W_\beta(-z)$:

$$\beta V(\theta) = \sup_{z\in\mathbb{R}^M}[-z^\top\theta - W_\beta(z)].$$

Then, clearly,

$$-W_\beta(\zeta_n)\le \beta V(\theta)+\zeta_n^\top\theta,\quad \forall\,\theta\in\Lambda^M\,,$$

and

$$V(\theta)=\log M+\sum_{j=1}^M\theta^{(j)}\log\theta^{(j)}\le\log M\,,\quad\forall\,\theta\in\Lambda^M\,.$$

The last two inequalities and the fact that $W_\beta(\zeta_0)=W_\beta(0)=0$ imply

$$\frac{E_n\big[W_\beta(\zeta_n)-W_\beta(\zeta_0)\big]}{n}\ge -\frac{\beta\log M}{n}-\min_{\theta\in\Lambda^M}\frac{E_n\big[\zeta_n^\top\theta\big]}{n}=-\frac{\beta\log M}{n}-\min_{1\le j\le M}A(e_j).\tag{2.13}$$

Combining (2.11), (2.12) and (2.13) gives the lemma. ∎

In view of Lemma 2.1, to prove Theorems 2.1 and 2.2, it remains to give appropriate upper bounds for S_1.

Proof of Theorem 2.1. Since $Q_1 = Q_2+Q_3$, with $-Q_3(z,\theta,\theta')\le R(z)$ for all $z\in\mathcal{Z},\theta,\theta'\in\Lambda^M$, the quantity S_1 can be bounded from above as follows

$$S_1\le\beta E_n\log\Big(\mathbb{E}\exp\Big[-\frac{Q_2(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Big]\Big)+E_n\big[R(Z_n)\big]\,.$$

Now since $-Q_2(z,\cdot)/\beta$ is exponentially concave on Λ^M for all $z\in\mathcal{Z}$, the Jensen inequality yields

$$\mathbb{E}\exp\Big[-\frac{Q_2(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Big]\le\exp\Big[-\frac{Q_2(Z_n,\mathbb{E}[\omega],\mathbb{E}[\omega])}{\beta}\Big]=1$$

Therefore $S_1\le E_n\big[R(Z_n)\big]$. This and Lemma 2.1 imply the result of the Theorem 2.1. ∎

PROOF OF THEOREM 2.2. Using the Jensen inequality twice, with the concave functions $\log(\cdot)$ and $\Psi_\beta(\cdot, \mathbb{E}[\omega])$, we get

$$\begin{aligned} S_1 &\le \beta E_{n-1} \log\Big(E\,\mathbb{E} \exp\Big[- \frac{Q_1(Z,\omega,\mathbb{E}[\omega])}{\beta}\Big]\Big) \\ &= \beta E_{n-1} \log\Big(\mathbb{E}\, E \exp\Big[- \frac{Q_1(Z,\omega,\mathbb{E}[\omega])}{\beta}\Big]\Big) \\ &\le \beta E_{n-1} \log\Big(\mathbb{E}\Psi_\beta(\omega, \mathbb{E}[\omega])\Big) \\ &\le \beta E_{n-1} \log\Big(\Psi_\beta(\mathbb{E}[\omega], \mathbb{E}[\omega])\Big) = 0, \end{aligned} \tag{2.14}$$

where the first equality is due to the Fubini theorem. Theorem 2.2 follows now from (2.14) and Lemma 2.1.

4. Examples

In this section we apply Theorems 2.1 and 2.2 to three common statistical problems (regression, classification and density estimation). All the loss functions considered below are twice differentiable. The following proposition gives a simple sufficient condition for exponential concavity.

PROPOSITION 2.1. *Let g be a twice differentiable function on Λ^M with gradient $\nabla g(\theta)$ and Hessian matrix $\nabla^2 g(\theta)$, $\theta \in \Lambda^M$. If there exists $\beta > 0$ such that for any $\theta \in \Lambda^M$, the matrix*

$$\beta \nabla^2 g(\theta) - \nabla g(\theta)(\nabla g(\theta))^\top,$$

is positive semi-definite then $-g(\cdot)/\beta$ is exponentially concave on the simplex Λ^M.

PROOF. Since g is twice differentiable $\exp(-g(\cdot)/\beta)$ is also twice differentiable with Hessian matrix

$$\mathcal{H}(\theta) = \frac{1}{\beta} \exp\Big(-\frac{g(\theta)}{\beta}\Big) \left[\frac{\nabla g(\theta)(\nabla g(\theta))^\top}{\beta} - \nabla^2 g(\theta) \right]. \tag{2.15}$$

For any $\lambda \in \mathbb{R}^M$, $\theta \in \Lambda^M$, we have

$$\lambda^\top \mathcal{H}(\theta) \lambda = \frac{1}{\beta} \exp\Big(-\frac{g(\theta)}{\beta}\Big) \left[\frac{(\lambda^\top \nabla g(\theta))^2}{\beta} - \lambda^\top [\nabla^2 g(\theta)] \lambda \right] \le 0.$$

Hence $\exp(-g(\cdot)/\beta)$ has a negative semi-definite Hessian and is therefore concave. ∎

4.1. Applications of Theorem 2.1.

We begin with the models that satisfy assumptions of Theorem 2.1.

1. REGRESSION WITH SQUARED LOSS. Let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a complete separable metric space equipped with its Borel σ-algebra. Consider a random variable $Z = (X, Y)$ with $X \in \mathcal{X}$ and $Y \in \mathbb{R}$. Assume that the conditional expectation $f(X) = E(Y|X)$ exists and define $\xi = Y - E(Y|X)$, so that

$$Y = f(X) + \xi, \tag{2.16}$$

where $X \in \mathcal{X}$ is a random variable with probability distribution P_X, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ is the regression function and ξ is a real valued random variable satisfying $E(\xi|X) = 0$. Assume that $E(Y^2) < \infty$ and $\|f\|_\infty \le L$ for some finite constant $L > 0$ where $\|\cdot\|_\infty$ denotes the $L_\infty(P_X)$-norm. We have M functions $f_1, \dots, f_M$ such that $\|f_j\|_\infty \le L$, for

any $j = 1, \dots, M$. Define the $L_2(P_X)$-norm by $\|f\|^2_{2,P_X} = \int_{\mathcal{X}} f^2(x) P_X(\mathrm{d}x)$. Our goal is to construct an aggregate that mimics the oracle

$$\min_{1 \le j \le M} \|f_j - f\|^2_{2,P_X} .$$

The aggregate is based on the i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ where (X_i, Y_i) have the same distribution as (X, Y). For this model, with $z = (x, y) \in \mathcal{X} \times \mathbb{R}$, define the loss function

$$Q(z, \theta) = \big(y - \theta^\top H(x)\big)^2, \ \forall\, \theta \in \Lambda^M ,$$

with $H(x) = \big(f_1(x), \dots, f_M(x)\big)^\top$. It yields for all $z \in \mathcal{Z}, \theta, \theta' \in \Lambda^M$,

$$Q_1(z, \theta, \theta') = Q(z, \theta) - Q(z, \theta') = 2y(\theta' - \theta)^\top H(x) + \big[\theta^\top H(x)\big]^2 - \big[\theta'^\top H(x)\big]^2 .$$

Consider positive constants b and B and assume that $\beta > (b/B)^2$. We now decompose Q_1 into the sum $Q_1 = Q_2 + Q_3$ where

$$Q_2(z, \theta, \theta') = 2y \mathbb{1}_{\{|y| < B\beta\}} (\theta' - \theta)^\top H(x) + \big[\theta^\top H(x)\big]^2 - \big[\theta'^\top H(x)\big]^2 + \frac{y^2}{B\beta} \big[(\theta' - \theta)^\top H(x)\big]^2 \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}},$$

and

$$Q_3(z, \theta, \theta') = 2y \mathbb{1}_{\{|y| \ge B\beta\}} (\theta' - \theta)^\top H(x) - \frac{y^2}{B\beta} \big[(\theta' - \theta)^\top H(x)\big]^2 \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}}.$$

We have

$$-Q_3(z, \theta, \theta') \le 4L|y| \mathbb{1}_{\{|y| \ge B\beta\}} + \frac{4L^2 y^2}{B\beta} \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}} \triangleq R_\beta(y) . \tag{2.17}$$

On the other hand, $Q_2(z, \theta, \theta) = 0, \forall \theta \in \Lambda^M, z \in \mathcal{Z}$ and we can prove that the mapping $\theta \mapsto -Q_2(z, \theta, \theta')/\beta$ is exponentially concave for any $z \in \mathcal{Z}, \theta' \in \Lambda^M$ when b and B are properly chosen. For all $\theta \in \Lambda^M$ and $z = (x, y)$, the gradient and Hessian of Q_2 are respectively given by

$$\begin{aligned} \nabla_\theta Q_2 &= \nabla_\theta Q_2(z, \theta, \theta') \\ &= -2\big(y \mathbb{1}_{\{|y| < B\beta\}} - \theta^\top H(x)\big) H(x) - 2\frac{y^2}{B\beta} \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}} \big[(\theta' - \theta)^\top H(x)\big] H(x) \end{aligned}$$

and

$$\nabla^2_{\theta\theta} Q_2 = \nabla^2_{\theta\theta} Q_2(z, \theta, \theta') = 2H(x)H(x)^\top + 2\frac{y^2}{B\beta} \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}} H(x)H(x)^\top .$$

We now prove that Proposition 2.1 applies for $g(\theta) = Q_2(z, \theta, \theta')$, for all $z = (x, y) \in \mathcal{Z}$ and $\theta' \in \Lambda^M$. For any $\lambda \in \mathbb{R}^M$, any $\theta, \theta' \in \Lambda^M$, and any $z \in \mathcal{Z}$,

$$\big(\lambda^\top \nabla_\theta Q_2\big)^2 \le \Big(2|y| \mathbb{1}_{\{|y| < B\beta\}} + 2L + \frac{4Ly^2}{B\beta} \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}}\Big)^2 \big[\lambda^\top H(x)\big]^2.$$

Note now that $|y| \le B\beta$ implies that $y^2/B\beta \le |y|$. Hence

$$\begin{aligned} \big(\lambda^\top \nabla_\theta Q_2\big)^2 &\le \Big(2|y| \mathbb{1}_{\{|y| \le b\sqrt{\beta}\}} + 2L + (4L + 2)|y| \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}}\Big)^2 \big[\lambda^\top H(x)\big]^2 \\ &\le \Big(8b^2\beta + 8L^2 + 2(4L + 2)^2 |y|^2 \mathbb{1}_{\{b\sqrt{\beta} < |y| < B\beta\}}\Big) \big[\lambda^\top H(x)\big]^2 . \end{aligned}$$

It yields

$$\frac{\big(\lambda^\top \nabla_\theta Q_2\big)^2}{\beta} - \lambda^\top (\nabla^2_{\theta\theta} Q_2)\lambda \le \Big(8b^2 + \frac{8L^2}{\beta} - 2 + \Big[2(4L+2)^2 - \frac{2}{B}\Big]\frac{|y|^2}{\beta}\mathbb{I}_{\{b\sqrt{\beta}<|y|<B\beta\}}\Big)\big[\lambda^\top H(x)\big]^2 .$$

If we choose $B \le (4L+2)^{-2}$ and $LB < b < 1/4$, the above quadratic form is smaller than or equal to 0 and Proposition 2.1 applies for any $\beta > (b/B)^2$. Now, since $A(\theta) = EQ(Z,\theta) = E\big(Y - \theta^\top H(X)\big)^2 = \|f - \theta^\top H\|^2_{2,P_X} + E(\xi^2)$ for all $\theta \in \Lambda^M$, we obtain the following corollary of Theorem 2.1.

COROLLARY 2.1. *Consider the regression model* (2.16) *where* $X \in \mathcal{X}$, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ *and* $\xi = Y - f(X)$ *is a real valued random variable satisfying* $E(\xi|X) = 0$. *Assume also that* $E\big(Y^2\big) < \infty$ *and* $\|f_j\|_\infty \le L$, $j = 1, \dots, M$, *for some finite constant* $L > 0$. *Then for any positive constants* $B \ge (4L+2)^{-2}, LB < b < 1/4$ *and any* $\beta \ge (b/B)^2$, *the aggregate estimator* $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by Algorithm* LMA, *satisfies*

$$E_{n-1}\|\tilde{f}_n - f\|^2_{2,P_X} \le \min_{1\le j\le M} \|f_j - f\|^2_{2,P_X} + \frac{\beta \log M}{n} + E\big[R_\beta(Y)\big], \tag{2.18}$$

where

$$R_\beta(y) = 4L|y|\mathbb{I}_{\{|y|\ge B\beta\}} + \frac{4L^2y^2}{B\beta}\mathbb{I}_{\{b\sqrt{\beta}<|y|<B\beta\}} .$$

This result improves an inequality obtained by Bunea and Nobel (2005). We note that the aggregate $\tilde{f}_n$ as in Corollary 2.1 is of the form suggested by Catoni (1999, 2004). If there exists a constant $L_0 > 0$ such that $|Y| \le L_0$ a.s., the last summand disappears for $\beta > 16L_0^2$, and (2.18) follows from Catoni (1999, 2004), under a more restrictive assumption on β.

An advantage of Corollary 2.1 is that no heavy assumption on the moments of ξ is needed to get reasonable bounds. Thus, the second moment assumption on Y is enough for a bound with the $n^{-1/2}$ rate. Indeed, choosing $\beta \sim (n/\log M)^{2/(2+s)}$, $s > 0$, in Corollary 2.1, we immediately get the following result.

COROLLARY 2.2. *Consider the regression model* (2.16) *where* $X \in \mathcal{X}$, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ *and* $\xi = Y - f(X)$ *is a real valued random variable satisfying* $E(\xi|X) = 0$. *Assume also that* $E(|Y|^s) \le m_s < \infty$ *for some* $s \ge 2$ *and* $\|f_j\|_\infty \le L$, $j = 1, \dots, M$, *for some finite constant* $L > 0$. *Then there exist constants* $C_1 > 0$ *and* $C_2 = C_2(m_s, L, C_1) > 0$ *such that the aggregate estimator* $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by Algorithm* LMA *with* $\beta = C_1(n/\log M)^{2/(2+s)}$, *satisfies*

$$E_{n-1}\|\tilde{f}_n - f\|^2_{2,P_X} \le \min_{1\le j\le M} \|f_j - f\|^2_{2,P_X} + C_2\left(\frac{\log M}{n}\right)^{s/(2+s)} . \tag{2.19}$$

It has been conjectured by Audibert (2006) that the rate $\left(\frac{\log M}{n}\right)^{s/(2+s)}$ in the right hand side of (2.19) is optimal in a minimax sense.

2. CLASSIFICATION. Consider the problem of binary classification. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, and set $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$. Consider $Z = (X, Y)$ where X is a random variable with values in $\mathcal{X}$ and Y is a random label with values in $\{-1, 1\}$. For a fixed convex twice differentiable function $\varphi : \mathbb{R} \to \mathbb{R}_+$, define the φ-risk of a real valued classifier $h : \mathcal{X} \to [-1, 1]$ as $E\varphi(-Yh(X))$. In our framework, we have M classifiers $h_1, \dots, h_M$ and

the goal is to mimic the oracle

$$\min_{1\le j\le M} E\varphi(-Yh_j(X))\,,$$

based on the i.i.d. sample $(X_1,Y_1),\dots,(X_n,Y_n)$ where the (X_i,Y_i) have the same distribution as (X,Y). For any $z=(x,y)\in\mathcal{X}\times\{-1,1\}$, we define the loss function

$$Q(z,\theta)=\varphi(-y\theta^\top H(x))\ge 0,\ \forall\,\theta\in\Lambda^M\,,$$

where $H(x)=\big(h_1(x),\dots,h_M(x)\big)^\top$. For such a function and for all $\theta\in\mathbb{R}^M, z=(x,y)\in\mathcal{X}\times\{-1,1\}$ we have

$$\nabla_\theta Q_1(z,\theta,\theta')=-y\varphi'\big(-y\theta^\top H(x)\big)H(x)$$
$$\nabla^2_{\theta\theta} Q_1(z,\theta,\theta')=\varphi''\big(-y\theta^\top H(x)\big)H(x)H(x)^\top\,.$$

Thus, from Proposition 2.1 the mapping $\theta\mapsto -Q_1(z,\theta,\theta')/\beta$ is exponentially concave for all z and θ' if $\beta\ge\beta_\varphi$ where β_φ is such that $\big[\varphi'(x)\big]^2\le\beta_\varphi\varphi''(x),\ \forall\,|x|\le 1$. Now, since $A(\theta)=EQ(Z,\theta)$ and $Q(Z,\theta)=\varphi(-Y\theta^\top H(X)),\ \forall\,\theta\in\mathbb{R}^M$, $Z=(X,Y)$, we obtain the following corollary of Theorem 2.1 applied with $Q_2=Q_1$ and $Q_3\equiv 0$.

Corollary 2.3. *Consider the binary classification problem as described above. Assume that the convex function φ is such that $\big[\varphi'(x)\big]^2\le\beta_\varphi\varphi''(x),\ \forall\,|x|\le 1$. Then the aggregate classifier $\tilde h_n(x)=\hat\theta_n^\top H(x)$, $x\in\mathcal{X}$, where $\hat\theta_n$ is obtained by Algorithm* LMA *with $\beta\ge\beta_\varphi$, satisfies*

$$E_n\varphi(-Y_n\tilde h_n(X_n))\le\min_{1\le j\le M}E\varphi(-Y\tilde h_j(X))+\frac{\beta\log M}{n}\,. \tag{2.20}$$

For example, inequality (2.20) holds with the exponential Boosting loss $\varphi_1(x)=\mathrm{e}^x$, for which $\beta_{\varphi_1}=\mathrm{e}$ and for the Logit-Boosting loss $\varphi_2(x)=\log_2(1+\mathrm{e}^x)$ (in that case $\beta_{\varphi_2}=\mathrm{e}\log 2$). For the squared loss $\varphi_3(x)=(1-x)^2$ and the 2-norm soft margin loss $\varphi_4(x)=\max\{0,1-x\}^2$ inequality (2.20) is satisfied with $\beta\ge 2$.

3. Nonparametric density estimation with Kullback-Leibler (KL) loss. Let X be a random variable with values in a measurable space $(\mathcal{X},\mathcal{F})$. Assume that the distribution of X admits a density p w.r.t. a σ-finite measure μ on $(\mathcal{X},\mathcal{F})$. Assume also that we have M probability densities p_j w.r.t. μ on $(\mathcal{X},\mathcal{F})$ (estimators of p). Let $X_1,\dots,X_n$ be i.i.d. sample, where the X_i take values in $\mathcal{X}$, and have the same distribution as X. Define the KL divergence between two probability densities p and q w.r.t. μ as

$$\mathcal{K}(p,q)\triangleq\int_{\mathcal{X}}\log\left(\frac{p(x)}{q(x)}\right)p(x)\mu(\mathrm{d}x)\,,$$

if the probability distribution corresponding to p is absolutely continuous w.r.t. the one corresponding to q, and $\mathcal{K}(p,q)=\infty$ otherwise. We assume that the entropy integral $\int p(x)\log p(x)\mu(\mathrm{d}x)$ is finite.

Our goal is to construct an aggregate that mimics the KL oracle defined by

$$\min_{1\le j\le M}\mathcal{K}(p,p_j)\,.$$

For $x\in\mathcal{X},\theta\in\Lambda^M$, we introduce the following loss function

$$Q(x,\theta)=-\log(\theta^\top H(x)),$$

where $H(x) = \big(p_1(x), \dots, p_M(x)\big)^\top$. We set $Z = X$. Then

$$A(\theta) = EQ(X, \theta) = -\int \log(\theta^\top H(x)) p(x) \mu(\mathrm{d}x)\,,$$

where the integral is finite if all the divergences $\mathcal{K}(p, p_j)$ are finite. In particular, $A(e_j) = \mathcal{K}(p, p_j) - \int p(x) \log p(x) \mu(\mathrm{d}x)$. Since, for all $x \in \mathcal{X}$,

$$\exp(-Q_1(x, \theta, \theta')/\beta) = \big(\theta^\top H(x)\big)^{1/\beta} \big(\theta'^\top H(x)\big)^{-1/\beta}\,,$$

the mapping $\theta \mapsto -Q_1(x, \theta, \theta')/\beta$ is exponentially concave on Λ^M for any $\beta \geq 1$. Hence, we can apply Theorem 2.1, again with $Q_2 = Q_1$ and $Q_3 \equiv 0$ and we obtain the following corollary.

COROLLARY 2.4. *Consider the density estimation problem with the KL loss as described above, such that* $\int p(x) |\log p(x)| \mu(\mathrm{d}x) < \infty$. *Then the aggregate estimator* $\tilde{p}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by Algorithm* LMA *with* $\beta = 1$, *satisfies*

$$E_{n-1} \mathcal{K}(p, \tilde{p}_n) \leq \min_{1 \leq j \leq M} \mathcal{K}(p, p_j) + \frac{\log M}{n}\,.$$

We note that the KL aggregate $\tilde{p}_n$ as in Corollary 2.4 coincides with the "progressive mixture rule" considered by Catoni (1997, 1999, 2004) and Yang (2000) and the oracle inequality of Corollary 2.4 is the one obtained in those papers. Extension of Corollary 2.4 to $\beta \geq 1$ is straightforward but the oracle inequality for the corresponding aggregate ("Gibbs estimator", cf. Catoni (2004)) is less interesting because it has obviously a larger remainder term.

4.2. Applications of Theorem 2.2.

4. REGRESSION WITH SQUARED LOSS AND FINITE EXPONENTIAL MOMENT. We consider here the regression model described in Corollary 2.1 under the additional assumption that, conditionally on X, the regression residual ξ admits an exponential moment, i.e., there exist positive constants b and D such that, P_X-a.s.,

$$E\big(\exp(b|\xi|)|X\big) \leq D.$$

Since $E(\xi|X) = 0$, this assumption is equivalent to the existence of positive constants b_0 and σ^2 such that, P_X-a.s.,

$$E\big(\exp(t\xi)|X\big) \leq \exp(\sigma^2 t^2/2), \quad \forall\ |t| \leq b_0, \tag{2.21}$$

(cf. Petrov, 1995, p. 56).

In this case, application of Corollary 2.1 leads to suboptimal rates because of the term $E\big[R_\beta(Y)\big]$ in (2.18). We show now that using Theorem 2.2 we can obtain an oracle inequality with optimal rate $(\log M)/n$.

To apply Theorem 2.2, we analyze the mapping $\theta \mapsto E \exp\big(-Q_1(Z, \theta, \theta')/\beta\big)$. For the regression model with squared loss as described above, we have $Z = (X, Y)$, $Q(Z, \theta) = (Y - \theta^\top H(X))^2$, and

$$E \exp\big(-Q_1(Z, \theta, \theta')/\beta\big) =$$

$$= E \exp\Big(-\frac{1}{\beta}\Big[(Y - H(X)^\top \theta)^2 - (Y - H(X)^\top \theta')^2\Big]\Big)$$
$$= E \exp\Big(-\frac{1}{\beta}\Big[-2\xi(U(X, \theta) - U(X, \theta')) + U^2(X, \theta) - U^2(X, \theta')\Big]\Big)\,,$$

where $U(X,\theta) \triangleq f(X) - H(X)^\top\theta$. Since $|2(U(X,\theta)-U(X,\theta'))| = 2|(\theta-\theta')^\top H(X)| \le 4L$, conditioning on X and using (2.21) we get that, for any $\beta \ge 4L/b_0$,

$$E\exp\big(-Q_1(Z,\theta,\theta')/\beta\big) \le \Psi_\beta(\theta,\theta')$$

where

$$\Psi_\beta(\theta,\theta') \triangleq E\exp\left(\frac{2\sigma^2}{\beta^2}\big[(\theta-\theta')^\top H(X)\big]^2 - \frac{1}{\beta}\big[U^2(X,\theta) - U^2(X,\theta')\big]\right).$$

Clearly, $\Psi_\beta(\theta,\theta) = 1$. Thus, to apply Theorem 2.2 it suffices now to specify $\beta_0 > 0$ such that the mapping

$$\theta \mapsto \bar{Q}(x,\theta,\theta') \triangleq \left(-\frac{1}{\beta} + \frac{2\sigma^2}{\beta^2}\right)(\theta^\top H(x))^2 - \frac{4\sigma^2}{\beta^2}(H(x)^\top\theta)(H(x)^\top\theta') + \frac{2}{\beta}f(x)(H(x)^\top\theta)$$

is exponentially concave for all $\beta \ge \beta_0$, $\theta' \in \Lambda^M$ and almost all $x \in \mathcal{X}$. Note that

$$\begin{aligned}\nabla_\theta\bar{Q}(x,\theta,\theta') &= \left(2\gamma(f(x) - H(x)^\top\theta) - \frac{4\sigma^2}{\beta^2}(f(x) - H(x)^\top\theta') + \frac{2}{\beta}f(x)\right)H(x),\\ \nabla^2_{\theta\theta}\bar{Q}(x,\theta,\theta') &= -2\gamma H(x)H(x)^\top\end{aligned}$$

where $\gamma = \frac{1}{\beta} - \frac{2\sigma^2}{\beta^2}$. Proposition 2.1 implies that $\bar{Q}$ is exponentially concave in θ if $\nabla^2_{\theta\theta}\bar{Q}(x,\theta,\theta') + \nabla_\theta\bar{Q}(x,\theta,\theta')(\nabla_\theta\bar{Q}(x,\theta,\theta'))^\top \le 0$. If we denote by $\tilde{L} = \sup_j \|f - f_j\|_\infty$, we obtain that the latter property holds for $\beta \ge \beta_0 \triangleq 2\sigma^2 + 2\tilde{L}^2$. Thus, Theorem 2.2 applies for $\beta \ge \max(2\sigma^2 + 2\tilde{L}^2,\, 4L/b_0)$ and we have proved the following result.

COROLLARY 2.5. *Consider the regression model* (2.16) *where* $X \in \mathcal{X}$, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ *and the random variable* $\xi = Y - f(X)$ *is such that there exist positive constants* b_0 *and* σ^2 *for which* (2.21) *holds* P_X*-a.s. Assume also that* $\|f - f_j\|_\infty \le \tilde{L}$ *and* $\|f_j\|_\infty \le L$, $j = 1,\dots,M$, *for some finite positive constants* $L, \tilde{L}$. *Then for any* $\beta \ge \max(2\sigma^2 + 2\tilde{L}^2,\, 4L/b_0)$ *the aggregate estimator* $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by Algorithm* LMA, *satisfies*

$$E_{n-1}\|\tilde{f}_n - f\|^2_{2,P_X} \le \min_{1\le j\le M}\|f_j - f\|^2_{2,P_X} + \frac{\beta\log M}{n}. \tag{2.22}$$

To see how good the constants are, we may compare this corollary with the results obtained in other papers for the particular case where ξ is conditionally Gaussian given X. In this case we have $b_0 = \infty$ and Corollary 2.5 yields the following result.

COROLLARY 2.6. *Consider the regression model* (2.16) *where* $X \in \mathcal{X}$, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ *and, conditionally on* X, *the random variable* $\xi = Y - f(X)$ *is Gaussian with zero mean and variance bounded by* σ^2. *Assume that* $\|f - f_j\|_\infty \le \tilde{L}$, *for some finite constant* $\tilde{L} > 0$. *Then for any* $\beta \ge 2\sigma^2 + 2\tilde{L}^2$ *the aggregate estimator* $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by Algorithm* LMA, *satisfies* (2.22).

When we also assume that f and all $f_j, j = 1,\dots,M$ are uniformly bounded by L, we have $\tilde{L} \le 2L$ and the condition on β in the corollary becomes $\beta \ge 2\sigma^2 + 8L^2$. It improves upon the best known bound $18.01\sigma^2 + 70.4L^2$ (Catoni, 2004, p. 89).

5. NONPARAMETRIC DENSITY ESTIMATION WITH THE L_2 LOSS. Let μ be a σ-finite measure on the measurable space $(\mathcal{X},\mathcal{F})$. In this whole example, densities are understood w.r.t μ and $\|\cdot\|_\infty$ denotes the $L_\infty(\mu)$-norm. Assume that we have M probability densities p_j, $\|p_j\|_\infty \le L$, $j = 1,\dots,M$. Let $X_1,\dots,X_n$ be an i.i.d. sample, where the X_i take values

in $\mathcal{X}$, and are distributed as a random variable X with unknown probability density p such that $\|p\|_\infty \le L$ for some positive constant L. Our goal is to mimic the oracle defined by

$$\min_{1\le j\le M} \|p_j - p\|_2^2 \,,$$

where $\|p\|_2^2 = \int p^2(x)\mu(\mathrm{d}x)$.

The corresponding loss function is defined, for any $x \in \mathcal{X}, \theta \in \Lambda^M$, by

$$Q(x,\theta) = \theta^\top G\theta - 2\theta^\top H(x)\,, \tag{2.23}$$

where $H(x) = \big(p_1(x), \dots, p_M(x)\big)^\top$ and G is an $M \times M$ positive semi-definite matrix with elements $G_{jk} = \int p_j p_k \mathrm{d}\mu$ and such that its largest eigenvalue is bounded from above by L^*. We set $Z = X$. Then $A(\theta) = EQ(X,\theta) = \|p - \theta^\top H\|_2^2 - \|p\|_2^2$. We now want to check conditions of Theorem 2.2, i.e., to show that for the loss function (2.23), the mapping $\theta \mapsto E\exp\big(-Q_1(X,\theta,\theta')/\beta\big)$ is concave on Λ^M, for any $\theta' \in \Lambda^M$ and for $\beta \ge \beta_0$ with some $\beta_0 > 0$ that will be specified below. Note first that

$$Q_1(x,\theta,\theta') = Q(x,\theta) - Q(x,\theta') = (\theta-\theta')^\top G(\theta+\theta') - 2(\theta-\theta')^\top H(x). \tag{2.24}$$

Fix $\theta' \in \Lambda^M$. Concavity of the above mapping can be checked by considering its Hessian $\tilde{\mathcal{H}}$ which, in view of (2.15), satisfies

$$\lambda^\top \tilde{\mathcal{H}}(\theta)\lambda = \frac{1}{\beta^2} E\Big\{ \exp\left(-\frac{Q_1(X,\theta,\theta')}{\beta}\right)\Big[(\lambda^\top \nabla_\theta Q_1(X,\theta,\theta'))^2 - \beta\lambda^\top \nabla^2_{\theta\theta} Q_1(X,\theta,\theta')\lambda\Big]\Big\},$$

for any $\lambda \in \mathbb{R}^M$, $\theta \in \Lambda^M$. Note that for any $x \in \mathcal{X}, \theta \in \Lambda^M$ we have

$$\nabla_\theta Q_1(x,\theta,\theta') = 2G\theta - 2H(x) \quad \text{and} \quad \nabla^2_{\theta\theta} Q_1(x,\theta,\theta') = 2G\,.$$

By (2.24) this yields, for any $\lambda \in \mathbb{R}^M$, $\theta, \theta' \in \Lambda^M$,

$$\begin{aligned}\lambda^\top \tilde{\mathcal{H}}(\theta)\lambda = -\frac{2}{\beta^2} E\Big\{ &\exp\left(-\frac{(\theta-\theta')^\top G(\theta+\theta') - 2(\theta-\theta')^\top H(X)}{\beta}\right) \\ &\Big[\beta\lambda^\top G\lambda - 2(\lambda^\top(G\theta - H(X)))^2\Big]\Big\} \\ \le -\frac{2}{\beta^2}\exp&\left(-\frac{(\theta-\theta')^\top G(\theta+\theta')}{\beta}\right) F(\lambda,\theta,\theta'),\end{aligned} \tag{2.25}$$

where

$$F(\lambda,\theta,\theta') = E\left\{\exp\left(\frac{2(\theta-\theta')^\top H(X)}{\beta}\right)\Big[\beta\lambda^\top G\lambda - 4(\lambda^\top G\theta)^2 - 4(\lambda^\top H(X))^2\Big]\right\}.$$

Next, for any $\theta \in \Lambda^M$,

$$\theta^\top G\theta \le L^*\theta^\top\theta \le L^*\,.$$

Using the Cauchy inequality, the previous remark yields

$$(\lambda^\top G\theta)^2 \le \lambda^\top G\lambda\, \theta^\top G\theta \le L^*\lambda^\top G\lambda, \ \ \forall\theta \in \Lambda^M. \tag{2.26}$$

Further,

$$E(\lambda^\top H(X))^2 = \int \big(\lambda^\top H(x)\big)^2 p(x)\mu(\mathrm{d}x) \le L\int \big(\lambda^\top H(x)\big)^2\mu(\mathrm{d}x) = L\lambda^\top G\lambda. \tag{2.27}$$

Using (2.26) and (2.27) and the fact that $\|\theta-\theta'\|_1 \le 2$ where $\|\cdot\|_1$ stands for the $\ell_1(\mathbb{R}^M)$ norm, we obtain

$$\begin{aligned} F(\lambda,\theta,\theta') &\ge (\beta-4L^*)\lambda^\top G\lambda E\exp\left(\frac{2(\theta-\theta')^\top H(X)}{\beta}\right) \\ &\qquad - 4E\left\{\exp\left(\frac{2(\theta-\theta')^\top H(X)}{\beta}\right)(\lambda^\top H(X))^2\right\} \\ &\ge (\beta-4L^*)\lambda^\top G\lambda\,\exp\left(-\frac{4L}{\beta}\right) - 4L\lambda^\top G\lambda\,\exp\left(\frac{4L}{\beta}\right) \ge 0\,, \end{aligned}$$

provided that

$$\frac{\beta-4L^*}{4L}\exp\left(-\frac{8L}{\beta}\right) \ge 1.$$

Note that the last inequality is guaranteed for $\beta \ge \beta_0 = 12\max(L,L^*)$. We conclude that for $\beta \ge 12\max(L,L^*)$, the Hessian $\tilde{\mathcal{H}}$ in (2.25) is negative semi-definite and therefore the mapping $\theta \mapsto E\exp\big(-Q_1(X,\theta,\theta')/\beta\big)$ is concave on Λ^M for any fixed $\theta' \in \Lambda^M$. Thus we have proved the following corollary of Theorem 2.2.

COROLLARY 2.7. *Let $H(x) = (p_1,\dots,p_M)^\top$ be a vector of functions in $L_2(\mathbb{R}^d)$ such that $\|p_j\|_\infty \le L$, for any $j = 1,\dots,M$ and such that the scalar product matrix G with elements $G_{jk} = \int p_j p_k \mathrm{d}\mu$ has its largest eigenvalue bounded from above by L^*. Then, for any $\beta \ge 12\max(L,L^*)$, the aggregate estimator $\tilde{p}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, where $\hat{\theta}_n$ is obtained by Algorithm* LMA*, satisfies*

$$E_{n-1}\|\tilde{p}_n - p\|_2^2 \le \min_{1\le j\le M}\|p_j - p\|_2^2 + \frac{\beta\log M}{n}\,.$$

6. PARAMETRIC ESTIMATION WITH KULLBACK-LEIBLER (KL) LOSS. Let $\mathcal{P} = \{P_a,\, a \in \mathcal{A}\}$ be a family of probability measures on a measurable space $(\mathcal{X},\mathcal{F})$ dominated by a σ-finite measure μ on $(\mathcal{X},\mathcal{F})$. Here $\mathcal{A} \subset \mathbb{R}^m$ is a bounded set of parameters. The densities relative to μ are denoted by $p(x,a) = (\mathrm{d}P_a/\mathrm{d}\mu)(x)$, $x \in \mathcal{X}$. Let X be a random variable with values in $\mathcal{X}$ distributed according to P_{a^*} where $a^* \in \mathcal{A}$ is the unknown true value of the parameter.

In the aggregation framework, we have M values $a_1,\dots,a_M \in \mathcal{A}$ (preliminary estimators of a) and an i.i.d. sample $X_1,\dots,X_n$ where the X_i take values in $\mathcal{X}$, and have the same distribution as X. Our goal is to construct an aggregate $\tilde{a}_n$ that mimics the parametric KL oracle defined by

$$\min_{1\le j\le M} K(a^*,a_j)\,,$$

where

$$K(a,b) \triangleq \mathcal{K}\big(p(\cdot,a),p(\cdot,b)\big)\,, \quad \forall\, a,b \in \mathcal{A}.$$

For $x \in \mathcal{X}, \theta \in \Lambda^M$, we introduce the corresponding loss function

$$Q(x,\theta) = -\log p(x,\theta^\top H),$$

where $H = (a_1,\dots,a_M)^\top$. We set $Z = X$. Then

$$A(\theta) = EQ(X,\theta) = -\int \log\big(p(x,\theta^\top H)\big)p(x,a^*)\mu(\mathrm{d}x)$$

and

$$A(e_j) = K(a^*,a_j) - \int p(x,a^*)\log\big(p(x,a^*)\big)\mu(\mathrm{d}x)\,.$$

Since, for all $x \in \mathcal{X}$, $\exp(-Q(x,\theta)/\beta) = \big(p(x,\theta^\top H)\big)^{1/\beta}$, to apply Theorem 2.2 we need the following assumption.

Assumption 4.1 *For some $\beta > 0$ and for any $a \in \mathcal{A}$ there exists a Borel function $\Psi_\beta : \Lambda^M \times \Lambda^M \to \mathbb{R}_+$ such that $\theta \mapsto \Psi_\beta(\theta,\theta')$ is concave on the simplex Λ^M for all $\theta' \in \Lambda^M$, $\Psi_\beta(\theta,\theta) = 1$ and*

$$\int \left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) \le \Psi_\beta(\theta,\theta')\,,$$

for all $\theta,\theta' \in \Lambda^M$.

Corollary 2.8. *Consider the parametric estimation problem with the KL loss as described above and let $\int p(x,a^*)|\log p(x,a^*)|\mu(\mathrm{d}x) < \infty$. Suppose that Assumption 4.1 is fulfilled for some $\beta > 0$. Then the aggregate estimator $\tilde{a}_n = \hat{\theta}_n^\top H$ of the parameter a^*, where $\hat{\theta}_n$ is obtained by Algorithm* LMA*, satisfies*

$$E_{n-1}K(a^*,\tilde{a}_n) \le \min_{1\le j\le M} K(a^*,a_j) + \frac{\beta\log M}{n}\,. \tag{2.28}$$

Aggregation procedures can be used to construct pointwise adaptive locally parametric estimators in nonparametric regression (cf. Belomestny and Spokoiny, 2004). In this case inequality (2.28) can be applied to prove the corresponding adaptive risk bounds. We now check that Assumption 4.1 is satisfied for several standard parametric families.

Univariate Gaussian distribution. Let μ be the Lebesgue measure on $\mathbb{R}$ and let $p(x,a) = (\sigma\sqrt{2\pi})^{-1}\exp(-(x-a)^2/(2\sigma^2))$ be the univariate Gaussian density with mean $a \in \mathcal{A} = [-L,L]$ and known variance $\sigma^2 > 0$. Replacing $f(x)$ by a^* and $H(x)$ by H in the proof of Corollary 2.6, and following exactly the same argument as there we find that Assumption 4.1 is satisfied for any $\beta \ge \beta_0 = 2\sigma^2 + 8L^2$. Hence, (2.28) also holds for such β. Note that in this case $K(a^*,a) = (a^*-a)^2/(2\sigma^2)$.

Bernoulli distribution. Let μ be the discrete measure on $\{0,1\}$ such that $\mu(0) = \mu(1) = 1$ and let $p(x,a) = a\mathbb{1}_{\{x=0\}} + (1-a)\mathbb{1}_{\{x=1\}}$ be the density of a Bernoulli random variable with parameter $a \in \mathcal{A} = (0,1)$. Then

$$\int \left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) = \left(\frac{H^\top\theta}{H^\top\theta'}\right)^{1/\beta} a + \left(\frac{1-H^\top\theta}{1-H^\top\theta'}\right)^{1/\beta}(1-a) \triangleq \Psi_\beta(\theta,\theta')\,.$$

This function is concave in θ for any $\theta' \in \Lambda^M$ if $\beta \ge 1$ and obviously $\Psi_\beta(\theta,\theta) = 1$. Therefore Assumption 4.1 is satisfied and Corollary 2.8 applies with $\beta = 1$.

Poisson distribution. Let μ be the counting measure on the set of the nonnegative integers $\mathbb{N}$: $\mu(k) = 1$, $\forall\, k \in \mathbb{N}$, and let $p(x,a) = \sum_{k=0}^{\infty} \frac{a^k}{k!}\mathrm{e}^{-a}\mathbb{1}_{\{x=k\}}$ be the density of a Poisson random variable with parameter $a \in \mathcal{A} = [\ell, L]$ where $0 < \ell < L < \infty$. Then

$$\int \left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) = \exp\left[a\left(\frac{H^\top\theta}{H^\top\theta'}\right)^{1/\beta} - a - \frac{H^\top(\theta-\theta')}{\beta}\right] \triangleq \Psi_\beta(\theta,\theta')\,. \tag{2.29}$$

Clearly, $\Psi_\beta(\theta,\theta) = 1$ and it is not hard to show that Ψ_β in (2.29) is concave as a function of θ for any $\theta' \in \Lambda^M$, provided that $\beta \geq 1 + L(1 + L/\ell)(L/\ell)^{1/(2L+1)}$. Therefore Assumption 4.1 is satisfied and Corollary 2.8 applies with $\beta \geq \beta_0 = 1 + L(1 + L/\ell)(L/\ell)^{1/(2L+1)}$.

CHAPTER 3

Optimal aggregation of density estimators

We study the problem of aggregation of M estimators of a probability density with respect to the mean integrated squared error. We provide procedures for the problems of linear, convex and model selection aggregation and we prove oracle inequalities for their risks. We also obtain lower bounds showing that these procedures are rate optimal in a minimax sense.

Contents

Most of the material in this chapter is a joint work with Alexandre Tsybakov (Rigollet and Tsybakov, 2006).

1. Introduction

Consider independent and identically distributed random vectors $X_1, \dots, X_n$ with values in $\mathbb{R}^d$ having an unknown common probability density $p \in L_2(\mathbb{R}^d)$ that we want to estimate. For an estimator $\hat{p}$ of p based on the sample $\mathbb{X}^n = (X_1, \dots, X_n)$, define the L_2-risk

$$R_n(\hat{p}, p) = E_p^n \|\hat{p} - p\|^2,$$

where E_p^n denotes the expectation with respect to the distribution P_p^n of $\mathbb{X}^n$ and, for a function $g \in L_2(\mathbb{R}^d)$,

$$\|g\| = \left(\int_{\mathbb{R}^d} g^2(x) \mathrm{d}x \right)^{1/2}.$$

Suppose that we have $M \geq 2$ estimators $\hat{p}_1, \dots, \hat{p}_M$ of the density p based on the sample $\mathbb{X}^n$. The problem that we study here is to construct a new estimator $\tilde{p}_n$ of p, called *aggregate*, which is approximately at least as good as the best linear or convex combination

of $\hat{p}_1, \ldots, \hat{p}_M$, or simply the best of them. For any $\lambda \in \mathbb{R}^M$, define the linear combination

$$\mathsf{p}_\lambda = \sum_{j=1}^{M} \lambda_j \hat{p}_j, \quad \lambda = (\lambda_1, \ldots, \lambda_M)\,.$$

The problems of linear, convex and model selection aggregation of density estimators under the L_2 loss can be stated as follows.

(1) **Problem (L): linear aggregation.** Find a *linear aggregate*, i.e., an estimator $\tilde{p}_n^{\mathbf{L}}$ which satisfies

$$R_n(\tilde{p}_n^{\mathbf{L}}, p) \le \inf_{\lambda \in \mathbb{R}^M} R_n(\mathsf{p}_\lambda, p) + \Delta_{n,M}^{\mathbf{L}}\,, \tag{3.1}$$

for every p belonging to a large class of densities $\mathcal{P}$ where $\Delta_{n,M}^{\mathbf{L}}$ is a sufficiently small remainder term that does not depend on p.

(2) **Problem (C): convex aggregation.** Find a *convex aggregate*, i.e., an estimator $\tilde{p}_n^{\mathbf{C}}$ which satisfies

$$R_n(\tilde{p}_n^{\mathbf{C}}, p) \le \inf_{\lambda \in H} R_n(\mathsf{p}_\lambda, p) + \Delta_{n,M}^{\mathbf{C}}\,, \tag{3.2}$$

for every p belonging to a large class of densities $\mathcal{P}$, where $\Delta_{n,M}^{\mathbf{C}}$ is a sufficiently small remainder term that does not depend on p, and H is a convex compact subset of $\mathbb{R}^M$. We will discuss in more detail the case $H = \Lambda^M$ where Λ^M is a simplex,

$$\Lambda^M = \Big\{\lambda \in \mathbb{R}^M : \ \lambda_j \ge 0, \sum_{j=1}^{M} \lambda_j \le 1\Big\}.$$

(3) **Problem (MS): model selection aggregation.** Find a *model selection aggregate*, i.e., an estimator $\tilde{p}_n^{\mathbf{MS}}$ which satisfies

$$R_n(\tilde{p}_n^{\mathbf{MS}}, p) \le \min_{1 \le j \le M} R_n(\hat{p}_j, p) + \Delta_{n,M}^{\mathbf{MS}}\,, \tag{3.3}$$

for every p belonging to a large class of densities $\mathcal{P}$, where $\Delta_{n,M}^{\mathbf{MS}}$ is a sufficiently small remainder term that does not depend on p.

Our aim is to find aggregates satisfying (3.1), (3.2) or (3.3) with the smallest possible remainder terms $\Delta_{n,M}^{\mathbf{L}}$, $\Delta_{n,M}^{\mathbf{C}}$ and $\Delta_{n,M}^{\mathbf{MS}}$. These remainder terms characterize the price to pay for aggregation.

The study of convergence properties of aggregation methods has been initiated by Nemirovski (2000); Catoni (1999, 2004) and Yang (2000). Most of the results were obtained for the regression and Gaussian white noise models (see a recent overview in Bunea *et al.*, 2004). Aggregation of density estimators has received less attention. The work on this subject is mainly devoted to model selection aggregation with the Kullback-Leibler divergence as a loss function (Catoni, 1999, 2004; Yang, 2000; Zhang, 2006), and is based on information-theoretical ideas close to the earlier papers of Barron (1987); Li and Barron (1999). Devroye and Lugosi (2001) developed a method of MS aggregation of density estimators satisfying certain complexity assumptions under the L_1 loss.

To our knowledge, linear aggregation of density estimators has not been previously studied. For convex aggregation, the only paper we are aware of is that of Birgé (2003) where this type of aggregation under the L_1 loss is considered, while we study here the L_2 loss. In his setup, Birgé (2003) proves an inequality which is weaker than (3.2), with the oracle risk on the right hand side multiplied by a constant which is much larger than 1.

We do not only suggest aggregates satisfying sharp oracle inequalities (3.1), (3.2) and (3.3) but also demonstrate their optimality. Namely, we introduce the notion of optimal rate of aggregation and show that our aggregates attain optimal rates. This extends to the density estimation context the results of the paper of Tsybakov (2003), where optimal rates of aggregation for the regression model have been obtained.

The main purpose of aggregation is to improve upon the initial set of estimators $\hat{p}_1, \dots, \hat{p}_M$. This is a general tool that applies to various kinds of estimators satisfying very mild conditions (we may only assume that they are square integrable). Consider, for example, the simplest case when we have only two estimators ($M = 2$), where $\hat{p}_1$ is a good parametric density estimator for some fixed regular parametric family and $\hat{p}_2$ is a nonparametric density estimator. If the underlying density p belongs to the parametric family, $\hat{p}_1$ is perfect: its risk converges with the parametric rate $O(1/n)$. But for densities which are not in this family it may not converge at all. As for $\hat{p}_2$, it converges with a slow nonparametric rate even if the underlying density is within the parametric family. Aggregation allows one to construct procedures that combine the advantages of both $\hat{p}_1$ and $\hat{p}_2$: the aggregates converge with the parametric rate $O(1/n)$ if p is within the parametric family, and with a nonparametric rate otherwise.

In this chapter, we deal with a "pure aggregation" framework as in most of the papers on the subject (cf., e.g., Nemirovski, 2000; Juditsky and Nemirovski, 2000; Tsybakov, 2003, for the regression problem), where the preliminary estimators are constructed from a frozen sample. This means that instead of the estimators $\hat{p}_1, \dots, \hat{p}_M$ we have fixed functions $p_1, \dots, p_M$. For more details about choosing the sample, see Chapter 5.

This chapter is organized as follows. In section 2, we prove oracle inequalities for the three problems of aggregation described above and the remainder terms of these inequalities are proved to be optimal in Section 3. Throughout the paper we denote by c_i finite positive constants.

2. Oracle inequalities

In this section, $p_1, \dots, p_M$ are fixed functions in $L_2(\mathbb{R}^d)$, not necessarily probability densities. From now on the notation p_λ for a vector $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$ is understood in the following sense:

$$\mathsf{p}_\lambda \triangleq \sum_{j=1}^{M} \lambda_j p_j,$$

and, since for any fixed $\lambda \in \mathbb{R}^M$, the function p_λ is non-random, we have

$$R_n(\mathsf{p}_\lambda, p) = \|\mathsf{p}_\lambda - p\|^2.$$

Denote by $\mathcal{P}_0$ the class of all densities on $\mathbb{R}^d$ bounded by the constant $L > 0$:

$$\mathcal{P}_0 \triangleq \left\{ p : \mathbb{R}^d \to \mathbb{R} \,:\, p \geq 0, \int_{\mathbb{R}^d} p(x)\mathrm{d}x = 1,\ \|p\|_\infty \leq L \right\},$$

where $\|\cdot\|_\infty$ stands for the $L_\infty(\mathbb{R}^d)$ norm. For any $L_0 > 0$, define

$$\mathcal{F}_\infty(L_0) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, \|f\|_\infty \leq L_0 \right\},$$

the set of function $f : \mathbb{R}^d \to \mathbb{R}$ that are uniformly bounded by L_0. We will see that optimal rates of convex aggregation differ whether we aggregate functions in $L_2(\mathbb{R}^d)$ or in $\mathcal{F}_\infty(L_0)$.

2.1. Linear aggregation. Denote by $\mathcal{L}$ the linear span of $p_1, \dots, p_M$. Let $\phi_1, \dots, \phi_{M'}$ with $M' \le M$, be an orthonormal basis of $\mathcal{L}$ in $L_2(\mathbb{R}^d)$. Define a linear aggregate

$$\tilde{p}_n^{\mathbf{L}}(x) = \sum_{j=1}^{M'} \hat{\lambda}_j^{\mathbf{L}} \phi_j(x), \quad x \in \mathbb{R}^d, \tag{3.4}$$

where

$$\hat{\lambda}_j^{\mathbf{L}} = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i).$$

THEOREM 3.1. *Assume that $p_1, \dots, p_M \in L_2(\mathbb{R}^d)$ and $p \in \mathcal{P}_0$. Then*

$$R_n(\tilde{p}_n^{\mathbf{L}}, p) \le \min_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 + \frac{LM}{n}, \tag{3.5}$$

for any integers $M \ge 2$ and $n \ge 1$.

PROOF. Consider the projection of p onto $\mathcal{L}$:

$$p_{\mathcal{L}}^* = \operatorname*{argmin}_{\mathsf{p}_\lambda \in \mathcal{L}} \|\mathsf{p}_\lambda - p\|^2 = \sum_{j=1}^{M'} \lambda_j^* \phi_j,$$

where $\lambda_j^* = (p, \phi_j)$, and $(\cdot, \cdot)$ is the scalar product in $L_2(\mathbb{R}^d)$. Using the Pythagorean theorem we get that, almost surely,

$$\|\tilde{p}_n^{\mathbf{L}} - p\|^2 = \sum_{j=1}^{M'} (\hat{\lambda}_j^{\mathbf{L}} - \lambda_j^*)^2 + \|p_{\mathcal{L}}^* - p\|^2.$$

To finish the proof it suffices to take expectations in the last equation and to note that $E_p^n(\hat{\lambda}_j^{\mathbf{L}}) = \lambda_j^*$. It yields

$$E_p^n\left[(\hat{\lambda}_j^{\mathbf{L}} - \lambda_j^*)^2\right] = \mathrm{Var}(\hat{\lambda}_j^{\mathbf{L}}) \le \frac{1}{n} \int_{\mathbb{R}^d} \phi_j^2(x) p(x) \mathrm{d}x \le \frac{L}{n}.$$

∎

2.2. Convex aggregation. The aim of convex aggregation is to mimic the *convex oracle* defined as $\lambda^* = \operatorname{argmin}_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2$ where H is a given convex compact subset of $\mathbb{R}^M$. Clearly,

$$\|\mathsf{p}_\lambda - p\|^2 = \|\mathsf{p}_\lambda\|^2 - 2 \int_{\mathbb{R}^d} \mathsf{p}_\lambda p + \|p\|^2.$$

Removing here the term $\|p\|^2$ independent of λ and estimating $\int_{\mathbb{R}^d} p_j p$ by $n^{-1} \sum_{i=1}^n p_j(X_i)$ we get the following estimate of the oracle

$$\hat{\lambda}^{\mathbf{C}} = \operatorname*{argmin}_{\lambda \in H} \left\{ \|\mathsf{p}_\lambda\|^2 - \frac{2}{n} \sum_{i=1}^{n} \mathsf{p}_\lambda(X_i) \right\}. \tag{3.6}$$

Now, we define a *convex aggregate* $\tilde{p}_n^{\mathbf{C}}$ by

$$\tilde{p}_n^{\mathbf{C}} = \sum_{j=1}^{M} \hat{\lambda}_j^{\mathbf{C}} p_j = \mathsf{p}_{\hat{\lambda}^{\mathbf{C}}}.$$

The following theorem contains two results. It first states that, for any functions $p_1, \dots, p_M$ in $L_2(\mathbb{R}^d)$, the convex aggregate $\tilde{p}_n^{\mathbf{C}}$ satisfies an oracle inequality similar to (3.2) with a remainder term of the order M/n. Moreover, if $p_j \in \mathcal{F}_\infty(L_0), j = 1, \dots, M$, the remainder

term in the oracle inequality can be of the order $\sqrt{\log(M)/n}$ which is smaller than M/n when M is relatively large compared to n.

THEOREM 3.2. *Fix two integers $M \geq 2$, $n \geq 1$ and let H be a convex compact subset of $\mathbb{R}^M$. Consider the functions $p_1, \ldots, p_M \in L_2(\mathbb{R}^d)$ and $p \in \mathcal{P}_0$. Then the convex aggregate $\tilde{p}_n^{\mathbf{C}}$ satisfies*

$$R_n(\tilde{p}_n^{\mathbf{C}}, p) \leq \min_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2 + \frac{4LM}{n}\,. \tag{3.7}$$

Moreover, if $p_j \in \mathcal{F}_\infty(L_0)$, $j = 1, \ldots, M$, when $H = \Lambda^M$, the convex aggregate $\tilde{p}_n^{\mathbf{C}}$ satisfies

$$R_n(\tilde{p}_n^{\mathbf{C}}, p) \leq \min_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 + 16L_0\sqrt{2\mathrm{e}}\sqrt{\frac{\log M}{n}}\,. \tag{3.8}$$

PROOF. We will write for brevity $\hat{\lambda} = \hat{\lambda}^{\mathbf{C}}$. First note that the mapping $\lambda \mapsto \|\mathsf{p}_\lambda\|^2 - \frac{2}{n}\sum_{i=1}^n \mathsf{p}_\lambda(X_i)$ is continuous, thus $\hat{\lambda}$ exists, and the oracle $\lambda^* = \operatorname{argmin}_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2$ also exists.

PROOF OF (3.7). The definition of $\hat{\lambda}$ implies that, for any $p \in \mathcal{P}_0$,

$$\|\mathsf{p}_{\hat{\lambda}} - p\|^2 \leq \|\mathsf{p}_{\lambda^*} - p\|^2 + 2T_n \tag{3.9}$$

where

$$T_n = \frac{1}{n}\sum_{i=1}^n \mathsf{p}_{\hat{\lambda}-\lambda^*}(X_i) - \int_{\mathbb{R}^d} \mathsf{p}_{\hat{\lambda}-\lambda^*} p.$$

Introduce the notation

$$\mathcal{Z}_n = \sup_{\mu \in \mathbb{R}^M : \|\mathsf{p}_\mu\| \neq 0} \frac{\left|\frac{1}{n}\sum_{i=1}^n \mathsf{p}_\mu(X_i) - E_p^n[\mathsf{p}_\mu(X_1)]\right|}{\|\mathsf{p}_\mu\|}.$$

Using the Cauchy-Schwarz inequality, the identity $\mathsf{p}_{\hat{\lambda}-\lambda^*} = \mathsf{p}_{\hat{\lambda}} - \mathsf{p}_{\lambda^*}$ and the elementary inequality $2\sqrt{xy} \leq ax + y/a$, $\forall\, x, y, a > 0$, we get

$$\begin{aligned} E_p^n|T_n| &\leq E_p^n\left(\mathcal{Z}_n\|\mathsf{p}_{\hat{\lambda}-\lambda^*}\|\right) \\ &\leq \sqrt{E_p^n(\mathcal{Z}_n^2)}\sqrt{E_p^n(\|\mathsf{p}_{\hat{\lambda}-\lambda^*}\|^2)} \\ &\leq \frac{a}{2}E_p^n(\|\mathsf{p}_{\hat{\lambda}} - \mathsf{p}_{\lambda^*}\|^2) + \frac{1}{2a}E_p^n(\mathcal{Z}_n^2), \quad \forall\, a > 0. \end{aligned} \tag{3.10}$$

Representing p_μ in the form $\mathsf{p}_\mu = \sum_{l=1}^{M'} \nu_l \phi_l$ where $\nu_l \in \mathbb{R}$ and $\{\phi_l\}$ is an orthonormal basis in $\mathcal{L}$ (cf. proof of Theorem 3.1) we find

$$\mathcal{Z}_n \leq \sup_{\nu \in \mathbb{R}^M \setminus \{0\}} \frac{|\sum_{l=1}^{M'} \nu_l \zeta_l|}{|\nu|} = \Big(\sum_{l=1}^{M'} \zeta_l^2\Big)^{1/2},$$

where $|\nu| = \Big(\sum_{l=1}^{M'} \nu_l^2\Big)^{1/2}$ and

$$\zeta_l = \frac{1}{n}\sum_{i=1}^n \phi_l(X_i) - E_p^n[\phi_l(X_1)].$$

Hence

$$E_p^n(\mathcal{Z}_n^2) \leq \frac{M'}{n} \max_{l=1,\ldots,M'} E_p^n[\phi_l^2(X_1)] \leq \frac{LM}{n}, \tag{3.11}$$

whenever $\|p\|_\infty \le L$. Since $\{\mathsf{p}_\lambda : \lambda \in H\}$ is a convex subset of $L_2(\mathbb{R}^d)$ and p_{λ^*} is the projection of p onto this set, we have

$$\|\mathsf{p}_\lambda - p\|^2 \ge \|\mathsf{p}_{\lambda^*} - p\|^2 + \|\mathsf{p}_\lambda - \mathsf{p}_{\lambda^*}\|^2, \quad \forall\, \lambda \in H,\ p \in L_2(\mathbb{R}^d). \tag{3.12}$$

Using (3.12) with $\lambda = \hat{\lambda}$, (3.10) and (3.11) we obtain

$$E_p^n|T_n| \le \frac{a}{2}\left\{E_p^n(\|\mathsf{p}_{\hat{\lambda}} - p\|^2 - \|\mathsf{p}_{\lambda^*} - p\|^2)\right\} + \frac{LM}{2an}\,.$$

This and (3.9) yield that, for any $0 < a < 1$,

$$E_p^n(\|\mathsf{p}_{\hat{\lambda}} - p\|^2) \le \|\mathsf{p}_{\lambda^*} - p\|^2 + \frac{LM}{a(1-a)n}\,.$$

Now, (3.7) follows by taking the infimum of the right hand side of this inequality over $0 < a < 1$.

Proof of (3.8). We essentially follow the proof of Juditsky and Nemirovski (2000, Theorem 2.1). For any $p \in \mathcal{P}_0$, we can rewrite (3.9) as

$$\|\mathsf{p}_{\hat{\lambda}} - p\|^2 \le \|\mathsf{p}_{\lambda^*} - p\|^2 + \frac{2}{n}\sum_{i=1}^n (\hat{\lambda} - \lambda^*)^\top \zeta_i \tag{3.13}$$

where for $i = 1, \dots, n$, the vector $\zeta_i \in \mathbb{R}^M$ has coordinates $\zeta_i^{(j)} = p_j(X_i) - E_p\big[p_j(X_1)\big]$. By Hölder's inequality,

$$\frac{2}{n}\sum_{i=1}^n (\hat{\lambda} - \lambda^*)^\top \zeta_i \le \frac{4}{n}\max_{1\le j\le M}\left|\sum_{i=1}^n \zeta_i^{(j)}\right|, \tag{3.14}$$

where we used the fact that $\sum_{j=1}^M |\hat{\lambda}_j - \lambda_j^*| \le 2$, since $\hat{\lambda}, \lambda^* \in \Lambda^M$. For $q = 2\log M$ and $z \in \mathbb{R}^M$ define the function $W : \mathbb{R}^M \to \mathbb{R}$,

$$W(z) = \frac{1}{2}\|z\|_q^2\,,$$

where for any $z = (z^{(1)}, \dots, z^{(M)}) \in \mathbb{R}^M$,

$$\|z\|_q = \Big(\sum_{j=1}^M |z^{(j)}|^q\Big)^{1/q}$$

and

$$\|z\|_\infty = \max_{1\le j\le M} |z^{(j)}|\,.$$

By virtue of Lemma A.7, we have

$$W\Big(\sum_{i=1}^{k+1}\zeta_i\Big) \le W\Big(\sum_{i=1}^{k}\zeta_i\Big) + (\zeta_{k+1})^\top \nabla W\Big(\sum_{i=1}^{k}\zeta_i\Big) + 4\mathrm{e}\log M\|\zeta_{k+1}\|_\infty^2\,. \tag{3.15}$$

Since

$$E_p^n\|\zeta_{k+1}\|_\infty^2 \le 4L_0^2\,, \quad k = 0, \dots, n-1\,,$$

taking expectations in (3.15) yields

$$E_p^n W\Big(\sum_{i=1}^{k+1}\zeta_i\Big) \le E_p^n W\Big(\sum_{i=1}^{k}\zeta_i\Big) + 16\mathrm{e}L_0^2\log M\,.$$

It follows that

$$E_p^n W\Big(\sum_{i=1}^{n}\zeta_i\Big) \le 16n\mathrm{e}L_0^2\log M\,.$$

Since $W(z) \geq \|z\|_\infty^2/2$, we eventually obtain

$$E_p^n \Big\| \sum_{i=1}^n \zeta_i \Big\|_\infty \leq 4L_0 \sqrt{2ne \log M}\,.$$

This inequality, combined with (3.13) and (3.14) gives the desired upper bound. ∎

2.3. Model selection aggregation. Let $\mathcal{E} = \{e_1, \ldots, e_M\}$ be the set of the vertices of the simplex Λ^M. Here, e_j is the jth coordinate unit vector in $\mathbb{R}^M$: $e_j = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^M$, where 1 appears in jth position.

To define a model selection aggregate, we use a result from Chapter 2. For any $x \in \mathbb{R}^d, \theta \in \Lambda^M$, define the loss function

$$Q(x, \theta) = \theta^\top G \theta - 2\theta^\top H(x)\,,$$

where $H(x) = \big(p_1(x), \ldots, p_M(x)\big)^\top$ and G is an $M \times M$ positive semi-definite matrix with elements $G_{jk} = \int_{\mathbb{R}^d} p_j(x) p_k(x) \mathrm{d}x$. Assume that G has its largest eigenvalue bounded from above by a constant $L^* > 0$.

We define the model selection aggregate $\tilde{p}_n^{\mathbf{MS}}$ by

$$\tilde{p}_n^{\mathbf{MS}} = \sum_{j=1}^M \hat{\lambda}_j^{\mathbf{MS}} p_j = \mathsf{p}_{\hat{\lambda}^{\mathbf{MS}}}\,,$$

where $\hat{\lambda}^{\mathbf{MS}}$ is the vector of weights output by the Algorithm LMA, defined in Chapter 2, Section 2 with $\beta \geq 12 \max(L, L^*)$.

THEOREM 3.3. *Assume that $p_1, \ldots, p_M \in \mathcal{F}_\infty(L)$ are such that the scalar product matrix G with elements $G_{jk} = \int p_j p_k$ has its largest eigenvalue bounded from above by a constant $L^* > 0$ and $p \in \mathcal{P}_0$. Then the MS aggregate $\tilde{p}_n^{\mathbf{MS}}$ satisfies*

$$R_n(\tilde{p}_n^{\mathbf{MS}}, p) \leq \min_{1 \leq j \leq M} \|p_j - p\|^2 + \frac{\beta \log M}{n}\,, \tag{3.16}$$

for any $\beta \geq 12 \max(L, L^*)$.

PROOF. See Corollary 2.7. ∎

Remark that the condition on the largest eigenvalue of G is satisfied if

$$G_{j,k} = \int_{\mathbb{R}^d} p_j(x) p_k(x) \mathrm{d}x \leq L^*, \quad \forall\ 1 \leq j, k \leq M\,.$$

A sufficient but not necessary condition for this condition to hold with $L = L^*$ is that the p_j's are probability densities in $\mathcal{P}_0$.

3. Lower bounds and optimal aggregation

We first recall the notion of *optimal rate of aggregation* for density estimation, similar to that for the regression problem given in Tsybakov (2003). It is related to the minimax behavior of the excess risk

$$R_n(\tilde{p}_n, p) - \inf_{\lambda \in H} \|\mathsf{p}_\lambda - p\|^2\,,$$

for a given class Θ of weights λ.

DEFINITION 3.1. *Let $\Theta \subseteq \mathbb{R}^M$ be a given class of weights, $\mathcal{P}$ be a given class of probability densities on $\mathbb{R}^d$ and let $\mathcal{F}$ be a subset of $L_2(\mathbb{R}^d)$. A sequence of positive numbers $\psi_n(M)$ is called* **optimal rate of aggregation** *for $(\Theta, \mathcal{P}, \mathcal{F})$ if*

- *for any functions $p_j \in \mathcal{F}, j = 1, \dots, M$, there exists an estimator $\tilde{p}_n$ of p (aggregate) such that*

$$\sup_{p\in\mathcal{P}} \left[R_n(\tilde{p}_n, p) - \inf_{\lambda\in\Theta} \|\mathsf{p}_\lambda - p\|^2 \right] \le C\psi_n(M), \tag{3.17}$$

for any integer $n \ge 1$ and for some constant $C < \infty$ independent of M and n,

and

- *there exist functions $p_j \in \mathcal{F}, j = 1, \dots, M$, such that for all estimators T_n of p, we have*

$$\sup_{p\in\mathcal{P}} \left[R_n(T_n, p) - \inf_{\lambda\in\Theta} \|\mathsf{p}_\lambda - p\|^2 \right] \ge c\psi_n(M), \tag{3.18}$$

for any integer $n \ge 1$ and for some constant $c > 0$ independent of M and n.

When (3.18) *holds, an aggregate $\tilde{p}_n$ satisfying* (3.17) *is called* **rate optimal aggregate** *for $(\Theta, \mathcal{P}, \mathcal{F})$.*

Note that this definition applies to aggregation of any functions p_j in $\mathcal{F} \subset L_2(\mathbb{R}^d)$, in particular, they are not necessarily supposed to be probability densities.

REMARK 3.1. *For $\mathcal{P} = \mathcal{P}_0$ there is a natural limitation on the value $c\psi_n(M)$ on the right hand side of* (3.18)*, whatever is Θ. In fact, we have*

$$\begin{aligned} \inf_{T_n} \sup_{p\in\mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda\in\Theta} \|\mathsf{p}_\lambda - p\|^2 \right] \le \inf_{T_n} \sup_{p\in\mathcal{P}_0} R_n(T_n, p) &\le \sup_{p\in\mathcal{P}_0} R_n(0, p) \\ &= \sup_{p\in\mathcal{P}_0} \|p\|^2 \\ &\le L. \end{aligned}$$

Therefore, we must have $c\psi_n(M) \le L$ where c is the constant in (3.18)*.*

3.1. Optimal rate of linear aggregation. We are going to prove (3.18) with $\psi_n(M) = LM/n$, $\mathcal{P} = \mathcal{P}_0$, $\mathcal{F} = L_2(\mathbb{R}^d)$ and $\Theta = \mathbb{R}^M$. For $\psi_n(M) = LM/n$, in view of Remark 3.1, only the values M such that $M \le c_0 n$ are allowed, where $c_0 > 0$ is a constant. The upper bounds of Theorems 3.1 and 3.2 are too rough (non-optimal) when $M = M_n$ depends on n and the condition $M \le c_0 n$ is not satisfied. In the sequel, we will apply those theorems with $M = M_n$ depending on n and satisfying $M_n/n \to 0$, as $n \to \infty$, so that the condition $M \le c_0 n$ will obviously hold with any finite c_0 for n large enough.

THEOREM 3.4. *Let the integers $M \ge 2$ and $n \ge 1$ be such that $M \le c_0 n$ where c_0 is a positive constant. Then there exist probability densities $p_j \in L_2(\mathbb{R}^d)$, $j = 1, \dots, M$, such that for all estimators T_n of p we have*

$$\sup_{p\in\mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda\in\mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 \right] \ge c\frac{LM}{n},$$

where $c > 0$ is a constant that depends only on c_0.

PROOF. We are going to apply Lemma A.3.Set $r = M - 1 \ge 1$ and fix $0 < a \le 1$. Consider the function $\tilde{g}$ defined for any $t \in \mathbb{R}$ by

$$\tilde{g}(t) = \frac{aL}{2} \mathbb{I}_{[0, \frac{1}{Lr}]}(t) - \frac{aL}{2} \mathbb{I}_{(\frac{1}{Lr}, \frac{2}{Lr})}(t),$$

where $\mathbb{1}_A(\cdot)$ denotes the indicator function of a set A. Let $\{\tilde{g}_j\}_{j=1}^r$ be the family of functions defined by $\tilde{g}_j(t) = \tilde{g}(t - 2(j-1)/Lr)$, $1 \le j \le r$. Define also the density $\tilde{f}(t) = (L/2)\mathbb{1}_{[0,2/L]}(t)$, $t \in \mathbb{R}$. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ consider the functions

$$f(x) = \tilde{f}(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k) \quad g_j(x) = \tilde{g}_j(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k), \ j = 1, \dots, r\,.$$

Define the probability densities p_j by $p_1 = f$, $p_{j+1} = f + g_j$, $j = 1, \dots, M-1$.

Consider now the set of functions

$$\mathcal{Q} = \{q_\delta : q_\delta = f + \sum_{j=1}^{r} \delta_j g_j,\ \delta = (\delta_1, \dots, \delta_r) \in \ \{0,1\}^r\}\,.$$

Clearly, for any $\delta \in \{0,1\}^r$, q_δ satisfies $\int_{\mathbb{R}^d} q_\delta(x)\mathrm{d}x = 1$, $q_\delta \ge 0$ and $\|q_\delta\|_\infty \le L$. Therefore $\mathcal{Q} \subset \mathcal{P}_0$. Also, $\mathcal{Q} \subset \{\mathsf{p}_\lambda, \lambda \in \mathbb{R}^M\}$. Thus,

$$\inf_{T_n} \sup_{p \in \mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 \right] \ge \inf_{T_n} \sup_{p \in \mathcal{Q}} R_n(T_n, p)\,.$$

To prove that $\inf_{T_n} \sup_{p \in \mathcal{Q}} R_n(T_n, p) \ge cLM/n$ we check conditions (A.1) of Lemma A.3. The first condition in (A.1) is obviously satisfied since

$$\|g_j\|^2 = \int_0^{\frac{2}{Lr}} \tilde{g}^2(t)\mathrm{d}t = \frac{a^2 L}{2r}\,, \quad j = 1, \dots, r.$$

To check the second condition in (A.1), note that for $j = 1, \dots, r$ we have

$$\begin{aligned}
h^2(f, f+g_j) &= \frac{1}{2} \int_0^{\frac{2}{Lr}} \left(\sqrt{L/2} - \sqrt{L/2 + \tilde{g}(t)} \right)^2 \mathrm{d}t \\
&= \frac{L}{4} \int_0^{\frac{2}{Lr}} \left(1 - \sqrt{1 + (2/L)\tilde{g}(t)} \right)^2 \mathrm{d}t \\
&= \frac{L}{4} \left[\frac{4}{Lr} - 2 \int_0^{\frac{2}{Lr}} \sqrt{1 + (2/L)\tilde{g}(t)} \mathrm{d}t \right] \\
&= \frac{1}{r} - \frac{1}{2r} \left(\sqrt{1+a} + \sqrt{1-a} \right) \le \frac{a^2}{2r}\,,
\end{aligned}$$

where we used the fact that $\sqrt{1+a} + \sqrt{1-a} \ge 2 - a^2$ for $|a| \le 1$. Define now $\tilde{c}_0 = \max(c_0, 3)$ and choose $a^2 = M/(\tilde{c}_0 n) \le 1$. Then $a^2/(2r) \le (\tilde{c}_0 n)^{-1}$ for $M \ge 2$. Applying Lemma A.3 with $\beta = (\tilde{c}_0 n)^{-1}$ and $\alpha = \dfrac{ML}{2\tilde{c}_0 nr}$ we get

$$\inf_{T_n} \sup_{p \in \mathcal{C}} R_n(T_n, p) \ge \frac{1}{8\tilde{c}_0} \left(1 - \sqrt{\frac{2}{\tilde{c}_0}} \right) \frac{LM}{n}\,.$$

∎

Theorems 3.1 and 3.4 imply the following result.

COROLLARY 3.1. *Let the integers $M \ge 2$ and $n \ge 1$ be such that $M \le c_0 n$ where c_0 is a positive constant. Then $\psi_n(M) = LM/n$ is an optimal rate of aggregation for $(\mathbb{R}^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$, and $\tilde{p}_n^{\mathbf{L}}$ defined in (3.4) is a rate optimal aggregate for $(\mathbb{R}^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$.*

3.2. Optimal rates of convex aggregation. We analyze here only the case $\Theta = \Lambda^M$. Other examples of convex sets Θ can be treated similarly.

We begin by the case $\mathcal{F} = L_2(\mathbb{R}^d)$.

THEOREM 3.5. *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 n$. Then there exist functions $p_j \in L_2(\mathbb{R}^d), j = 1, \dots, M$, such that for all estimators T_n of p we have*

$$\sup_{p \in \mathcal{P}_0} \Big[R_n(T_n, p) - \inf_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 \Big] \geq c_1 \frac{LM}{n} \,,$$

where $c_1 > 0$ is a constant that depends only on c_0.

PROOF. Consider the same family of densities $\mathcal{Q}$ as defined in the proof of Theorem 3.4. We may rewrite it in the form $\mathcal{Q} = \{q_\delta : q_\delta = \lambda_1 M f + \sum_{j=1}^r \lambda_{j+1} M \delta_j g_j, \, \delta = (\delta_1, \dots, \delta_r) \in \{0,1\}^r\}$ where $\lambda_j = 1/M$, $j = 1, \dots M$. Define now $p_1 = Mf$, $p_{j+1} = M(f + g_j)$, $j = 1, \dots, M-1$. Since $\sum_{j=1}^M \lambda_j = 1$ we have $\mathcal{Q} \subset \{\mathsf{p}_\lambda, \lambda \in \Lambda^M\}$. The rest of the proof is identical to that of Theorem 3.4. ∎

Remark that the functions p_j in the proof of Theorem 3.5 are not bounded. A similar result exists when $\mathcal{F} = \mathcal{F}_\infty(L_0)$. In that case $\psi_n(M) = L_0 \sqrt{\log(M)/n}$ and from Remark 3.1, only the values of M such that $M \leq c_0 \mathrm{e}^n$ make sense, where c_0 is a positive constant depending only on L and L_0.

THEOREM 3.6. *Let the integers $M \geq 8$ and $n \geq 1$ be such that $M \leq c_0 \mathrm{e}^n$. Then there exist functions $p_j \in \mathcal{F}_\infty(L_0)$, $j = 1, \dots, M$, such that for all estimators T_n of p we have*

$$\sup_{p \in \mathcal{P}_0} \Big[R_n(T_n, p) - \inf_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 \Big] \geq \begin{cases} c_2 \tilde{L} \sqrt{\dfrac{1}{n} \log\Big(\dfrac{M}{\sqrt{n}} + 1\Big)} & \text{if } M \geq \sqrt{n} \\ c_2' \tilde{L} \dfrac{M}{n} & \text{if } M < \sqrt{n} \end{cases}$$

where $\tilde{L} = \min(L, L_0)$ and c_2, c_2' are positive absolute constants. Moreover, if $M \geq n^{\frac{1+\alpha}{2}}$ for some fixed $\alpha > 0$, the lower bound becomes

$$\sup_{p \in \mathcal{P}_0} \Big[R_n(T_n, p) - \inf_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 \Big] \geq c_3 \tilde{L} \sqrt{\frac{\log(M)}{n}} \,,$$

where $c_3 > 0$ depends only on $\alpha > 0$.

PROOF. We are going to apply Lemma A.4. It amounts to finding a collection of probability densities on $\mathbb{R}^d$ that are (i) far from each other for the L_2 distance and (ii) close to each other for the Kullback-Leibler divergence defined in the appendix. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, and a tuning parameter $0 < \gamma < 1$, set

$$p_1(x) = \tilde{L} \mathbb{1}_{\left[0, \frac{2}{\tilde{L}}\right]}(x_1) \prod_{k=2}^d \mathbb{1}_{[0,1]}(x_k)$$

and

$$p_j(x) = \frac{\gamma \tilde{L}}{2} h_j(x_1 \tilde{L}/2) \prod_{k=2}^d \mathbb{1}_{[0,1]}(x_k), \; j = 2, \dots, M,$$

where

$$h_j(t) = \cos(4\pi j t)\big(\mathbb{1}_{[0,1/2]}(t) - \mathbb{1}_{[1/2,1]}(t)\big) \,, \quad t \in [0,1] \,.$$

The h_j are orthogonal in $L_2([0,1]^d)$.

Consider first the case $M \geq \sqrt{n}$ and define an integer

$$m = 2 \left\lfloor c_4 \Big\lfloor n / \log\Big(\frac{M}{\sqrt{n}} + 1\Big) \Big\rfloor^{1/2} \right\rfloor + 1 \geq 3 \,, \tag{3.19}$$

for a constant $c_4 > 0$ chosen in such a way that $M \geq 3(m-1)$, where $\lceil \cdot \rceil$ is the ceiling function. Define the set

$$\Omega_1 = \left\{ \omega = (\omega_2, \dots, \omega_M) \in \{0,1\}^{M-1}, \sum_{j=2}^{M} \omega_j = m-1 \right\},$$

and consider the family of densities, $\mathcal{C}_1 = \{q_\omega, \omega \in \Omega_1\}$, where

$$q_\omega = \frac{1}{2} p_1 + \frac{1}{2(m-1)} \sum_{j=2}^{M} \omega_j p_j \,.$$

Clearly, $\mathcal{C}_1 \subset \mathcal{P}_0$ and $\min_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 = 0$ for any $p \in \mathcal{C}_1$. We now bound from below the supremum $\sup_{p \in \mathcal{C}_1} R(T_n, p)$ uniformly over all estimators T_n. Let Ω_1' be a subset of Ω_1 extracted using Lemma A.2 to which the point $(0, \dots, 0) \in \{0,1\}^{M-1}$ has been added. Consider the subset $\mathcal{C}_1' = \{q_\omega, \omega \in \Omega_1'\}$ of $\mathcal{C}_1$. There exist positive constants c_5 and c_6 such that

$$\log(\mathrm{card}(\mathcal{C}_1')) \geq c_5 \frac{m-1}{2} \log\left(\frac{2M}{m-1} + 1 \right) \geq c_6 \frac{n}{m-1} \,,$$

where we used equation (3.19) for the definition of m. Moreover, for any two $\omega_1, \omega_2 \in \Omega_1'$,

$$\|q_{\omega_1} - q_{\omega_2}\|^2 = \frac{\gamma^2 \tilde{L}}{8(m-1)^2} \rho(\omega_1, \omega_2) \geq c_7 \frac{\tilde{L}}{m-1} \,,$$

for a constant $c_7 > 0$ and where $\rho(\cdot, \cdot)$ is the Hamming distance. Using the definition of the Kullback-Leibler divergence $K(\cdot, \cdot)$ given in Appendix, we obtain for any $\omega \in \Omega_1'$,

$$\begin{aligned}
K(q_0, q_\omega) &= -\int_0^1 \log\left(1 + \frac{\gamma}{2(m-1)} \sum_{j=2}^{M} \omega_j h_j(x) \right) \mathrm{d}x \\
&= -\int_0^1 \log\left(1 - \frac{\gamma^2}{4(m-1)^2} \Big(\sum_{j=2}^{M} \omega_j \cos(4\pi j x) \Big)^2 \right) \mathrm{d}x \\
&\leq \frac{c_8 \gamma^2}{(m-1)^2} \int_0^1 \left(\sum_{j=2}^{M} \omega_j \cos(4\pi j x) \right)^2 \mathrm{d}x \\
&\leq \frac{c_9 \gamma^2}{(m-1)} \leq \frac{\log(\mathrm{card}(\mathcal{C}_1'))}{16 n} \,,
\end{aligned}$$

for $\gamma < 1$ such that $\gamma^2 \leq c_6/(16 c_9)$. Thus, in virtue of Lemma A.4, we get,

$$\inf_{T_n} \sup_{p \in \mathcal{C}_1'} R_n(T_n, p) \geq c_{10} \tilde{L} \sqrt{\frac{\log(\frac{M}{\sqrt{n}} + 1)}{n}} \,.$$

Consider now the case $M \leq \sqrt{n}$. Define the family of densities, $\mathcal{C}_2 = \left\{ q_\omega, \omega \in \{0,1\}^{M-1} \right\}$, where

$$q_\omega = \frac{1}{2} p_1 + \frac{1}{2\sqrt{n}} \sum_{j=2}^{M} \omega_j p_j \,.$$

Clearly, $\mathcal{C}_2 \subset \mathcal{P}_0$ and since $M \leq \sqrt{n}$, $\mathcal{C}_2 \subset \{\mathsf{p}_\lambda : \lambda \in \Lambda^M\}$. Therefore, we have $\min_{\lambda \in \Lambda^M} \|\mathsf{p}_\lambda - p\|^2 = 0$ for any $p \in \mathcal{C}_2$. We now bound from below the supremum $\sup_{p \in \mathcal{C}_2} R(T_n, p)$ uniformly over all estimators T_n. Let Ω_2' be a subset of $\{0,1\}^{M-1}$ extracted using Lemma A.1 and such that $(0, \dots, 0) \in \Omega''$. Consider the subset $\mathcal{C}_2' = \{q_\omega, \omega \in$

$\Omega_2'\}$ of $\mathcal{C}_2$. It holds

$$\log(\mathrm{card}(\mathcal{C}_2')) \geq \frac{(M-1)\log 2}{8}.$$

Moreover, for any two $\omega_1, \omega_2 \in \Omega_2'$,

$$\|q_{\omega_1} - q_{\omega_2}\|^2 = \frac{\gamma^2 \tilde{L}}{8n}\rho(\omega_1,\omega_2) \geq c_{11}\frac{\tilde{L}M}{n}\,,$$

for a constant $c_{11} > 0$. For any $\omega \in \Omega_2'$, we have

$$\begin{aligned}
K(q_0, q_\omega) &= -\int_0^1 \log\Big(1 + \frac{\gamma}{2\sqrt{n}}\sum_{j=2}^M \omega_j h_j(x)\Big)\mathrm{d}x \\
&= -\int_0^1 \log\Big(1 - \frac{\gamma^2}{4n}\big(\sum_{j=2}^M \omega_j \cos(4\pi j x)\big)^2\Big)\mathrm{d}x \\
&\leq \frac{c_{12}\gamma^2}{n}\int_0^1 \Big(\sum_{j=2}^M \omega_j \cos(4\pi j x)\Big)^2 \mathrm{d}x \\
&\leq \frac{c_{13}\gamma^2(M-1)}{n} \leq \frac{\log(\mathrm{card}(\mathcal{C}_2'))}{16n}\,,
\end{aligned}$$

for $\gamma < 1$ such that $\gamma^2 \leq \log 2/(8c_{13})$. Thus, in virtue of Lemma A.4, we get,

$$\inf_{T_n}\sup_{p\in\mathcal{C}_2'} R_n(T_n, p) \geq c_{14}\frac{\tilde{L}M}{n}\,.$$

We have proved the first assertion of the theorem. The second assertion follow immediately. ∎

When $\mathcal{F} = L_2(\mathbb{R}^d)$, Theorems 3.2 and 3.5 imply the following result.

COROLLARY 3.2. *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 n$. Then $\psi_n(M) = LM/n$ is an optimal rate of aggregation for $(\Lambda^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$, and $\tilde{p}_n^{\mathbf{C}}$ is a rate optimal aggregate for $(\Lambda^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$.*

When $\mathcal{F} = \mathcal{F}_\infty(L_0)$, Theorems 3.2, 3.5 and 3.6 imply the following result.

COROLLARY 3.3. *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 \mathrm{e}^n$. Then when $M \geq n^{\frac{1+\alpha}{2}}$ for some fixed $\alpha > 0$, $\psi_n(M) = \sqrt{\log(M)/n}$ is an optimal rate of aggregation for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$, and $\tilde{p}_n^{\mathbf{C}}$ is a rate optimal aggregate for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$. When $M \leq \sqrt{n}$, $\psi_n(M) = M/n$ is an optimal rate of aggregation for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$, and $\tilde{p}_n^{\mathbf{C}}$ is a rate optimal aggregate for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$.*

3.3. Optimal rates of model selection aggregation. We only consider the case $\mathcal{F} = \mathcal{F}_\infty(L)$. The proof of Theorem 3.4 can be easily adapted to obtain the following lower bound.

THEOREM 3.7. *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 \mathrm{e}^n$. Then there exist functions $p_j \in \mathcal{F}_\infty(L), j = 1, \ldots, M$, such that for all estimators T_n of p we have*

$$\sup_{p\in\mathcal{P}_0}\Big[R_n(T_n, p) - \min_{1\leq j\leq M}\|p_j - p\|^2\Big] \geq c_{15}L\frac{\log M}{n}\,,$$

where $c_{15} > 0$ is a constant that depends only on c_0.

PROOF. We follow the same proof as for Theorem 3.4. Assume for simplicity that there exists $r \geq 1$ such that $M = 2^r$ and fix $0 < a \leq 1$. Consider the function $\tilde{g}$ defined for any $t \in \mathbb{R}$ by

$$\tilde{g}(t) = \frac{aL}{2} \mathbb{1}_{[0, \frac{1}{Lr}]}(t) - \frac{aL}{2} \mathbb{1}_{(\frac{1}{Lr}, \frac{2}{Lr})}(t),$$

where $\mathbb{1}_A(\cdot)$ denotes the indicator function of a set A. Let $\{\tilde{g}_j\}_{j=1}^r$ be the family of functions defined by $\tilde{g}_j(t) = \tilde{g}(t - 2(j-1)/Lr)$, $1 \leq j \leq r$. Define also the density $\tilde{f}(t) = (L/2)\mathbb{1}_{[0,2/L]}(t)$, $t \in \mathbb{R}$. For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ consider the functions

$$f(x) = \tilde{f}(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k) \quad g_j(x) = \tilde{g}_j(x_1) \prod_{k=2}^{d} \mathbb{1}_{[0,1]}(x_k), \ j = 1, \ldots, r\,.$$

Consider now the set of functions

$$\mathcal{Q} = \{q_\delta : q_\delta = f + \sum_{j=1}^{r} \delta_j g_j, \, \delta = (\delta_1, \ldots, \delta_r) \in \ \{0,1\}^r\}\,.$$

Clearly, for any $\delta \in \{0,1\}^r$, q_δ satisfies $\int_{\mathbb{R}^d} q_\delta(x)\mathrm{d}x = 1$, $q_\delta \geq 0$ and $\|q_\delta\|_\infty \leq L$. Therefore $\mathcal{Q} \subset \mathcal{P}_0$. Define the probability densities p_j by $p_j = q_{\delta(j-1)}$, $j = 1, \ldots, 2^r$, where $\delta(j)$ denotes the binary decomposition of j. Since $M = 2^r$, we have $\mathcal{Q} = \{p_1, \ldots, p_M\}$. Thus,

$$\inf_{T_n} \sup_{p \in \mathcal{P}_0} \left[R_n(T_n, p) - \inf_{\lambda \in \mathbb{R}^M} \|\mathsf{p}_\lambda - p\|^2 \right] \geq \inf_{T_n} \sup_{p \in \mathcal{Q}} R_n(T_n, p)\,.$$

To prove a lower bound for $\inf_{T_n} \sup_{p \in \mathcal{Q}} R_n(T_n, p)$ we check conditions (A.1) of Lemma A.3. The first condition in (A.1) is obviously satisfied since

$$\|g_j\|^2 = \int_0^{\frac{2}{Lr}} \tilde{g}^2(t)\mathrm{d}t = \frac{a^2 L}{2r}\,, \quad j = 1, \ldots, r.$$

To check the second condition in (A.1), note that for $j = 1, \ldots, r$ we have

$$\begin{aligned}
h^2(f, f+g_j) &= \frac{1}{2} \int_0^{\frac{2}{Lr}} \left(\sqrt{L/2} - \sqrt{L/2 + \tilde{g}(t)} \right)^2 \mathrm{d}t \\
&= \frac{L}{4} \int_0^{\frac{2}{Lr}} \left(1 - \sqrt{1 + (2/L)\tilde{g}(t)} \right)^2 \mathrm{d}t \\
&= \frac{L}{4} \left[\frac{4}{Lr} - 2 \int_0^{\frac{2}{Lr}} \sqrt{1 + (2/L)\tilde{g}(t)} \mathrm{d}t \right] \\
&= \frac{1}{r} - \frac{1}{2r} \left(\sqrt{1+a} + \sqrt{1-a} \right) \leq \frac{a^2}{2r}\,.
\end{aligned}$$

Define now $\tilde{c}_0 = \max(c_0 \log 2, 3)$ and choose $a^2 = r/(\tilde{c}_0 n) \leq 1$. Then $a^2/(2r) \leq (\tilde{c}_0 n)^{-1}$ for $M \geq 2$. Applying Lemma A.3 with $\beta = (\tilde{c}_0 n)^{-1}$ and $\alpha = \dfrac{L}{2\tilde{c}_0}$ we get

$$\inf_{T_n} \sup_{p \in \mathcal{C}} R_n(T_n, p) \geq \frac{1}{8\tilde{c}_0 \log 2} \left(1 - \sqrt{\frac{2}{\tilde{c}_0}} \right) \frac{L \log M}{n} = c_{15} \frac{L \log M}{n}\,.$$

∎

When $\mathcal{F} = \mathcal{F}_\infty(L)$, Theorems 3.3 and 3.7 imply the following result.

COROLLARY 3.4. *Let the integers $M \geq 2$ and $n \geq 1$ be such that $M \leq c_0 \mathrm{e}^n$. Then $\psi_n(M) = L(\log M)/n$ is an optimal rate of aggregation for $(\mathcal{E}, \mathcal{P}_0, \mathcal{F}_\infty(L))$, and $\tilde{p}_n^{\mathbf{MS}}$ is a rate optimal aggregate for $(\mathcal{E}, \mathcal{P}_0, \mathcal{F}_\infty(L))$.*

4. Conclusion

We now summarize and comment the results obtained in this chapter.

(1) **Problem (L): linear aggregation.** The optimal rate of aggregation for $(\mathbb{R}^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$ and for $(\mathbb{R}^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$ is the same and is given by

$$\psi_n^{\mathbf{L}}(M) = \frac{LM}{n}\,.$$

(2) **Problem (C): convex aggregation.** The optimal rate of aggregation for $(\Lambda^M, \mathcal{P}_0, L_2(\mathbb{R}^d))$ is given by

$$\psi_n^{\mathbf{C},L_2(\mathbb{R}^d)}(M) = \frac{LM}{n}\,,$$

and the optimal rate of aggregation for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$ is given by

$$\psi_n^{\mathbf{C},\mathcal{F}_\infty(L_0)}(M) = \begin{cases} \frac{M}{n} & \text{if } M \le \sqrt{n}\,, \\ \sqrt{\frac{\log M}{n}} & \text{if } M \ge n^{(1+\alpha)/2}, \alpha > 0. \end{cases} \tag{3.20}$$

(3) **Problem (MS): model selection aggregation.** The optimal rate of aggregation for $(\mathcal{E}, \mathcal{P}_0, \mathcal{F}_\infty(L))$ is given by

$$\psi_n^{\mathbf{MS}}(M) = \frac{L \log M}{n}\,.$$

Remark that in the case of convex aggregation there is a gap between the range $M \le \sqrt{n}$ and $M \ge n^{\frac{1+\alpha}{2}}$ for some $\alpha > 0$. In fact the optimal rate of convex aggregation for $(\Lambda^M, \mathcal{P}_0, \mathcal{F}_\infty(L_0))$ is given by

$$\sqrt{\frac{1}{n}\log\Big(\frac{M}{\sqrt{n}} + 1\Big)}\,,$$

when $M \ge \sqrt{n}$. However, it cannot be attained by the procedure described here. To overcome this problem, a solution consists in using *concentrated aggregation* as in Juditsky and Nemirovski (2000) and implicitly in Tsybakov (2003).

Part 2

Oracle inequalities and adaptation

CHAPTER 4

Sharp minimax estimation of a probability density

We consider the problem of estimation of a probability density with respect to the mean integrated squared error (MISE). For a Sobolev class of densities with non integer smoothness parameter $\beta > 1/2$, we evaluate the minimax rate of convergence, the exact constant and provide an asymptotically exact estimator.

Contents

The results of this chapter belong to Golubev (1991, 1992) who stated them essentially without proofs. Since some details of the proofs are not straightforward, we give them for completeness.

1. Introduction

We study the problem of estimating a density in $L_2(\mathbb{R}^d)$, with usually $d = 1$ except when generalization to greater d is straightforward. Consider independent and identically distributed random vectors $X_1, \dots, X_n$ with values in $\mathbb{R}^d$ having an unknown common probability density $p \in L_2(\mathbb{R}^d)$ that we want to estimate. For an estimator $\hat{p}$ of p based on the sample $\mathbb{X}^n = (X_1, \dots, X_n)$, define the mean integrated squared error by

$$R_n(\hat{p}, p) = E_p^n \|\hat{p} - p\|^2,$$

where E_p^n denotes the expectation with respect to the distribution P_p^n of $\mathbb{X}^n$ and, for a function $g \in L_2(\mathbb{R}^d)$,

$$\|g\| = \left(\int_{\mathbb{R}^d} g^2(x) \mathrm{d}x \right)^{1/2}.$$

For any function $h \in L_2(\mathbb{R}^d)$, let $\mathcal{F}[h]$ be its Fourier transform defined by $\mathcal{F}[h](\omega) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\omega^\top x} h(x) \mathrm{d}x,\ \omega \in \mathbb{R}^d$ (the integral is understood in the *limit in mean* sense). For any $\beta > 0,\ Q > 0$ and any integer $d \geq 1$ define the Sobolev classes of densities on $\mathbb{R}^d$ by

$$\Theta(\beta, Q) = \left\{ p : \mathbb{R}^d \to \mathbb{R} \,\middle|\, p \geq 0, \int_{\mathbb{R}^d} p(x) \mathrm{d}x = 1,\ \int_{\mathbb{R}^d} \|t\|_d^{2\beta} |\varphi(t)|^2 \mathrm{d}t \leq Q \right\},$$

where $\|\cdot\|_d$ denotes the Euclidean norm in $\mathbb{R}^d$ and $\varphi = \mathcal{F}[p]$ is the characteristic function of X_1. We now recall the definition of a sharp minimax estimator.

DEFINITION 4.1. *The estimator $\hat{p}_n$ of the density p is called* sharp minimax *(or* asymptotically exact*) on a class of densities $\mathcal{P}$ if it satisfies*

$$\sup_{p\in\mathcal{P}}\left[\psi_n^{-1}R_n(\hat{p}_n,p)\right]=\inf_{T_n}\sup_{p\in\mathcal{P}}\left[\psi_n^{-1}R_n(T_n,p)\right](1+o(1)),\quad n\to\infty, \tag{4.1}$$

where the infimum is taken over all possible estimators of p and ψ_n is the minimax rate of convergence over the class of densities $\mathcal{P}$, i.e.,

$$c\le\liminf_{n\to\infty}\inf_{T_n}\sup_{p\in\mathcal{P}}\left[\psi_n^{-1}R_n(T_n,p)\right]\le\limsup_{n\to\infty}\inf_{T_n}\sup_{p\in\mathcal{P}}\left[\psi_n^{-1}R_n(T_n,p)\right]\le C, \tag{4.2}$$

for some finite positive constants c and C.

In what follows, we specify ψ_n that satisfies (4.2) and provide an estimator $\hat{p}_n$ that satisfies (4.1).

2. Minimax lower bound

We begin by giving minimax lower bounds in the case $d=1$. Fix $d\ge 1, \beta>d/2, Q>0$ and define the constant

$$C^*=\frac{[Q(2\beta+d)]^{\frac{d}{2\beta+d}}}{d(2\pi)^d}\left(\frac{\beta S_d}{\beta+d}\right)^{\frac{2\beta}{2\beta+d}}, \tag{4.3}$$

where $S_d=2\pi^{d/2}/\Gamma(d/2)$ is the surface of a sphere of radius 1 in $\mathbb{R}^d$ and $\Gamma(\cdot)$ denotes the Gamma function. For $d=1$, the constant C^* equals the Pinsker constant (Pinsker, 1980; Tsybakov, 2004a, Chapter 3).

THEOREM 4.1. *Fix $d=1$ and $\beta>1/2$, $Q>0$. Then*

$$\liminf_{n\to\infty}\inf_{T_n}\sup_{p\in\Theta(\beta,Q)}\left[n^{\frac{2\beta}{2\beta+1}}R_n(T_n,p)\right]\ge C^*,$$

where the infimum is taken over all estimators of p.

This result can be found in Golubev (1992) without proof. The first proof of Theorem 4.1 should be attributed to Schipper (1996) who considered however, only integer values of β. Dalelane (2005a) gives a proof of Theorem 4.1 which essentially follows that of Schipper (1996). Efromovich (2000) proved another result on sharp minimax lower bounds in the multivariate case extending the one-dimensional setting of Efroĭmovich and Pinsker (1982). However, it cannot be applied to our setup in full generality since it treats the densities with respect to a finite measure.

In this proof, we need a theorem by Golubev (1991) which uses *local asymptotic normality* (LAN). Even though theory has been clearly established in the parametric case (see, e.g., Ibragimov and Khas′minskiĭ, 1981), there exist several formulations of the LAN property in a nonparametric framework. One of them, due to Ibragimov and Khas′minskiĭ (1991), is a direct generalization of the results of Ibragimov and Khas′minskiĭ (1981). Le Cam (1986) has a different approach to define the same property. Golubev (1991) gives a definition of LAN which is a particular case of the one given by Ibragimov and Khas′minskiĭ (1991). Before giving the definition of the LAN property, we describe the framework of Golubev (1991). Let $\left\{\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),P_p^n,p\in\Theta\right\}$, $\Theta\subset L_2(\mathbb{R})$, be the statistical experiment of estimating the density $p\in\Theta$ from the observations $\mathbb{X}^n=(X_1,\dots,X_n)$. Here the number of observations n is a large parameter ($n\to\infty$). For an estimator $\hat{p}_n$ of

p, define its risk by

$$R_n\big(p_n^*, B_r(p_0)\big) = \sup_{p\in B_r(p_0)} R_n(p_n^*, p) = \sup_{p\in B_r(p_0)} E_p^n \int_{\mathbb{R}} \big(p(t) - p_n^*(t)\big)^2 \mathrm{d}t\,,$$

where E_p^n is the expectation corresponding to the measure P_p^n, and $\{B_r(p_0), r>0\}$ is a family of neighborhoods of p_0 in Θ shrinking to the singleton $\{p_0\}$, $p_0\in\Theta$, when $r\to 0$. Let P and Q be positive numbers such that $P<Q$ and fix $\beta>1/2$. Denote by $\|\cdot\|_\infty$ the L_∞- norm, i.e., $\|u\|_\infty = \sup_{t\in\mathbb{R}}|u(t)|$ and let K be a compact subset of $\mathbb{R}$. Consider then the family of neighborhoods $B_r(p_0) = B_r(p_0,\beta,P,K)$, where $B_r(p_0,\beta,P,K)$ is the set of all functions p in $L_2(\mathbb{R})$ such that:

(1) $p(t) = p_0(t),\ t\notin K$,
(2) $\int [p(t)-p_0(t)]\mathrm{d}t = 0$,
(3) $\|p-p_0\|_\infty < r$,
(4) $\int |\omega|^{2\beta}|\mathcal{F}[p](\omega) - \mathcal{F}[p_0](\omega)|^2\mathrm{d}\omega < P, \qquad \beta>1/2,\ 0<P<Q$,

Moreover, let

$$D = \left\{ v\in C^\infty(\mathbb{R}), v(t) = 0,\ \forall\ |t|>1/2, \int_{\mathbb{R}} v(t)\mathrm{d}t = 0 \right\}.$$

We can now state the definition of the *LAN property*. Let $(g_n)_n$ be a non-increasing sequence of positive numbers.

Definition 4.2. *The family of probability measures* $\{P_p^n,\ p\in\Theta\}$ *obeys* LAN with normalization g_n *at the point* (p_0,s), $p_0\in\Theta$, $s\in\mathbb{R}$, *if there exists* $I(p_0,s)>0$ *such that the following representation holds for any function* $v\in D$*:*

$$\frac{\mathrm{d}P^n_{p_0+v_s^n}}{\mathrm{d}P^n_{p_0}}(\mathbb{X}^n) = \exp\left(L^n[v] - \frac{\|v\|^2}{2} + \Phi^n[v]\right), \tag{4.4}$$

where

- v_s^n *is defined as follows*

$$v_s^n(t) = \frac{1}{n(I(p_0,s)g_n)^{1/2}} v\left(\frac{t-s}{g_n}\right), \tag{4.5}$$

- L^n *is a linear functional on* D *such that* $L^n[v]$ *converges weakly under* $P^n_{p_0}$ *to a Gaussian random variable with zero mean and variance* $\|v\|^2$ *when* $n\to\infty$,
- Φ^n *is a functional on* D *such that, for any* $R>0$ *and* $\delta>0$,

$$\lim_{n\to\infty}\sup_{\|v\|_\infty<R} P^n_{p_0}\{|\Phi^n[v]|>\delta\} = 0.$$

The main result of Golubev (1991), applied to our framework, is the following.

Lemma 4.1. *Fix* $p_0\in\Theta$. *Let* K *be a compact set in* $\mathbb{R}$ *and* $g_n = n^{-\frac{1}{2\beta+1}}$. *Moreover, let the following conditions hold:*

(i) $B_r(p_0)\subset\Theta$ *for sufficiently small* r,
(ii) *the family* $\{P_p^n,\ p\in\Theta\}$ *obeys LAN with normalization* g_n *at point* (p_0,s), *for all* $s\in K$,
(iii) $s\mapsto 1/I(p_0,s)>0$ *is a continuous function on* K.

Then,

$$\liminf_{r\to 0}\liminf_{n\to\infty}\inf_{T_n}\left[ng_nR_n\big(T_n, B_r(p_0)\big)\right] \geq \Delta(\beta)\left[\int_K \frac{\mathrm{d}s}{I(p_0,s)}\right]^{\frac{2\beta}{2\beta+1}},$$

where the infimum is taken over all possible estimators of p and

$$\Delta(\beta) = \frac{1}{\pi}\left[\frac{P\,(2\beta+1)}{2}\right]^{\frac{1}{2\beta+1}}\left(\frac{\beta}{\beta+1}\right)^{\frac{2\beta}{2\beta+1}}.$$

In the sequel we will need the following simple lemma.

LEMMA 4.2. *If $\beta > 1/2$, then every density $p \in \Theta(\beta, Q)$ is continuous.*

Proof is straightforward and is therefore omitted.

PROOF OF THEOREM 4.1. We are going to apply Lemma 4.1. We first show that for a proper choice of K, the family $\left\{P_p^n, p \in \Theta(\beta, Q)\right\}$ obeys LAN, i.e. we check *(ii)* of Lemma 4.1. Let K_1 be a compact set in $\mathbb{R}$ with non empty interior and such that $K_1 \subset \{p_0(s) > 0\}$.

LEMMA 4.3. *Let $p_0 \in \Theta(\beta, Q), \beta > 1/2, Q > 0$. Then $\left\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_p^n, p \in \Theta(\beta, Q)\right\}$, the statistical experiment of estimating the density $p \in \Theta(\beta, Q)$ from the observation $\mathbb{X}^n = (X_1, \ldots, X_n)$ obeys LAN at all points (p_0, s) such that $p_0(s) > 0$, $s \in K_1$, with $I(p_0, s) = 1/p_0(s)$ and normalization $g_n = n^{-\frac{1}{2\beta+1}}$.*

PROOF. Fix s such that $p_0(s) > 0$. With $I(p_0, s) = 1/p_0(s)$ in (4.5), we have

$$v_s^n(t) = \sqrt{\frac{p_0(s)}{ng_n}}\, v\left(\frac{t-s}{g_n}\right).$$

Further, note that for the sample $\mathbb{X}^n$ with marginal density p_0, we have almost surely $p_0(X_j) > 0$ for all $j = 1, \ldots, n$. A Taylor expansion gives

$$\begin{aligned}\frac{\mathrm{d}P_{p_0+v_s^n}^n}{\mathrm{d}P_{p_0}^n}(\mathbb{X}^n) &= \prod_{j=1}^n \left(1 + \frac{\sqrt{p_0(s)}}{p_0(X_j)\sqrt{ng_n}}\, v\left(\frac{X_j - s}{g_n}\right)\right) \\ &= \exp\left[\frac{1}{\sqrt{ng_n}}\sum_{j=1}^n \frac{\sqrt{p_0(s)}}{p_0(X_j)}\, v\left(\frac{X_j - s}{g_n}\right) - \frac{1}{2ng_n}\sum_{j=1}^n \frac{p_0(s)}{p_0^2(X_j)}\, v^2\left(\frac{X_j - s}{g_n}\right) + r_n\right],\end{aligned}$$

where

$$r_n = \sum_{j=1}^n \int_0^{b_{n,j}} \frac{(b_{n,j} - t)^3}{3(1+t)^3}\mathrm{d}t\,,$$

with

$$b_{n,j} = \frac{\sqrt{p_0(s)}}{p_0(X_j)\sqrt{ng_n}}\, v\left(\frac{X_j - s}{g_n}\right), \quad j = 1, \ldots, n\,.$$

To prove (4.4), it is sufficient to show that for any function $v \in D$, we have the following properties:

$$S_n = \frac{1}{\sqrt{ng_n}}\sum_{j=1}^n \frac{\sqrt{p_0(s)}}{p_0(X_j)}\, v\left(\frac{X_j - s}{g_n}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \|v\|^2\right), \tag{4.6}$$

and for any $R > 0,\ \delta > 0$, we have

$$\sup_{\|v\|_\infty < R} P_{p_0}^n\left\{\left|\frac{\|v\|^2}{2} - \frac{1}{2ng_n}\sum_{j=1}^n \frac{p_0(s)}{p_0^2(X_j)}\, v^2\left(\frac{X_j - s}{g_n}\right) + r_n\right| > \delta\right\} \xrightarrow[n\to\infty]{} 0. \tag{4.7}$$

Proof of (4.6)**.** First, note that if $p_0(s) > 0$, when $g_n = n^{-\frac{1}{2\beta+1}}$, for sufficiently large n, $p_0(s + ug_n) > 0$ for any fixed $u \in \mathbb{R}$, by continuity of p_0 (cf. Lemma 4.2). Set then

$$Y_j = \frac{1}{\sqrt{g_n}} \frac{\sqrt{p_0(s)}}{p_0(X_j)} v\left(\frac{X_j - s}{g_n}\right) \left(p_0(s) \int \frac{v^2(u)}{p_0(s+ug_n)} \mathrm{d}u\right)^{-1/2} \left(\int v^2(u)\mathrm{d}u\right)^{1/2}.$$

The random variables $Y_j, j = 1, \ldots, n$ are independent and since $v \in D$,

$$E^n_{p_0}[Y_j] = \left(\int v^2(u)\mathrm{d}u\right)^{1/2} \left(p_0(s) \int \frac{v^2(u)}{p_0(s+ug_n)} \mathrm{d}u\right)^{-1/2} \frac{\sqrt{p_0(s)}}{\sqrt{g_n}} \int v\left(\frac{t-s}{g_n}\right) \mathrm{d}t = 0.$$

Moreover,

$$\begin{aligned} E^n_{p_0}[Y_j^2] &= \left(\int v^2(u)\mathrm{d}u\right) \left(p_0(s) \int \frac{v^2(u)}{p_0(s+ug_n)} \mathrm{d}u\right)^{-1} \frac{p_0(s)}{g_n} \int v^2\left(\frac{t-s}{g_n}\right) \frac{1}{p_0(t)} \mathrm{d}t \\ &= \int v^2(u)\mathrm{d}u = \|v\|^2. \end{aligned}$$

It follows from continuity of p_0 at point s that

$$\left(p_0(s) \int \frac{v^2(u)}{p_0(s+ug_n)} \mathrm{d}u\right)^{-1/2} \left(\int v^2(u)\mathrm{d}u\right)^{1/2} \to 1, \quad n \to \infty\,.$$

Thus, by the central limit theorem combined to Slutsky's lemma, we get the following convergence in distribution,

$$S_n \to \mathcal{N}\left(0, \|v\|^2\right), \quad n \to \infty.$$

Proof of (4.7). We have

$$0 \le r_n \le \frac{1}{3} \sum_{j=1}^n b_{n,j}^4 \le \frac{1}{3} \sum_{j=1}^n \frac{p_0^2(s)}{p_0^4(X_j) n^2 g_n^2} v^4\left(\frac{X_j - s}{g_n}\right).$$

By continuity of p_0 on the compact set K_1, there exist two constants $c_1 > 0$ and $c_2 > 0$ such that $c_1 \le p_0(s) \le c_2$, for any $s \in K_1$. Furthermore, since v vanishes outside of $[-1/2, 1/2]$,

$$v^4\left(\frac{X_j - s}{g_n}\right) = 0 \quad \text{when} \quad |X_j - s| > \frac{g_n}{2}.$$

If $|X_j - s| \le g_n/2$, for sufficiently large n, we have $p_0(X_j) \ge c_1/2$. Therefore, there exists a constant $c_3 > 0$ such that, for $\|v\|_\infty < R$,

$$0 \le r_n \le \frac{1}{3} \sum_{j=1}^n \frac{16 c_2^2}{c_1^4 n^2 g_n^2} R^4 \le \frac{c_3}{n g_n^2}.$$

Since $n g_n^2 = n^{\frac{2\beta-1}{2\beta+1}} \to +\infty,\ n \to +\infty$, for any $\beta > 1/2$ we have almost surely $r_n \to 0$ when n tends to infinity. To end the proof, remark that the law of large numbers yields

$$\left| \frac{\|v\|^2}{2} - \frac{1}{2ng_n} \sum_{j=1}^n \frac{p_0(s)}{p_0^2(X_j)} v^2\left(\frac{X_j - s}{g_n}\right) \right| \to 0, \quad n \to \infty, \quad \text{a.s.}$$

∎

We now check conditions *(i)* and *(iii)* of Lemma 4.1. By Lemma 4.2, condition *(iii)* of Lemma 4.1 is satisfied for $I(p_0, s) = 1/p_0(s)$ if $\beta > 1/2$. Fix $\beta > 1/2$ and $Q > 0$. Next, we choose p_0 to be the density of $\mathcal{N}(0, \sigma^2)$ with a variance $\sigma^2 = \sigma^2(\beta, Q)$ chosen

large enough to ensure that $p_0 \in \Theta(\beta, (Q-P)/2)$. For $K_\varepsilon = [-1/\varepsilon, 1/\varepsilon],\ 0 < \varepsilon < 1$, we have $B_r(p_0, \beta, P, K_\varepsilon) \subset \Theta(\beta, Q)$. Indeed, for every $p \in B_r(p_0, \beta, P, K_\varepsilon)$ we have on the one hand,

$$|\mathcal{F}[p](\omega) - \mathcal{F}[p_0](\omega)| = \left| \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}t\omega} \left(p(t) - p_0(t)\right) \mathrm{d}t \right| \le \int_{K_\varepsilon} |p(t) - p_0(t)| \mathrm{d}t \le |K_\varepsilon| r\,,$$

where $|K_\varepsilon| = 2/\varepsilon$ is the Lebesgue measure of the interval K_ε. On the other hand,

$$\begin{aligned}
\int_{\mathbb{R}} |\omega|^{2\beta} |\mathcal{F}[p](\omega)|^2 \mathrm{d}\omega &= \int_{\mathbb{R}} |\omega|^{2\beta} |\mathcal{F}[p](\omega) - \mathcal{F}[p_0](\omega) + \mathcal{F}[p_0](\omega)|^2 \mathrm{d}\omega \\
&\le \int_{\mathbb{R}} |\omega|^{2\beta} |\mathcal{F}[p](\omega) - \mathcal{F}[p_0](\omega)|^2 \mathrm{d}\omega + \int_{\mathbb{R}} |\omega|^{2\beta} \left[\mathcal{F}[p_0](\omega)\right]^2 \mathrm{d}\omega \\
&\qquad + 2 \int_{\mathbb{R}} |\omega|^{2\beta} |\mathcal{F}[p](\omega) - \mathcal{F}[p_0](\omega)| \left[\mathcal{F}[p_0](\omega)\right] \mathrm{d}\omega \\
&\le P + \frac{Q-P}{2} + 2|K_\varepsilon| r \int_{\mathbb{R}} |\omega|^{2\beta} \mathcal{F}[p_0](\omega) \mathrm{d}\omega.
\end{aligned} \tag{4.8}$$

Since $\mathcal{F}[p_0](\omega) = \mathrm{e}^{-\frac{\sigma^2\omega^2}{2}}$,

$$\int_{\mathbb{R}} |\omega|^{2\beta} \mathcal{F}[p_0](\omega) \mathrm{d}\omega < \infty,$$

Hence there exists $r_0 > 0$ such that

$$2|K_\varepsilon| r \int_{\mathbb{R}} |\omega|^{2\beta} \mathcal{F}[p_0](\omega) \mathrm{d}\omega \le \frac{Q-P}{2}, \quad \forall\, r \le r_0\,.$$

It follows that for any $r \le r_0$,

$$\int_{\mathbb{R}} |\omega|^{2\beta} |\mathcal{F}[p](\omega)|^2 \mathrm{d}\omega \le Q,$$

that is, $p \in \Theta(\beta, Q)$ and condition *(i)* of Lemma 4.1 is satisfied for $K = K_\varepsilon$ and sufficiently small r. Thus all the conditions of Lemma 4.1 are satisfied. This lemma implies that

$$\begin{aligned}
\liminf_{n\to\infty} \inf_{T_n} \sup_{p \in \Theta(\beta,Q)} \left[n^{\frac{2\beta}{2\beta+1}} R_n(T_n, p)\right] &\ge \liminf_{r\to 0} \liminf_{n\to\infty} \inf_{T_n} \sup_{p \in B_r(p_0)} \left[n^{\frac{2\beta}{2\beta+1}} R_n(T_n, p)\right] \\
&\ge \Delta(\beta) \int_{K_\varepsilon} p_0(s) \mathrm{d}s.
\end{aligned}$$

Since the last inequality is valid for any $P < Q$, we have

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{p \in \Theta(\beta,Q)} \left[n^{\frac{2\beta}{2\beta+1}} R_n(T_n, p)\right] \ge C^* \int_{K_\varepsilon} p_0(s) \mathrm{d}s,$$

where C^* is defined in (4.3). Finally, letting $\varepsilon \to 0$, we conclude the proof of the theorem. ∎

3. Attainability of the lower bound

A widely used nonparametric density estimator is the kernel density estimator defined by

$$\hat{p}_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big), \quad \forall\, x \in \mathbb{R}^d\,. \tag{4.9}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is an integrable function satisfying $K(x) = K(-x),\ \forall\, x \in \mathbb{R}^d$ and $\int K(u)\mathrm{d}u = 1$.

Denote by φ_n the empirical characteristic function (e.c.f.) defined by

$$\varphi_n(\omega) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}\omega^\top X_k}, \quad \forall\, \omega \in \mathbb{R}^d.$$

Then, the Fourier transform of a kernel estimator defined in (4.9) presented in terms of its e.c.f. is:

$$\mathcal{F}[\hat{p}_{n,h}](\omega) = \varphi_n(\omega)\mathcal{F}[K](h\omega).$$

A kernel density estimator is therefore completely determined by the Fourier transform of the kernel K and the bandwidth parameter, $h > 0$. Consider the Pinsker kernel K_β, i.e., the kernel having the Fourier transform

$$\mathcal{F}[K_\beta](\omega) = \left(1 - \|\omega\|_d^\beta\right)_+, \quad \omega \in \mathbb{R}^d,$$

where $x_+ = \max(x, 0)$. For integer values of β and $d = 1$, the kernel K_β can be computed explicitly using inverse Fourier transform and is given by

$$K_\beta(x) = \frac{\beta!}{\pi}\sum_{j=1}^{\beta} \frac{\sin^{(j)}(x)}{(\beta - j)!x^{j+1}}.$$

Fix $\beta > d/2, Q > 0$, and define the *Pinsker type kernel density estimator* $\tilde{p}_{n,h^*}$ as a kernel density estimator of the type (4.9) with kernel $K = K_\beta$ and bandwidth parameter

$$h^* = D^* n^{-\frac{1}{2\beta+d}} \quad \text{where} \quad D^* = \left(\frac{\beta S_d}{Q(\beta + d)(2\beta + d)}\right)^{\frac{1}{2\beta+d}}. \tag{4.10}$$

We now give the following upper bound for the general case $d \geq 1$ and $\beta > 0$.

THEOREM 4.2. *For any integer $d \geq 1$ and any $\beta > d/2,\ Q > 0$, the Pinsker type kernel density estimator $\hat{p}_{n,h^*}$ satisfies*

$$\sup_{p \in \Theta(\beta,Q)} R_n(\tilde{p}_{n,h^*}, p) \leq C^* n^{-\frac{2\beta}{2\beta+d}}(1 + o(1)),$$

where C^* is defined in (4.3).

PROOF. We will use the following Fourier representation for the MISE of kernel density estimators that can be easily obtained from Plancherel's formula -it is a multivariate extension of the representation for $d = 1$ given, e.g., in Golubev (1992) and in Wand and Jones (1995, p. 55). For any kernel density estimator $\hat{p}_{n,h}$ with kernel K and bandwidth parameter $h > 0$, we have

$$R_n(\hat{p}_{n,h}, p) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left(|1 - \mathcal{F}[K](ht)|^2|\varphi(t)|^2 + \frac{1}{n}\left(1 - |\varphi(t)|^2\right)\left|\mathcal{F}[K](ht)\right|^2\right)\mathrm{d}t. \tag{4.11}$$

Using (4.11) and the fact that $0 \leq \mathcal{F}[K_\beta](t) \leq 1,\ \forall\, t \in \mathbb{R}^d$, we get for the Pinsker type kernel density estimator

$$\begin{aligned} R_n(\tilde{p}_{n,h}, p) &\leq \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left(|1 - \mathcal{F}[K_\beta](ht)|^2|\varphi(t)|^2 + \frac{1}{n}\left|\mathcal{F}[K_\beta](ht)\right|^2\right)\mathrm{d}t \\ &\leq \frac{1}{(2\pi)^d}\left(Qh^{2\beta} + \frac{1}{n}\int_{\mathbb{R}^d}\left|\mathcal{F}[K_\beta](ht)\right|^2\mathrm{d}t\right), \forall\, h > 0,\ p \in \Theta(\beta, Q). \end{aligned} \tag{4.12}$$

Note now that h^* is a solution of the equation

$$\int_{\mathbb{R}^d} \|t\|_d^\beta \mathcal{F}[K_\beta](ht)\mathrm{d}t = Qnh^\beta. \tag{4.13}$$

With h^* satisfying (4.13), inequality (4.12) becomes

$$\begin{aligned} R_n(\tilde{p}_{n,h^*}, p) &\leq \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} \mathcal{F}[K_\beta](h^*t) \left[\mathcal{F}[K_\beta](h^*t) + \|h^*t\|_d^\beta\right] \mathrm{d}t \\ &= \frac{1}{(2\pi)^d n (h^*)^d} \int_{\mathbb{R}^d} \mathcal{F}[K_\beta](t)\mathrm{d}t \\ &= \frac{1}{(2\pi)^d n (h^*)^d} \int_0^1 \left(1 - r^\beta\right) r^{d-1} S_d \mathrm{d}r \\ &= C^* n^{-\frac{2\beta}{2\beta+d}}. \end{aligned}$$

∎

From Theorems 4.1 and 4.2 we obtain the following corollary.

COROLLARY 4.1. *Fix $d = 1, \beta > 1/2$ and $Q > 0$. Then $\psi_n = n^{-\frac{2\beta}{2\beta+1}}$ is the minimax rate of convergence over the Sobolev class of densities $\Theta(\beta, Q)$, with exact constant C^* defined in* (4.3)*. Moreover, the Pinsker type kernel density estimator $\tilde{p}_{n,h^*}$ is sharp minimax on the Sobolev class of densities $\Theta(\beta, Q)$, i.e.,*

$$\sup_{p \in \Theta(\beta,Q)} \left[n^{\frac{2\beta}{2\beta+1}} R_n(\tilde{p}_{n,h^*}, p)\right] = \inf_{T_n} \sup_{p \in \Theta(\beta,Q)} \left[n^{\frac{2\beta}{2\beta+1}} R_n(T_n, p)\right] (1 + o(1)), \quad n \to \infty,$$

where the infimum is taken over all possible estimators of p.

We conjecture that the same corollary holds for $d \geq 1$. However, the proof of the lower bound becomes more cumbersome, even tough it should not introduce new mathematical aspects and could be solved using the same technique as in the proof of Theorem 4.1 as suggested by Golubev (1991).

CHAPTER 5

From aggregation to adaptation

In this chapter, we apply general results to aggregation of multivariate kernel density estimators with different bandwidths. We show that linear and convex aggregates described in Chapter 3 mimic the kernel oracles in an asymptotically exact sense for a large class of kernels including Gaussian, Silverman's and Pinsker's ones. We prove that, for Pinsker's kernel, the proposed aggregates are sharp asymptotically minimax simultaneously over a large scale of Sobolev classes of densities. Finally, we provide simulations demonstrating performance of the convex aggregation procedure.

Contents

The material of this chapter is a joint work with Alexandre Tsybakov (Rigollet and Tsybakov, 2006).

1. Introduction

Besides being in themselves precise finite sample results, exact oracle inequalities are a very useful tool in nonparametric estimation. In particular, they allow one to prove that an aggregate estimator is sharp minimax adaptive in several cases. We present here a method that solves the problem of adaptation to the unknown smoothness of a probability density using aggregation of kernel density estimators. Of course, there exists a large variety of other methods of adaptation to unknown smoothness. In the numerical examples that we consider, our aggregates are comparable to benchmarks, and show somewhat more stable behavior for densities with highly inhomogeneous smoothness (cf. Section 5). It is important to note that aggregation can be used for adaptation to other characteristics than smoothness, for example, to the dimension of the subspace where the data effectively lie, under dimension reduction models (cf. Samarov and Tsybakov, 2005). Even though results are presented here in the particular context of nonparametric density estimation to avoid the introduction of cumbersome notations, the general method for deriving adaptivity from aggregation can be extended to other popular statistical models such as nonparametric regression or single index model.

Specifically, we deal with aggregation of multivariate kernel density estimators with different bandwidths. Here the number $M = M_n$ of estimators depends on n and satisfies

$M_n/n \to 0$, as $n \to \infty$. We show in Corollary 5.2, that linear and convex aggregates mimic the kernel oracles in a sharp asymptotic sense. This corollary is in the spirit of Stone's (1984) theorem on asymptotic optimality of cross-validation, but it is more powerful in several aspects because it is obtained under weaker conditions on p and covers kernels with unbounded support including Gaussian, Silverman's and Pinsker's kernels. Another application of our results is that, for Pinsker's kernel, we construct aggregates that are sharp asymptotically minimax simultaneously over a large scale of Sobolev classes of densities in the multidimensional case.

2. Sample splitting and averaged aggregates

We first recall the framework of Chapter 3. Consider i.i.d. random vectors $X_1, \dots, X_n$ with values in $\mathbb{R}^d$ having an unknown common probability density $p \in L_2(\mathbb{R}^d)$ that we want to estimate. For an estimator $\hat{p}$ of p based on the sample $\mathbb{X}^n = (X_1, \dots, X_n)$, define the L_2-risk

$$R_n(\hat{p}, p) = E_p^n \|\hat{p} - p\|^2,$$

where E_p^n denotes the expectation with respect to the distribution P_p^n of $\mathbb{X}^n$ and, for a function $g \in L_2(\mathbb{R}^d)$,

$$\|g\| = \left(\int_{\mathbb{R}^d} g^2(x) \mathrm{d}x \right)^{1/2}.$$

To perform aggregation, we use a sample splitting scheme. The sample $\mathbb{X}^n$ is split into two independent subsamples $\mathbb{X}_1^m$ (training sample) and $\mathbb{X}_2^\ell$ (validation sample) of sizes m and ℓ respectively where $m + \ell = n$ and usually $m \gg \ell$. The first subsample $\mathbb{X}_1^m$ is used to construct estimators $\hat{p}_j = \hat{p}_{m,j}$, $j = 1, \dots, M$, while the second subsample $\mathbb{X}_2^\ell$ is used to aggregate them, i.e., to construct $\tilde{p}_n$ (thus, $\tilde{p}_n$ is measurable w.r.t. the whole sample $\mathbb{X}^n$). The sample splitting scheme is illustrated in Figure 1.

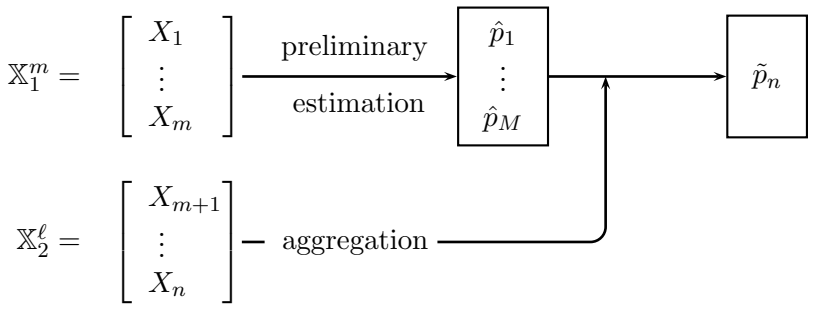

FIGURE 1. Sample splitting scheme

For given $m < n$ the two subsamples can be obtained by different splits. The choice of split being arbitrary, it may influence the result of estimation. In order to avoid the arbitrariness, we use a jackknife type procedure that averages the aggregates over different splits. Define a *split* $\mathcal{S}$ of the initial sample $\mathbb{X}^n$ as a mapping

$$\mathcal{S} \ : \ \mathbb{X}^n \mapsto \left(\mathbb{X}_1^m, \mathbb{X}_2^\ell \right).$$

Denote by $\mathbb{X}_{1,\mathcal{S}}^m, \mathbb{X}_{2,\mathcal{S}}^\ell$ subsamples obtained for a fixed split $\mathcal{S}$ and consider an arbitrary set of splits $\mathbb{S}$. It can be, for example, the set of all splits. Define $\tilde{p}_n^{\mathcal{S}}$ as a linear or convex

aggregate ($\tilde{p}_n^{\mathbf{L}}$ or $\tilde{p}_n^{\mathbf{C}}$ respectively, using notation of Chapter 3) based on the validation sample $\mathbb{X}_{2,\mathcal{S}}^{\ell}$ and on the initial set of estimators $p_j = \hat{p}_{m,j}^{\mathcal{S}}, j = 1, \dots, M$, where each of the $\hat{p}_{m,j}^{\mathcal{S}}$'s is constructed from the training sample $\mathbb{X}_{1,\mathcal{S}}^{m}$. Introduce the following *averaged aggregate* estimator:

$$\tilde{p}_n^{\mathbb{S}} \triangleq \frac{1}{\mathrm{card}(\mathbb{S})} \sum_{\mathcal{S} \in \mathbb{S}} \tilde{p}_n^{\mathcal{S}}. \tag{5.1}$$

Let H be either $\mathbb{R}^M$ or a convex compact subset of $\mathbb{R}^M$. Define

$$\Delta_{\ell,M} = \begin{cases} LM/\ell & \text{if } H = \mathbb{R}^M, \\ 4LM/\ell & \text{if } H \text{ is a convex compact subset of } \mathbb{R}^M. \end{cases}$$

We get the following corollary of Theorems 3.1 and 3.2.

COROLLARY 5.1. *Let $m < n$, $\ell = n - m$, and let H be either $\mathbb{R}^M$ or a convex compact subset of $\mathbb{R}^M$. Let $\mathbb{S}$ be an arbitrary set of splits. Assume that $\hat{p}_{m,1}^{\mathcal{S}}, \dots, \hat{p}_{m,M}^{\mathcal{S}} \in L_2(\mathbb{R}^d)$ for fixed $\mathbb{X}_{1,\mathcal{S}}^{m}$, $\forall \mathcal{S} \in \mathbb{S}$, and that $p \in \mathcal{P}_0$. Then the averaged aggregate* (5.1) *satisfies*

$$R_n(\tilde{p}_n^{\mathbb{S}}, p) \le \inf_{\lambda \in H} R_m\Big(\sum_{j=1}^{M} \lambda_j \hat{p}_{m,j}^{\mathcal{S}}, p\Big) + \Delta_{\ell,M} \tag{5.2}$$

for any integers $M \ge 2$, $n \ge 1$ and any split $\mathcal{S}$.

PROOF. For any fixed $\mathcal{S} \in \mathbb{S}$ and for a fixed training subsample $\mathbb{X}_{1,\mathcal{S}}^{m}$ inequalities (3.5) and (3.7) imply

$$E_p^{\ell,\mathcal{S}} \|\tilde{p}_n^{\mathcal{S}} - p\|^2 \le \min_{\lambda \in H} \Big\| \sum_{j=1}^{M} \lambda_j \hat{p}_{m,j}^{\mathcal{S}} - p \Big\|^2 + \Delta_{\ell,M}, \quad \forall\, p \in \mathcal{P}_0, \tag{5.3}$$

where $E_p^{\ell,\mathcal{S}}$ denotes the expectation w.r.t. the distribution of the validation sample $\mathbb{X}_{2,\mathcal{S}}^{\ell}$ when the true density is p. Taking expectations of both sides of (5.3) w.r.t. the training sample $\mathbb{X}_{1,\mathcal{S}}^{m}$ we get

$$R_n(\tilde{p}_n^{\mathcal{S}}, p) \le \inf_{\lambda \in H} R_m\Big(\sum_{j=1}^{M} \lambda_j \hat{p}_{m,j}^{\mathcal{S}}, p\Big) + \Delta_{\ell,M}. \tag{5.4}$$

The right hand side here does not depend on $\mathcal{S}$. By Jensen's inequality,

$$R_n(\tilde{p}_n^{\mathbb{S}}, p) \le \frac{1}{\mathrm{card}(\mathbb{S})} \sum_{\mathcal{S} \in \mathbb{S}} R_n(\tilde{p}_n^{\mathcal{S}}, p).$$

This and (5.4) yield (5.2). ∎

3. Kernel aggregates for density estimation

Let $\hat{p}_{m,h}$ denote a kernel density estimator based on $\mathbb{X}_1^m$ with $m \le n$,

$$\hat{p}_{m,h}(x) \triangleq \frac{1}{mh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \mathbb{I}_{\{\mathbb{X}_1^m\}}(X_i), \quad x \in \mathbb{R}^d, \tag{5.5}$$

where $h > 0$ is a bandwidth and $K \in L_2(\mathbb{R}^d)$ is a kernel. In order to cover such examples as the sinc kernel we will not assume that K is integrable.

Define $h_0 = (n \log n)^{-1/d}$, $a_n = a_0 / \log n$, where $a_0 > 0$ is a constant, and M such that

$$M - 2 = \max\left\{ j \in \mathbb{N} : h_0(1 + a_n)^j < 1 \right\}.$$

It is easy to see that $M \le c_4(\log n)^2$, where $c_4 > 0$ is a constant depending only on a_0 and d. Consider a grid $\mathcal{H}$ on $[0,1]$ with a weakly geometrically increasing step:

$$\mathcal{H} \triangleq \{h_0, h_1, \dots, h_{M-1}\} ,$$

where $h_j = (1+a_n)^j h_0$, $j = 1, \dots, M-2$, and $h_{M-1} = 1$. Fix now an arbitrary family of splits $\mathbb{S}$ such that, for $n \ge 3$,

$$m = \lfloor n\left(1 - (\log n)^{-1}\right)\rfloor \quad \text{and} \quad \ell = n - m \ge \frac{n}{\log n} ,$$

where $\lfloor x \rfloor$ denotes the integer part of x.

Define $\tilde{p}_n^{\mathbb{S},K}$ as the linear or convex (with $H = \Lambda^M$) averaged aggregate $\tilde{p}_n^{\mathbb{S}}$ where the initial estimators are taken in the form $p_j = \hat{p}_{m,h_{j-1}}, j = 1, \dots, M$, with $\hat{p}_{m,h}$ given by (5.5).

REMARK 5.1. *The kernel density estimator p_1 for instance is not bounded when $n \to \infty$ and we cannot apply results concerning aggregation of bounded estimators (cf. Theorems 3.1 and 3.3). Therefore, we focus on linear and convex aggregation of kernel density estimators.*

Since $\Delta_{\ell,M} \le 4LM/\ell$, we get from (5.2) that, under the assumptions of Corollary 5.1,

$$R_n(\tilde{p}_n^{\mathbb{S},K}, p) \le \min_{h\in\mathcal{H}} R_m(\hat{p}_{m,h}, p) + \Delta_{\ell,M} \le \min_{h\in\mathcal{H}} R_m(\hat{p}_{m,h}, p) + \frac{4c_4(\log n)^3}{n} . \tag{5.6}$$

We now give a theorem that extends (5.6) to the n-sample oracle risk $\inf_{h>0} R_n(\hat{p}_{n,h}, p)$ instead of $\min_{h\in\mathcal{H}} R_m(\hat{p}_{m,h}, p)$. Denote by $\mathcal{F}[f]$ the Fourier transform defined for $f \in L_2(\mathbb{R}^d)$ and normalized in such a way that its restriction to $f \in L_2(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ has the form $\mathcal{F}[f](t) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}x^T t} f(x)\mathrm{d}x$, $t \in \mathbb{R}^d$. In the sequel $\varphi = \mathcal{F}[p]$ denotes the characteristic function associated to p.

THEOREM 5.1. *Assume that p satisfies $\|p\|_\infty \le L$ with $0 < L < \infty$ and let $K \in L_2(\mathbb{R}^d)$ be a kernel such that a version of its Fourier transform $\mathcal{F}[K]$ takes values in $[0,1]$ and satisfies the monotonicity condition $\mathcal{F}[K](h't) \ge \mathcal{F}[K](ht)$, $\forall\, t \in \mathbb{R}^d$, $h > h' > 0$. Then there exists an integer $n_0 = n_0(L, \|K\|) \ge 4$ such that for $n \ge n_0$ the averaged aggregate $\tilde{p}_n^{\mathbb{S},K}$ satisfies the oracle inequality*

$$R_n(\tilde{p}_n^{\mathbb{S},K}, p) \le \left(1 + c_5(\log n)^{-1}\right) \inf_{h>0} R_n(\hat{p}_{n,h}, p) + c_6\frac{(\log n)^3}{n} , \tag{5.7}$$

where c_5 is a positive constant depending only on d and a_0, and $c_6 > 0$ depends only on $L, \|K\|, d$ and a_0.

PROOF. Assume throughout that $n \ge 4$. First note that (5.7) deduces from (5.6) and from the following two inequalities that we are going to prove below:

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{h>0} R_n(\hat{p}_{n,h}, p) + \|K\|^2\frac{\log n}{n} , \tag{5.8}$$

$$\min_{j=1,\dots,M} R_m(\hat{p}_{m,h_{j-1}}, p) \le \left(1 + c_5(\log n)^{-1}\right) \inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) + \frac{c_5 L}{n\log n} . \tag{5.9}$$

In turn, (5.8) follows if we show that

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{0<h<h_0} R_n(\hat{p}_{n,h}, p), \tag{5.10}$$

$$\inf_{h\in[h_0,h_{M-1}]} R_n(\hat{p}_{n,h}, p) \le \inf_{h>h_{M-1}} R_n(\hat{p}_{n,h}, p) + \|K\|^2\frac{\log n}{n} . \tag{5.11}$$

Thus, it remains to prove (5.9) – (5.11). Recall the Fourier representation for the MISE of kernel density estimators that was obtained from Plancherel's formula in (4.11)

$$R_n(\hat{p}_{n,h},p)=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\Big(|1-\mathcal{F}[K](ht)|^2|\varphi(t)|^2+\frac{1}{n}\left(1-|\varphi(t)|^2\right)\left|\mathcal{F}[K](ht)\right|^2\Big)\mathrm{d}t. \tag{5.12}$$

Furthermore, using again Plancherel's formula we get

$$\begin{aligned}&\int_{\mathbb{R}^d}|\varphi(t)|^2\mathrm{d}t=(2\pi)^d\int_{\mathbb{R}^d}p^2(x)\mathrm{d}x\le(2\pi)^dL,\\ &\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\mathcal{F}[K](ht)|^2\mathrm{d}t=h^{-d}\|K\|^2,\ \forall\, h>0.\end{aligned} \tag{5.13}$$

Proof of (5.10). Using (5.12), (5.13) and the fact that $0\le\mathcal{F}[K](t)\le 1$, $\forall t\in\mathbb{R}^d$, for any $h<h_0=(n\log n)^{-1/d}$ we obtain

$$R_n(\hat{p}_{n,h},p)\ge\frac{1}{n(2\pi)^d}\int_{\mathbb{R}^d}\left(1-|\varphi(t)|^2\right)\left|\mathcal{F}[K](ht)\right|^2\mathrm{d}t\ge\frac{\|K\|^2}{nh^d}-\frac{L}{n}\ge\|K\|^2\log n-\frac{L}{n}. \tag{5.14}$$

On the other hand, since $h_{M-1}=1$ we get

$$R_n(\hat{p}_{n,h_{M-1}},p)\le\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left(|1-\mathcal{F}[K](t)|^2|\varphi(t)|^2+\frac{1}{n}\left|\mathcal{F}[K](t)\right|^2\right)\mathrm{d}t\le L+\frac{\|K\|^2}{n}. \tag{5.15}$$

The right hand side of (5.14) is larger than that of (5.15) for $n\ge n_0$, where n_0 depends only on L and $\|K\|$. Thus, (5.10) is valid for $n\ge n_0$.

Proof of (5.11). Clearly, (5.11) follows if we show that

$$R_n(\hat{p}_{n,h'},p)\le\inf_{h>h_{M-1}}R_n(\hat{p}_{n,h},p)+\|K\|^2\frac{\log n}{n}$$

for $h'=(\log n)^{-1/d}\in[h_0,h_{M-1}]$. To prove this inequality, first note that, by the monotonicity of $h\mapsto\mathcal{F}[K](ht)$, we have

$$\int_{\mathbb{R}^d}|1-\mathcal{F}[K](ht)|^2|\varphi(t)|^2\mathrm{d}t\ge\int_{\mathbb{R}^d}|1-\mathcal{F}[K](h't)|^2|\varphi(t)|^2\mathrm{d}t,\quad\forall h>h_{M-1}.$$

This, together with (5.12) and the second equality in (5.13), yields that, for any $h>h_{M-1}$,

$$\begin{aligned}R_n(\hat{p}_{n,h},p)&\ge R_n(\hat{p}_{n,h'},p)-\frac{1}{n(2\pi)^d}\int_{\mathbb{R}^d}\left(1-|\varphi(t)|^2\right)\left|\mathcal{F}[K](h't)\right|^2\mathrm{d}t\\ &\ge R_n(\hat{p}_{n,h'},p)-\|K\|^2\frac{\log n}{n}\,.\end{aligned}$$

Proof of (5.9). We will show that for any $h\in[h_0,h_{M-1}]$ one has

$$R_m(\hat{p}_{m,\overline{h}},p)\le\left(1+c_5(\log n)^{-1}\right)R_n(\hat{p}_{n,h},p)+\frac{c_5L}{n\log n} \tag{5.16}$$

where $\overline{h}\triangleq\max\{h_j:h_j\le h\}$. Clearly, this implies (5.9). To prove (5.16), note that if $h_j\le h<h_{j+1}$ we have $\overline{h}=h_j$, $h/h_j\le 1+a_n=1+a_0/\log n$. Therefore, (5.12) and the

monotonicity of $h \mapsto \mathcal{F}[K](ht)$ imply

$$\begin{aligned} R_m(\hat{p}_{m,h_j}, p) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left([1 - \mathcal{F}[K](h_j t)]^2 |\varphi(t)|^2 + \frac{1}{m} [\mathcal{F}[K](h_j t)]^2 \right) \mathrm{d}t \\ &\qquad - \frac{1}{(2\pi)^d m} \int_{\mathbb{R}^d} |\varphi(t)|^2 [\mathcal{F}[K](h_j t)]^2 \, \mathrm{d}t \\ &\le \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big([1 - \mathcal{F}[K](ht)]^2 |\varphi(t)|^2 + \frac{1}{n} [\mathcal{F}[K](ht)]^2 \frac{nh^d}{mh_j^d} \Big) \mathrm{d}t \\ &\qquad - \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} |\varphi(t)|^2 [\mathcal{F}[K](ht)]^2 \, \mathrm{d}t \\ &\le \frac{nh^d}{mh_j^d} R_n(\hat{p}_{n,h}, p) + \left(\frac{nh^d}{mh_j^d} - 1 \right) \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} |\varphi(t)|^2 [\mathcal{F}[K](ht)]^2 \, \mathrm{d}t. \end{aligned}$$

Using here the fact that $(n/m)(h/h_j)^d \le (1 - (\log n)^{-1} - n^{-1})(1 + a_0/\log n)^d \le 1 + c_5 (\log n)^{-1}$ for $n \ge 4$ and for a constant $c_5 > 0$ depending only on d, a_0, and applying (5.13) we get (5.16). ∎

COROLLARY 5.2. *Let the assumptions of Theorem 5.1 be satisfied, and let the condition* $\inf_{h>0} R_n(\hat{p}_{n,h}, p) \ge c n^{-1+\alpha}$ *hold for some* $c > 0, \alpha > 0$. *Then*

$$R_n(\tilde{p}_n^{\mathbb{S},K}, p) \le \inf_{h>0} R_n(\hat{p}_{n,h}, p)(1 + o(1)), \qquad n \to \infty. \tag{5.17}$$

Using the argument as in Stone (1984) it is not hard to check that the assumption of Corollary 5.2 is valid for any non-negative kernel. In the one-dimensional case it also holds for any kernel satisfying the conditions of Lemma A.6.

Theorem 5.1 and Corollary 5.2 show that linear or convex aggregate $\tilde{p}_n^{\mathbb{S},K}$ mimics the best kernel estimator, without being itself in the class of kernel estimators with data-driven bandwidth. Another method with such a property in the one-dimensional case is described in Chapter 6; it is based on a block Stein procedure in the Fourier domain.

The results of this section can be compared to the work on optimality of bandwidth selection in the L_2 sense for kernel density estimation. A key reference is the theorem of Stone (1984) establishing that, under some assumptions,

$$\lim_{n\to\infty} \frac{\|\hat{p}_{n,h_n} - p\|^2}{\inf_{h>0} \|\hat{p}_{n,h} - p\|^2} = 1, \quad \text{with probability 1,}$$

where h_n is a data-dependent bandwidth chosen by cross-validation. Our results are of a different type, because they treat convergence of the expected risk rather than almost sure convergence. In addition, we provide oracle inequalities with precisely defined remainder terms that hold under mild assumptions on the density and on the kernel. Unlike Stone (1984), we do not require the one-dimensional marginals of the density p to be uniformly bounded. Wegkamp (1999) considers a model selection approach to the choice of the bandwidth for kernel density estimation. His main result is of the form of (5.17) with a model selection kernel estimator in place of $\tilde{p}_n^{\mathbb{S},K}$, but it is valid for bounded, nonnegative, Lipschitz kernels with compact support (similar assumptions on K are imposed by Stone, 1984). Our result covers kernels with unbounded support, for example, the Gaussian and Silverman's kernels that are often implemented, and Pinsker's kernel that gives sharp minimax adaptive estimators on Sobolev classes (cf. Chapter 4). In a recent work (Dalelane, 2004, 2005b), the choice of bandwidth and of the kernel by cross-validation is investigated

for the one-dimensional case ($d = 1$). It provides an oracle inequality similar to (5.7) with a remainder term of the order $n^{\delta-1},\ 0 < \delta < 1$, instead of $(\log n)^3/n$ that we have here.

All these papers consider the model selection approach, i.e., they study estimators with a single data-driven bandwidth chosen from a set of candidate bandwidths. Our approach is different since we estimate the density by a linear or convex combination of kernel estimators with bandwidths in the candidate set. Simulations (see Section 5 below) show that in most cases one of these estimators gets highly dominant weight in the resulting mixture. However, inclusion of other estimators with some smaller weights allows one to treat more efficiently densities with inhomogeneous smoothness.

4. Sharp minimax adaptivity of kernel aggregates

In this section we show that there exist kernel density estimators such that the resulting kernel aggregate is sharp minimax adaptive over a scale of Sobolev classes of densities. The definition of such classes is given in Chapter 4 and is reproduced here for convenience.

For any $\beta > 0,\ Q > 0$ and any integer $d \geq 1$ define the Sobolev classes of densities on $\mathbb{R}^d$ by

$$\Theta(\beta, Q) \triangleq \left\{ p : \mathbb{R}^d \to \mathbb{R} \;\middle|\; p \geq 0, \int_{\mathbb{R}^d} p(x)\mathrm{d}x = 1,\ \int_{\mathbb{R}^d} \|t\|_d^{2\beta} |\varphi(t)|^2 \mathrm{d}t \leq Q \right\},$$

where $\|\cdot\|_d$ denotes the Euclidean norm in $\mathbb{R}^d$ and $\varphi = \mathcal{F}[p]$ is the characteristic function. In Chapter 4 we proved that for fixed $d = 1, \beta > 1/2$ and $Q > 0$, the Pinsker type kernel density estimator is sharp minimax over $\Theta(\beta, Q)$. Recall that the Pinsker kernel K_β is the kernel having the Fourier transform

$$\mathcal{F}[K_\beta](t) = \left(1 - \|t\|_d^\beta\right)_+, \quad t \in \mathbb{R}^d,$$

where $x_+ = \max(x, 0)$.

Corollary 5.3. *For any integer $d \geq 1$ and any $\beta > d/2,\ Q > 0$, the averaged linear or convex kernel aggregate $\tilde{p}_n^{\mathbb{S},K_\beta}$ defined in Section 3 satisfies*

$$\sup_{p \in \Theta(\beta,Q)} R_n(\tilde{p}_n^{\mathbb{S},K_\beta}, p) \leq C^* n^{-\frac{2\beta}{2\beta+d}}(1 + o(1)), \qquad n \to \infty,$$

where C^* is defined in (4.3).

Proof. Denote by $\tilde{p}_{n,h}$ the Pinsker type kernel density estimator defined in (5.5) with $m = n$, $K = K_\beta$ and bandwidth $h > 0$. From Theorem 4.2, it holds

$$\inf_{h>0} R_n(\tilde{p}_{n,h}, p) \leq R_n(\tilde{p}_{n,h^*}, p) \leq C^* n^{-\frac{2\beta}{2\beta+d}}, \quad \forall\ p \in \Theta(\beta, Q)\,, \tag{5.18}$$

where h^* is defined in (4.10). Note that the kernel $K = K_\beta$ satisfies the conditions of Theorem 5.1, and it is easy to see that for $\beta > d/2$ there exists a constant $0 < L < \infty$ such that $\|p\|_\infty \leq L$ for all $p \in \Theta(\beta, Q)$. Thus, (5.7) holds, and to prove the corollary it suffices to take suprema of both sides of (5.7) over $p \in \Theta(\beta, Q)$ and to use (5.18). ∎

Along with Corollary 5.3, for any $\beta > d/2,\ Q > 0$ the following lower bound holds:

$$\inf_{T_n} \sup_{p \in \Theta(\beta,Q)} R_n(T_n, p) \geq C^* n^{-\frac{2\beta}{2\beta+d}}(1 + o(1)), \qquad n \to \infty, \tag{5.19}$$

where C^* is defined in (4.3) and $\inf_{T_n}$ denotes the infimum over all estimators of p. For $d = 1$, the lower bound (5.19) is proved in Theorem 4.1. For $d > 1$ a proof of (5.19) can be

found for a slightly different but essentially analogous minimax setup in Efromovich (2000). Corollary 5.3 and the lower bound (5.19) imply that the estimator $\tilde{p}_n^{\mathbb{S},K_\beta}$ is asymptotically minimax in the exact sense (with the constant) over the Sobolev class of densities $\Theta(\beta, Q)$ and is adaptive to Q for any given β. However, $\tilde{p}_n^{\mathbb{S},K_\beta}$ is not adaptive to the unknown smoothness β since the Pinsker kernel K_β depends on β

To get adaptation to β, we need to push aggregation one step forward: we will aggregate kernel density estimators not only for different bandwidths but also for different kernels. To this end, we refine the notation $\hat{p}_{n,h}$ of (5.5) to $\hat{p}_{n,h,K}$, indicating the dependence of the density estimator both on kernel K and bandwidth h. For a family of $N \geq 2$ kernels, $\mathcal{K} = \{K_{(1)}, \dots, K_{(N)}\}$, define $\tilde{p}_n^{\mathbb{S},\mathcal{K}}$ as the linear or convex averaged aggregate where the initial estimators are taken in the collection of kernel density estimators $\{\hat{p}_{n,h,K}, K \in \mathcal{K}, h \in \mathcal{H}\}$. Thus, we aggregate now NM estimators instead of M. The following corollary is obtained by the same argument as Theorem 5.1, by merely inserting the minimum over $K \in \mathcal{K}$ in the oracle inequality and by replacing $\|K\|$ with its upper or lower bounds in the remainder terms.

COROLLARY 5.4. *Assume that p satisfies $\|p\|_\infty \leq L$ with $0 < L < \infty$ and let $\mathcal{K} = \{K_{(1)}, \dots, K_{(N)}\}$ be a family of kernels satisfying the assumptions of Theorem 5.1 and such that there exist constants $0 < \underline{c} < \bar{c} < \infty$ with $\underline{c} < \|K_{(j)}\| < \bar{c}$, $j = 1, \dots, N$. Then there exists an integer $n_1 = n_1(L, \underline{c}, \bar{c}) \geq 4$ such that for $n \geq n_1$, the averaged aggregate $\tilde{p}_n^{\mathbb{S},\mathcal{K}}$ satisfies the oracle inequality*

$$R_n(\tilde{p}_n^{\mathbb{S},\mathcal{K}}, p) \leq \left(1 + c_5(\log n)^{-1}\right) \min_{K \in \mathcal{K}} \inf_{h>0} R_n(\hat{p}_{n,h,K}, p) + c_7 \frac{N(\log n)^3}{n}, \tag{5.20}$$

where $c_5 > 0$ is the same constant as in Theorem 5.1, and $c_7 > 0$ depends only on $L, \underline{c}, \bar{c}, d$ and a_0.

Consider now a particular family of kernels $\mathcal{K}$. Define $\mathcal{B} = \{\beta_1, \dots, \beta_N\}$ where $\beta_1 = d/2$, $\beta_j = \beta_{j-1} + N^{-1/2}$, $j = 2, \dots, N$, and let $\mathcal{K}_\mathcal{B} = \{K_b,\ b \in \mathcal{B}\}$ be a family of Pinsker kernels indexed by $b \in \mathcal{B}$. We will later assume that $N = N_n \to \infty$, as $n \to \infty$, but for the moment assume that $N \geq 2$ is fixed. Note that $\mathcal{K} = \mathcal{K}_\mathcal{B}$ satisfies the assumptions of Corollary 5.4. In fact,

$$\|K_\beta\|^2 = S_d Q_d(\beta) \quad \text{where} \quad Q_d(\beta) = \frac{1}{d} - \frac{2}{\beta + d} + \frac{1}{2\beta + d},$$

and

$$\frac{1}{6d} \leq Q_d(\beta) \leq \frac{1}{d}, \quad \forall\ \beta \geq d/2. \tag{5.21}$$

Thus, the oracle inequality (5.20) holds with $\mathcal{K} = \mathcal{K}_\mathcal{B}$. We will now prove that, under the assumptions of Corollary 5.4 the linear or convex aggregate $\tilde{p}_n^{\mathbb{S},\mathcal{K}_\mathcal{B}}$ with the initial estimators in $\{\hat{p}_{n,h,K}, K \in \mathcal{K}_\mathcal{B}, h \in \mathcal{H}\}$ satisfies the following inequality where β in the oracle risk varies continuously:

$$R_n(\tilde{p}_n^{\mathbb{S},\mathcal{K}_\mathcal{B}}, p) \leq \left(1 + \frac{c_5}{\log n}\right)\left(1 + \frac{6}{\sqrt{N}}\right) \inf_{\substack{h>0 \\ d/2 < \beta < \beta_N}} R_n(\hat{p}_{n,h,K_\beta}, p) + c_8 \frac{N(\log n)^3}{n}. \tag{5.22}$$

Fix $\beta \in (d/2, \beta_N), Q > 0$ and $p \in \Theta(\beta, Q)$. Define $\bar{\beta} = \min\{\beta_j \in \mathcal{B} : \beta_j > \beta\}$. In view of (5.20) with $\mathcal{K} = \mathcal{K}_\mathcal{B}$, to prove (5.22) it is sufficient to show that for any $h > 0$ one has

$$R_n(\hat{p}_{n,h,K_{\bar{\beta}}}, p) \leq (1 + 6N^{-1/2}) \left(R_n(\hat{p}_{n,h,K_\beta}, p) + \frac{L}{n}\right). \tag{5.23}$$

Using (5.12) and the inequality $\bar{\beta} > \beta$ we get

$$R_n(\hat{p}_{n,h,K_{\bar{\beta}}}, p) \le R_n(\hat{p}_{n,h,K_\beta}, p) + \mathcal{I}(\bar{\beta}) - \mathcal{I}(\beta) \tag{5.24}$$

where

$$\mathcal{I}(\beta) \triangleq \frac{1}{(2\pi)^d n} \int_{\mathbb{R}^d} (1 - \|ht\|_d^\beta)_+^2 \mathrm{d}t = \frac{\|K_\beta\|^2}{(2\pi)^d n h^d} = \frac{S_d}{(2\pi)^d n h^d} Q_d(\beta).$$

Now, $Q_d(\bar{\beta}) = Q_d(\beta) + (\bar{\beta} - \beta) Q_d'(b_0)$ for some $b_0 \in [\beta, \bar{\beta}]$. Using (5.21) and the inequality $|Q_d'(\beta)| \le 1/d^2$ valid for all $\beta > d/2$, we find that

$$Q_d(\bar{\beta}) \le Q_d(\beta) + 6(\bar{\beta} - \beta) Q_d(\beta) \le (1 + 6N^{-1/2}) Q_d(\beta)\,.$$

Therefore,

$$\mathcal{I}(\bar{\beta}) \le (1 + 6N^{-1/2}) \mathcal{I}(\beta). \tag{5.25}$$

Also, in view of (5.12) and (5.13) we have

$$\mathcal{I}(\beta) \le R_n(\hat{p}_{n,h,K_\beta}, p) + \frac{L}{n}. \tag{5.26}$$

Combining (5.24), (5.25) and (5.26) we obtain (5.23), thus proving (5.22).

COROLLARY 5.5. *Assume that* $\mathrm{Card}(\mathcal{K}_\mathcal{B}) = N_n$ *where* $\lim_{n\to\infty} N_n = \infty$ *and that* $\limsup_{n\to\infty} N_n/(\log n)^\nu < \infty$ *for some* $\nu > 0$. *Then for any integer* $d \ge 1$ *and any* $\beta > d/2$, $Q > 0$, *the averaged linear or convex kernel aggregate* $\tilde{p}_n^{\mathbb{S},\mathcal{K}_\mathcal{B}}$ *satisfies*

$$\sup_{p \in \Theta(\beta,Q)} R_n(\tilde{p}_n^{\mathbb{S},\mathcal{K}_\mathcal{B}}, p) \le C^* n^{-\frac{2\beta}{2\beta+d}} (1 + o(1)), \qquad n \to \infty,$$

where C^* *is defined in* (4.3).

PROOF. Fix $\beta > d/2, Q > 0$. Let n be large enough to guarantee that $\beta < \beta_{N_n}$. Then the infimum on the right hand side of (5.22) is smaller or equal to $C^* n^{-\frac{2\beta}{2\beta+d}}$ for all $p \in \Theta(\beta, Q)$ [cf. Theorem 4.2]. To conclude the proof, it suffices to take suprema of both sides of (5.22) over $p \in \Theta(\beta, Q)$ and then pass to the limit as $n \to \infty$. ∎

Corollary 5.5 and the lower bound (5.19) imply that the aggregate $\tilde{p}_n^{\mathbb{S},\mathcal{K}_\mathcal{B}}$ is asymptotically minimax in the exact sense (with the constant) over all Sobolev classes of densities with $\beta > d/2$, $Q > 0$, and thus it is sharp adaptive (recall that its construction does not depend on the parameters Q and β of the class).

5. Simulations

Here we discuss the results of simulations for the averaged convex kernel aggregate with $H = \Lambda^M$ in the one-dimensional case. We focus on convex aggregation because simulations of linear aggregates show less numerical stability. The set of splits $\mathbb{S}$ is reduced to 10 random splits of the sample since we observed that the estimator is already stable for this number (cf. Figure 4). In the default simulations each sample is divided into two subsamples of equal sizes. The samples are drawn from 6 densities that can be classified in the following three groups.

- Common reference densities: the standard Gaussian density and the standard exponential density.
- Gaussian mixtures from Marron and Wand (1992) that are known to be difficult to estimate. We consider the Claw density and the Smooth Comb density.

- Densities with highly inhomogeneous smoothness. We consider two densities referenced to as dens1 and dens2 that are both mixtures of the standard Gaussian density $\varphi(\cdot)$ and of an oscillating density. They are defined as

$$0.5\varphi(\cdot) + 0.5\sum_{i=1}^{T} \mathbb{I}_{\left(\frac{2(i-1)}{T}, \frac{2i-1}{T}\right]}(\cdot)\,,$$

  where $T = 14$ for dens1 and $T = 10$ for dens2.

We used the procedure defined in Section 3 to aggregate 6 kernel density estimators constructed with the Gaussian $\mathcal{N}(0,1)$ kernel K and with bandwidths h from the set $\mathcal{H} = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. This procedure is further called *pure kernel aggregation* and quoted as `AggPure`. Another estimator that we analyze is `AggStein` procedure: it aggregates 7 estimators, namely the same 6 kernel estimators as for `AggPure` to which we add the block Stein density estimator described in Chapter 6. The optimization problem (3.6) that provides aggregates is solved numerically by a quadratic programming solver under linear constraints: here we used the package `quadprog` of R. Our simulation study shows that `AggPure` and `AggStein` have a good performance for moderate sample sizes and are reasonable competitors to kernel density estimators with common bandwidth selectors.

We start the simulation by a comparison of the Monte-Carlo mean integrated squared squared error (MISE) of `AggPure` and `AggStein` with benchmarks. The MISE has been computed by averaging integrated squared errors of 200 aggregate estimators calculated from different samples of size 50, 100, 200 and 500. We compared the performance of the convex aggregates and kernel estimators with common data-driven bandwidth selectors and Gaussian $\mathcal{N}(0,1)$ kernel. The following bandwidth selectors are taken from the default package `stats` of the R software.

- DPI that implements the direct plug-in method of Sheather and Jones (1991) to select the bandwidth using pilot estimation of derivatives.
- UCV and BCV that implement unbiased and biased cross-validation respectively (see, e.g., Wand and Jones, 1995).
- Nrd0 that implements Silverman's rule-of-thumb (cf. Silverman, 1986, p. 48). It defaults the choice of bandwidth to 0.9 times the minimum of the standard deviation and the inter-quartile range divided by 1.34 times the sample size to the negative one-fifth power.

These descriptions correspond to the function `bandwidth` in R which also allows for another choice of rule-of-thumb called `Nrd`. It is a modification of `Nrd0` given by Scott (1992), using factor 1.06 instead of 0.9. In our case, on the tested densities and sample sizes, this always leads to a MISE greater than that of `Nrd0` except for the Gaussian density for which it is tailored. For this density, the performance of `Nrd` is presented instead of that of `Nrd0`.

The results are reported in Tables 1 to 5 where we included also the MISE of the block Stein density estimator described in Chapter 6 and the oracle risk which is defined as the minimum MISE of kernel density estimators over the grid $\mathcal{H}$. It is, in general, greater than the convex oracle risk, that is why it sometimes slightly exceeds the MISE of convex aggregates or of other estimators that mimic more powerful oracles for specific densities (such as DPI or `Nrd` for the Gaussian density).

It is well known (see, e.g., Wand and Jones, 1995) that bandwidth selection by cross-validation (UCV) is unstable and leads too often to undersmoothing. The DPI and BCV methods were proposed in order to bypass the problem of undersmoothing. However,

	50	100	200	500	50	100	200	500
AggPure	0.020	0.011	0.006	0.002	0.084	0.057	0.039	0.025
AggStein	0.017	0.009	0.005	0.002	0.085	0.057	0.039	0.025
Stein	0.016	0.010	0.005	0.003	0.073	0.056	0.041	0.027
DPI	0.011	0.006	0.004	0.002	0.075	0.060	0.045	0.033
UCV	0.015	0.008	0.005	0.002	0.072	0.052	0.038	0.023
BCV	0.009	0.006	0.003	0.002	0.108	0.083	0.058	0.036
Nrd	0.010	0.006	0.003	0.002	0.085	0.072	0.061	0.051
Oracle	0.008	0.005	0.004	0.003	0.067	0.047	0.035	0.022

TABLE 1. *MISE for the Gaussian (left) and the exponential (right) densities*

	50	100	200	500	50	100	200	500
AggPure	0.058	0.041	0.029	0.014	0.064	0.042	0.029	0.017
AggStein	0.056	0.041	0.025	0.010	0.061	0.042	0.028	0.017
Stein	0.061	0.035	0.018	0.009	0.057	0.041	0.028	0.017
DPI	0.059	0.052	0.048	0.043	0.070	0.054	0.042	0.029
UCV	0.063	0.043	0.026	0.012	0.057	0.038	0.026	0.016
BCV	0.058	0.052	0.050	0.046	0.101	0.083	0.055	0.027
Nrd0	0.058	0.051	0.048	0.043	0.088	0.078	0.069	0.057
Oracle	0.058	0.037	0.025	0.012	0.064	0.038	0.025	0.016

TABLE 2. *MISE for the claw (left) and the smooth comb (right) densities*

	50	100	200	500	50	100	200	500
AggPure	0.145	0.125	0.100	0.067	0.142	0.119	0.093	0.061
AggStein	0.148	0.124	0.102	0.067	0.148	0.141	0.092	0.060
Stein	0.152	0.143	0.138	0.132	0.154	0.143	0.137	0.132
DPI	0.149	0.142	0.137	0.132	0.147	0.140	0.136	0.132
UCV	0.153	0.148	0.136	0.116	0.154	0.142	0.126	0.074
BCV	0.149	0.143	0.139	0.134	0.146	0.141	0.138	0.134
Nrd0	0.149	0.141	0.137	0.133	0.146	0.140	0.136	0.132
Oracle	0.148	0.144	0.133	0.067	0.145	0.128	0.101	0.062

TABLE 3. *MISE for dens1 (left) and dens2 (right)*

sometimes they lead to oversmoothing as in the case of the Claw density while convex aggregation works well. For the normal density DPI, BCV and Nrd are better, which comes as no surprise since these estimators are designed to estimate this density well. For the other densities that are more difficult to estimate these data driven bandwidth selectors do not provide good estimators whereas the aggregation procedures remain stable. The block Stein estimator performs well in all the cases except for the highly inhomogeneous densities (cf. Table 5). In conclusion, the estimators `AggPure` and `AggStein` are very robust, as compared to other tested procedures: they are not far from the best performance

for the densities that are easy to estimate and they are clear winners for densities with inhomogeneous smoothness for which other procedures fail.

`AggStein` is slightly better than `AggPure` for the Claw density and outperforms the other tested estimators in almost all the considered cases, so we studied this procedure in more detail. We focused on the Claw and Smooth Comb densities and a sample of size 500. Figure 2 gives a visual comparison of the `AggStein` procedure and the DPI procedure. It

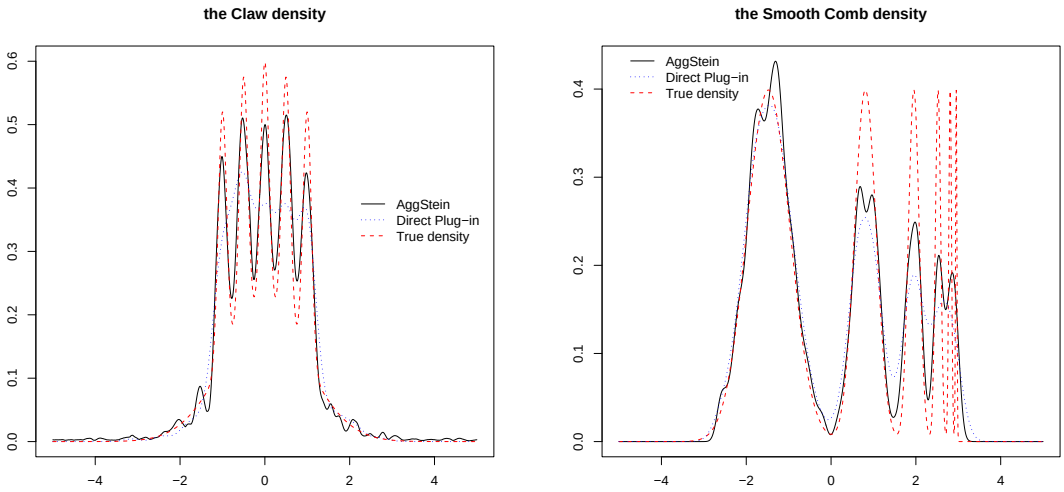

FIGURE 2. The Claw and Smooth Comb densities

illustrates the oversmoothing effect of the DPI procedure and the fact that the `AggStein` procedure adapts to inhomogeneous smoothness. We finally comment on two other aspects of the `AggStein` procedure:

- the distribution of weights that are allocated to the aggregated estimators,
- the robustness to the number and size of the splits.

The boxplots represented in Figure 3 give the distributions of weights allocated to 7 estimators to be aggregated, the 6 kernel density estimators and the block Stein estimator. The boxplots are constructed from 2000 values of the vector of the weights (200 samples times 10 splits). We immediately notice that for the Claw density a median weight greater than 0.65 is allocated to the block Stein estimator. This can be explained by the fact that the block Stein estimator performs better than kernel density estimators on this density [cf. MISE of `AggPure` and Stein in Table 2 (left)], and the `AggStein` procedure takes advantage of it. On the other hand, for the Smooth Comb density, the block Stein estimator does not perform significantly better than the kernel density estimators [see Table 2 (right)] and the `AggStein` procedure usually gives small weight to it.

A free parameter of the aggregation procedures is the set of splits. In this study we choose random splits and we only have to specify their number and sizes. Obviously, we are interested in having less splits in order to make the procedure less time consuming. Figure 4 gives the sensibility of MISE both to the number of splits and to the size of the training sample in the case of dens1 and dens2 with the overall sample size 200. Two important conclusions are: (i) there exists a size of the training sample that achieves the minimum MISE, and (ii) there is essentially nothing to gain by producing more than 20

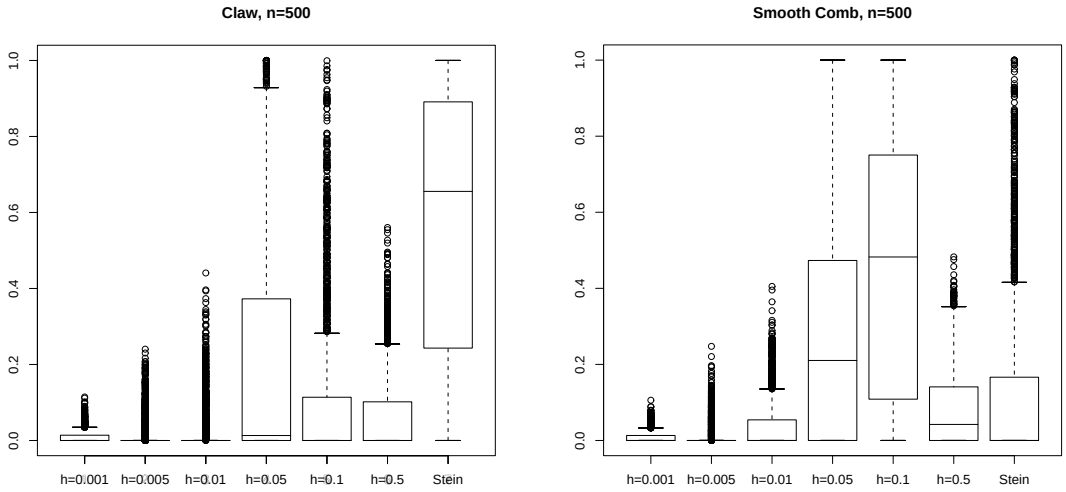

FIGURE 3. Boxplots for the Claw and Smooth Comb densities

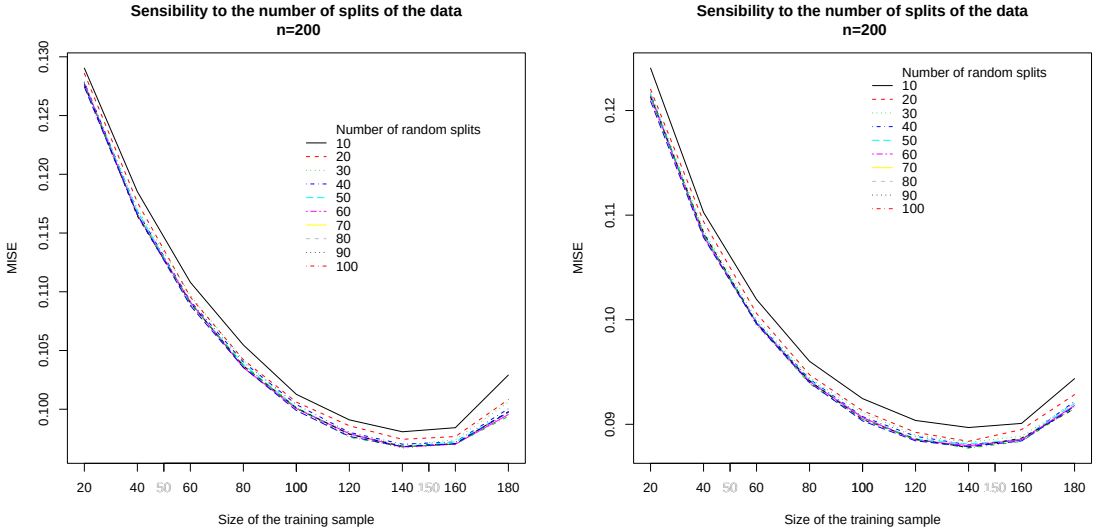

FIGURE 4. Sensibility to the number of splits for dens1 (left) and dens2 (right)

splits. Similar results are obtained for `AggPure`, and they are valid on the whole set of tested densities.

CHAPTER 6

Adaptive density estimation using the blockwise Stein method

We study the problem of nonparametric estimation of a probability density of unknown smoothness in $L_2(\mathbb{R})$. Expressing mean integrated squared error (MISE) in the Fourier domain, we show that it is close to mean squared error in the Gaussian sequence model. Then applying a modified version of Stein's blockwise method, we obtain a linear monotone oracle inequality. Two consequences of this oracle inequality are that the proposed estimator is sharp minimax adaptive on a scale of Sobolev classes of densities, and that its MISE is asymptotically smaller than or equal to that of kernel density estimators with any bandwidth provided that the kernel belongs to a large class of functions including many standard kernels. Simulations in classical cases are also provided and confirm a good performance of the estimator for a finite number of observations.

Contents

Most of the material in this chapter has been published (see Rigollet, 2006a). As compared to this publication, we give more detailed proofs and add a simulation study (see Rigollet, 2004).

1. Introduction

A Stein weakly geometrically increasing (WGI) blockwise shrinkage estimator, employing a classical Stein blockwise shrinkage together with WGI blocks, has recently been proposed and studied for the Gaussian white noise model in Cavalier and Tsybakov (2001)

and Tsybakov (2002). It has been established that the estimator possesses several very nice statistical properties. This chapter suggests a Stein WGI estimator for the problem of probability density estimation and then studies its properties via an oracle inequality. It also shows how to use an oracle inequality to obtain Stone type results for kernel estimates.

1.1. Setup. Consider independent and identically distributed random variables $X_1, \ldots, X_n$ having an unknown common probability density $p \in L_2(\mathbb{R})$. We study the estimation of p based on the sample $\mathbb{X}^n = (X_1, \ldots, X_n)$. Let $\hat{p}_n$ be an estimator of p. We measure the performance of $\hat{p}_n$ by its mean integrated squared error (MISE):

$$E_p^n \left\| \hat{p}_n - p \right\|^2 = E_p^n \int_{\mathbb{R}} \left(\hat{p}_n(x) - p(x) \right)^2 \mathrm{d}x,$$

where E_p^n denotes the expectation with respect to $\mathbb{X}^n$.

Define the characteristic function

$$\varphi(\omega) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\omega x} p(x) \mathrm{d}x, \quad \omega \in \mathbb{R}.$$

The empirical characteristic function (e.c.f.) is

$$\varphi_n(\omega) = \frac{1}{n} \sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}\omega X_k}, \quad \omega \in \mathbb{R}.$$

The e.c.f. is a fairly good estimator of φ. Indeed, it satisfies the following properties:

$$\begin{aligned} &(i) \qquad E_p^n \left[\varphi_n(\omega) \right] = \varphi(\omega) \\ &(ii) \qquad E_p^n \left| \varphi_n(\omega) \right|^2 = \left(1 - \frac{1}{n} \right) \left| \varphi(\omega) \right|^2 + \frac{1}{n}, \end{aligned} \tag{6.1}$$

i.e., it is an unbiased and $\sqrt{n}$-consistent estimator of $\varphi(t)$ for any fixed $t \in \mathbb{R}$.

For any function $h \in L_2(\mathbb{R})$, let $\omega \mapsto \mathcal{F}[h](\omega) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\omega x} h(x) \mathrm{d}x$ be its *Fourier transform* (the integral is understood in the "limit in mean" sense). A well-known estimation method uses a *kernel density estimator* or *Parzen-Rosenblatt's estimator* defined as

$$\hat{p}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h} \right), \tag{6.2}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an even integrable function such that $\int K(u) \mathrm{d}u = 1$ and $h > 0$ is the bandwidth parameter. Then, the Fourier transform of a kernel estimator defined in (6.2) presented in terms of its e.c.f. is:

$$\mathcal{F}[\hat{p}_{n,h}](\omega) = \varphi_n(\omega) \mathcal{F}[K](h\omega). \tag{6.3}$$

Denote by K_h the normalized kernel

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h} \right), \quad x \in \mathbb{R}.$$

Using the Plancherel equality, one easily gets

$$\begin{aligned} E_p^n \left\| \hat{p}_{n,h} - p \right\|^2 &= \frac{1}{2\pi} \int_{\mathbb{R}} \left(|1 - \mathcal{F}[K](h\omega)|^2 |\varphi(\omega)|^2 + \frac{1}{n} |\mathcal{F}[K](h\omega)|^2 \right) \\ &\qquad - \frac{1}{n} |\varphi(\omega)|^2 \left| \mathcal{F}[K](h\omega) \right|^2 \mathrm{d}\omega \\ &= \frac{1}{2\pi} \Big(\Delta_n \left(\mathcal{F}[K_h], |\varphi|^2 \right) - \overline{r}_n \left(\mathcal{F}[K_h], |\varphi|^2 \right) \Big), \end{aligned}$$

where

$$\Delta_n\left(g,|\varphi|^2\right)=\int_{\mathbb{R}}\left(|1-g(\omega)|^2|\varphi(\omega)|^2+\frac{1}{n}|g(\omega)|^2\right)\mathrm{d}\omega,\qquad\forall g\in L_2(\mathbb{R}) \tag{6.4}$$

and

$$\overline{r}_n\left(g,|\varphi|^2\right)=\frac{1}{n}\int_{\mathbb{R}}|\varphi(\omega)|^2|g(\omega)|^2\ \mathrm{d}\omega,\qquad\forall g\in L_2(\mathbb{R}).$$

Remark that (6.3) can be generalized to a linear estimator of the characteristic function φ defined by

$$\hat{\varphi}_\lambda(\omega)=\varphi_n(\omega)\lambda(\omega), \tag{6.5}$$

where $\omega\mapsto\lambda(\omega)$ is a *weight function* in $L_2(\mathbb{R})$. We define a density estimator $\hat{p}_\lambda$ as the inverse Fourier transform of $\hat{\varphi}_\lambda$. The performance of this new estimator is measured by its MISE, which, by the Plancherel equality can be written

$$E_p^n\left\|\hat{p}_\lambda-p\right\|^2=\frac{1}{2\pi}E_p^n\int_{\mathbb{R}}\left|\hat{\varphi}_\lambda(\omega)-\varphi(\omega)\right|^2\mathrm{d}\omega=\frac{1}{2\pi}R_n\left(\hat{\varphi}_\lambda,\varphi\right).$$

Thus, instead of considering the MISE of $\hat{p}_\lambda$, it is sufficient to study the MISE of the estimator of the characteristic function $\hat{\varphi}_\lambda$. We may write

$$\begin{aligned}R_n\left(\hat{\varphi}_\lambda,\varphi\right)&=\int_{\mathbb{R}}\left(|1-\lambda(\omega)|^2|\varphi(\omega)|^2+\frac{1}{n}|\lambda(\omega)|^2\right)\mathrm{d}\omega-\frac{1}{n}\int_{\mathbb{R}}|\varphi(\omega)|^2|\lambda(\omega)|^2\mathrm{d}\omega\\&=\Delta_n\left(\lambda,|\varphi|^2\right)-\overline{r}_n\left(\lambda,|\varphi|^2\right)\end{aligned}$$

where λ is such that the integrals are finite.

Lemma 6.1. *The MISE of the linear estimator defined in* (6.5) *is given by*

$$R_n\left(\hat{\varphi}_\lambda,\varphi\right)=\int_{\mathbb{R}}\left(|1-\lambda(\omega)|^2|\varphi(\omega)|^2+\frac{1}{n}|\lambda(\omega)|^2\right)\mathrm{d}\omega-\frac{1}{n}\int_{\mathbb{R}}|\varphi(\omega)|^2|\lambda(\omega)|^2\mathrm{d}\omega.$$

The MISE of the linear estimator is the sum of two terms: the Δ-*risk*, $\Delta_n\left(\lambda,|\varphi|^2\right)$ and a residual term $\overline{r}_n\left(\lambda,|\varphi|^2\right)$. Clearly $\overline{r}_n\left(\lambda,|\varphi|^2\right)=O\left(n^{-1}\right)$ uniformly in λ and φ such that the integral is finite and therefore, $\overline{r}_n$ is usually small compared to Δ_n. This suggests that, for sufficiently large n, the linear oracle on a class $\mathcal{H}$ can be imitated by

$$\lambda_{\mathcal{H}}^{\text{oracle}}=\operatorname*{argmin}_{\lambda\in\mathcal{H}}R_n\left(\hat{\varphi}_\lambda,\varphi\right)\approx\operatorname*{argmin}_{\lambda\in\mathcal{H}}\left[\Delta_n\left(\lambda,|\varphi|^2\right)\right].$$

The study of kernel density estimation by means of Fourier analysis goes back to Parzen (1958, 1962) and Watson and Leadbetter (1963). Cline (1988) uses Watson-Leadbetter's results to prove that *admissible* kernels have necessarily symmetric non-negative Fourier transform. The expression of the Δ-risk in (6.4) is similar to that for the mean squared error of a linear estimator in the Gaussian sequence model. We now recall some basics about the Gaussian sequence model that give insights on the choice of the class $\mathcal{H}$.

1.2. The Gaussian sequence model: a brief overview. Consider the Gaussian sequence model

$$y_k=\theta_k+\varepsilon\xi_k,\quad k=1,2,\ldots, \tag{6.6}$$

where y_k are the observations, ξ_k are independent and identically distributed random variables with distribution $\mathcal{N}(0,1)$, $0<\varepsilon<1$, and $\theta=(\theta_1,\theta_2,\ldots)\in\ell_2$ is the parameter to be estimated. Consider then the class of linear estimators:

$$\hat{\theta}=\hat{\theta}(h)=(\hat{\theta}_1,\hat{\theta}_2,\ldots),\ \hat{\theta}_k=h_ky_k,\quad k=1,2,\ldots,$$

where $h = (h_1, h_2, \ldots) \in \ell_2$ is an arbitrary sequence of weights in ℓ_2. The mean squared error of the linear estimator is:

$$R_\varepsilon(h, \theta) = \mathbb{E}_\theta \left\| \hat{\theta}(h) - \theta \right\|_2^2 = \sum_{k=1}^{\infty} \left\{ (1 - h_k)^2 \theta_k^2 + \varepsilon^2 h_k^2 \right\}. \tag{6.7}$$

Here $\| \, . \, \|_2$ is the ℓ_2-norm and $\mathbb{E}_\theta$ denotes the expectation with respect to $y = (y_1, y_2, \ldots)$ satisfying (6.6).

For a class of weights $\mathcal{H}$, let us define the *linear oracle on* $\mathcal{H}$.

DEFINITION 6.1. *Let $\mathcal{H}$ be a class of weights, $\Theta \subseteq \ell_2$ a set such that for any $\theta \in \Theta$ there exists $h^{\mathcal{H}}(\theta)$ satisfying*

$$h^{\mathcal{H}}(\theta) = \operatorname*{argmin}_{h \in \mathcal{H}} R_\varepsilon(h, \theta).$$

The function $\theta \mapsto h^{\mathcal{H}}(\theta)$ defined on Θ is called **linear oracle on $\mathcal{H}$**.

Assume that, for a given class $\mathcal{H}$, one can find a sequence of data-dependent weights $h^* = h^*(y) = (h_1^*(y), h_2^*(y), \ldots)$ such that the linear estimator $\theta^* = \hat{\theta}(h^*)$ satisfies

$$\mathbb{E}_\theta \left\| \hat{\theta}(h^*) - \theta \right\|_2^2 \le C \, (1 + o(1)) \inf_{h \in \mathcal{H}} \mathbb{E}_\theta \left\| \hat{\theta}(h) - \theta \right\|_2^2, \qquad \varepsilon \to 0 \tag{6.8}$$

for every $\theta \in \Theta \subseteq \ell_2$, where C is a positive constant. When $C = 1$, inequality (6.8) means that θ^* "mimics" the linear oracle on $\mathcal{H}$ for any sequence $\theta \in \Theta$. In this case, inequality (6.8) is called *exact oracle inequality*. When $C > 1$, the estimator θ^* only mimics the rate of convergence of the linear oracle and inequality (6.8) is simply called *oracle inequality* or *approximate oracle inequality*.

The Gaussian sequence model has been studied by many authors in the past decades and oracle inequalities have been widely used, although initially in an implicit form, to derive adaptation (see Shibata, 1981; Efroĭmovich and Pinsker, 1984; Li, 1987; Golubev, 1990, 1992; Golubev and Nussbaum, 1992; Polyak and Tsybakov, 1992; Kneip, 1994; Birgé and Massart, 2001; Cavalier *et al.*, 2002). Most of these papers use the Mallows C_p or its modifications to derive estimators that mimic the best estimator in various subclasses of linear estimators (i.e. the oracle).

A well-known idea of choosing the data-driven weights $h_k^* = h_k^*(y)$ is based on minimizing the unbiased estimator of the risk which is also called C_p-criterion (Mallows, 1973; Akaike, 1973; Stein, 1981) over a particular class of weights $\mathcal{H}$. The class of blockwise constant weights turns out to be a convenient choice (see, e.g., Tsybakov, 2004a). We now define this class.

Let $N \ge 2$ be a fixed integer. Consider a partition of $\{1, \ldots, N\}$ in J blocs B_j, $j = 1, 2, \ldots, J$, satisfying $\min\{k \in B_j\} > \max\{k \in B_{j-1}\}$.

The *blockwise Stein method* is obtained by minimizing an unbiased estimator of the risk over the class of blockwise constant sequences

$$\mathcal{H}^* = \left\{ h \ : \ h_k = \sum_{j=1}^{J} t_j \mathbb{1}_{B_j}(k), \ 0 \le t_j \le 1, \ j = 1, 2, \ldots, J \right\},$$

where $\mathbb{1}_{B_j}$ denotes the indicator of the set B_j. The minimum is obtained (see Tsybakov, 2004a) for

$$\tilde{\theta}_k = \begin{cases} \left(1 - \frac{\varepsilon^2 |B_j|}{\|y\|^2_{(j)}}\right)_+ y_k, & k \in B_j,\ j = 1, 2, \dots, J, \\ 0, & k > N, \end{cases} \tag{6.9}$$

where $|B_j|$ denotes the cardinality of the block B_j and $\|y\|^2_{(j)} = \sum_{k \in B_j} y_k^2$.

DEFINITION 6.2. *The estimator $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \dots)$, where $\tilde{\theta}_k$ is defined by* (6.9)*, is called* ***blockwise Stein estimator*** *or* ***block Stein estimator****.*

Blockwise constant weights show particularly good statistical properties and have been widely discussed in the statistical literature, first by Efroĭmovich and Pinsker (1984), and more recently by Nemirovski (2000) and Efromovich (2004) who consider block estimators different from Stein's one. Tsybakov (2004a) considers Stein's estimator with a particular system of blocks, namely, the WGI blocks. Cavalier and Tsybakov (2001) improve the previous results by using a penalized version of the block Stein estimator.

As for the wavelet framework, the subject has been discussed by Donoho and Johnstone (1994, 1995) and Härdle *et al.* (1998). In the same setting, Cai (1999) and Efromovich (2000) use block thresholding type estimators that both satisfy oracle inequalities within the class of blockwise linear estimators. These estimators exhibit good performance in simulations.

Goldenshluger and Tsybakov (2001) apply the block Stein estimator with WGI blocks to the Gaussian regression problem with infinitely many parameters. They show that it is sharp minimax adaptive over a scale of Sobolev ellipsoids in ℓ_2. Tsybakov (2002) discusses in particular the anisotropic multidimensional white noise model. He shows that the block Stein estimator, again with WGI blocks, is adaptive simultaneously with respect to the real dimension, direction and smoothness of the parameter over a scale of Sobolev ellipsoids. In both papers, adaptation is derived from oracle inequalities.

Using the similarity between (6.7) and (6.4), it seems natural to extend the results for the Gaussian sequence model discussed in Section 1.2 to nonparametric density estimation. The similarity between kernel density estimation and the Gaussian sequence model based on Fourier analysis have been examined by Golubev (1992) and later by Boĭko and Golubev (2000). Golubev and Levit (1996) consider the problem of the second-order minimax adaptive estimation of an unknown distribution function over Sobolev ellipsoids. They develop techniques that are also useful for the density estimation problem considered here.

Whereas the Gaussian sequence model has been extensively studied, there are few results concerning blockwise density estimation. A discussion of blockwise density estimates can be found in Efroĭmovich (1985) and Efromovich (2000), where the Efromovich-Pinsker shrinkage procedure together with polynomial blocks is explored, and in Hall *et al.* (1998), where a block-thresholding shrinkage procedure employing small logarithmic blocks is explored.

Cavalier and Tsybakov (2001) obtained powerful oracle inequalities for penalized Stein estimates in the context of the Gaussian sequence model. Under certain hypotheses, they lead to adaptive properties in the minimax sense - in particular, over any ellipsoid in ℓ_2 with monotone decreasing coefficients.

This chapter is devoted to developing a Stein WGI estimator for a density estimation setting. The estimator employs a classical Stein blockwise shrinkage which uses a zero thresholding (imposes no penalty). Let us recall that this shrinkage procedure has been

very attractive for filtering problems; see the discussion in Tsybakov (2002). The WGI blocks employed were also recommended in Tsybakov (2002).

The primary complication in the development and study of a density estimate, based on a known analogue for Gaussian sequence models, consists in the fact that the observations are not Gaussian and precise results such as Stein's lemma do not apply. In Section 3, we use unbiased estimation of the risk to derive several oracle inequalities for the proposed estimator. We then give two corollaries of these results. First, in Section 4 we show that it is sharp minimax adaptive over a scale of Sobolev classes of densities. Second, in Section 5 we show that its MISE is asymptotically smaller than or equal to that of kernel density estimators with any bandwidth provided that the kernel belongs to a large class of functions including many standard kernels.

2. Application of the blockwise Stein method to density estimation

2.1. Estimation of the Δ-risk. Blockwise Stein methods in the Gaussian sequence model are related to the *unbiased estimation of the risk*. For density estimation, we will consider only an asymptotically unbiased estimator of the risk. It follows from (6.1) that $\left[|\varphi_n(\omega)|^2 - 1/n\right]$ is an asymptotically unbiased estimator of $\left|\varphi(\omega)\right|^2$. Thus we can define an asymptotically unbiased estimator of $\Delta_n\left(\lambda, |\varphi|^2\right)$ by

$$\begin{aligned}\tilde{l}_n(\lambda) &= \int_{\mathbb{R}} \left(|1-\lambda(\omega)|^2 \left[\left|\varphi_n(\omega)\right|^2 - \frac{1}{n} \right] + \frac{1}{n}|\lambda(\omega)|^2 \right) \mathrm{d}\omega \\ &= \int_{\mathbb{R}} \left(\left[|\lambda(\omega)|^2 - 2\mathrm{Re}\left(\lambda(\omega)\right) + 1 \right] \left[\left|\varphi_n(\omega)\right|^2 - \frac{1}{n} \right] + \frac{1}{n}|\lambda(\omega)|^2 \right) \mathrm{d}\omega,\end{aligned}$$

for λ such that the integrals are finite. Then dropping the term independent of λ, we get the functional

$$l_n(\lambda) = \int_{\mathbb{R}} \left(\left[|\lambda(\omega)|^2 - 2\mathrm{Re}\left(\lambda(\omega)\right) \right] \left|\varphi_n(\omega)\right|^2 + \frac{2}{n}\mathrm{Re}\left[\lambda(\omega)\right] \right) \mathrm{d}\omega. \tag{6.10}$$

The idea is now to choose a weight function λ that minimizes l_n over a certain class $\mathcal{H}$. To define a reasonable $\mathcal{H}$, we shall restrict the possible weight functions λ to *admissible* ones. Cline (1988) proves that if λ is an arbitrary complex-valued function, it is possible to find a real, non-negative function $\overline{\lambda}$, bounded by one such that the risk corresponding to $\overline{\lambda}$ is smaller than that corresponding to λ. All such $\overline{\lambda}$ will be called *admissible*. For all admissible λ, the functional $l_n(\lambda)$ defined in (6.10) becomes

$$l_n(\lambda) = \int_{\mathbb{R}} \left(\left[\lambda^2(\omega) - 2\lambda(\omega) \right] \left|\varphi_n(\omega)\right|^2 + \frac{2}{n}\lambda(\omega) \right) \mathrm{d}\omega.$$

We also impose the restriction that λ is even. This assumption is satisfied whenever λ is admissible and is the Fourier transform of some even function as in (6.3). From now on, let $\mathcal{H}_0$ be the class of all even, square-integrable functions on $\mathbb{R}$ taking values in $[0,1]$.

In the next section, we study a simple class of weight functions which leads to the definition of an analogue of the Stein estimator for density estimation. For this estimator, a first oracle inequality is given. Then the construction is generalized to a slightly more complex class of weight functions to define the blockwise Stein estimator.

2.2. Stein's estimators applied to density estimation.

2.2.1. *Stein's estimator on a set.* Consider a particularly simple class of weight functions

$$\mathcal{H}_A = \left\{ \lambda \in \mathcal{H}_0 : \lambda(\omega) = t\mathbb{1}_A(\omega),\ t \in [0,1] \right\},$$

where A is a finite union of bounded, non-trivial intervals on $\mathbb{R}$ (typically, the union of two intervals symmetric about 0, which we will call *symmetrized intervals*). The *Stein estimator on* A is the solution of the minimization problem

$$\lambda_A^* = \operatorname*{argmin}_{\lambda \in \mathcal{H}_A} l_n(\lambda)\,,$$

which can be explicitly written as

$$\lambda_A^*(\omega) = \left(1 - \frac{|A|}{n\int_A \left|\varphi_n(\omega)\right|^2 \mathrm{d}\omega}\right)_+ \mathbb{1}_A(\omega) = t_A^* \mathbb{1}_A(\omega) \quad \omega \in \mathbb{R}. \tag{6.11}$$

where $|A|$ is the Lebesgue measure of A. Then let $\lambda_A^{\text{oracle}}$ be the *linear oracle on* $\mathcal{H}_A$, defined by

$$\lambda_A^{\text{oracle}} = \operatorname*{argmin}_{\lambda \in \mathcal{H}_A} R_n^A\left(\hat{\varphi}_\lambda, \varphi\right),$$

where

$$R_n^A\left(\hat{\varphi}_\lambda, \varphi\right) = E_p^n \int_A \left|\hat{\varphi}_\lambda(\omega) - \varphi(\omega)\right|^2 \mathrm{d}\omega.$$

It is easy to see that

$$\lambda_A^{\text{oracle}}(\omega) = \left(\frac{\int_A |\varphi|^2}{\int_A |\varphi|^2 + n^{-1}\int_A (1 - |\varphi|^2)}\right) \mathbb{1}_A(\omega) = t_A^{\text{oracle}} \mathbb{1}_A(\omega). \tag{6.12}$$

2.2.2. *The block Stein estimator.* Now introduce a constant $b_0 > 0$ and a finite value Ω_n depending only on n and consider a partition $\{B_j\}_{j=0}^J$ of $[-\Omega_n, \Omega_n]$, such that $B_0 = (-b_0, b_0)$ and,

$$\forall\ 1 \le j \le J, \quad B_j = -B_j' \cup B_j', \quad B_j' = [b_{j-1}, b_j),\ -B_j' = (-b_j, -b_{j-1}], \quad 0 < b_{j-1} < b_j.$$

Let $\mathcal{H}^*$ be the class of weight functions given by

$$\mathcal{H}^* = \left\{ \lambda \in \mathcal{H}_0 : \lambda(\omega) = \sum_{j=0}^J t_j \mathbb{1}_{B_j}(\omega),\ 0 \le t_j \le 1,\ j = 0, \ldots, J \right\} \subset \mathcal{H}_0\,.$$

Minimization of l_n over $\mathcal{H}^*$ follows directly from the minimization over $\mathcal{H}_{B_j}$. Indeed, the function $\tilde{\lambda} = \operatorname{argmin}_{\lambda \in \mathcal{H}^*} l_n(\lambda)$ is constant on each B_j,

$$\tilde{\lambda}(\omega) = \sum_{j=0}^J \lambda_{B_j}^* \mathbb{1}_{B_j}(\omega),$$

where $\lambda_{B_j}^*$ is defined in (6.11). Define *blockwise Stein estimator on the system* $\{B_j\}_{j=0}^J$ by

$$\tilde{\lambda}(\omega) = \sum_{j=0}^J \left(1 - \frac{|B_j|}{n\int_{B_j} \left|\varphi_n(\omega)\right|^2 \mathrm{d}\omega}\right)_+ \mathbb{1}_{B_j}(\omega)\,.$$

3. Oracle inequalities

We first give an oracle inequality satisfied by Stein's estimator on a set.

THEOREM 6.1. *Let* $1 \le |A| \le 4n$ *and let* φ *satisfy* $\int_A |\varphi(\omega)| \mathrm{d}\omega \le G$, *for some* $G < \infty$. *Then, there exist an absolute constant* $C > 0$ *and a constant* $D_1 > 0$ *that depends only on* G *such that for any* $\mu_n > C$, *the Stein estimator on the set* A *satisfies the following oracle*

inequality

$$R_n^A\left(\hat{\varphi}_{\lambda_A^*},\varphi\right) \le \frac{1}{1-C\mu_n^{-1}}\left(R_n^A\left(\hat{\varphi}_{\lambda_A^{\text{oracle}}},\varphi\right) + D_1\frac{(\log n)^4\,\mu_n}{n}\right). \tag{6.13}$$

The proof of Theorem 6.1 is given in Section 8.

From (6.13) we obtain another oracle inequality for the blockwise Stein estimator. Indeed, if every $B_j, j=0,\dots,J$, satisfies $1\le|B_j|\le 4n$, by Theorem 6.1, for $\mu_n > C$ and $\int|\varphi|\le G$ we have

(6.14)

$$\begin{aligned} R_n(\hat{\varphi}_{\tilde{\lambda}},\varphi) &= \sum_{j=0}^{J} R_n^{B_j}\left(\hat{\varphi}_{\lambda^*_{B_j}},\varphi\right) + \int_{|\omega|>\Omega_n}\left|\varphi(\omega)\right|^2\mathrm{d}\omega \\ &\le \frac{1}{1-C\mu_n^{-1}}\left(\sum_{j=0}^{J} R_n^{B_j}\left(\hat{\varphi}_{\lambda^{\text{oracle}}_{B_j}},\varphi\right) + JD_1\frac{(\log n)^4\,\mu_n}{n}\right) + \int_{|\omega|>\Omega_n}\left|\varphi(\omega)\right|^2\mathrm{d}\omega. \end{aligned}$$

Let $\lambda^{\text{oracle}}_{\mathcal{H}^*}$ be the linear blockwise constant oracle defined by

$$\lambda^{\text{oracle}}_{\mathcal{H}^*} = \operatorname*{argmin}_{\lambda\in\mathcal{H}^*} R_n\left(\hat{\varphi}_\lambda,\varphi\right).$$

Since

$$\sum_{j=0}^{J} R_n^{B_j}\left(\hat{\varphi}_{\lambda^{\text{oracle}}_{B_j}},\varphi\right) + \int_{|\omega|>\Omega_n}\left|\varphi(\omega)\right|^2\mathrm{d}\omega = R_n(\hat{\varphi}_{\lambda^{\text{oracle}}_{\mathcal{H}^*}},\varphi),$$

equation (6.14) implies the following result.

THEOREM 6.2. *Let $1\le|B_j|\le 4n$ for any $j=0,\dots,J$ and let φ satisfy $\int_{B_j}|\varphi(\omega)|\mathrm{d}\omega\le G$, for any $j=0,\dots,J$ and some $G<\infty$. Then there exist an absolute constant $C>0$ and a constant $D_1>0$ that depends only on G such that for any $\mu_n>C$, the blockwise Stein estimator on the system $\{B_j\}_{j=0}^J$ satisfies the oracle inequality*

$$R_n(\hat{\varphi}_{\tilde{\lambda}},\varphi) \le \frac{1}{1-C\mu_n^{-1}}\left(R_n(\hat{\varphi}_{\lambda^{oracle}_{\mathcal{H}^*}},\varphi) + JD_1\frac{(\log n)^4\,\mu_n}{n}\right). \tag{6.15}$$

In what follows we prove that the oracle on $\mathcal{H}^*$ is close to oracles on classes that are larger than $\mathcal{H}^*$, namely the classes of *monotone* weight functions.

Consider the two classes of monotone weight functions

$$\mathcal{H}^{\infty}_{mon} = \{\lambda\in\mathcal{H}_0 : \lambda(\omega_1)\le\lambda(\omega_2),\ 0\le\omega_2\le\omega_1\}$$

and

$$\mathcal{H}^{\Omega_n}_{\text{mon}} = \mathcal{H}^{\infty}_{\text{mon}}\cap\{\lambda:\ \lambda(\omega)=0,\ |\omega|>\Omega_n\}.$$

The space $L_2(\mathbb{R})$ equipped with the $\|\cdot\|$-norm is a reflexive Banach space, $\mathcal{H}^{\Omega_n}_{\text{mon}}$, $\Omega_n\le+\infty$, is a closed convex subset of $L_2(\mathbb{R})$. Moreover, the functional $F:\lambda\mapsto R_n(\hat{\varphi}_\lambda,\varphi)$ is quadratic and coercive (i.e. $F(\lambda)\to+\infty$, $\|\lambda\|\to\infty$) and thus strictly convex and continuous. These remarks and Ekeland and Temam (1976, Proposition 1.2, p. 35) show that one can uniquely define $\lambda^{\Omega_n}_{\text{mon}}$ as the solution of the minimization problem

$$\lambda^{\Omega_n}_{\text{mon}} = \operatorname*{argmin}_{\lambda\in\mathcal{H}^{\Omega_n}_{\text{mon}}} R_n(\hat{\varphi}_\lambda,\varphi).$$

We call it *linear monotone oracle*. In the Gaussian sequence model, under some assumptions on the system of blocs, the blockwise Stein estimator is almost as good as the linear

monotone oracle (Tsybakov, 2004a). An analogue of this result for density estimation is given below. Let the system of symmetrized intervals satisfy the next assumption. Suppose now $\Omega_n < +\infty$.

Assumption A *The inequality*

$$\max_{0\le j\le J-1} \frac{|B_{j+1}|}{|B_j|} \le 1+\eta$$

holds for some $\eta > 0$.

LEMMA 6.2. *Under Assumption A, for all* $\varphi \in L_2(\mathbb{R})$,

$$\begin{aligned} \min_{\lambda\in\mathcal{H}^*} R_n(\hat\varphi_\lambda,\varphi) &\le \min_{\lambda\in\mathcal{H}^*\cap\mathcal{H}^{\Omega_n}_{\mathrm{mon}}} R_n(\hat\varphi_\lambda,\varphi) \\ &\le (1+\eta)\min_{\lambda\in\mathcal{H}^{\Omega_n}_{\mathrm{mon}}} R_n(\hat\varphi_\lambda,\varphi) + \frac{1}{n}\left(|B_0| + 3(1+\eta)\|\varphi\|^2\right). \end{aligned} \tag{6.16}$$

PROOF. It is sufficient to show that for any $\lambda \in \mathcal{H}^{\Omega_n}_{\mathrm{mon}}$, there exists $\overline{\lambda} \in \mathcal{H}^* \cap \mathcal{H}^{\Omega_n}_{\mathrm{mon}}$ such that

$$R_n(\hat\varphi_{\overline\lambda},\varphi) \le (1+\eta)R_n(\hat\varphi_\lambda,\varphi) + \frac{1}{n}\left(|B_0| + 3(1+\eta)\|\varphi\|^2\right). \tag{6.17}$$

Fix $\lambda \in \mathcal{H}^{\Omega_n}_{\mathrm{mon}}$ and define $\tilde\lambda(\omega) = \min\left[\lambda(\omega), \left(1+n^{-1/2}\right)^{-1}\right]$. Inequality (6.17) holds for

$$\overline{\lambda}(\omega) = \sum_{j=0}^{J} \overline{\lambda}_{(j)} \mathbb{I}_{\{\omega\in B_j\}},$$

where $\overline{\lambda}_{(j)} = \sup_{f\in B_j}\tilde\lambda(f)$. Indeed,

$$\begin{aligned} R_n(\hat\varphi_{\overline\lambda},\varphi) &= \int_{\mathbb{R}} \left((1-\overline{\lambda}(\omega))^2|\varphi(\omega)|^2 + \frac{1}{n}\overline{\lambda}^2(\omega)\right)\mathrm{d}\omega - \frac{1}{n}\int_{\mathbb{R}}|\varphi(\omega)|^2\overline{\lambda}^2(\omega)\mathrm{d}\omega \\ &\le \int_{\mathbb{R}} \left((1-\tilde\lambda(\omega))^2|\varphi(\omega)|^2 + \frac{1}{n}\overline{\lambda}^2(\omega)\right)\mathrm{d}\omega - \frac{1}{n}\int_{\mathbb{R}}|\varphi(\omega)|^2\tilde\lambda^2(\omega)\mathrm{d}\omega. \end{aligned}$$

But $\overline{\lambda}$ satisfies

$$\int_{\mathbb{R}}\overline{\lambda}^2(\omega)\mathrm{d}\omega = \int_{-\Omega_n}^{\Omega_n}\overline{\lambda}^2(\omega)\mathrm{d}\omega \le |B_0| + (1+\eta)\int_{\mathbb{R}}\tilde\lambda^2(\omega)\mathrm{d}\omega.$$

Since,

$$\int_{\mathbb{R}}\left((1-\tilde\lambda(\omega))^2 - \frac{(\tilde\lambda(\omega))^2}{n}\right)|\varphi(\omega)|^2\mathrm{d}\omega \ge 0,$$

it follows that

$$R_n(\hat\varphi_{\overline\lambda},\varphi) \le (1+\eta)R_n(\hat\varphi_{\tilde\lambda},\varphi) + \frac{|B_0|}{n}. \tag{6.18}$$

One the other hand,

$$\begin{aligned} R_n(\hat\varphi_{\tilde\lambda},\varphi) &= \int_{\mathbb{R}}(1-\tilde\lambda(\omega))^2|\varphi(\omega)|^2\mathrm{d}\omega + \frac{1}{n}\int_{\mathbb{R}}\left(1-|\varphi(\omega)|^2\right)\tilde\lambda^2(\omega)\mathrm{d}\omega \\ &\le \int_{\mathbb{R}}(1-\tilde\lambda(\omega))^2|\varphi(\omega)|^2\mathrm{d}\omega + \frac{1}{n}\int_{\mathbb{R}}\left(1-|\varphi(\omega)|^2\right)\lambda^2(\omega)\mathrm{d}\omega. \end{aligned}$$

But, if we note $\int f = \int f(\omega)\mathrm{d}\omega$ for a function f, the first term of the right-hand side becomes

$$\begin{aligned}\int_{\mathbb{R}}(1-\tilde{\lambda})^2|\varphi|^2 &= \int_{\mathbb{R}}(1-\lambda)^2|\varphi|^2 + \int_{\mathbb{R}}(\lambda-\tilde{\lambda})^2|\varphi|^2 + 2\int_{\mathbb{R}}(1-\lambda)(\lambda-\tilde{\lambda})|\varphi|^2 \\ &\leq \int_{\mathbb{R}}(1-\lambda)^2|\varphi|^2 + \frac{\|\varphi\|^2}{(1+\sqrt{n})^2} + 2\frac{\|\varphi\|^2}{(1+\sqrt{n})^2} \\ &= \int_{\mathbb{R}}(1-\lambda)^2|\varphi|^2 + 3\frac{\|\varphi\|^2}{(1+\sqrt{n})^2}\end{aligned}$$

Therefore,

$$R_n(\hat{\varphi}_{\tilde{\lambda}},\varphi) \leq R_n(\hat{\varphi}_{\lambda},\varphi) + \frac{3\|\varphi\|^2}{n}. \tag{6.19}$$

The combination of (6.18) and (6.19) proves the lemma. ∎

THEOREM 6.3. *Let the system $\{B_j\}_{j=0}^J$ satisfy Assumption (A) and let the conditions of Theorem 6.2 hold. Then, there exist an absolute constant $C>0$ and a constant $D_1>0$ that depends only on G such that for any $\mu_n > C$, the blockwise Stein estimator on the system $\{B_j\}_{j=0}^J$ satisfies the following oracle inequality*

(6.20)

$$R_n(\hat{\varphi}_{\tilde{\lambda}},\varphi) \leq \frac{1}{1-C\mu_n^{-1}}\left((1+\eta)R_n(\hat{\varphi}_{\lambda^{\Omega_n}_{mon}},\varphi) + \frac{1}{n}\left(|B_0| + 3(1+\eta)G + JD_1\left(\log n\right)^4\mu_n\right)\right)$$

PROOF. The proof follows directly from Theorem 6.2 and Lemma 6.2. ∎

The next lemma allows us to extend the oracle inequality (6.20) to the class $\mathcal{H}^{\infty}_{mon}$ of monotone weight functions that do not necessarily have a compact support. Set $\Omega_n = n(\log n)^2$, so that for sufficiently large n, $\Omega_n \geq Gn\log n$.

LEMMA 6.3. *Assume that $\|\varphi\|^2 \leq G$. For $n \geq n_0(G) > 0$, there exist positive constants $\kappa_1 = \kappa_1(n_0)$ and $\kappa_2 = \kappa_2(n_0)$ such that*

$$\min_{\lambda\in\mathcal{H}^{\Omega_n}_{\mathrm{mon}}} R_n(\hat{\varphi}_{\lambda},\varphi) \leq \left(1+\frac{\kappa_1}{\log n}\right)\min_{\lambda\in\mathcal{H}^{\infty}_{\mathrm{mon}}} R_n(\hat{\varphi}_{\lambda},\varphi) + \kappa_2\frac{G}{n},$$

PROOF. Define:

$$\Omega_n^0 = \max\left\{|\omega| : |\lambda(\omega)| \geq (\log n)^{-1/2}\right\}.$$

Then, setting $\lambda = \lambda^{\infty}_{\mathrm{mon}}$ and $\lambda_0 \equiv 0 \in \mathcal{H}^{\infty}_{\mathrm{mon}}$,

$$\|\varphi\|^2 = R_n(\hat{\varphi}_{\lambda_0},\varphi) \geq R_n(\hat{\varphi}_{\lambda},\varphi) \geq \frac{1}{n}\int_{|\omega|\leq\Omega_n^0}\frac{1}{\log n}\mathrm{d}\omega - \frac{\|\varphi\|^2}{n} = \frac{2\Omega_n^0}{n\log n} - \frac{\|\varphi\|^2}{n}$$

Thus $\Omega_n^0 \leq \|\varphi\|^2 n\log n \leq \Omega_n$. Now define $\lambda_{\Omega_n}(\omega) = \lambda(\omega)\mathbb{I}_{\{|\omega|\leq\Omega_n\}}$, therefore $\lambda_{\Omega_n} \in \mathcal{H}^{\Omega_n}_{\mathrm{mon}}$ and,

$$\begin{aligned}R_n(\hat{\varphi}_{\lambda_{\Omega_n}},\varphi) &\leq \int_{-\Omega_n}^{\Omega_n}\left[(1-\lambda(\omega))^2|\varphi(\omega)|^2 + \frac{\lambda^2(\omega)}{n}\right]\mathrm{d}\omega + \int_{|\omega|\geq\Omega_n}|\varphi(\omega)|^2\mathrm{d}\omega \\ &\leq \left(1-\frac{1}{\sqrt{\log n}}\right)^{-2}\int_{\mathbb{R}}\left[(1-\lambda(\omega))^2|\varphi(\omega)|^2 + \frac{\lambda^2(\omega)}{n}\right]\mathrm{d}\omega \\ &\leq \left(1+\frac{\kappa_1}{\log n}\right)R_n(\hat{\varphi}_{\lambda},\varphi) + \kappa_2\frac{G}{n}, \quad \text{for } n \geq n_0.\end{aligned}$$

∎

An important question is how to construct systems of symmetrized intervals $\{B_j\}_{j=0}^J$ satisfying the assumptions of Theorem 6.3 and such that the residual term on the right-hand side of inequality (6.20) is asymptotically negligible with respect to the principal term $(1+\eta)R_n(\hat{\varphi}_{\lambda^{\Omega_n}_{\mathrm{mon}}}, \varphi)$ under rather general conditions on φ. We now give an example of such a construction. In what follows, set $\Omega_n = n^\alpha (\log n)^{\alpha'}$, where $\alpha \geq 1/2$, $\alpha' \geq 0$. Let ν_n be a deterministic quantity such that $\nu_n \to \infty$ when $n \to \infty$.

Set $\eta_n = 1/\nu_n$ and define the system of symmetrized intervals $\{B_j\}_{j=0}^J$ with the size $|B_j|$ of each symmetrized interval B_j:

$$
\begin{aligned}
&|B_0| = \nu_n, \\
&|B_j| = (1+\eta_n)^j \nu_n, \qquad j = 1, 2, \ldots, J-1, \\
&|B_J| = \Omega_n - \sum_{j=0}^{J-1} |B_j|,
\end{aligned}
\tag{6.21}
$$

where

$$
J = \min\Big\{ m \ : \ \sum_{j=0}^{m} (1+\eta_n)^j \nu_n \geq \Omega_n \Big\}.
$$

Clearly, this system satisfies Assumption A with $\eta = \eta_n$. We call the system $\{B_j\}_{j=0}^J$ a *weakly geometrically increasing system of symmetrized intervals* or *WGI system*. The corresponding blockwise Stein estimator is called *the Stein WGI estimator*. For $n \geq 2$,

$$
\sum_{j=0}^{J-1} (1+\eta_n)^j \nu_n \leq \Omega_n. \tag{6.22}
$$

Solving inequality (6.22) with respect to J, we find that there exist $n_0 \geq 2$ and $C = C(n_0, \alpha, \alpha')$ such that $J \leq C (\log n)\, \nu_n$, for $n \geq n_0$. For the Stein WGI estimator, by Plancherel's identity, inequality (6.20) yields the following theorem.

THEOREM 6.4. *Assume that $\int_{\mathbb{R}} |\varphi| \leq G$. Then, there exist an absolute constant $C > 0$ and a constant $D_2 > 0$ that depends only on G such that for any $\mu_n > C$ and sufficiently large n, the Stein WGI estimator satisfies the oracle inequality:*

$$
R_n(\hat{p}_{\lambda_{\Omega_n}}, p) \leq \frac{1}{1 - C\mu_n^{-1}} \left((1+\nu_n^{-1}) R_n(\hat{p}_{\lambda^{\Omega_n}_{\mathrm{mon}}}, p) + \frac{\tau_n(D_2)}{2\pi} \right), \tag{6.23}
$$

where the residual

$$
\tau_n(D_2) = \frac{D_2 \nu_n}{n} \left(1 + (\log n)^5 \, \mu_n \right).
$$

4. Application to sharp minimax adaptation

DEFINITION 6.3. *Consider the scale of classes $\mathcal{A} = \{\Theta_\gamma\}_{\gamma \in V}$, for a set of indexes V (typically, $V \subset \mathbb{R}^d$). An estimator $\hat{p}_n$ of p is called* sharp minimax adaptive *on the scale of classes $\mathcal{A}$ if it is asymptotically sharp minimax simultaneously on all the classes Θ_γ, $\gamma \in V$.*

We will show that for a properly chosen system of blocs, the block Stein estimator is sharp minimax adaptive on a wide scale of Sobolev classes of densities. The definition of such classes is given in Chapter 4 and is reproduced here for convenience.

For any $\beta > 0$ and $Q > 0$ define the Sobolev classes of densities on $\mathbb{R}$ by

$$\Theta(\beta, Q) = \left\{ p : \mathbb{R} \to \mathbb{R} \;\Big|\; p \geq 0, \int_{\mathbb{R}^d} p(x)\mathrm{d}x = 1, \; \int_{\mathbb{R}} \|\omega\|_1^{2\beta} |\varphi(\omega)|^2 \mathrm{d}\omega \leq Q \right\},$$

where $\|\cdot\|_1$ denotes the Euclidean norm in $\mathbb{R}^1$ and $\varphi = \mathcal{F}[p]$. In Chapter 4 we proved that for fixed $\beta > 1/2$ and $Q > 0$, the Pinsker type kernel density estimator is sharp minimax over $\Theta(\beta, Q)$. Recall that the Pinsker kernel K_β is the kernel having the Fourier transform

$$\mathcal{F}[K_\beta](\omega) = \Big(1 - \|\omega\|^\beta\Big)_+, \quad \omega \in \mathbb{R}, \tag{6.24}$$

where $x_+ = \max(x, 0)$. Moreover, the bandwidth parameter of the Pinsker type density estimator is taken equal to

$$h^* = D^* n^{-\frac{1}{2\beta+1}} \quad \text{where} \quad D^* = \left(\frac{2\beta}{Q(\beta+1)(2\beta+1)} \right)^{\frac{1}{2\beta+d}}. \tag{6.25}$$

Set $\beta > 1/2$ and $Q > 0$. From (6.24) and (6.25), it is obvious that for sufficiently large n, the weight function $\ell^*(\omega) = \mathcal{F}[K_\beta](h^*\omega)$ belongs to $\mathcal{H}^{\sqrt{n}}_{\mathrm{mon}}$. Now, if $p \in \Theta(\beta, Q)$, for $\beta > 1/2$ we have $\int_{\mathbb{R}} |\varphi| \leq C(\beta, Q)$, where $C(\beta, Q) < \infty$ is a positive constant that depends only on β and Q. Taking then the supremum over a Sobolev class of densities $\Theta(\beta, Q)$ of both sides of (6.23) , we get

$$\sup_{p \in \Theta(\beta,Q)} R_n(\hat{p}_{\tilde{\lambda}}, p) \leq \frac{1}{1 - C\mu_n^{-1}} \left((1 + \nu_n^{-1}) \sup_{p \in \Theta(\beta,Q)} R_n(\hat{p}_{\lambda^{\Omega_n}_{\mathrm{mon}}}, p) + \frac{\tau_n(D(Q,\beta))}{2\pi} \right), \tag{6.26}$$

where $D(Q, \beta) < \infty$ depends only on Q and β. But $\ell \in \mathcal{H}^{\sqrt{n}}_{\mathrm{mon}}$, so for every $p \in \Theta(\beta, Q)$ with Fourier transform φ, we have that $R_n(\hat{\varphi}_{\lambda^{\sqrt{n}}_{\mathrm{mon}}}, \varphi) \leq R_n(\hat{\varphi}_\ell, \varphi)$, which implies that $R_n(\hat{p}_{\lambda^{\sqrt{n}}_{\mathrm{mon}}}, p) \leq R_n(\hat{p}_\ell, p)$. For $\Omega_n = \sqrt{n}$, inequality (6.26) with, for instance, $\mu_n = \nu_n = \log(\log n)$ yields

$$\begin{aligned} \sup_{p \in \Theta(\beta,Q)} R_n(\hat{p}_{\tilde{\lambda}}, p) &\leq \frac{1}{1 - C\mu_n^{-1}} \left((1 + \nu_n^{-1}) \sup_{p \in \Theta(\beta,Q)} R_n(\hat{p}_\ell, p) + \frac{\tau_n(D(Q,\beta))}{2\pi} \right) \\ &\leq C^* n^{-\frac{2\beta}{2\beta+1}} (1 + o(1)), \qquad n \to +\infty, \end{aligned} \tag{6.27}$$

where C^* is the Pinsker constant defined in (4.3) with $d = 1$, i.e,

$$C^* = (2\beta + 1) \left[\frac{\pi \, (2\beta + 1) \, (\beta + 1)}{\beta} \right]^{-\frac{2\beta}{2\beta+1}} Q^{\frac{1}{2\beta+1}}.$$

The last inequality in (6.27) is a known upper bound for the Pinsker type density estimator that is proved by Schipper (1996) for integer β, although the proof remains valid for any positive β (see Theorem 4.2). On the other hand, the following lower bound holds.

$$\inf_{T_n} \sup_{p \in \Theta(\beta,Q)} n^{\frac{2\beta}{2\beta+1}} \mathbb{E}_p \|T_n - p\|^2 \geq C^*(1 + o(1)), \qquad n \to +\infty, \tag{6.28}$$

where the infimum is taken over all estimators of p (see Theorem 4.1). From (6.27) and (6.28), we conclude that the Stein WGI estimator $\hat{p}_{\tilde{\lambda}}$ is sharp minimax adaptive over the scale of Sobolev classes of densities $\{\Theta(\beta, Q), \; \beta > 1/2, \; Q > 0\}$ in the sense of standard definitions given in the introduction of the present manuscript.

5. Application to kernel density estimation

For $h > 0$, define the kernel density estimator $\hat{p}_{n,h}$ as in (6.2), where the kernel belongs to the class $\mathcal{K}_0$, the class of kernels K that admit a version of Fourier transform $\mathcal{F}[K]$ symmetric about 0, decreasing on $\mathbb{R}_+$ and taking its values in $[0,1]$. Set $\Omega_n = n(\log n)^2$. Lemma 6.3, Theorem 6.4 and equality (6.3) together lead to the following theorem.

THEOREM 6.5. *Assume that $\int_{\mathbb{R}} |\varphi| \le G$. Then, there exist an absolute constant $C > 0$ and a constant $D_3 > 0$ that depends only on G such that for any $\mu_n > C$ and sufficiently large n, the Stein WGI estimator satisfies the kernel oracle inequality*

(6.29)
$$R_n(\hat{p}_{\tilde{\lambda}}, p) \le \frac{1}{1 - C\mu_n^{-1}} \left((1 + \nu_n^{-1})(1 + \kappa_1 (\log n)^{-1}) \inf_{K \in \mathcal{K}_0} \inf_{h>0} R_n(\hat{p}_{n,h}, p) + \frac{\tau_n(D_3)}{2\pi} \right).$$

Note that we do not suppose that $K \in L_1(\mathbb{R})$. This can be interpreted to mean that the infimum over $\mathcal{K}_0$ on the right-hand side of (6.29) is not attained for such kernels.

COROLLARY 6.1. *Let $K \in \mathcal{K}_0$ be kernel satisfying the conditions of Lemma A.6. Then for every fixed probability density $p \in L_2(\mathbb{R})$*

$$R_n(\hat{p}^*_{\tilde{\lambda}}, p) \le (1 + o(1)) \inf_{h>0} R_n(\hat{p}_{n,h}, p), \quad n \to \infty. \tag{6.30}$$

Here $o(1)$ depends on K and p.

Proof is straightforward from Theorem 6.5 and Lemma A.6.

6. Concluding remarks

(1) It is important to note that the Stein WGI estimator mimics an oracle that is more powerful than any kernel oracle in a large class of kernels. The main difference from previous results of Stone's type (Stone, 1984; Devroye and Penrod, 1984; Wegkamp, 1999) is that one and the same estimator is shown to be simultaneously as powerful or even more powerful asymptotically than the kernel oracles corresponding to various kernels.

(2) Here are some examples of admissible kernels covered by Theorem 6.5 and Corollary 6.1: the triangular kernel, the biweight kernel, Silverman's kernel, Fejer's kernel, the Gaussian kernel and the *sinc* kernel (the last one covered only by Theorem 6.5 and not by Corollary 6.1). Note that in this list, which is not exhaustive, only the triangular and biweight kernels obey the conditions in Stone (1984). Of course, Theorem 6.5 says nothing about the kernels whose Fourier transform does not take values in $[0,1]$. These kernels are not admissible (see Cline, 1988) because they have higher MISE. This is the case of the parabolic kernel (Epanechnikov, 1969) and of the rectangular kernel.

(3) Inequality (6.29) may not be a valid oracle inequality if the residual $\tau_n(D_3)$ is of order greater than the MISE of the kernel oracle. Indeed, Ibragimov and Khas′minskiĭ (1982) prove that for all densities p with Fourier transform vanishing outside of a compact set, the oracle MISE $\inf_{K \in \mathcal{K}_0} \inf_{h>0} R_n(\hat{p}_{n,h}, p)$ is of order n^{-1}. In that case, the residual term $\tau_n(D_3)$ is of order greater than the oracle MISE.

7. Numerical results

This section presents a simulation study of the Stein WGI estimator. The free parameters have been chosen as follows:

- The size of the support of the Stein WGI estimator is $\Omega_n = \sqrt{n}$.
- The symmetrized intervals $\{B_j\}_{j=0}^J$ in the WGI-system are chosen as in (6.21) with parameters

$$\nu_n = \log n \quad \text{and} \quad \eta_n = \nu_n^{-1} = (\log n)^{-1},$$

As there exists no universal method of testing the efficiency of nonparametric density estimators, we use the approach which consists in simulating the estimator on statistically popular densities and on pathological examples where usual estimators do not perform well. We take the test densities as in Butucea (2001) and in Boĭko and Golubev (2000) following the earlier suggestions in Marron and Wand (1992) and Devroye (1997) (see also Wand and Jones, 1995). Instead of choosing an estimator computed from an arbitrary sample, we found more relevant to give confidence intervals. The given confidence intervals are honest in the sense that they are not centered about the true density p. Figures 4 to 15 represent simulated symmetric confidence intervals computed at level 90% from 1000 samples of size $n = 1000$ for each density. Our procedure is then as follows.

- Draw 1000 samples of size $n = 1000$ with the considered density.
- Compute the Stein WGI estimator on a grid provided by Fast Fourier transform (FFT) algorithm.
- For each point x of the grid, sort increasingly the values of the estimators at point x. Note $S(x)$ the corresponding vector. Then set $I_u(x)$ equal to the 5% upper quantile and $I_l(x)$ equal to the 5% lower quantile, i.e. $I_u(x)$ is the 950^{th} coordinate of $S(x)$ and $I_l(x)$ is the 50^{th} coordinate of $S(x)$.

The algorithm has been implemented on `Scilab` and used the FFT algorithm to calculate the estimator of the density from the estimator of the characteristic function. The FFT calculated the estimator of the density on a grid of 500 points.

We see that the numerical performance of the estimator is quite nice: in particular, it adapts to inhomogeneous smoothness as in Figures 5, 12, 13 and 15.

To study the behavior of the Stein WGI estimator computed from a small size sample, we estimate its MISE by Monte-Carlo techniques for the Claw and Smooth comb densities. We also present the case of the standard Gaussian density to show competitiveness of the Stein WGI estimator even in easy cases. For each n from 20 to 1000 with an incremental step-size of 20, we consider 500 samples of size n. From these 500 samples we estimate the MISE of the estimator by its empirical counterpart. Figures 1 and 2 below summarizes the results. The Stein WGI estimator is compared to common data-driven bandwidth selectors and Gaussian $\mathcal{N}(0,1)$ kernel. The bandwidth selectors are chosen as in Chapter 5. We give here for completeness the description given by the R software.

- `nrd0` implements a rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator. It defaults to 0.9 times the minimum of the standard deviation and the inter-quartile range divided by 1.34 times the sample size to the negative one-fifth power (Silverman, 1986, page 48, eqn (3.31)).
- `nrd` is the more common variation given by Scott (1992), using factor 1.06 instead of 0.9 in `nrd`.

- `ucv` and `bcv` implement unbiased and biased cross-validation respectively (see, e.g., Wand and Jones, 1995).
- `SJ-dpi` and `SJ-ste` implement respectively the *direct plug-in* and *solve the equation* methods of Sheather and Jones (1991) to select the bandwidth using pilot estimation of derivatives

We see that the empirical MISE of the Stein WGI estimator is comparable to that of the proposed competitors for the standard Gaussian density. As for the Claw and Smooth comb densities, only `ucv` stands the comparison.

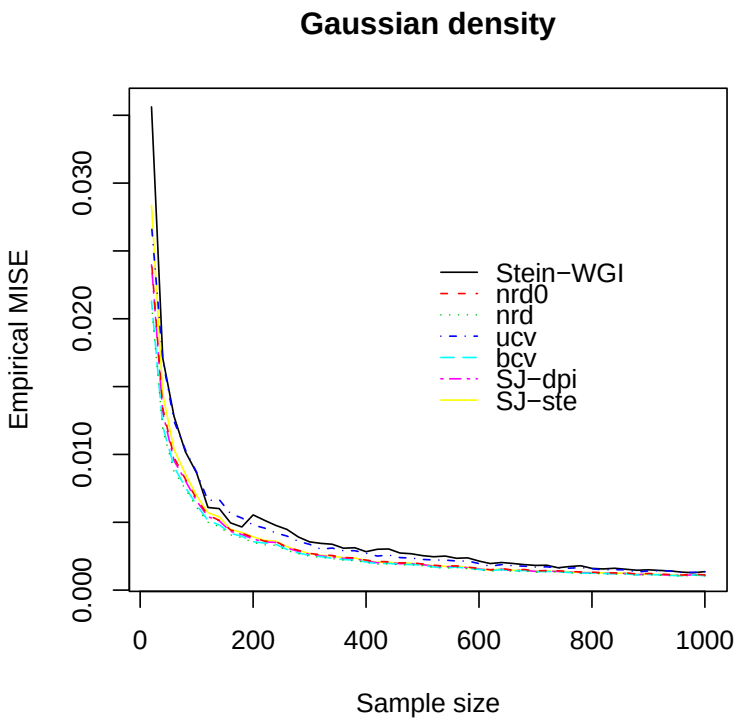

FIGURE 1. Comparison of empirical MISE for the standard Gaussian density

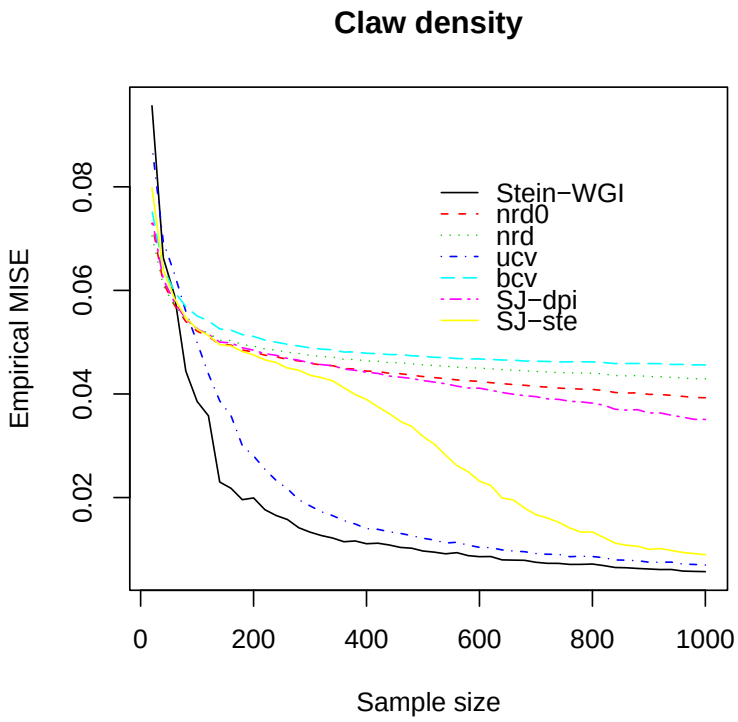

FIGURE 2. Comparison of empirical MISE for the Claw density

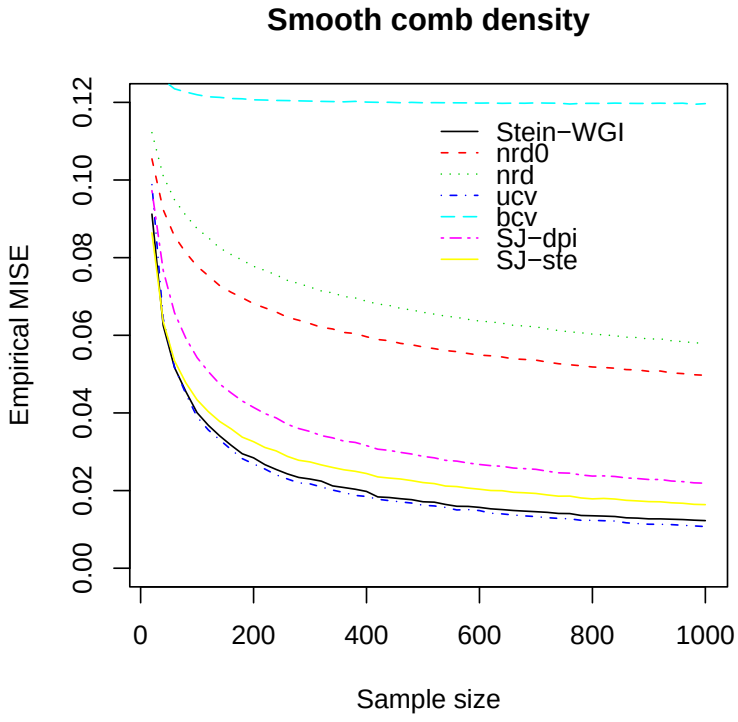

FIGURE 3. Comparison of empirical MISE for the Smooth comb density

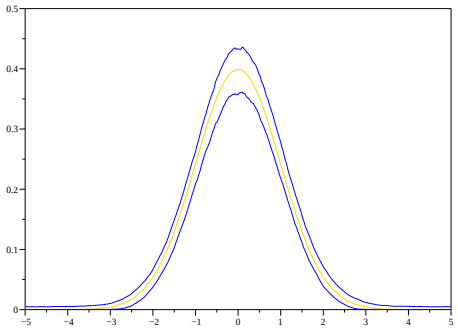

FIGURE 4. Gaussian density

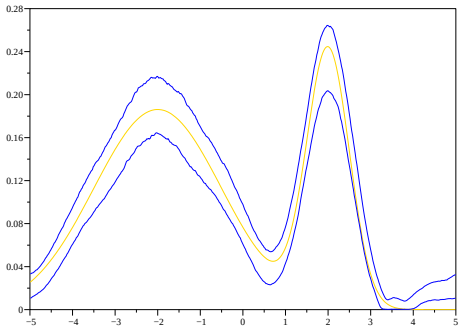

FIGURE 5. Mixed Gaussian density

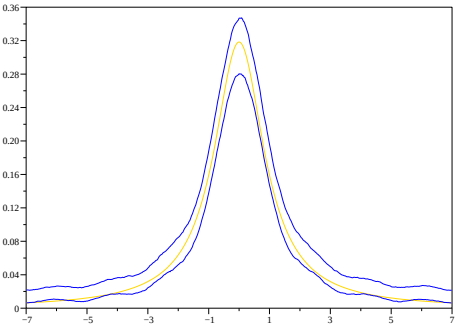

FIGURE 6. Cauchy density

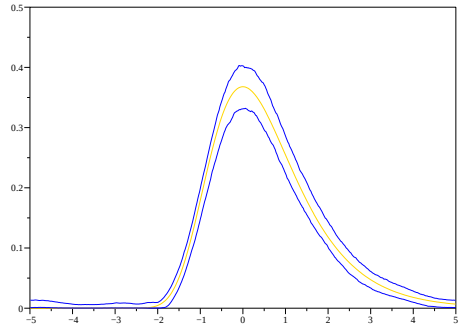

FIGURE 7. Extreme Value density

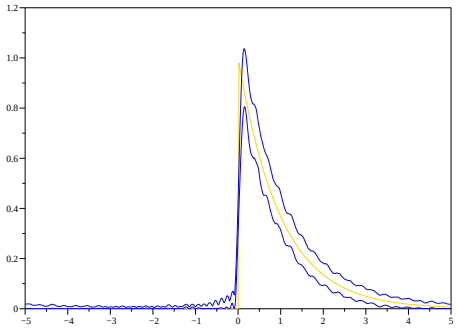

FIGURE 8. Exponential density

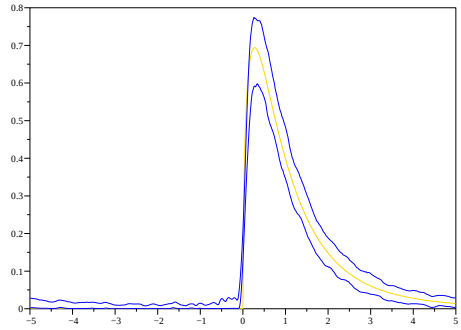

FIGURE 9. Fisher's density

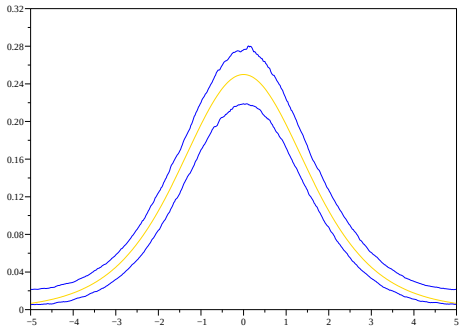

FIGURE 10. Logistic density

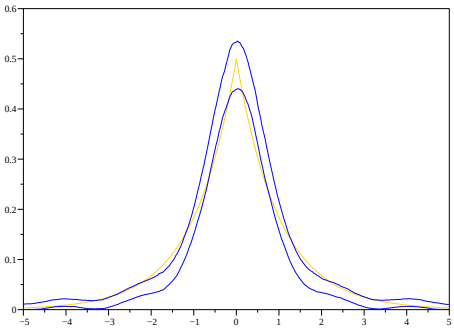

FIGURE 11. Laplace density

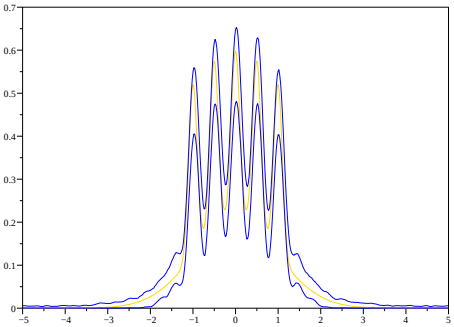

FIGURE 12. Claw density

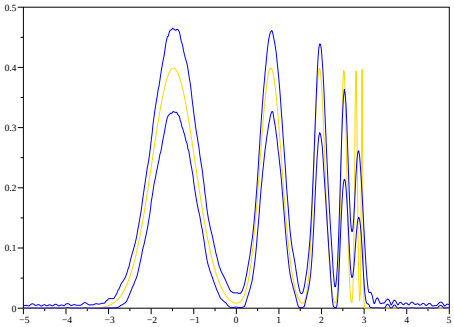

FIGURE 13. Smooth comb density

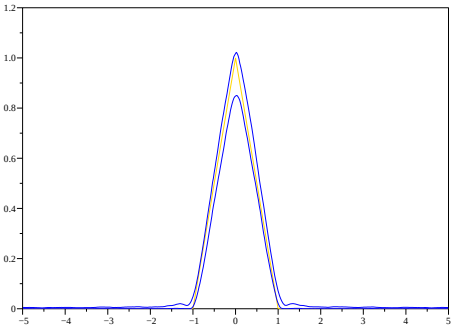

FIGURE 14. Triangular density

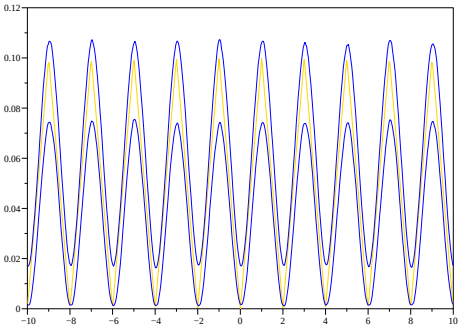

FIGURE 15. Saw Tooth density

8. Proofs of main results

Let ζ_n be the empirical process

$$\zeta_n(\omega) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(\mathrm{e}^{\mathrm{i}\omega X_k} - \varphi(\omega) \right) = \sqrt{n} \left(\varphi_n(\omega) - \varphi(\omega) \right).$$

The supremum of this process has been extensively studied by Csörgő and Totik (1983) Csörgő (1985). Yet the bounds for the supremum are not of a small enough order for our use. Thus we need to investigate the expectation of the process ζ_n. Several properties of this process have been given by Golubev and Levit (1996). We reproduce their results adapted to our framework in the next section.

8.1. Properties of the process ζ_n. The following two lemmas are proved in Golubev and Levit (1996) for any φ such that $\int_{\mathbb{R}} |\varphi| \leq G$.

LEMMA 6.4. *Let $n \geq 1$ be a fixed integer and b be a real function such that*

$$\int_{\mathbb{R}} |b(\omega)| \, \mathrm{d}\omega \leq 2\sqrt{n} \|b\|, \tag{6.31}$$

where $\|\,.\,\|$ denotes the $L_2(\mathbb{R})$ norm. Then, for any $k \geq 1$, there exists a constant $C > 0$ such that

$$\left| E_p^n \int_{\mathbb{R}} b(\omega) \zeta_n(\omega) \mathrm{d}\omega \right|^{2k} \leq (Ck)^{4k} \|b\|^{2k}.$$

If, moreover, b satisfies

$$\max_{\substack{l \in \mathbb{N} \\ l \geq 3}} \|b\|^{-l} \int_{\mathbb{R}} |b(\omega)|^l \, \mathrm{d}\omega \leq 1, \tag{6.32}$$

then, for any integer $k \geq 1$, there exists a constant $C' > 0$ such that

$$E_p^n \left(\int_{\mathbb{R}} b(\omega) \left(|\zeta_n(\omega)|^2 - E_p^n |\zeta_n(\omega)|^2 \right) \mathrm{d}\omega \right)^{2k} \leq \left(C'k \right)^{4k} \|b\|^{2k}.$$

LEMMA 6.5. *Let $\tilde{J} \in \{1, \ldots, N\}$ be a random index and $(\xi(\omega), \omega \in \mathbb{R})$ be a random process satisfying*

$$E_p^n \left| \int_{\mathbb{R}} b_j(\omega) \xi(\omega) \mathrm{d}\omega \right|^{2k} \leq (Dk)^{4k} \|b_j\|^{2k}, \qquad k \in \mathbb{N}^*, j = 1, \ldots, N,$$

for some constant $D > 0$. Then there exists a constant $D' > 0$ such that

$$E_p^n \left| \int_{\mathbb{R}} b_{\tilde{J}}(\omega) \xi(\omega) \mathrm{d}\omega \right| \leq \left(D' \log N \right)^2 \left(E_p^n \|b_{\tilde{J}}\|^2 \right)^{1/2}.$$

8.2. Proof of Theorem 6.1. The following lemma, stated by Golubev (1992), gives two important formulae that are used to construct oracle inequalities.

LEMMA 6.6. *Let A be a subset of $\mathbb{R}$ and*

$$R_n^A \left(\hat{\varphi}_\lambda, \varphi \right) = E_p^n \int_A \left| \hat{\varphi}_\lambda(\omega) - \varphi(\omega) \right|^2 \mathrm{d}\omega.$$

Then for all λ such that $\lambda(\omega) = h\mathbb{1}_A(\omega)$, $h = h(\mathbb{X}^n) \in \mathbb{R}$,

(6.33)

$$R_n^A\left(\hat{\varphi}_\lambda, \varphi\right) = E_p^n\left[l_n(\lambda)\right] + \int_A |\varphi(\omega)|^2 \mathrm{d}\omega + \frac{2}{n} E_p^n\left[\int_A h\left(|\zeta_n(\omega)|^2 - 1\right)\mathrm{d}\omega\right] + \frac{2}{\sqrt{n}}\mathrm{Re}\left[E_p^n \int_A (h-1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right],$$

and

(6.34)

$$E_p^n\left[l_n(\lambda)\right] = R_n^A\left(\hat{\varphi}_\lambda, \varphi\right) - \left(1 - \frac{E_p^n[h^2]+1}{n}\right)\int_A |\varphi(\omega)|^2\mathrm{d}\omega + \frac{1}{n}E_p^n\left[\int_A (1-h)^2\left(|\zeta_n(\omega)|^2 - 1\right)\mathrm{d}\omega\right] + \frac{2}{\sqrt{n}}\mathrm{Re}\left[E_p^n\int_A (1-h)^2\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right].$$

Proof. The result is stated by Golubev (1992) without proof, we give it here for completeness. For a function f, we note $\int f = \int f(\omega)\mathrm{d}\omega$.

$$\begin{aligned}
R_n^A\left(\hat{\varphi}_\lambda, \varphi\right) &= E_p^n \int_A |\hat{\varphi}_\lambda - \varphi|^2 \\
&= E_p^n \int_A |\varphi - \varphi_n + \varphi_n - h\varphi_n|^2 \\
&= E_p^n \int_A \left\{|\varphi - \varphi_n|^2 + (1-h)^2|\varphi_n|^2 + 2(1-h)\mathrm{Re}\left[(\varphi - \varphi_n)\,\overline{\varphi_n}\right]\right\} \\
&= E_p^n \int_A \left\{\frac{|\zeta_n|^2}{n} + (h^2 - 2h)|\varphi_n|^2 + |\varphi_n|^2 - 2(1-h)\mathrm{Re}\left[\frac{\zeta_n}{\sqrt{n}}\overline{\varphi_n}\right]\right\} \\
&= E_p^n\left[l_n(\lambda)\right] + E_p^n \int_A \left\{\frac{|\zeta_n|^2}{n} - \frac{2h}{n} + |\varphi_n|^2 - 2(1-h)\mathrm{Re}\left[\frac{\zeta_n}{\sqrt{n}}\overline{\varphi_n}\right]\right\} \\
&= E_p^n\left[l_n(\lambda)\right] + \int_A |\varphi|^2 \\
&\quad + E_p^n \int_A \left\{\frac{|\zeta_n|^2}{n} - \frac{2h}{n} + |\varphi_n|^2 - |\varphi|^2 - 2(1-h)\mathrm{Re}\left[\frac{\zeta_n}{\sqrt{n}}\overline{\varphi_n}\right]\right\}.
\end{aligned}$$

But,

$$|\varphi_n|^2 - |\varphi|^2 = \frac{|\zeta_n|^2}{n} + 2\mathrm{Re}\left[\overline{\varphi}\,(\varphi_n - \varphi)\right]$$

Thus,

$$\begin{aligned}
R_n^A\left(\hat{\varphi}_\lambda,\varphi\right) &= E_p^n\left[l_n(\lambda)\right] + \int_A |\varphi|^2 \\
&\qquad + E_p^n \int_A \left\{ 2\frac{|\zeta_n|^2}{n} - \frac{2h}{n} + 2\mathrm{Re}\left[\frac{\zeta_n}{\sqrt{n}}\overline{\varphi}\right] - 2(1-h)\mathrm{Re}\left[\frac{\zeta_n}{\sqrt{n}}\overline{\varphi_n}\right]\right\} \\
&= E_p^n\left[l_n(\lambda)\right] + \int_A |\varphi|^2 + \frac{2}{n}E_p^n\int_A h\left(|\zeta_n|^2-1\right) \\
&\qquad\qquad + 2\mathrm{Re}\left[E_p^n\int_A (1-h)\left(\frac{|\zeta_n|^2}{n} - \frac{\zeta_n}{\sqrt{n}}\overline{\varphi_n}\right) + \frac{\zeta_n}{\sqrt{n}}\overline{\varphi}\right] \\
&= E_p^n\left[l_n(\lambda,\mathbb{X}^n)\right] + \int_A |\varphi|^2 + \frac{2}{n}E_p^n\int_A h\left(|\zeta_n|^2-1\right) \\
&\qquad\qquad + 2\mathrm{Re}\left[E_p^n\left\{\int_A (h-1)\frac{\zeta_n}{\sqrt{n}}\overline{\varphi}\right\}\right].
\end{aligned}$$

Which is (6.33). Next, we have

$$\begin{aligned}
E_p^n\left[l_n(\lambda)\right] &= E_p^n\int_A\left[(h^2-2h)|\varphi_n|^2 + \frac{2}{n}h\right] \\
&= E_p^n\int_A\left[(1-h)^2|\varphi_n|^2 - |\varphi_n|^2 + \frac{2}{n}h\right] \\
&= E_p^n\int_A\left[(1-h)^2|\varphi|^2 + (1-h)^2(|\varphi_n|^2-|\varphi|^2) - |\varphi_n|^2 + \frac{2}{n}h\right] \\
&= R_n^A\left(\hat{\varphi}_\lambda,\varphi\right) + E_p^n\int_A\left[(1-h)^2(|\varphi_n|^2-|\varphi|^2) - |\varphi_n|^2 + \frac{2}{n}h - \frac{1}{n}h^2 + \frac{1}{n}h^2|\varphi|^2\right] \\
&= R_n^A\left(\hat{\varphi}_\lambda,\varphi\right) - \left(1 - \frac{E_p^n[h^2]+1}{n}\right)\int_A|\varphi|^2 + E_p^n\int_A(1-h)^2\left(|\varphi_n|^2-|\varphi|^2-\frac{1}{n}\right) \\
&= R_n^A\left(\hat{\varphi}_\lambda,\varphi\right) - \left(1 - \frac{E_p^n[h^2]+1}{n}\right)\int_A|\varphi|^2 + \frac{1}{n}E_p^n\int_A(1-h)^2\left(|\zeta_n|^2-1\right) \\
&\qquad\qquad + 2\mathrm{Re}\left[E_p^n\int_A(1-h)^2\overline{\varphi}\frac{\zeta_n}{\sqrt{n}}\right].
\end{aligned}$$

And (6.34) is proved.

∎

Using the definitions of t_A^* and t_A^{oracle} given respectively in (6.11) and in (6.12), the following lemma holds.

LEMMA 6.7. *The Stein estimator on the set A satisfies the inequality:*

$$\begin{aligned}
R_n^A\left(\hat{\varphi}_{\lambda_A^*},\varphi\right) &\le R_n^A\left(\hat{\varphi}_{\lambda_A^{\text{oracle}}},\varphi\right) + \frac{2}{n}\int_A|\varphi(\omega)|^2\mathrm{d}\omega \\
&\qquad + \frac{2}{n}E_p^n\int_A t_A^*\left(|\zeta_n(\omega)|^2 - E_p^n|\zeta_n(\omega)|^2\right)\mathrm{d}\omega \\
&\qquad + \frac{2}{\sqrt{n}}\mathrm{Re}\left[E_p^n\int_A(t_A^*-1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right].
\end{aligned} \tag{6.35}$$

PROOF. By definition of λ_A^* we have,

$$E_p^n\left[l_n(\lambda_A^*)\right] \le E_p^n\left[l_n(\lambda_A^{\text{oracle}})\right],$$

Remark that for any $\omega \in \mathbb{R}$,

$$E_p^n|\zeta_n(\omega)|^2 = 1 - |\varphi(\omega)|^2 \quad \text{and} \quad E_p^n[\zeta_n(\omega)] = 0\,.$$

Using (6.34) it follows that

$$\begin{aligned} E_p^n\left[l_n(\lambda_A^{\text{oracle}})\right] &= R_n^A\left(\hat{\varphi}_{\lambda_A^{\text{oracle}}}, \varphi\right) - \left(1 - \frac{(t_A^{\text{oracle}})^2 + 1}{n}\right)\int_A |\varphi(\omega)|^2 \mathrm{d}\omega \\ &\quad - \frac{1}{n}\int_A (1 - t_A^{\text{oracle}})^2 |\varphi(\omega)|^2 \mathrm{d}\omega \\ &= R_n^A\left(\hat{\varphi}_{\lambda_A^{\text{oracle}}}, \varphi\right) - \left(1 - \frac{2t_A^{\text{oracle}}}{n}\right)\int_A |\varphi(\omega)|^2 \mathrm{d}\omega\,. \end{aligned} \tag{6.36}$$

Moreover, using (6.33), we obtain

$$\begin{aligned} R_n^A\left(\hat{\varphi}_{\lambda_A^*}, \varphi\right) &= E_p^n\left[l_n(\lambda_A^*)\right] + \int_A |\varphi(\omega)|^2 \mathrm{d}\omega \\ &\quad + \frac{2}{n} E_p^n \int_A t_A^* \left(|\zeta_n(\omega)|^2 - 1\right) \mathrm{d}\omega + \frac{2}{\sqrt{n}} \mathrm{Re}\left[E_p^n \int_A (t_A^* - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right] \end{aligned}$$

Hence, by Lemma 6.1 and (6.36), we get

$$\begin{aligned} R_n^A\left(\hat{\varphi}_{\lambda_A^*}, \varphi\right) &\le R_n^A\left(\hat{\varphi}_{\lambda_A^{\text{oracle}}}, \varphi\right) + \frac{2}{n}\int_A |\varphi(\omega)|^2 \mathrm{d}\omega \\ &\quad + \frac{2}{n} E_p^n \int_A t_A^* \left(|\zeta_n(\omega)|^2 - 1\right) \mathrm{d}\omega \\ &\quad + \frac{2}{\sqrt{n}} \mathrm{Re}\left[E_p^n \int_A (t_A^* - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right]. \end{aligned}$$

∎

In view of Lemma 6.7, to prove Theorem 6.1, it is sufficient to bound from above the last two summands on the right-hand side of (6.35), that is,

$$E_p^n \int_A t_A^* \left(|\zeta_n(\omega)|^2 - E_p^n|\zeta_n(\omega)|^2\right) \mathrm{d}\omega \quad \text{and} \quad E_p^n \int_A (t_A^* - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\,.$$

For this purpose, we will use Lemmas 6.4 and 6.5 with the additional assumption that A is symmetric about 0 as in Section 2.

Set $n \ge 1$ and let $|A| \ge 1$ be the length of A. Let us now cover the interval $[0, 1]$ by $N = \lfloor n|A| \rfloor + 1 \le 2n|A|$ disjoint intervals, $\Delta_1, \ldots, \Delta_N$, with centres $t_1, \ldots, t_N$ and lengths $1/(n|A|)$. Here $\lfloor x \rfloor$ denotes the integer part of x. Let $\tilde{t}_A$ be the projection of t_A^* on $\{t_1, \ldots, t_N\}$, i.e.

$$\tilde{t}_A = \operatorname*{argmin}_{t \in \{t_1, \ldots, t_N\}} \left|t - t_A^*\right|.$$

Thus $\left|\tilde{t}_A - t_A^*\right| \le 1/n|A|$ and we have

$$\begin{aligned} \left|E_p^n \int_A (t_A^* - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right| &\le \left|E_p^n \int_A (\tilde{t}_A - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right| \\ &\quad + \left|E_p^n \int_A (t_A^* - \tilde{t}_A)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right|. \end{aligned}$$

Next, we have,

$$\left|E_p^n \int_A (t_A^* - \tilde{t}_A)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega\right| \leq \frac{1}{n|A|}\int_A E_p^n|\varphi(\omega)||\zeta_n(\omega)|\mathrm{d}\omega$$
$$\leq \frac{1}{2n|A|}\int_A E_p^n\left(|\varphi(\omega)|^2 + |\zeta_n(\omega)|^2\right)\mathrm{d}\omega = \frac{1}{2n}.$$

By the same argument,

$$\left|E_p^n \int_A (t_A^* - \tilde{t}_A)\left(|\zeta_n(\omega)|^2 - E_p^n|\zeta_n(\omega)|^2\right)\mathrm{d}\omega\right| \leq \frac{2}{n}.$$

Note now that $b_j^{\mathrm{Re}}(\omega) = (t_j - 1)\mathrm{Re}\left(\overline{\varphi(\omega)}\right)\mathbb{1}_A(\omega)$ satisfies (6.31), for $|A| \leq 4n$ and $j = 1, \ldots, N$. Indeed, by the Cauchy-Schwarz inequality

$$\int_{\mathbb{R}} |b_j^{\mathrm{Re}}(\omega)|\mathrm{d}\omega = \int_{\mathbb{R}} |(t_j - 1)\mathrm{Re}\left(\varphi(\omega)\right)\mathbb{1}_A(\omega)|\mathrm{d}\omega$$
$$\leq (1 - t_j)\sqrt{|A|}\left(\int_A |\mathrm{Re}\left(\varphi(\omega)\right)|^2\mathrm{d}\omega\right)^{\frac{1}{2}}$$
$$\leq 2\sqrt{n}(1 - t_j)\left(\int_A |\mathrm{Re}\left(\varphi(\omega)\right)|^2\mathrm{d}\omega\right)^{\frac{1}{2}},$$

and

$$\|b_j^{\mathrm{Re}}\| = \left(\int_{\mathbb{R}} |(1 - t_j)\mathrm{Re}\left(\varphi(\omega)\right)\mathbb{1}_A(\omega)|^2\mathrm{d}\omega\right)^{\frac{1}{2}} = (1 - t_j)\left(\int_A |\mathrm{Re}\left(\varphi(\omega)\right)|^2\mathrm{d}\omega\right)^{\frac{1}{2}},$$

hence

$$\int_{\mathbb{R}} |b_j^{\mathrm{Re}}(\omega)|\mathrm{d}\omega \leq 2\sqrt{n}\|b_j^{\mathrm{Re}}\|.$$

In the same manner, $b_j^{\mathrm{Im}}(\omega) = (t_j - 1)\mathrm{Im}\left(\overline{\varphi(\omega)}\right)\mathbb{1}_A(\omega)$ satisfies (6.31).

Moreover, $b_j'(\omega) = t_j\mathbb{1}_A(\omega)$ satisfies (6.31) and (6.32) for $1 \leq |A| \leq 4n$ and $j = 1, \ldots, N$. Indeed,

$$\int_{\mathbb{R}} |b_j'(\omega)|\mathrm{d}\omega = \int_{\mathbb{R}} |t_j\mathbb{1}_A(\omega)|\mathrm{d}\omega = t_j|A| \leq 2\sqrt{n}t_j\sqrt{|A|},$$

and

$$\|b_j'\| = \left(\int_{\mathbb{R}} |t_j\mathbb{1}_A(\omega)|^2\mathrm{d}\omega\right)^{\frac{1}{2}} = t_j\sqrt{|A|},$$

hence

$$\int_{\mathbb{R}} |b_j'(\omega)|\mathrm{d}\omega \leq 2\sqrt{n}\|b_j'\|.$$

Therefore $b_j'(\omega) = t_j\mathbb{1}_A(\omega)$ satisfies (6.31). On the other hand,

$$\|b_j'\| = t_j\sqrt{|A|} \quad \text{and} \quad \int_{\mathbb{R}} |b_j'(\omega)|^l\mathrm{d}\omega = (t_j)^l|A|.$$

Therefore, for all $l \geq 2$ and for $|A| \geq 1$,

$$\|b_j'\|^{-l}\int_{\mathbb{R}} |b_j'(\omega)|^l\mathrm{d}\omega = |A|^{1-\frac{l}{2}} \leq 1.$$

Thus, $b'_j(\omega) = t_j \mathbb{1}_A(\omega)$ satisfies (6.32). By Lemma 6.4 for any integer $k \geq 1$, any $j = 1, \ldots, N$ and any A such that $1 \leq |A| \leq 4n$, we have the following upper bounds:

$$E_p^n \left| \int_A (\tilde{t}_A - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega \right|^{2k} \leq (Ck)^{4k} \left((1-t_j)^2 \int_A |\varphi(\omega)|^2 \mathrm{d}\omega \right)^k \tag{6.37}$$

and

$$E_p^n \left(\int_A t_j \left(|\zeta_n(\omega)|^2 - E_p^n |\zeta_n(\omega)|^2 \right) \mathrm{d}\omega \right)^{2k} \leq \left(C'k\right)^{4k} \left(t_j \sqrt{|A|} \right)^{2k}. \tag{6.38}$$

By (6.37) and (6.38), Lemma 6.5 can be applied to processes $\xi(\omega) = \zeta_n(\omega)$ and $\xi(\omega) = |\zeta_n(\omega)|^2 - E_p^n|\zeta_n(\omega)|^2$. We obtain the inequalities

$$E_p^n \left| \int_A (\tilde{t}_A - 1)\overline{\varphi(\omega)}\zeta_n(\omega)\mathrm{d}\omega \right| \leq C \left(\log N\right)^2 \left(E_p^n \left[(1-\tilde{t}_A)^2 \right] \int_A |\varphi(\omega)|^2 \mathrm{d}\omega \right)^{1/2} \tag{6.39}$$

and

$$E_p^n \left| \int_A \tilde{t}_A \left(|\zeta_n(\omega)|^2 - E_p^n |\zeta_n(\omega)|^2 \right) \mathrm{d}\omega \right| \leq C' \left(\log N\right)^2 \left(E_p^n \left[{\tilde{t}_A}^2 \right] |A| \right)^{1/2}. \tag{6.40}$$

Furthermore, for $|A| \geq 1$ and any $\mu_n > 0$,

$$\begin{aligned}
\frac{2}{\sqrt{n}} \left(E_p^n \left[(1-\tilde{t}_A)^2 \right] \int_A |\varphi|^2 \right)^{1/2} &\leq \frac{(\log n)^2 \mu_n}{n} + \frac{1}{(\log n)^2 \mu_n} E_p^n \left[(1-\tilde{t}_A)^2 \right] \int_A |\varphi|^2 \\
&\leq \frac{(\log n)^2 \mu_n}{n} + \frac{1}{(\log n)^2 \mu_n} E_p^n \left[(1-t_A^*)^2 \right] \int_A |\varphi|^2 \mathrm{d} \\
&\quad + \frac{1}{n^2 (\log n)^2 \mu_n} + \frac{2}{n (\log n)^2 \mu_n} \\
&\leq C \frac{(\log n)^2 \mu_n}{n} + \frac{1}{(\log n)^2 \mu_n} E_p^n \left[(1-t_A^*)^2 \right] \int_A |\varphi|^2.
\end{aligned}$$

And, by the same argument

$$\begin{aligned}
\frac{2}{n} \left(E_p^n \left[{\tilde{t}_A}^2 \right] |A| \right)^{1/2} \leq C' \frac{(\log n)^2 \mu_n}{n} &+ \frac{1}{n (\log n)^2 \mu_n} E_p^n \left[(t_A^*)^2 \right] \int_A \left(1 - |\varphi|^2 \right) \\
&+ \frac{1}{n (\log n)^2 \mu_n} \int_A |\varphi|^2.
\end{aligned}$$

Therefore, using the two preceding inequalities, Lemma 6.7, (6.39), (6.40) and the residuals of order $1/n$ due to discretization, one obtains, for strictly positive constants c_1, c_2 and c_3,

$$\begin{aligned}
R_n^A \left(\hat{\varphi}_{\lambda_A^*}, \varphi \right) \leq R_n^A \left(\hat{\varphi}_{\lambda_A^{\text{oracle}}}, \varphi \right) &+ \frac{2}{n} \int_A |\varphi(\omega)|^2 \mathrm{d}\omega + \frac{1}{n \left(\log n\right)^2 \mu_n} \int_A |\varphi(\omega)|^2 \mathrm{d}\omega \\
&+ c_1 \frac{\left(\log N\right)^2}{\left(\log n\right)^2 \mu_n} R_n^A \left(\hat{\varphi}_{\lambda_A^*}, \varphi \right) + c_2 \frac{\left(\log N\right)^2 \left(\log n\right)^2 \mu_n}{n} + c_3 \frac{1}{n}.
\end{aligned}$$

Then, noting that

$$\log N \leq \log(2|A|n) \leq \log(8n^2) \leq 5 \log n, \quad \text{for } n \geq 2,$$

we find a constant $c_4 > 0$ such that for $n \geq 3$ and $\mu_n \geq 1$,

$$\left(1 - c_3 \frac{1}{\mu_n} \right) R_n^A \left(\hat{\varphi}_{\lambda_A^*}, \varphi \right) \leq R_n^A \left(\hat{\varphi}_{\lambda_A^{\text{oracle}}}, \varphi \right) + \frac{3}{n} \int_A |\varphi(\omega)|^2 \mathrm{d}\omega + c_4 \frac{\left(\log n\right)^4 \mu_n}{n}.$$

and Theorem 6.1 is proved.

Part 3

Oracle inequalities and excess criteria

CHAPTER 7

Excess risk bounds in semi-supervised classification under the cluster assumption

We consider semi-supervised classification when part of the available data is unlabeled. These unlabeled data can be useful for the classification problem when we make an assumption relating the behavior of the regression function to that of the marginal distribution. Seeger (2000) proposed the well-known *cluster assumption* as a reasonable one. We propose a mathematical formulation of this assumption and a method based on density level sets estimation that takes advantage of it to achieve fast rates of convergence both in the number of unlabeled examples and the number of labeled examples.

Contents

This chapter is an extended version of Rigollet (2006b).

1. Introduction

Semi-supervised classification has been of growing interest over the past few years and many methods have been proposed. The methods try to give an answer to the question: "How to improve classification accuracy using unlabeled data together with the labeled data?". Unlabeled data can be used in different ways depending on the assumptions on

the model. There are mainly two approaches to solve this problem. The first one consists in using the unlabeled data to reduce the *complexity* of the problem in a broad sense. For instance, assume that we have a set of potential classifiers and we want to aggregate them. In that case, unlabeled data is used to measure the *compatibility* between the classifiers and reduces the complexity of the set of candidate classifier (see, e.g., Balcan and Blum, 2005; Blum and Mitchell, 1998). Unlabeled data can also be used to reduce the dimension of the problem (see, e.g., Belkin and Niyogi, 2004), which is another way to reduce complexity. In the cited paper, it is assumed that the data actually live on a submanifold of low dimension.

The second approach is the one that we use here. It assumes that the data contains clusters that have homogeneous labels and the unlabeled observations are used to identify these clusters. This is the so-called *cluster assumption*. This idea can be put in practice in several ways giving rise to various methods. The simplest is the one presented here: estimate the clusters, then label each cluster uniformly. Most of these methods use Hartigan's (Hartigan, 1975) definition of clusters, namely the connected components of the density level sets. However, they use a parametric (usually mixture) model to estimate the underlying density which can be far from reality. Moreover, no generalization error bounds are available for such methods. In the same spirit, Tipping (1999) and Rattray (2000) propose methods that learn a distance using unlabeled data in order to have intra-cluster distances smaller than inter-clusters distances. The whole family of graph-based methods aims also at using unlabeled data to learn the distances between points. The edges of the graphs reflect the proximity between points. For a detailed survey on graph methods we refer to Zhu (2005). Finally, we mention kernel methods, where unlabeled data are used to build the kernel. Recalling that the kernel measures proximity between points, such methods can also be viewed as learning a distance using unlabeled data (see Bousquet *et al.*, 2004; Chapelle and Zien, 2005; Chapelle *et al.*, 2006).

The cluster assumption can be interpreted in another way, i.e., as the requirement that the decision boundary has to lie in low density regions. This interpretation has been widely used in learning since it can be used in the design of standard algorithms such as Boosting (d'Alché Buc *et al.*, 2001; Hertz *et al.*, 2004) or SVM (Bousquet *et al.*, 2004; Chapelle and Zien, 2005), which are closely related to kernel methods mentioned above. In these algorithms, a greater penalization is given to decision boundaries that cross a cluster. For more details, see, e.g., Seeger (2000); Zhu (2005); Chapelle *et al.* (2006). Although most methods make, sometimes implicitly, the cluster assumption, no formulation in probabilistic terms has been provided so far. The formulation that we propose in this paper remains very close to its original text formulation and allows to derive generalization error bounds. We also discuss what can and cannot be done using unlabeled data. One of the conclusions is that considering the whole excess-risk is too ambitious and we need to concentrate on a smaller part of it to observe the improvement of semi-supervised classification over standard classification.

Outline of the paper. After describing the model, we formulate the cluster assumption and discuss why and how it can improve classification performance in the Section 2. The main result of this section is Proposition 7.1 which essentially states that the effect of unlabeled data on the rates of convergence cannot be observed observed on the whole excess-risk. We therefore introduce the *cluster excess-risk* which corresponds to a part of the excess-risk that is interesting for this problem. In Section 3, we study the population

when the clusters are perfectly known, to get an idea of our target. Indeed, such a population case corresponds in some way to the case when the amount of unlabeled data is infinite. Section 4 contains the main result: after having defined the clusters in terms of density level sets, we propose an algorithm for which we derive rates of convergence for the cluster excess-risk as a measure of performance. An example of consistent density level set estimators is given in Section 5. Section 6 is devoted to discussion on the choice of λ and possible improvements. Proofs of the results are gathered in Section 7.

Notation. Throughout the paper, we denote positive constants by c_j. We write Γ^c for the complement of the set Γ. For two sequences $(u_p)_p$ and $(v_p)_p$ (in that paper, p will be m or n), we write $u_p = O(v_p)$ if there exists a constant $C > 0$ such that $u_p \leq C v_p$ and we write $u_p = \widetilde{O}(v_p)$ if $u_p \leq C(\log p)^\alpha v_p$ for some constants $\alpha > 0, C > 0$. Moreover, we write $u_p = o(v_p)$, if there exists a non negative sequence $(\varepsilon_p)_p$ that tends to 0 when p tends to infinity and such that $|u_p| \leq \varepsilon_p |v_p|$. Thus, if $u_p = \widetilde{O}(v_p)$, we have $u_p = o(v_p p^\beta)$, for any $\beta > 0$.

2. The model

Let (X, Y) be a random couple with joint distribution P, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of d features and $Y \in \{0, 1\}$ is a label indicating the class to which X belongs. The distribution P of the random couple (X, Y) is completely determined by the pair (P_X, η) where P_X is the marginal distribution of X and η is the regression function of Y on X, i.e., $\eta(x) \triangleq P(Y = 1 | X = x)$. The goal of classification is to predict the label Y given the value of X, i.e., to construct a measurable function $g : \mathcal{X} \to \{0, 1\}$ called a *classifier*. The performance of g is measured by the average classification error

$$R(g) \triangleq P\left(g(X) \neq Y\right) .$$

A minimizer of the risk $R(g)$ over all classifiers is given by the *Bayes classifier* $g^\star(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}$, where $\mathbb{1}_{\{.\}}$ denotes the indicator function. Assume that we have a sample of n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are independent copies of (X, Y). An empirical classifier is a random function $\hat{g}_n : \mathcal{X} \to \{0, 1\}$ constructed on the basis of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Since $g^\star$ is the best possible classifier, we measure the performance of an empirical classifier $\hat{g}_n$ by its *excess-risk*

$$\mathcal{E}(\hat{g}_n) = \mathbb{E}_n R(\hat{g}_n) - R(g^\star) ,$$

where $\mathbb{E}_n$ denotes the expectation with respect to the joint distribution of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. We denote hereafter by $\mathbb{P}_n$ the corresponding probability.

In many applications, a large amount of unlabeled data is available as well as a small set of labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the goal of semi-supervised classification is to use unlabeled data to improve the performance of classifiers. Thus, we observe two independent samples $\mathbb{X}_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\mathbb{X}_u = \{X_{n+1}, \ldots, X_{n+m}\}$, where n is rather small and typically $m \gg n$. Most existing theoretical studies of supervised classification use empirical processes theory (Devroye *et al.*, 1996; Vapnik, 1998; van de Geer, 2000; Boucheron *et al.*, 2005) to obtain rates of convergence for the excess-risk that are polynomial in n. Typically these rates are of the order $O(1/\sqrt{n})$ and can be as small as $\widetilde{O}(1/n)$ under some low noise assumptions (cf., e.g., Tsybakov, 2004b; Audibert and Tsybakov, 2005). However, simulations indicate that much faster rates should be attainable when the unlabeled data is used to identify homogeneous clusters. Of course, it is well known that in order to make use of the additional unlabeled observations, we

have to make an assumption on the dependence between the marginal distribution of X and the joint distribution of (X, Y). Seeger (2000) formulated the rather intuitive *cluster assumption* as follows[1]

> Two points $x, x' \in \mathcal{X}$ should have the same label y if there is a path between them which passes only through regions of relatively high P_X.

This assumption, in its raw formulation cannot be exploited in the probabilistic model since (i) the labels are random variables Y, Y' so that the expression "should have the same label" is meaningless unless η takes values in $\{0, 1\}$ and (ii) it is not clear what "regions of relatively high P_X" are. To match the probabilistic framework, we propose the following modifications

(i) $P[Y = Y'|X, X' \in C] \geq P[Y \neq Y'|X, X' \in C]$, where C is a cluster.
(ii) Define "regions of relatively high P_X" in terms of *density level sets*.

Assume for the moment that we know what the clusters are, so that we do not have to define them in terms of density level sets. This will be done in Section 4. Let $T_1, T_2, \ldots,$ be a countable family of subsets of $\mathcal{X}$. We now make the assumption that the T_j's are clusters of homogeneous data.

> **Cluster Assumption (CA1):** Let $T_j, j = 1, 2, \ldots,$ be a collection of measurable sets such that $T_j \subset \mathcal{X}, j = 1, 2, \ldots$ Then the function $x \in \mathcal{X} \mapsto \mathbb{I}\{\eta(x) \geq 1/2\}$ takes a constant value on each of the $T_j, j = 1, 2, \ldots$.

It is not hard to see that the cluster assumption **(CA1)** is equivalent to the following assumption.

Let $T_j, j = 1, 2, \ldots,$ be a collection of measurable sets such that $T_j \subset \mathcal{X}, j = 1, 2, \ldots$ Then, for any $j = 1, 2, \ldots,$ we have

$$P[Y = Y'|X, X' \in T_j] \geq P[Y \neq Y'|X, X' \in T_j]\,.$$

A question remains: what happens outside of the clusters? Define the union of the clusters,

$$\mathcal{C} = \bigcup_{j \geq 1} T_j \tag{7.1}$$

and assume that we are in the problematic case, $P_X(\mathcal{C}^c) > 0$ such that the question makes sense. Since the cluster assumption **(CA1)** says nothing about what happens outside of the set $\mathcal{C}$, we can only perform supervised classification on $\mathcal{C}^c$. Consider a classifier $\hat{g}_{n,m}$ built from labeled and unlabeled samples $(\mathbb{X}_l, \mathbb{X}_u)$ pooled together. The excess-risk of $\hat{g}_{n,m}$ can be written (see Devroye *et al.*, 1996),

$$\mathcal{E}(\hat{g}_{n,m}) = \mathbb{E}_{n,m} \int_{\mathcal{X}} |2\eta(x) - 1| \mathbb{I}_{\{\hat{g}_{n,m}(x) \neq g^\star(x)\}} p(x) \mathrm{d}x\,,$$

where $\mathbb{E}_{n,m}$ denotes the expectation with respect to the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$. We denote hereafter by $\mathbb{P}_{n,m}$ the corresponding probability. Since, the unlabeled sample is of no help to classify points in $\mathcal{C}^c$, any reasonable classifier should be based on the sample $\mathbb{X}_l$ so that $\hat{g}_{n,m}(x) = \hat{g}_n(x),\ \forall\, x \in \mathcal{C}^c$, and we have

$$\mathcal{E}(\hat{g}_{n,m}) \geq \mathbb{E}_n \int_{\mathcal{C}^c} |2\eta(x) - 1| \mathbb{I}_{\{\hat{g}_n(x) \neq g^\star(x)\}} p(x) \mathrm{d}x\,. \tag{7.2}$$

[1] the notation is adapted to the present framework

Since we assumed $P_X(\mathcal{C}^c) \neq 0$, the RHS of (7.2) is bounded from below by the optimal rates of convergence that appear in supervised classification.

Recall that the distribution P of the random couple (X, Y) is completely characterized by the couple (P_X, η) where P_X is the marginal distribution of X and η is the regression function of Y on X. Thus, any class of distributions $\mathcal{D}$ can be decomposed as $\mathcal{D} = \mathcal{M} \times \Xi$ where $\mathcal{M}$ is a class of marginal distributions on $\mathcal{X}$ and Ξ is a class of regression functions on $\mathcal{X}$ with values in $[0, 1]$.

PROPOSITION 7.1. *Fix $n, m \geq 1$ and let $\mathcal{C}$ be a measurable subset of $\mathcal{X}$. Assume that the class $\mathcal{D} = \mathcal{M} \times \Xi$ is such that for any $\eta \in \Xi$ and any $x \in \mathcal{C}^c$ the value of $\eta(x)$ is independent of P_X. Then, for any marginal distribution $P_X^0 \in \mathcal{M}$, we have*

$$\inf_{T_n} \sup_{\eta \in \Xi} \mathbb{E}_n \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_n \neq g^\star\}} \mathrm{d}P_X^0 \leq \inf_{T_{n,m}} \sup_{P \in \mathcal{D}} \mathbb{E}_{n,m} \int_{\mathcal{C}^c} |2\eta - 1| \mathbb{1}_{\{T_{n,m} \neq g^\star\}} \mathrm{d}P_X \,, \tag{7.3}$$

where $\inf_{T_{n,m}}$ denotes the infimum over all classifiers based on the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$ and $\inf_{T_n}$ denotes the infimum over all classifiers based only on the labeled sample $\mathbb{X}_l$.

The main consequence of Proposition 7.1 is that even when the cluster assumption **(CA1)** is valid the unlabeled data are useless to improve the rates of convergence. If the class $\mathcal{M}$ is reasonably large and satisfies $P_X^0(\mathcal{C}^c) > 0$, the left hand side in (7.3) can be bounded from below by the minimax rate of convergence with respect to n, over the class $\mathcal{D}$. Indeed a careful check of the proofs of minimax lower bounds reveals that they are constructed using a single marginal P_X^0 that is well chosen. These rates are typically of the order $n^{-\alpha}, 0 < \alpha \leq 1$ (see e.g. Mammen and Tsybakov (1999); Tsybakov (2004b); Audibert and Tsybakov (2005) and Boucheron *et al.* (2005) for a comprehensive survey).

Thus, unlabeled data do not improve the rate of convergence of this part of the excess-risk. To observe the effect of unlabeled data on the rates of convergence, we have to consider the *cluster excess-risk* of a classifier $\hat{g}_{n,m}$ defined by

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_{n,m}) \triangleq \mathbb{E}_{n,m} \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{\hat{g}_{n,m}(x) \neq g^\star(x)\}} p(x) \mathrm{d}x \,. \tag{7.4}$$

We will therefore focus on this measure of performance. The cluster excess-risk can also be expressed in terms of an excess-risk. To that end, define the set $\mathcal{G}$ of all classifiers restricted to $\mathcal{C}$:

$$\mathcal{G} = \big\{ g : \mathcal{C} \to \{0, 1\},\ g \text{ measurable} \big\} \,.$$

The performance of a classifier $g \in \mathcal{G}$ is measured by the average classification error on $\mathcal{C}$

$$R_{\mathcal{C}}(g) = P\big(g(X) \neq Y, X \in \mathcal{C}\big)$$

A minimizer of $R_{\mathcal{C}}(\cdot)$ over $\mathcal{G}$ is given $g^\star_{|\mathcal{C}}(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}$, $x \in \mathcal{C}$, i.e., the restriction of the Bayes classifier to $\mathcal{C}$. Now it can be easily shown that for any classifier $g \in \mathcal{G}$ we have,

$$R_{\mathcal{C}}(g) - R_{\mathcal{C}}(g^\star_{|\mathcal{C}}) = \int_{\mathcal{C}} |2\eta(x) - 1| \mathbb{1}_{\{g(x) \neq g^\star_{|\mathcal{C}}(x)\}} p(x) \mathrm{d}x \,. \tag{7.5}$$

Taking expectations on both sides of (7.5) with $g = \hat{g}_{n,m}$, it follows that

$$\mathbb{E}_{n,m} R_{\mathcal{C}}(\hat{g}_{n,m}) - R_{\mathcal{C}}(g^\star_{|\mathcal{C}}) = \mathcal{E}_{\mathcal{C}}(\hat{g}_{n,m}) \,.$$

Therefore, cluster excess-risk equals the excess-risk of classifiers in $\mathcal{G}$. In the sequel, we only consider classifiers $\hat{g}_{n,m} \in \mathcal{G}$, i.e., classifiers that are defined on $\mathcal{C}$.

We now propose a method to obtain good upper bounds on the cluster excess-risk, taking advantage of the cluster assumption **(CA1)**. The idea is to estimate the regions where the sign of $(\eta - 1/2)$ is constant and make a majority vote on each region.

3. Results for known clusters

Consider the ideal situation where the family $T_1, T_2, \ldots$, is known and we observe only the labeled sample $\mathbb{X}_l = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Define

$$\mathcal{C} = \bigcup_{j \geq 1} T_j \, .$$

Under the cluster assumption **(CA1)**, the function $x \mapsto \eta(x) - 1/2$ has constant sign on each T_j. Thus a simple and intuitive method for classification is to perform a majority vote on each T_j.

For any $j \geq 1$, define $\delta_j \geq 0$, $\delta_j \leq 1$ by

$$\delta_j = \int_{T_j} |2\eta(x) - 1| P_X(\mathrm{d}x) \, .$$

We now define our classifier based on the sample $\mathbb{X}_l$. For any $j \geq 1$, define the random variable

$$Z_n^j = \sum_{i=1}^{n} (2Y_i - 1) \, \mathbb{1}_{\{X_i \in T_j\}} \, ,$$

and denote by $\hat{g}_n^j$ the function $\hat{g}_n^j(x) = \mathbb{1}_{\{Z_n^j > 0\}}$ for all $x \in T_j$. Consider the classifier defined on $\mathcal{C}$ by

$$\hat{g}_n(x) = \sum_{j \geq 1} \hat{g}_n^j(x) \mathbb{1}_{\{x \in T_j\}}, \quad x \in \mathcal{C} \, .$$

The following theorem gives an exponential rate of convergence for the cluster excess-risk of the classifier $\hat{g}_n$ under **(CA1)**.

THEOREM 7.1. *Let $T_j, j \geq 1$ be a family of measurable sets that satisfy Assumption* **(CA1)**. *Then, the classifier $\hat{g}_n$ defined above satisfies*

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_n) \leq 2 \sum_{j \geq 1} \delta_j \mathrm{e}^{-n\delta_j^2/2} \, . \tag{7.6}$$

Moreover, if there exists $\delta > 0$ such that $\delta = \inf_j \{\delta_j : \delta_j > 0\}$, we obtain an exponential rate of convergence:

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_n) \leq 2\mathrm{e}^{-n\delta^2/2} \, . \tag{7.7}$$

In a different framework, Castelli and Cover (1995, 1996) have proved that exponential rates of convergence were attainable for semi-supervised classification. A rapid overview of the proof shows that the rate of convergence $\mathrm{e}^{-n\delta^2/2}$ cannot be improved without further assumption. It will be our target in semi-supervised classification. However, we need estimators of the clusters $T_j, j = 1, 2, \ldots$. In the next section we provide the main result on semi-supervised learning, that is when the clusters are unknown but we can estimate them using the unlabeled sample $\mathbb{X}_u$.

4. Main result

We now deal with a more realistic case where the clusters $T_1, T_2, \ldots$, are unknown and we have to estimate them using the unlabeled sample $\mathbb{X}_u = \{X_1, \ldots, X_m\}$. We begin by

giving a definition of the clusters in terms of density level sets. In this section, we assume that $\mathcal{X}$ is *connected* (see definition below) and has finite Lebesgue measure.

4.1. Definition of the clusters. Following Hartigan (1975), we propose a definition of clusters that is also compatible with the expression "regions of relatively high P_X" proposed by Seeger (2000).

Assume that P_X admits a density p with respect to the Lebesgue measure on $\mathbb{R}^d$ denoted hereafter by Leb_d. For a fixed $\lambda > 0$, the λ-level set of the density p is defined by

$$\Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\} \ . \tag{7.8}$$

On these sets, the density is relatively high. The cluster assumption involves also a notion of connectedness of a set. A set $C \subset \mathbb{R}^d$ is said to be *connected* (or pathwise connected) if, for any $x, x' \in C$, there exists a continuous map $f : [0,1] \to C$, such that $f(0) = x$ and $f(1) = x'$. A direct consequence of this definition is that a connected set cannot be defined up to a set of null Lebesgue measure. Indeed, consider for example the case $d = 1$ and $C = [0,1]$. This set is obviously connected (take the map f equal to the identity on $[0,1]$) but the set $\widetilde{C} = C \setminus \{1/2\}$ is no more connected even though C and $\widetilde{C}$ only differ by a set of null Lebesgue measure. In our setup we want to impose connectedness on certain subsets of the λ-level set of the density p which is actually defined up to a set of null Lebesgue measure. To overcome this problem, we introduce the following notions.

Let $\mathcal{B}(x,r)$ be the d-dimensional closed ball of center $x \in \mathbb{R}^d$ and radius $r > 0$, defined by

$$\mathcal{B}(x,r) = \left\{ y \in \mathbb{R}^d : \|y - x\| \leq r \right\} \ ,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

Definition 7.1. *Fix $r_0 > 0$, $c_0 > 0$ and let $\bar{d}$ be an integer such that $\bar{d} \geq d$. Let C be a measurable subset of $\mathcal{X}$. For two points $x, x' \in C$, we say that x* is r_0-connected to x' in C *and we write $x \overset{r_0}{\underset{C}{\longleftrightarrow}} x'$ if there exists a continuous map $f : [0,1] \to \mathcal{X}$ such that $f(0) = x, f(1) = x'$ and for any $t \in [0,1]$ and any $0 < r \leq r_0$, we have*

$$\mathrm{Leb}_d\big(\mathcal{B}(f(t), r) \cap C\big) \geq c_0 r^{\bar{d}} \, .$$

Moreover, we say that C is a standard *set, if for any $x \in C$, we have $x \overset{r_0}{\underset{C}{\longleftrightarrow}} x$.*

Remark that the definition of a standard set has been introduced by Cuevas and Fraiman (1997). This definition ensures that the set C has no "flat" parts which allows to exclude pathological cases such as the one presented on the left hand side of Figure 1. Remark also that the path f takes values in $\mathcal{X}$ which is connected, so that removing sets of null Lebesgue measure from C does not affect the r_0-connectedness of its elements, contrary to the usual notion of connectedness defined above.

When C is standard, the following lemma holds.

Lemma 7.1. *If the set C is standard, then the binary relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$ is an equivalence relation and C can be partitioned into its classes of equivalence.*

Before considering the classes of equivalence of the relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$, for some set $C \subset \mathcal{X}$ we make sure that there is only a finite number of them. To that end, we introduce the notion of s_0-separated sets.

Define the pseudo-distance distance d_∞, between two sets C_1 and C_2 by

$$d_\infty(C_1, C_2) = \inf_{\substack{x \in C_1 \\ y \in C_2}} \|x - y\|$$

We say that two sets $C_1, C_2 \ldots,$ are s_0-*separated* if $d_\infty(C_1, C_2) \geq s_0$, for some $s_0 \geq 0$. On the right hand side of Figure 1, we show an example of two sets that are not s_0-separated.

PROPOSITION 7.2. *Fix $r_0 > 0, s_0 > 0$ and assume that C is a standard set such that the classes of equivalence of the relation $\overset{r_0}{\underset{C}{\longleftrightarrow}}$ are two by two s_0-separated. Then there exists a partition $C_1, \ldots C_J$ of C, where the C_j are such that*

- *For any $j = 1, \ldots, J$ and any $x, x' \in C_j$, we have $x \overset{r_0}{\underset{C}{\longleftrightarrow}} x'$ and*
- *For any $j \neq j'$ and any $x \in C_j, x' \in C_{j'}$, x is not r_0-connected to x' in C.*

We call $C_1, \ldots, C_J$ the r_0-connected components *of C.*

We now formulate the cluster assumption when the clusters are defined in terms of density level sets. In the rest of the section, fix $\lambda > 0$ and let Γ denote the λ-level set of the density p.

Cluster Assumption (CA2): Fix $s_0 > 0, r_0 > 0, c_0 > 0$ and assume that Γ admits a version that is standard and such that the classes of equivalence of the relation $\overset{r_0}{\underset{\Gamma}{\longleftrightarrow}}$ are two by two s_0-separated. Denote by $T_1, \ldots, T_J$ the r_0-connected components of this version of Γ. Then the function $x \in \mathcal{X} \mapsto \mathbb{1}\{\eta(x) \geq 1/2\}$ takes a constant value on each of the $T_j, j = 1, \ldots J$.

4.2. Estimation of the clusters. Assume that p is uniformly bounded by a constant $L(p)$ and that $\mathrm{Leb}_d(\mathcal{X}) < \infty$. Denote by $\mathbb{P}_m$ and $\mathbb{E}_m$ respectively the probability and the expectation w.r.t the sample $\mathbb{X}_u$ of size m. Assume that we use the sample $\mathbb{X}_u$ to construct an estimator $\hat{G}_m$ of Γ satisfying

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] \to 0, \quad m \to +\infty\,, \tag{7.9}$$

where $\triangle$ is the sign for the symmetric difference. We call such estimators *consistent* estimators of Γ. However, for any $r_0 > 0$, the r_0-connected components of a consistent estimator of Γ are not in general consistent estimators of the r_0-connected components of Γ. To ensure componentwise consistency, we make assumptions on the estimator $\hat{G}_m$. Note that the performance of a density level estimator $\hat{G}_m$ is measured by the quantity

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big] + \mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big]\,. \tag{7.10}$$

For some estimators, such as the penalized plug-in density level sets estimators presented in Section 5, we can prove that the dominant term in the RHS of (7.10) is $\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m^c \cap \Gamma)\big]$. It yields that the probability of having Γ included in the consistent estimator $\hat{G}_m$ is negligible. We now give a precise definition of such estimators.

DEFINITION 7.2. *Let $\hat{G}_m$ be an estimator of Γ and fix $\alpha > 0$. We say that the estimator $\hat{G}_m$ is* consistent from inside *at rate $m^{-\alpha}$ if it satisfies*

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \triangle \Gamma)\big] = \widetilde{O}(m^{-\alpha})$$

and

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\hat{G}_m \cap \Gamma^c)\big] = \widetilde{O}(m^{-2\alpha})$$

For a fixed $\alpha > 0$, let $\hat{G}_m \subset \mathcal{X}$ be an estimator of Γ that is consistent from inside at rate $m^{-\alpha}$ and recall that we want to estimate the r_0-connected components of Γ. To this end, we apply the following transformations to the estimator $\hat{G}_m$:

1. **Clipping:** In this step we remove some elements from $\hat{G}_m$ and obtain a clipped set $\tilde{G}_m$. Since $\hat{G}_m$ is an estimator of Γ that is consistent from inside, it ensures that any connected subset of $\tilde{G}_m$ is included in one of the r_0-connected components of Γ except on an event of negligible probability. In other words, it ensures that we do not estimate the union of two r_0-connected components of Γ by a single r_0-connected component of $\hat{G}_m$.
2. **Merging:** In this step we want to prevent ourselves from estimating a single r_0-connected component of Γ by several closer and closer disjoint connected sets. The idea used here is to estimate the r_0-connected components of Γ by the connected components of the clipped set $\tilde{G}_m$. When two connected components of $\tilde{G}_m$ are too close we merge them by taking their union.

We now describe the clipping step in more details. Define the set

$$\mathrm{Clip}(\hat{G}_m) = \big\{x \in \hat{G}_m : \mathrm{Leb}_d\big(\hat{G}_m \cap \mathcal{B}(x, (\log m)^{-1})\big) \leq \frac{(\log m)^{-d}}{m^{\alpha}}\big\}.$$

Since for sufficiently large m, we have $(\log m)^{-d} m^{-\alpha} \leq r_0$ eventually, $\mathrm{Clip}(\hat{G}_m)$ is such that none of its elements is r_0-connected to itself in $\hat{G}_m$. In the sequel, we will only consider the clipped version of $\hat{G}_m$ defined by $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$.

PROPOSITION 7.3. *Fix $\alpha > 0$. Assume that $\mathrm{Leb}_d(\mathcal{X}) < \infty$ and let $\hat{G}_m$ be an estimator of Γ that is consistent from inside at rate $m^{-\alpha}$. Then, the clipped estimator $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ is also consistent from inside a rate $m^{-\alpha}$ and has a finite number $\tilde{J}_m$ of connected components.*

Denote by $\tilde{T}_1, \ldots, \tilde{T}_{\tilde{J}_m}$ the connected components of $\tilde{G}_m$, where $\tilde{J}_m$ is the number of connected components of $\tilde{G}_m$. This number depends on $\mathbb{X}_u$ and is therefore random.

We now describe the merging step. For simplicity we present it in terms of a recursive pseudo-algorithm. For any $j = 1, \ldots, \tilde{J}_m$, define the set of integers

$$\mathcal{N}(j) = \big\{k \in \{1, \ldots, \tilde{J}_m\} : d_\infty(\tilde{T}_j, \tilde{T}_k) \leq 2(\log m)^{-1}\big\},$$

Consider the following pseudo-algorithm.

MERGING

- Initialize: $\mathcal{Z} = \{1, \ldots, \tilde{J}_m\}$, $j = 0$, $k = 0$.
- While $\mathcal{Z} \neq \emptyset$, do:

$$\begin{aligned} j &\leftarrow \min(\mathcal{Z}), \\ k &\leftarrow k + 1, \\ \tilde{H}_k &= \bigcup_{l \in \mathcal{N}(j)} \tilde{T}_j, \\ \mathcal{Z} &\leftarrow \mathcal{Z} \setminus \mathcal{N}(j), \end{aligned}$$

- $\tilde{K}_m = k$.

Remark that since $j \in \mathcal{N}(j)$, the pseudo-algorithm stops after at most $\tilde{J}_m$ iterations. The family of sets $\tilde{H}_1, \ldots, \tilde{H}_{\tilde{K}_m}$ is such that $d_\infty(\tilde{H}_k, \tilde{H}_{k'}) > 2(\log m)^{-1}$, $\forall\, k \neq k'$. The $\tilde{H}_k$'s correspond to the estimators of the r_0-connected components of Γ. The next proposition states that that the $\tilde{H}_k$'s are consistent estimators of the r_0-connected components of $\Gamma(\lambda)$.

Let $\mathcal{J}$ be a subset of $\{1, \ldots, J\}$. Define $\kappa(j) = \{k = 1, \ldots, \tilde{K}_m : \tilde{H}_k \cap T_j \neq \emptyset\}$ and let $D(\mathcal{J})$ be the event on which the sets $\kappa(j), j \in \mathcal{J}$ are reduced to singletons $\{k(j)\}$ that are disjoint, i.e.,

$$
\begin{aligned}
D(\mathcal{J}) &= \Big\{\kappa(j) = \{k(j)\}, k(j) \neq k(j'),\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\} \\
&= \Big\{\kappa(j) = \{k(j)\},\ (T_j \cup \tilde{H}_{k(j)}) \cap (T_{j'} \cup \tilde{H}_{k(j')}) = \emptyset,\ \forall\, j, j' \in \mathcal{J}, j \neq j'\Big\}.
\end{aligned} \tag{7.11}
$$

In other words, on the event $D(\mathcal{J})$, there is a one-to-one correspondence between the collection $\{T_j\}_{j\in\mathcal{J}}$ and the collection $\big\{\{\tilde{H}_k\}_{k\in\kappa(j)}\big\}_{j\in\mathcal{J}}$. Componentwise convergence of $\tilde{G}_m$ to Γ, is ensured when $D(\{1, \ldots, J\})$ has asymptotically overwhelming probability. The following proposition gives an upper bound on the probability of the complement of the event $D(\mathcal{J})$.

PROPOSITION 7.4. *Fix $r_0 > 0, s_0 > 0$ and assume that there exists a version of Γ that admits a decomposition into a number $J \geq 1$ of r_0-connected components $\Gamma = \bigcup_{j=1}^{J} T_j$ where the $\{T_j\}_{j=1,\ldots,J}$ are two by two s_0-separated. Consider an estimator $\hat{G}_m$ of Γ that is consistent from inside at rate $m^{-\alpha}$. Denote by $\{\tilde{H}_k\}_{k=1,\ldots,\tilde{K}_m}$ the family of sets obtained by the clipping and merging steps described above. Then, for any $\mathcal{J} \subset \{1, \ldots, J\}$, we have*

$$
\mathbb{P}_m\big(D^c(\mathcal{J})\big) = \widetilde{O}\big(m^{-\alpha}\big)\,,
$$

where $\mathcal{D}(\mathcal{J})$ is defined in (7.11).

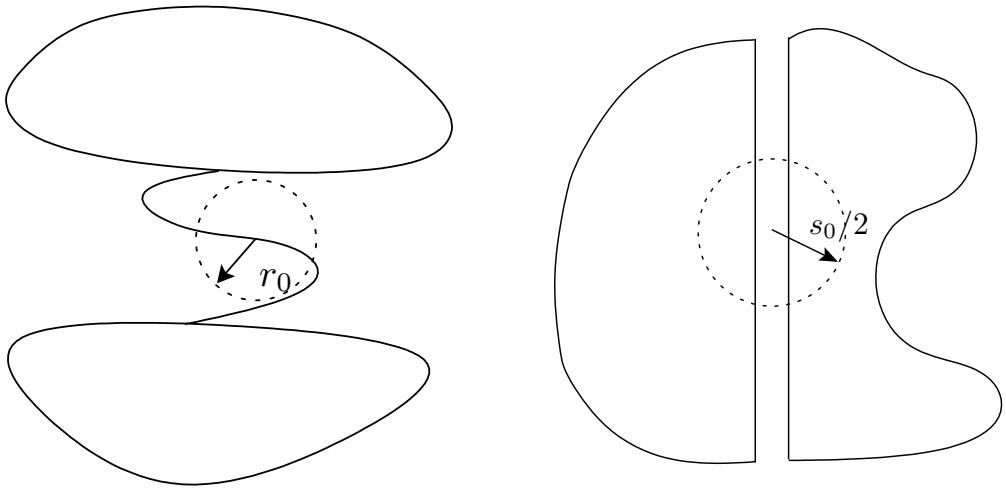

FIGURE 1. Set that is 0-connected but not r_0-connected for any $r_0 > 0$ (left) and non-separated connected components (right).

4.3. Labeling the clusters. To estimate the homogeneous regions, we will simply estimate the connected components of Γ and apply the clipping and merging steps described above. Then we make a majority vote on each homogeneous region. It yields the following procedure.

THREE-STEP PROCEDURE

(1) Use the unlabeled data $\mathbb{X}_u$ to construct an estimator $\hat{G}_m$ of Γ that is consistent from inside at rate $m^{-\alpha}$.
(2) Define homogeneous regions as the unions of the connected components of $\tilde{G}_m = \hat{G}_m \setminus \mathrm{Clip}(\hat{G}_m)$ (clipping step) that are closer than $2(\log m)^{-1}$ for the distance d_∞ using pseudo-algorithm MERGING.
(3) Assign a single label to each estimated homogeneous region by majority vote on labeled data.

This method translates into two distinct error terms, one term in m and another term in n. We apply our three-step procedure to build a classifier $\tilde{g}_{n,m}$ based on the pooled sample $(\mathbb{X}_l, \mathbb{X}_u)$. Fix $\alpha > 0$ and let $\hat{G}_m$ be an estimator of the density level set Γ, that is consistent from inside at rate $m^{-\alpha}$. For any $1 \leq k \leq \tilde{K}_m$, define the random variable

$$Z_{n,m}^k = \sum_{i=1}^{n} (2Y_i - 1)\, \mathbb{1}_{\{X_i \in \tilde{H}_k\}}\,,$$

where $\tilde{H}_k$ is obtained by the clipping and merging steps defined in the previous subsection. Denote by $\tilde{g}_{n,m}^k$ the function $\tilde{g}_{n,m}^k(x) = \mathbb{1}_{\{Z_{n,m}^k > 0\}}$ for all $x \in \tilde{H}_k$ and consider the classifier defined on $\mathcal{X}$ by

$$\tilde{g}_{n,m}(x) = \sum_{k=1}^{\tilde{K}_m} \tilde{g}_{n,m}^k(x) \mathbb{1}_{\{x \in \tilde{H}_k\}}, \quad x \in \mathcal{X}\,. \tag{7.12}$$

Note that the classifier $\tilde{g}_{n,m}$ assigns label 0 to any x outside of $\tilde{G}_m$. This is a notational convention and we can assign any value to x on this set since we are only interested in the cluster excess-risk. Nevertheless, it is more appropriate to assign a label referring to a rejection, e.g., the values "2"or "R" (or any other value different from $\{0, 1\}$). The rejection meaning that this point should be classified using labeled data only. However, when the amount of labeled data is too small, it might be more reasonable not to classify this point at all. This modification is of particular interest in the context of classification with a rejection option when the cost of rejection is smaller than the cost of misclassification (see, e.g., Herbei and Wegkamp, 2006). Remark that when there is only a finite number of clusters, there exists $\delta > 0$ such that

$$\delta = \min_{j=1,\ldots,J} \delta_j\,. \tag{7.13}$$

THEOREM 7.2. *Fix $\alpha > 0, r_0 > 0$ and assume that* **(CA2)** *holds. Consider an estimator $\hat{G}_m$ of Γ, based on $\mathbb{X}_u$ that is consistent from inside at rate $m^{-\alpha}$. Then, the classifier $\tilde{g}_{n,m}$ defined in* (7.12) *satisfies*

$$\mathcal{E}_\Gamma\left(\tilde{g}_{n,m}\right) \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + \sum_{j=1}^{J} \delta_j \mathrm{e}^{-n(\theta\delta_j)^2/2}\,, \leq \widetilde{O}\left(\frac{m^{-\alpha}}{1-\theta}\right) + \mathrm{e}^{-n(\theta\delta)^2/2} \tag{7.14}$$

for any $0 < \theta < 1$ and where $\delta > 0$ is defined in (7.13).

Note that, since we often have $m \gg n$, the first term in the RHS of (7.14) can be considered negligible so that we achieve an exponential rate of convergence in n which is

almost the same (up to the constant θ in the exponent) as in the case where the clusters are completely known. The constant θ seems to be natural since it balances the two terms.

5. Plug-in rules for density level sets estimation

Fix $\lambda > 0$ and recall that our goal is to use the unlabeled sample $\mathbb{X}_u$ of size m to construct an estimator $\hat{G}_m$ of $\Gamma = \Gamma(\lambda) = \{x \in \mathcal{X} : p(x) \geq \lambda\}$, that is consistent from inside at rate $m^{-\alpha}$ for some $\alpha > 0$ that should be as large as possible. A simple and intuitive way to achieve this goal is to use *plug-in estimators* of Γ defined by

$$\hat{\Gamma} = \hat{\Gamma}(\lambda) = \{x \in \mathcal{X} : \hat{p}_m(x) \geq \lambda\} \ ,$$

where $\hat{p}_m$ is some estimator of p. A straightforward generalization are the *penalized plug-in estimators* of $\Gamma(\lambda)$, defined by

$$\tilde{\Gamma}_\ell = \tilde{\Gamma}_\ell(\lambda) = \{x \in \mathcal{X} : \hat{p}_m(x) \geq \lambda + \ell\} \ ,$$

where $\ell > 0$ is a penalization. Clearly, we have $\tilde{\Gamma}_\ell \subset \hat{\Gamma}$. Keeping in mind that we want estimators that are consistent from inside we are going to consider sufficiently large penalization $\ell = \ell(m)$.

Plug-in rules is not the only choice for density level set estimation. Direct methods such as empirical excess mass maximization (see, e.g., Polonik, 1995; Tsybakov, 1997; Steinwart *et al.*, 2005) are also popular. One advantage of plug-in rules over direct methods is that once we have an estimator $\hat{p}_m$, we can compute the whole collection $\{\tilde{\Gamma}_\ell(\lambda), \lambda > 0\}$, which might be of interest for the user who wants to try several values of λ. Note also that a wide range of density estimators is available in usual software. A density estimator can be parametric, typically based on a mixture model, or nonparametric such as histograms or kernel density estimators.

The next assumption has been introduced in Polonik (1995). It is an analog of the margin assumption formulated in Mammen and Tsybakov (1999) and Tsybakov (2004b) but for arbitrary level λ in place of $1/2$.

Definition 7.3. *For any $\lambda, \gamma \geq 0$, a function $f : \mathcal{X} \to \mathbb{R}$ is said to have* γ-exponent at level λ *if there exists a constant $c^\star > 0$ such that, for all $\varepsilon > 0$,*

$$\mathrm{Leb}_d \{x \in \mathcal{X} : |f(x) - \lambda| \leq \varepsilon\} \leq c^\star \varepsilon^\gamma \, .$$

When $\gamma > 0$ it ensures that the function f has no flat part at level λ.

The next theorem gives fast rates of convergence for penalized plug-in rules when $\hat{p}_m$ satisfies an exponential inequality and p has γ-exponent at level λ. Moreover, it ensures that when the penalization ℓ is suitably chosen, the plug-in estimator is consistent from inside.

Theorem 7.3. *Fix $\lambda > 0, \gamma > 0$ and $\Delta > 0$. Let $\hat{p}_m$ be an estimator of the density p based on the sample $\mathbb{X}_u$ of size $m \geq 1$ and let $\mathcal{P}$ be a class of densities on $\mathcal{X}$. Assume that there exist positive constants c_1, c_2 and $a \leq 1$, such that for P_X-almost all $x \in \mathcal{X}$, we have*

$$\sup_{p \in \mathcal{P}} \mathbb{P}_m \left(|\hat{p}_m(x) - p(x)| \geq \delta \right) \leq c_1 \mathrm{e}^{-c_2 m^a \delta^2} \ , \ m^{-a/2} < \delta < \Delta \, . \tag{7.15}$$

Assume further that p has γ-exponent at level λ for any $p \in \mathcal{P}$ and that the penalty ℓ is chosen as

$$\ell = \ell(m) = m^{-\frac{a}{2}} \log m \, . \tag{7.16}$$

Then the plug-in estimator $\tilde{\Gamma}_\ell$ is consistent from inside at rate $m^{-\frac{\gamma a}{2}}$ for any $p \in \mathcal{P}$.

Consider a kernel density estimator $\hat{p}_m^K$ based on the sample $\mathbb{X}_u$ defined by

$$\hat{p}_m^K(x) = \frac{1}{mh^d} \sum_{i=n+1}^{n+m} K\left(\frac{X_i - x}{h}\right), \quad x \in \mathcal{X}, \tag{7.17}$$

where $h > 0$ is the bandwidth parameter and $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel. If p is assumed to have Hölder smoothness parameter $\beta > 0$ and if K and h are suitably chosen, it is a standard exercise to prove inequality of type (7.15) with $a = 2\beta/(2\beta + d)$. In that case, it can be shown that the rate $m^{-\frac{\gamma a}{2}}$ is optimal in a minimax sense (see Chapter 8).

6. Discussion

We proposed a formulation of the cluster assumption in probabilistic terms. This formulation relies on Hartigan's (Hartigan, 1975) definition of clusters but it can be modified to match other definitions of clusters.

We also proved that there is no hope to improve the classification performance outside of these clusters. Based on these remarks, we defined the cluster excess-risk which can be easily generalized to the setup of general clusters defined above. Finally we proved that when we have consistent estimators of the clusters, it is possible to achieve exponential rates of convergence for the cluster excess-risk. The theory developed here can be extended to any definition of clusters as long as they can be consistently estimated.

Note that our definition of clusters is parametrized by λ which is left to the user, depending on his trust in the cluster assumption. Indeed, density level sets have the monotonicity property: $\lambda \geq \lambda'$, implies $\Gamma(\lambda) \subset \Gamma(\lambda')$. In terms of the cluster assumption, it means that when λ decreases to 0, the assumption **(CA2)** concerns bigger and bigger sets $\Gamma(\lambda)$ and in that sense, it becomes more and more restrictive. As a result, the parameter λ can be considered as a level of confidence characterizing to which extent the cluster assumption is valid for the distribution P and its choice is left to the user.

The choice of λ can be made by fixing $P_X(\mathcal{C})$, where $\mathcal{C}$ is defined in (7.1), the probability of the rejection region. We refer to Cuevas *et al.* (2001) for more details. Note that data-driven choices of λ could be easily derived if we impose a condition on the purity of the clusters, i.e., if we are given the δ in (7.13). Such a choice could be made by decreasing λ until the level of purity is attained. However, any data-driven choice of λ has to be made using the labeled data. It would therefore yield much worse bounds when $n \ll m$.

This paper is an attempt to give a proper mathematical framework for the cluster assumption proposed in Seeger (2000). As mentioned above, the definition of clusters we use here is one among several available and it could be interesting to modify the formulation of the cluster assumption to match other definitions of cluster. In particular, the definition of cluster as r_0-connected components of the λ-level set of the density leaves the problem of choosing λ correctly.

7. Proofs

7.1. Proof of Proposition 7.1. Since the distribution of the unlabeled sample $\mathbb{X}_u$ does not depend on η, we have for any marginal distribution P_X,

$$\begin{aligned}\sup_{\eta\in\Xi}\mathbb{E}_{n,m}\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X &= \sup_{\eta\in\Xi}\mathbb{E}_m\mathbb{E}_n\Big[\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X\big|\mathbb{X}_u\Big]\\ &= \mathbb{E}_m\sup_{\eta\in\Xi}\mathbb{E}_n\Big[\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_{n,m}\neq g^\star\}}\mathrm{d}P_X\big|\mathbb{X}_u\Big]\\ &\geq \inf_{T_n}\sup_{\eta\in\Xi}\mathbb{E}_n\int_{\mathcal{C}^c}|2\eta-1|\mathbb{1}_{\{T_n\neq g^\star\}}\mathrm{d}P_X\,,\end{aligned}$$

where in the last inequality, we used the fact that conditionally on $\mathbb{X}_u$, the classifier $T_{n,m}$ only depends on $\mathbb{X}_l$ and can therefore be written T_n.

7.2. Proof of Theorem 7.1. We can decompose $\mathcal{E}_{\mathcal{C}}(\hat{g}_n)$ into

$$\mathcal{E}_{\mathcal{C}}(\hat{g}_n)=\mathbb{E}_n\sum_{j\geq 1}\int_{T_j}|2\eta(x)-1|\mathbb{1}_{\{\hat{g}_n^j(x)\neq g^\star(x)\}}p(x)\mathrm{d}x\,.$$

Fix $j\in\{1,2,\ldots\}$ and assume w.l.o.g. that $\eta\geq 1/2$ on T_j. It yields $g^\star(x)=1,\ \forall\, x\in T_j$, and since $\hat{g}_n$ is also constant on T_j, we get

$$\begin{aligned}\int_{T_j}|2\eta(x)-1|\mathbb{1}_{\{\hat{g}_n^j(x)\neq g^\star(x)\}}p(x)\mathrm{d}x &= \mathbb{1}_{\{Z_n^j\leq 0\}}\int_{T_j}(2\eta(x)-1)p(x)\mathrm{d}x\\ &\leq \delta_j\mathbb{1}_{\left\{|\delta_j-\frac{Z_n^j}{n}|\geq\delta_j\right\}}\,.\end{aligned}\tag{7.18}$$

Taking expectation $\mathbb{E}_n$ on both sides of (7.18) we get

$$\begin{aligned}\mathbb{E}_n\int_{T_j}|2\eta(x)-1|\mathbb{1}_{\{\hat{g}_n^j(x)\neq g^\star(x)\}}p(x)\mathrm{d}x &\leq \delta_j\mathbb{P}_n\Big[\big|\delta_j-\frac{Z_n^j}{n}\big|\geq\delta_j\Big]\\ &\leq 2\delta_j\mathrm{e}^{-n\delta_j^2/2}\,,\end{aligned}\tag{7.19}$$

where we used Hoeffding's inequality to get the last bound. Summing now over j yields the theorem.

7.3. Proof of Lemma 7.1. We have to prove three points: reflexivity, symmetry and transitivity. Reflexivity is obvious from the definition of a standard set. Next, remark that if $x\xleftrightarrow[C]{r_0}x'$, there exists a continuous map $f_1:[0,1]\to\mathcal{X}$ such that $f_1(0)=x, f_1(1)=x'$ and for any $t\in[0,1]$ and any $0<r\leq r_0$, we have

$$\mathrm{Leb}_d\big(\mathcal{B}(f_1(t),r)\cap C\big)\geq c_0 r^{\bar{d}}\,.$$

To prove symmetry, it is sufficient to consider the continuous map $\tilde{f}_1$ defined by $\tilde{f}_1(t)=f_1(1-t)$ for any $t\in[0,1]$. We now prove transitivity. Assume also that $x'\xleftrightarrow[C]{r_0}x''$, i.e., there exists a continuous map $f_2:[0,1]\to\mathcal{X}$ such that $f_2(0)=x', f_2(1)=x''$ and for any $t\in[0,1]$ and any $0<r\leq r_0$, we have

$$\mathrm{Leb}_d\big(\mathcal{B}(f_2(t),r)\cap C\big)\geq c_0 r^{\bar{d}}\,.$$

Define now the map $f:[0,1]\to\mathcal{X}$ by:

$$f(t)=\begin{cases}f_1(2t) & \text{if } t\in[0,1/2]\\ f_2(2t-1) & \text{if } t\in[1/2,1]\end{cases}$$

This map is obviously continuous on $[0,1]$ and satisfies $f(0)=x, f(1)=x''$. Moreover, for any $t\in[0,1]$, we have

$$\mathrm{Leb}_d\big(\mathcal{B}(f(t),r)\cap C\big)\geq c_0 r^{\bar{d}}\,.$$

The second assertion in the lemma is trivial.

7.4. Proof of Proposition 7.2. From Lemma 7.1, we know that C can be decomposed into is classes of equivalence. Fix $k\geq 1$ and assume that there is at least k classes of equivalence that we denote by $C_1,\ldots,C_k$. Since the classes are assumed to be s_0-separated, for any $1\leq j\leq k$, for any $x_j\in C_j$, the Euclidean balls $\mathcal{B}(x_j,s_0/2)$ are disjoint. Thus

$$\infty>\mathrm{Leb}_d(\mathcal{X})\geq\sum_{j=1}^{k}\mathrm{Leb}_d\big[\mathcal{B}(x_j,s_0/2)\cap\mathcal{X}\big]\geq\sum_{j=1}^{k}\mathrm{Leb}_d\big[\mathcal{B}(x_j,s_0/2)\cap C\big]\geq ck\,,$$

for a positive constant c. Thus we must have a finite decomposition.

7.5. Proof of Proposition 7.3. Consider a regular grid $\mathcal{G}$ on $\mathbb{R}^d$ with step size $1/\log(m)$ and let $c_1>0$ be a constant such that the Euclidean balls of centers in $\tilde{\mathcal{G}}=\mathcal{G}\cap\mathrm{Clip}(\hat{G}_m)$ cover the set $\hat{G}_m$. Since $\mathrm{Leb}_d(\mathcal{X})<\infty$, there exists a constant $c_2>0$ such that $\mathrm{card}\{\tilde{\mathcal{G}}\}\leq c_2(\log m)^d$. Therefore

$$\mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m))\leq\sum_{x\in\tilde{\mathcal{G}}}\mathrm{Leb}_d\big(\mathcal{B}(x,1/\log(m))\cap\hat{G}_m\big)\leq\frac{c_2(\log m)^{\bar{d}-d}}{m^\alpha}\,.$$

Therefore, the rate of convergence $\tilde{G}_m$ is the same as that of $\hat{G}_m$. We conclude the proof by observing that $\tilde{G}_m\subset\hat{G}_m$, so that $\tilde{G}_m$ is also consistent from inside.

7.6. Proof of Proposition 7.4. Define $m_0=\exp(1/(r_0\wedge s_0))$ and denote $D(\mathcal{J})$ by D. For any $j=1,\ldots,J$, the r_0 connectedness of T_j yields on the one hand,

$$\begin{aligned}A_1(j)=\{\mathrm{card}[\kappa(j)]=0\}\quad&\subset\quad\big\{\mathrm{Leb}_d\big[\tilde{G}_m\triangle\Gamma\big]>c_0(\log m)^{-\bar{d}}\big\}\,,\\ A_2(j)=\{\mathrm{card}[\kappa(j)]\geq 2\}\quad&\subset\quad\big\{\mathrm{Leb}_d\big[\tilde{G}_m\triangle\Gamma\big]>c_0(\log m)^{-\bar{d}}\big\}\,.\end{aligned}$$

The previous inclusions are illustrated in Figure 2.
On the other hand, $\kappa(j)\cap\kappa(j')\neq\emptyset$ for some $j'\neq j$ when either (i) $\exists\, l$ s.t. $\tilde{T}_l\cap T_j\neq\emptyset$, $\tilde{T}_l\cap T_{j'}\neq\emptyset$ or (ii) $\exists\, l\neq l'$ s.t. $\tilde{T}_l\cap T_j\neq\emptyset$, $\tilde{T}_{l'}\cap T_{j'}\neq\emptyset$ and $d_\infty(\tilde{T}_l,\tilde{T}_{l'})<2(\log m)^{-1}$. Both cases yield the existence of $x\in\Gamma^c\cap\tilde{G}_m$ such that $\mathcal{B}(x,(\log m)^{-1})\subset\Gamma^c$ for $m\geq m_0$. Therefore

$$\mathrm{Leb}_d(\tilde{G}_m\cap\Gamma^c)\geq\mathrm{Leb}_d(\tilde{G}_m\cap\mathcal{B}(x,(\log m)^{-1}))\,.$$

Since $x\in\tilde{G}_m$, we have $\mathrm{Leb}_d(\hat{G}_m\cap\mathcal{B}(x,(\log m)^{-1}))\geq m^{-\alpha}(\log m)^{-d}$. On the other hand, we have

$$\begin{aligned}\mathrm{Leb}_d(\tilde{G}_m\cap\mathcal{B}(x,\frac{1}{\log m}))&=\mathrm{Leb}_d(\hat{G}_m\cap\mathcal{B}(x,\frac{1}{\log m}))-\mathrm{Leb}_d(\mathrm{Clip}(\hat{G}_m)\cap\mathcal{B}(x,\frac{1}{\log m}))\\&\geq m^{-\alpha}(\log m)^{-d}-\mathrm{Leb}_d(\hat{G}_m\cap\Gamma^c)\\&\geq m^{-\alpha}(\log m)^{-d}-c_3m^{-2\alpha}\\&\geq c_4m^{-\alpha}(\log m)^{-d}\,,\end{aligned}$$

where we used the fact that $\hat{G}_m$ is consistent from inside at rate $m^{-\alpha}$. Hence,

$$A_3(j)=\bigcup_{j'\neq j}\{\kappa(j)\cap\kappa(j')\neq\emptyset\}\subset\big\{\mathrm{Leb}_d(\tilde{G}_m\cap\Gamma^c)\geq c_5m^{-\alpha}(\log m)^{-d}\big\}\,.$$

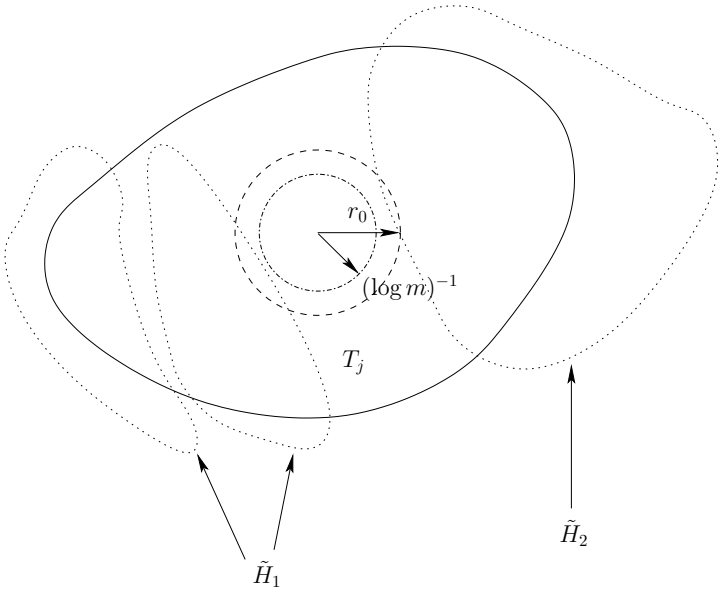

FIGURE 2. By construction, $\tilde{H}_1$ and $\tilde{H}_2$ are separated by a ball of radius $(\log m)^{-1}$, which is included in $\mathcal{B}(x, r_0)$ when $m \geq m_0$. So if $\{1, 2\} \subset \kappa(j)$ or $\kappa(j) = \emptyset$, this ball is included in in $\tilde{G}_m \triangle \Gamma$.

Both cases are illustrated in Figure 3.

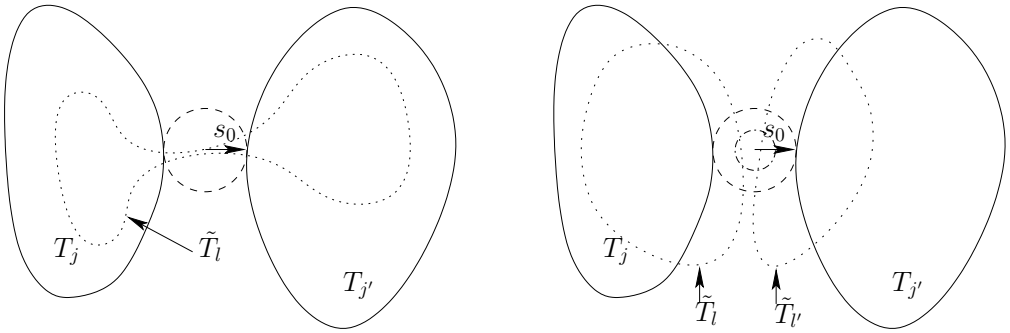

FIGURE 3. Case (i) (left) and case (ii) (right).

Now, since

$$D^c = \bigcup_{j=1}^{J} A_1(j) \cup A_2(j) \cup A_3(j),$$

we get

$$\mathbb{P}_m(D^c) \leq \mathbb{P}_m\left\{\mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > c_0(\log m)^{-\bar{d}}\right\} + \mathbb{P}_m\left\{\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq c_5 m^{-\alpha}(\log m)^{-d}\right\}.$$

Using the Markov inequality for both terms we obtain

$$\mathbb{P}_m\left\{\mathrm{Leb}_d\left[\tilde{G}_m \triangle \Gamma\right] > c_0(\log m)^{-\bar{d}}\right\} = \widetilde{O}\left(m^{-\alpha}\right).$$

and

$$\mathbb{P}_m\left\{\mathrm{Leb}_d(\tilde{G}_m \cap \Gamma^c) \geq c_5 m^{-\alpha}(\log m)^{-d}\right\} = \widetilde{O}\left(m^{-\alpha}\right)$$

where we used the fact that $\tilde{G}_m$ is consistent from inside with rate $m^{-\alpha}$. It yields the statement of the proposition.

7.7. Proof of Theorem 7.2. The cluster excess-risk $\mathcal{E}_\Gamma(\tilde{g}_{n,m})$ can be decomposed w.r.t the event D and its complement. It yields

$$\mathcal{E}_\Gamma(\tilde{g}_{n,m}) \leq \mathbb{E}_m\Big[\mathbb{1}_D \mathbb{E}_n\Big(\int_\Gamma |2\eta(x)-1| \mathbb{1}_{\{\tilde{g}_{n,m}(x)\neq g^\star(x)\}} p(x)\mathrm{d}x \Big| \mathbb{X}_u\Big)\Big] + \mathbb{P}_m\left(D^c\right) .$$

We now treat the first term of the RHS of the above inequality, i.e., on the event D. Fix $j \in \{1,\dots,J\}$ and assume w.l.o.g. that $\eta \geq 1/2$ on T_j. Simply write Z^k for $Z^k_{m,n}$. By definition of D, there is a one-to-one correspondence between the collection $\{T_j\}_j$ and the collection $\{\tilde{H}_k\}_k$. We denote by $\tilde{H}_j$ the unique element of $\{\tilde{H}_k\}_k$ such that $\tilde{H}_j \cap T_j \neq \emptyset$. On D, for any $j = 1,\dots,J$, we have,

$$\begin{aligned}
&\mathbb{E}_n\Big(\int_{T_j} |2\eta(x)-1| \mathbb{1}_{\{\tilde{g}^j_{n,m}(x)\neq g^\star(x)\}} p(x)\mathrm{d}x \Big| \mathbb{X}_u\Big) \\
&\qquad \leq \int_{T_j\setminus\tilde{G}_m} (2\eta-1)\mathrm{d}P_X + \mathbb{E}_n\Big(\mathbb{1}_{\{Z^j\leq 0\}} \int_{T_j\cap\tilde{H}_j} (2\eta-1)\mathrm{d}P_X \Big| \mathbb{X}_u\Big) \\
&\qquad \leq L(p)\mathrm{Leb}_d(T_j\setminus\tilde{G}_m) + \delta_j \mathbb{P}_n\big(Z^j \leq 0 | \mathbb{X}_u\big) .
\end{aligned}$$

On the event D, for any $0<\theta<1$, it holds

$$\begin{aligned}
\mathbb{P}_n\big(Z^j \leq 0|\mathbb{X}_u\big) &= \mathbb{P}_n\Big(\int_{T_j} (2\eta-1)\mathrm{d}P_X - Z^j \geq \delta_j | \mathbb{X}_u\Big) \\
&\leq \mathbb{P}_n\Big(\big|Z^j - \int_{\tilde{H}_j} (2\eta-1)\mathrm{d}P_X\big| \geq \theta\delta_j | \mathbb{X}_u\Big) \\
&\qquad + \mathbb{1}_{\Big\{P_X\big[T_j\triangle\tilde{H}_j\big]\geq(1-\theta)\delta_j\Big\}} .
\end{aligned}$$

Using Hoeffding's inequality to control the first term, we get

$$\mathbb{P}_n\big(Z^j \leq 0|\mathbb{X}_u\big) \leq 2\mathrm{e}^{-n(\theta\delta_j)^2/2} + \mathbb{1}_{\Big\{P_X\big[T_j\triangle\tilde{H}_j\big]\geq(1-\theta)\delta_j\Big\}} .$$

Taking expectations, and summing over j, the cluster excess-risk is upper bounded by

$$\mathcal{E}_\Gamma(\tilde{g}_{n,m}) \leq \frac{2L(p)}{1-\theta}\mathbb{E}_m\Big[\mathrm{Leb}_d(\Gamma\triangle\tilde{G}_m)\Big] + 2\sum_{j=1}^J \delta_j \mathrm{e}^{-n(\theta\delta_j)^2/2} + \mathbb{P}_m\left(D^c\right) ,$$

where we used the fact that on D,

$$\sum_{j=1}^J \mathrm{Leb}_d\big[T_j\triangle\tilde{H}_j\big] \leq \mathrm{Leb}_d\big[\Gamma\triangle\tilde{G}_m\big] .$$

From Proposition 7.4, we have $\mathbb{P}_m\left(D^c\right) = \widetilde{O}\left(m^{-\alpha}\right)$ and $\mathbb{E}_m\Big[\mathrm{Leb}_d(\Gamma\triangle\tilde{G}_m)\Big] = \widetilde{O}\left(m^{-\alpha}\right)$ and the theorem is proved.

7.8. Proof of Theorem 7.3. Recall that

$$\tilde{\Gamma}_\ell \triangle \Gamma = \Big(\tilde{\Gamma}_\ell \cap \Gamma^c\Big) \cup \Big(\tilde{\Gamma}^c_\ell \cap \Gamma\Big) .$$

We begin by the first term. We have

$$\tilde{\Gamma}_\ell \cap \Gamma^c = \big\{x\in\mathcal{X} : \hat{p}_m(x) \geq \lambda+\ell, p(x)<\lambda\big\} \subset \big\{x\in\mathcal{X} : |\hat{p}_m(x)-p(x)| \geq \ell\big\} .$$

The Fubini theorem yields

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell\cap\Gamma^c)\big] \leq \mathrm{Leb}_d(\mathcal{X}) \sup_{x\in\mathcal{X}} \mathbb{P}_m\left[|\hat{p}_m(x)-p(x)| \geq \ell\right] \leq c_6\mathrm{e}^{-c_2 m^a \ell^2} ,$$

where the last inequality is obtained using (7.15) and $c_6 = c_1 \mathrm{Leb}_d(\mathcal{X}) > 0$. Taking ℓ as in (7.16) yields for $m \geq \exp(\gamma a / c_2)$,

$$\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\big] \leq c_6 m^{-\gamma a}. \tag{7.20}$$

We now prove that $\mathbb{E}_m\big[\mathrm{Leb}_d(\tilde{\Gamma}_\ell \cap \Gamma^c)\big] = \widetilde{O}\big(m^{-\frac{\gamma a}{2}}\big)$. Consider the following decomposition where we drop the dependence in x for notational convenience,

$$\tilde{\Gamma}_\ell^c \cap \Gamma = B_1 \cup B_2,$$

where

$$B_1 = \big\{\hat{p}_m < \lambda + \ell, p \geq \lambda + 2\ell\big\} \subset \big\{|\hat{p}_m - p| \geq \ell\big\}$$

and

$$B_2 = \big\{\hat{p}_m < \lambda + \ell, \lambda \leq p(x) < \lambda + 2\ell\big\} \subset \big\{|p - \lambda| \leq \ell\big\}.$$

Using (7.15) and (7.16) in the same fashion as above we get $\mathbb{E}_m\big[\mathrm{Leb}_d(B_1)\big] = \widetilde{O}\big(m^{-\gamma a}\big)$. The term corresponding to B_2 is controlled using the γ-exponent of density p at level λ. Indeed, we have

$$\mathrm{Leb}_d(B_2) \leq c^\star \ell^\gamma = c^\star (\log m)^\gamma m^{-\frac{\gamma a}{2}} = \widetilde{O}\big(m^{-\frac{\gamma a}{2}}\big).$$

The previous upper bounds for $\mathrm{Leb}_d(B_1)$ and $\mathrm{Leb}_d(B_2)$ together with (7.20) yield the consistency from inside.

CHAPTER 8

Fast rates for plug-in estimators of density level sets

In the context of density level set estimation, we recall the notion of γ-exponent of a density at a certain level. This notion is similar to Tsybakov's margin assumption and allows us to prove fast rates of convergence for general plug-in methods, up to order n^{-1} when the density is supposed to be smooth in a neighborhood of the level under consideration. Lower bounds proving optimality of the rates in a minimax sense are also provided. Finally, when the density jumps around the level under consideration, we show that exponential rates of convergence are attainable.

Contents

The material in this chapter is a joint work with Régis Vert.

1. Introduction

Let Q be a positive σ-finite measure on $\mathcal{X} \subseteq \mathbb{R}^d$. Consider i.i.d random vectors $(X_1, \ldots, X_n)$ with distribution P, having an unknown probability density p with respect to the measure Q. For a fixed $\lambda > 0$, we are interested in the estimation of the λ-*level set* of the density p:

$$\Gamma_p(\lambda) \triangleq \{x \in \mathcal{X} \ : \ p(x) \geq \lambda\} \ .$$

Throughout the chapter we fix $\lambda > 0$ and when no confusion is possible we use the notation $\Gamma(\lambda)$ or simply Γ instead of $\Gamma_p(\lambda)$. When Q is the Lebesgue measure on $\mathbb{R}^d$, density level sets typically correspond to minimum volume sets of given P-probability mass, as shown in Polonik (1997). More generally, if Q is an arbitrary probability distribution on $\mathbb{R}^d$, density level sets correspond to the critical regions of likelihood ratio (Neyman-Pearson) tests, which are known to be optimal for testing the hypothesis P versus the alternative Q.

Here are two possible applications of density level set estimation.

Anomaly detection: The goal is to detect an abnormal observation from a sample (see Schölkopf *et al.*, 2001). One way to deal with that problem is to assume that abnormal observations do not belong to a group of concentrated observations. In this framework, observations are considered as abnormal when they do not belong to $\Gamma(\lambda)$ for some fixed $\lambda \geq 0$. The special case $\lambda = 0$, which corresponds to support estimation has been examined by Devroye and Wise (1980). In the general case, λ can be considered as a tolerance level for anomalies: the smaller λ, the fewer observations are considered as being abnormal. In particular, it is often the case that the user has a fixed budget, allowing him to qualify only a limited fraction of the data as outliers.

Unsupervised or semi-supervised classification: These problems amount to identify areas where the observations are concentrated with possible use of some available labels for the semi-supervised case. More precisely it can be assumed that the connected components of $\Gamma(\lambda)$, for a fixed λ, are clusters of homogeneous observations as described in Hartigan (1975). Again, the bigger the λ, the smaller the clusters.

REMARK 8.1. *In both applications, the choice of λ can be left to the user depending on its tolerance to anomalies or the desired number of clusters.*

There are essentially two approaches towards estimating density level sets from the sample $(X_1, \ldots, X_n)$: *plug-in* methods where the density p in the expression for $\Gamma_p(\lambda)$ is replaced by its estimate computed from the sample, and *direct* methods which are based on empirical excess-mass maximization (see Hartigan, 1987; Müller and Sawitzki, 1987).

While local versions of direct methods have been deeply analyzed and proved to be optimal in a minimax sense, over a certain family of well-behaved distributions (see Tsybakov, 1997), and although reasonable implementations have been recently proposed (see for instance Steinwart *et al.*, 2005), they are still not very easy to use for practical purposes, compared to plug-in methods. Indeed, it is common to estimate density level sets for different level values -typically when the goal is to compute a density level set of prespecified probability mass (or acceptance rate) and unknown density level. In that case, using direct methods, one has to run an optimization procedure several times, one for different density level values, then choose a posteriori the most suited level according to the desired rejection rate. Plug-in methods do not involve such a complex process: the density estimation step is only performed once and the construction of a density level set estimate simply amounts to thresholding the density estimate at the desired level.

On the other hand, in the related context of binary classification where more theoretical advances have been developed, the different analyses proposed so far have mainly supported a belief in the superiority of direct methods. Yang (1999) shows that, under general assumptions, plug-in rules cannot achieve a classification error risk convergence rate faster than $O\big(1/\sqrt{n}\big)$ (where n is the size of the data sample), and suffer from the curse of dimensionality. On the contrary, under slightly different assumptions, direct methods achieve this rate $O\big(1/\sqrt{n}\big)$ whatever the dimensionality (see e.g. Vapnik, 1998; Devroye *et al.*, 1996; Tsybakov, 2004b), and can even reach faster convergence rates- up to $O\big(1/n\big)$- under *Tsybakov's margin assumption* (see Mammen and Tsybakov, 1999; Tsybakov, 2004b; Tsybakov and van de Geer, 2005; Tarigan and van de Geer, 2006). This contributed to raising some pessimism concerning plug-in methods. Nevertheless such a comparison between plug-in methods and direct methods is far from being legitimate, since

the aforementioned analyzes of both plug-in methods and direct ones have been carried out under the different sets of assumptions (those sets are not disjoint, but none of them is included in the other).

Recently, Audibert and Tsybakov (2005) have introduced a new type of assumption dealing with the smoothness of the *regression function* in the standard classification framework, under which they derive fast convergence rates- even faster than $O(1/n)$ in some situations- for plug-in classification rules based on local polynomial estimators. This new result reveals that plug-in methods should not be considered as inferior to direct methods and, more importantly, that this new type of assumption on the regression function is a critical point in the general analysis of classification procedures.

In this chapter we extend such positive results to the density level set estimation (DLSE) framework: we revisit the analysis of plug-in density level set estimators, and show that they can be also very efficient under smoothness assumptions on the underlying density function p. Related papers are Baíllo *et al.* (2001) and Baíllo (2003), who investigate plug-in rules based on a certain type of kernel density estimates. Baíllo (2003) derives almost sure rates of convergence for a quantity different from the one studied here. It is interesting to observe that she introduces a condition similar to the γ-exponent used here. The particular case $\lambda = 0$, corresponds to estimation of the support of density p and is often applied to anomaly detection. Following the pioneer paper of Devroye and Wise (1980), this problem has received more attention than the general case $\lambda \geq 0$ and has been treated using plug-in methods for example by Cuevas and Fraiman (1997). Unlike the previously cited papers, we derive fast rates of convergence and prove that these rates are optimal in a minimax sense.

A general plug-in approach has been studied previously by Molchanov (1998), where a result on the asymptotic distribution of the Hausdorff distance is given. In a recent paper, Cuevas *et al.* (2006) study general plug-in estimators of the level sets. Under very general assumptions they derive consistency with respect to the Hausdorff metric and the measure of the symmetric difference. However, this very general framework does not allow them to derive rates of convergence.

This chapter is organized as follows. Section 2 introduces the notation and definitions. Section 3 presents the main result, that is a new bound on the error of penalized plug-in estimators based on general density estimators that satisfy a certain exponential inequality. As an example we prove that kernel density estimators are valid for this method. The upper bounds are proved to be optimal in Section 4, where we prove the corresponding minimax lower bounds. Finally, we prove in Section 5 that exponential rates are attainable when the density jumps around the level under consideration.

2. Notation and Setup

For any vector $x \in \mathbb{R}^d$, denote by $x^{(j)}$ its jth coordinate, $j = 1, \ldots, d$. Denote by $\mathcal{B}_\alpha(x, r)$ the closed ball in $\mathcal{X}$ centered at $x \in \mathcal{X}$ and of radius $r > 0$ with respect to the norm $\|\cdot\|_\alpha$, $1 \leq \alpha < \infty$ defined by

$$\|x\|_\alpha \triangleq \Big(\sum_{i=1}^{d} |x^{(j)}|^\alpha \Big)^{1/\alpha}.$$

The probability and expectation with respect to the joint distribution of $(X_1, \ldots, X_n)$ are denoted by $\mathbb{P}$ and $\mathbb{E}$ respectively. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$ the sup-norm of f and by $\|f\| = \big(\int_{\mathbb{R}^d} f^2(x) \mathrm{d}x \big)^{1/2}$ its L_2-norm.

Throughout the chapter, we denote by c_j positive constants and by A^c the complement of the set A.

2.1. Penalized plug-in rules. For a fixed $\lambda > 0$, the plug-in estimator of $\Gamma(\lambda)$ is defined by

$$\hat{\Gamma}(\lambda) = \{\hat{p}_n \geq \lambda\} \ ,$$

where $\hat{p}_n$ is a nonparametric estimator of p. For example, $\hat{p}_n$ can be a kernel density estimator of p,

$$\hat{p}_n(x) = \hat{p}_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \ x \in \mathcal{X} \, ,$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a suitably chosen kernel and $h > 0$ is the bandwidth parameter.

It is sometimes the case, for some applications that $\hat{G}amma$ is required to be included in Γ with high probability (see, e.g., Rigollet, 2006b). For this reason, we consider in this chapter the more general family of *penalized plug-in rules* $\tilde{G}amma_{\ell_n}$ defined as follows:

$$\tilde{\Gamma}_{\ell_n} = \tilde{\Gamma}_{\ell_n}(\lambda) = \hat{\Gamma}(\lambda + \ell_n) = \{\hat{p}_n \geq \lambda + \ell_n\} \, ,$$

where (ℓ_n) is a non-negative sequence that typically tends to 0 as n tends to infinity. This family includes in particular the estimator $\hat{\Gamma}$ when ℓ_n is taken equal to 0.

2.2. Measures of performance. Recall that Q is a positive σ-finite measure on $\mathcal{X}$ and define the measure $\tilde{Q}_\lambda$ that has density $|p(\cdot) - \lambda|$ with respect to Q. To assess the performance of a density level set estimator, we use the two pseudo-distances between closed sets G_1 and $G_2 \subseteq \mathcal{X}$:

(i) The Q-measure of the symmetric difference between G_1 and G_2:

$$d_\triangle(G_1, G_2) = Q(G_1 \triangle G_2) \, .$$

(ii) The $\tilde{Q}_\lambda$-measure of the symmetric difference between G_1 and G_2:

$$d_H(G_1, G_2) = \tilde{Q}_\lambda(G_1 \triangle G_2) = \int_{G_1 \triangle G_2} |p(x) - \lambda| \mathrm{d}Q(x) \, .$$

The quantity $d_\triangle(G_1, G_2)$ is a standard and natural way to measure the distance between two sets G_1 and G_2. Another way to measure the quality of $\hat{G}$ is to compute its *excess-mass* $H(\hat{G})$ defined as follows (Hartigan, 1987; Müller and Sawitzki, 1987):

$$H\big(\hat{G}\big) = P\big(\hat{G}\big) - \lambda Q\big(\hat{G}\big) \ .$$

Excess-mass measures how the P-probability mass concentrates in the region $\hat{G}$, and it is maximized by $\Gamma = \Gamma(\lambda)$. Hence, it acts as a risk functional in the DLSE framework and it is natural to measure the performance of an estimator $\hat{G}$ by its *excess-mass deficit* $H(\Gamma) - H(\hat{G}) \geq 0$. Further justifications for the well-foundedness of the excess mass criterion can be found in Polonik (1995). Note that for any measurable set $G \subseteq \mathcal{X}$, the excess-mass $H(G)$ can be written

$$H(G) = \int_G (p(x) - \lambda) \, \mathrm{d}Q(x) \, .$$

Thus, we can rewrite,

$$\begin{aligned} H(\Gamma) - H(\hat{G}) &= \int_{\mathcal{X}} \left(\mathbb{1}_{\{p(\cdot) \geq \lambda\}}(x) - \mathbb{1}_{\hat{G}}(x) \right) (p(x) - \lambda) \, \mathrm{d}Q(x) \\ &= \int_{\Gamma \triangle \hat{G}} |p(x) - \lambda| \mathrm{d}Q(x) = d_H(\hat{G}, \Gamma) \, . \end{aligned}$$

This explains the notation d_H. The next definition allows us to link d_H to $d_\triangle$.

DEFINITION 8.1. *For any $\lambda, \gamma \geq 0$, a function $f : \mathcal{X} \to \mathbb{R}$ is said to have γ-**exponent at level** λ with respect to Q if there exist constants $c_0 > 0$ and $\varepsilon_0 > 0$ such that, for all $0 < \varepsilon \leq \varepsilon_0$,*

$$Q\{x \in \mathcal{X} : |f(x) - \lambda| \leq \varepsilon\} \leq c_0 \varepsilon^\gamma .$$

The exponent γ controls the slope of the density around level λ. When $\gamma = 0$, the condition is loose and the density is allowed to have flat parts at level λ. When γ is positive, the function f has no flat part at level λ. A standard case corresponds to $\gamma = 1$, arising for instance in the case where the gradient of f has a coordinate bounded away from 0 in a neighborhood of $\{f = \lambda\}$. If $\gamma < 1$, we call λ a *critical level*. To illustrate such cases, assume that Q is the Lebesgue measure on $\mathbb{R}^d$ denoted here by Leb_d. Let q be a positive number such that $q > d$ and consider a density p on $\mathbb{R}^d$ such that

- $p(x) = \lambda + \|x - x_0\|_2^q$, for all $x \in \mathbb{R}^d$ in a neighborhood of x_0 and
- $p(x) \leq \lambda/2$ for any $x \in \mathbb{R}^d$ outside of this neighborhood.

Then, $\mathrm{Leb}_d\{|p - \lambda| \leq \varepsilon\} = c_1 \varepsilon^{d/q}$, for some constant $c_1 > 0$ and for ε small enough.

We now show that the pseudo-distances $d_\triangle$ and d_H are linked when the density p has γ-exponent at level λ with $\gamma > 0$. Remark first that when $\|p\|_\infty \leq L_0 < \infty$, $d_\triangle$ dominates d_H in the sense that for any $G_1, G_2 \subseteq \mathcal{X}$, we have

$$d_H(G_1, G_2) \leq \max(\lambda, L_0) d_\triangle(G_1, G_2) .$$

This inequality is valid whatever $\gamma \geq 0$. When $\gamma > 0$, the following proposition holds.

PROPOSITION 8.1. *Fix $\lambda > 0, \gamma > 0$ and $L_Q > 0$. The two following statements are equivalent.*

(i) *$\exists\, c > 0$ and $\varepsilon_0 > 0$, such that for any $0 < \varepsilon \leq \varepsilon_0$, we have*

$$Q\{x \in \mathcal{X} : |p(x) - \lambda| \leq \varepsilon\} \leq c\varepsilon^\gamma .$$

(ii) *$\exists\, c' > 0$ and $\varepsilon_1 > 0$, such that for any $0 < \varepsilon \leq \varepsilon_1$, we have*

$$Q\{x \in \mathcal{X} : |p(x) - \lambda| \leq \varepsilon\} \leq L_Q$$

and for all $C \subseteq \mathcal{X}$ satisfying $Q(C) \leq L_Q$, we have

$$Q(C) \leq c' \left(\int_C |p(x) - \lambda| \mathrm{d}Q(x) \right)^{\frac{\gamma}{1+\gamma}} . \tag{8.1}$$

In particular, taking $C = G_1 \triangle G_2$ with G_1 and G_2 two closed subsets of $\mathcal{X}$ such that $Q(G_1) + Q(G_2) \leq L_Q$, if the density p has γ-exponent at level λ w.r.t Q, we have

$$d_\triangle(G_1, G_2) \leq c' \left(d_H(G_1, G_2)\right)^{\frac{\gamma}{1+\gamma}} .$$

PROOF. We first prove (i) $\Rightarrow$ (ii). Define

$$\varepsilon_1 = \min\left[\varepsilon_0, \left(\frac{L_Q}{c(1+\gamma)}\right)^{1/\gamma}\right] .$$

Remark that for any $0 < \varepsilon \leq \varepsilon_1$, we have

$$Q\{x \in \mathcal{X} : |p(x) - \lambda| \leq \varepsilon\} \leq c\varepsilon^\gamma \leq c\varepsilon_1^\gamma = \frac{L_Q}{1+\gamma} \leq L_Q .$$

Define $\mathcal{A}_\varepsilon = \{x : |p(x) - \lambda| > \varepsilon\}$, for all $0 < \varepsilon \leq \varepsilon_0$. For any measurable set $C \subset \mathcal{X}$, we have

$$\begin{aligned}\int_C |p(x) - \lambda| \mathrm{d}Q(x) &\geq \varepsilon Q(C \cap \mathcal{A}_\varepsilon) \\ &\geq \varepsilon \left[Q(C) - Q(\mathcal{A}_\varepsilon^c)\right] \\ &\geq \varepsilon \left[Q(C) - c\varepsilon^\gamma\right] ,\end{aligned}$$

where the last inequality is obtained using (i). Maximizing the last term w.r.t $\varepsilon > 0$, we get

$$\Big(\int_C |p(x) - \lambda| \mathrm{d}Q(x)\Big)^{\frac{\gamma}{1+\gamma}} \geq Q(C)\Big(\frac{\gamma}{1+\gamma}\Big)^{\frac{\gamma}{1+\gamma}}\Big(\frac{1}{1+\gamma}\Big)^{\frac{1}{1+\gamma}} c^{-1/(1+\gamma)} .$$

This yields (8.1) with $c' = \mathrm{e}^{-2/\mathrm{e}} c^{1/(1+\gamma)}$. Note that the maximum is obtained for $\varepsilon = \Big(\frac{Q(C)}{c(1+\gamma)}\Big)^{1/\gamma} \leq \varepsilon_0$ and (i) is valid for this particular ε.

We now prove that (ii) $\Rightarrow$ (i). Consider $\varepsilon_1 > 0$ such that $Q(\mathcal{A}_\varepsilon^c) \leq L_Q$ for any $0 < \varepsilon \leq \varepsilon_1$ and $c' > 0$ such that (8.1) is satisfied for any $C \subseteq \mathcal{X}$, $Q(C) \leq L_Q$. Taking $C = \mathcal{A}_\varepsilon^c$ in (8.1) yields

$$\begin{aligned}Q\left\{x : |p(x) - \lambda| \leq \varepsilon\right\} &= Q\left(\mathcal{A}_\varepsilon^c\right) \\ &\leq c' \left(\int_{\mathcal{A}_\varepsilon^c} |p(x) - \lambda| \mathrm{d}Q(x)\right)^{\frac{\gamma}{1+\gamma}} \\ &\leq c' \left(\varepsilon Q\left(\mathcal{A}_\varepsilon^c\right)\right)^{\frac{\gamma}{1+\gamma}} .\end{aligned}$$

Therefore,

$$Q\left\{x : |p(x) - \lambda| \leq \varepsilon\right\} \leq (c')^{1+\gamma} \varepsilon^\gamma .$$

This inequality yields (i) with $\varepsilon_0 = \varepsilon_1$ and $c = (c')^{1+\gamma}$. ∎

2.3. Classes of densities. Fix $\beta > 0$ and $\lambda > 0$. For any d-tuples $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and $x = (x_1, \ldots, x_d) \in \mathcal{X}$, we define $|s| = s_1 + \ldots + s_d$, $s! = s_1! \ldots s_d!$ and $x^s = x_1^{s_1} \ldots x_d^{s_d}$. Let D^s denote the differential operator

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}} .$$

Denote by $\lfloor \beta \rfloor$ the maximal integer that is strictly smaller than β and fix $x_0 \in \mathcal{X}$. For any real valued function g on $\mathcal{X}$ that is $\lfloor \beta \rfloor$-times continuously differentiable at point x_0, we denote by g_{x_0} its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x_0:

$$g_{x_0}^{(\beta)}(x) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(x - x_0)^s}{s!} D^s g(x_0) .$$

Let $L > 0$ and denote by $\Sigma(\beta, L, x_0)$ the set of functions $g : \mathcal{X} \to \mathbb{R}$ that are $\lfloor \beta \rfloor$-times continuously differentiable at point x_0 and satisfy

$$|g(x) - g_{x_0}^{(\beta)}(x)| \leq L \|x - x_0\|_2^\beta, \quad \forall\, x \in \mathcal{B}(x_0, r) ,$$

for some $r > 0$. The set $\Sigma(\beta, L, x_0)$ is called (β, L, x_0)-*locally Hölder class* of functions.

We now define the class of densities that are considered in this chapter.

Definition 8.2. *Fix $\beta > 0, L > 0, \lambda > 0$ and $\gamma \geq 0$. Let $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ denote the class of all probability densities p on $\mathcal{X}$ such that*

(i) $\exists\, \eta > 0$ *such that* $p \in \Sigma(\beta, L, x_0)$ *for all* $x_0 \in \mathcal{D}(\eta) = p^{-1}\left(]\lambda - \eta, \lambda + \eta[\right)$, *apart from a set of null measure* Q.
(ii) $\exists\, \beta' > 0$ *such that* $p \in \Sigma(\beta', L, x_0)$, *for all* $x_0 \notin \mathcal{D}(\eta)$, *apart from a set of null measure* Q.
(iii) p *has* γ*-exponent at level* λ *with respect to* Q.
(iv) p *is uniformly bounded by a constant* L^*.

The class $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ is the class of uniformly bounded (condition (iv)) densities that have γ-exponent at level λ with respect to Q (condition (iii)) and that are smooth in the neighborhood of the level under consideration (condition (i)). Remark that the parameters β' in condition (ii) and L^* in condition (iv) do not appear in the notation of the class. Indeed $\beta' > 0$ can be arbitrary close to 0 and this will not affect the rates of convergence. Actually, the role of condition (ii) is to ensure that any density from the class can be consistently estimated at any point with an arbitrary slow polynomial rate. In the same manner, the constant L^* does not appear in the rates of convergence and only affects the constants.

To estimate a density from this class we can use a *kernel density estimator* defined by:

$$\hat{p}_n(x) = \hat{p}_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{8.2}$$

where $h > 0$ is the bandwidth parameter and $K : \mathcal{X} \to \mathbb{R}$ is a kernel. In the sequel, a specific family of kernels is considered.

Definition 8.3. *Let* K *be a real-valued function defined on* $\mathbb{R}^d$. *Fix* $\beta > 0$, *and let* $\lfloor \beta \rfloor$ *denote the maximal integer that is strictly less than* β. *The function* $K(\cdot)$ *is said to be a* β*-valid kernel if it satisfies* $\int K = 1$, $\int |K|^p < \infty$ *for any* $p \geq 1$, $\int \|t\|^\beta |K(t)|\, \mathrm{d}t < \infty$, *and, in case* $\lfloor \beta \rfloor \geq 1$, *it satisfies* $\int t^s K(t) \mathrm{d}t = 0$ *for any* $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ *such that* $1 \leq s_1 + \ldots + s_d \leq \lfloor \beta \rfloor$.

Example 8.1. *Let* $\beta > 0$. *For any* β*-valid kernel* K *defined on* $\mathbb{R}$, *consider the following product kernel*

$$\tilde{K}(x) = K(x_1) K(x_2) \ldots K(x_d),$$

for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. *Then it can be easily shown that* $\tilde{K}$ *is a* β*-valid kernel on* $\mathbb{R}^d$. *Now, for any* $\beta > 0$, *an example of a 1-dimensional* β*-valid kernel is given in (Tsybakov, 2004a, section 1.2.2), the construction of which is based on Legendre polynomials. This eventually proves the existence of a multivariate* β*-valid kernel, for any given* $\beta > 0$.

The following proposition holds.

Proposition 8.2. *Fix* $\beta > 0$. *If* K *is a* β*-valid kernel, then* K *is also a* β'*-valid kernel for any* $0 < \beta' \leq \beta$.

Proof. Fix β and β' such that $0 < \beta' \leq \beta$. Remark that $\lfloor \beta' \rfloor \leq \lfloor \beta \rfloor$. For any β-valid kernel K, we have $\int t^s K(t) \mathrm{d}t = 0$ for any $s = (s_1, \ldots, s_d)$ such that $1 \leq s_1 + \ldots + s_d \leq \lfloor \beta' \rfloor$. It remains to check that

$$\int_{\mathbb{R}^d} \|t\|^{\beta'} |K(t)| \mathrm{d}t < \infty. \tag{8.3}$$

Consider the decomposition

$$\int_{\mathbb{R}^d} \|t\|^{\beta'} |K(t)| \mathrm{d}t = \int_{\|t\|\leq 1} \|t\|^{\beta'} |K(t)| \mathrm{d}t + \int_{\|t\|\geq 1} \|t\|^{\beta'} |K(t)| \mathrm{d}t$$
$$\leq \int_{\mathbb{R}^d} |K(t)| \mathrm{d}t + \int_{\|t\|\geq 1} \|t\|^{\beta} |K(t)| \mathrm{d}t \,.$$

To prove (8.3), remark that since K is a β-valid kernel, we have $\int_{\mathbb{R}^d} |K(t)| \mathrm{d}t < \infty$ and

$$\int_{\|t\|\geq 1} \|t\|^{\beta} |K(t)| \mathrm{d}t \leq \int_{\mathbb{R}^d} \|t\|^{\beta} |K(t)| \mathrm{d}t < \infty \,.$$

∎

Intuitively, parameters γ and β are conflicting. Indeed, the parameter β ensures that the density p has a relatively small slope around level λ and the parameter γ requires p to have a slope that is not too small around level λ. The following proposition gives an explicit constraint on the possible parameters γ and β.

PROPOSITION 8.3. *When Q is the Lebesgue measure Leb_d, if $\gamma(1 \wedge \beta) > d$, there is no density $p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ such that λ is in the interior of $p(\mathcal{X})$.*

PROOF. Let p be a density in $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ with λ in the interior of $p(\mathcal{X})$. There exists x_0 such that $p(x_0) = \lambda$ and for any x in a neighborhood of x_0 we can decompose

$$|p(x) - \lambda| \leq |p(x) - p_{x_0}^{(\beta)}(x)| + |p_{x_0}^{(\beta)}(x) - \lambda| \,.$$

Since p is (β, L, x_0)-locally Hölder the first term can be bounded from above by $L\|x - x_0\|_2^\beta$. Now, if $\beta \leq 1$, the second term is null. If $\beta > 1$, this term is a polynomial of degree $\lfloor \beta \rfloor \geq 1$ with no constant term. Hence

$$|p_{x_0}^{(\beta)}(x) - \lambda| = |p_{x_0}^{(\beta)}(x) - p(x_0)| \leq c_2 \max_{1\leq j\leq d} |x^{(j)} - x_0^{(j)}| \leq c_2 \|x - x_0\|_2 \,.$$

Since $\beta > 1$ and $\|x - x_0\|_2$ is bounded by a constant, $\|x - x_0\|_2^\beta \leq c_3 \|x - x_0\|_2$. Thus, for any $\beta > 0$, if p has γ-exponent at level λ with respect to Q,

$$c\varepsilon^\gamma \geq Q(x : |p(x) - \lambda| \leq \varepsilon) \geq Q(x : c_4 \|x - x_0\|_2^{\beta\wedge 1} \leq \varepsilon) \geq c_5 \varepsilon^{d/(\beta\wedge 1)} \,.$$

∎

3. Fast rates for penalized plug-in rules

The first theorem states that rates of convergence for penalized plug-in rules can be obtained using exponential inequalities for the corresponding nonparametric density estimator $\hat{p}_n$.

THEOREM 8.1. *Fix $\lambda > 0$ and $\Delta > 0$. Let $\hat{p}_n$ be an estimator of the density p such that $Q(\hat{p}_n \geq \lambda) \leq C$, almost surely for some positive constant C and let $\mathcal{P}$ be class of densities on $\mathcal{X}$. Assume that there exists positive constants $\eta, c_6, c_7, c_8, c_9, c_\delta, c'_\delta, a$ and b, such that*

- *for Q-almost all $x \in \mathcal{D}(\eta) = p^{-1}(]\lambda - \eta, \lambda + \eta[)$ and for any δ such that $c_\delta n^{-a/2} < \delta < \Delta$, we have*

$$\sup_{p\in\mathcal{P}} \mathbb{P}\left(|\hat{p}_n(x) - p(x)| \geq \delta\right) \leq c_6 \mathrm{e}^{-c_7 n^a \delta^2} \,,\ n \geq 1 \,. \tag{8.4}$$

- *and for Q-almost all $x \in \mathcal{X} \setminus \mathcal{D}(\eta)$, for any δ such that $c'_\delta n^{-b/2} \le \delta \le \Delta$, we have*

$$\sup_{p\in\mathcal{P}} \mathbb{P}\left(|\hat{p}_n(x) - p(x)| \ge \delta\right) \le c_8 \mathrm{e}^{-c_9 n^a \delta^2}\,,\ n \ge 1\,. \tag{8.5}$$

Then if p has γ-exponent at level λ for any $p \in \mathcal{P}$ and if $\ell_n = O\left(n^{-a/2}\right)$, the following upper bound holds,

$$\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \tilde{\Gamma}_{\ell_n})\right] \le c_{10} n^{-\frac{(1+\gamma)a}{2}}\,, \tag{8.6}$$

$$\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \tilde{\Gamma}_{\ell_n})\right] \le c_{11} n^{-\frac{\gamma a}{2}}\,, \tag{8.7}$$

for $n \ge n_0 = n_0(\lambda, \eta, a, b, \varepsilon_0, c_\delta, c'_\delta)$ and where $c_{10} > 0$ and $c_{11} > 0$ depend only on c_6, c_7, c_8, c_9, C, a, b,γ and λ.

Proof. For any measurable function f on $\mathcal{X}$ and any set $A \subset f(\mathcal{X})$, we write for simplicity $\{x \in \mathcal{X} : f(x) \in A\} = \{f \in A\}$. Note first that the conditions of Proposition 8.1 are satisfied. Indeed, $Q(\hat{p}_n \ge \lambda) + Q(p \ge \lambda) \le C + \lambda^{-1}$ and we choose $L_Q = C + \lambda^{-1}$. Therefore, (8.7) is a direct consequence of (8.6) and we prove the latter using the same scheme as in the proof of Audibert and Tsybakov (2005, Theorem 3.1).

Recall that $\tilde{\Gamma}_{\ell_n} \triangle \Gamma = (\tilde{\Gamma}_{\ell_n} \cap \Gamma^c) \cup (\tilde{\Gamma}^c_{\ell_n} \cap \Gamma)$. It yields

$$\mathbb{E}\left[d_H\big(\Gamma, \tilde{\Gamma}_{\ell_n}\big)\right] = \mathbb{E}\int_{\tilde{\Gamma}_{\ell_n}\cap\Gamma^c} |p(x) - \lambda| \mathrm{d}Q(x) + \mathbb{E}\int_{\tilde{\Gamma}^c_{\ell_n}\cap\Gamma} |p(x) - \lambda| \mathrm{d}Q(x)\,.$$

Define two sequences

$$\ell_n^a = n^{-a/2} \quad \text{and} \quad \ell_n^b = \left(\frac{2c_9 n^{a\wedge b}}{(1+\gamma)a \log n}\right)^{-1/2}\,.$$

Let n_0 be a positive integer such that $\ell_n^a < \ell_n^b < \min(\eta, \varepsilon_0) = r$ and $\ell_n^b > c'_\delta n^{-b/2}$ for all $n \ge n_0$. Recall that $\mathcal{D}(r) = p^{-1}\left(]\lambda - r, \lambda + r[\right)$. Consider the following decomposition:

$$\tilde{\Gamma}_{\ell_n} \cap \Gamma^c = \{\hat{p}_n \ge \lambda + \ell_n,\, p < \lambda\} = A_1 \cup A_2 \cup A_3\,, \tag{8.8}$$

where,

$$\begin{aligned} A_1 &= \{\hat{p}_n \ge \lambda + \ell_n,\ \lambda - \ell_n^a \le p < \lambda\}\,,\\ A_2 &= \{\hat{p}_n \ge \lambda + \ell_n,\ \lambda - \ell_n^b \le p < \lambda - \ell_n^a\}\,,\\ A_3 &= \{\hat{p}_n \ge \lambda + \ell_n,\ p < \lambda - \ell_n^b\}\,. \end{aligned}$$

Remark that $A_1 \subseteq \{|p - \lambda| \le \ell_n^a\}$. It yields for $n \ge n_0$,

$$\mathbb{E}\int_{A_1} |p(x) - \lambda| \mathrm{d}Q(x) \le \ell_n^a Q(A_1) \le c_0 (\ell_n^a)^{1+\gamma} = c_0 n^{-\frac{(1+\gamma)a}{2}}\,, \tag{8.9}$$

where in the last inequality we used the γ-exponent of p. Then when $n \ge n_0$, we can decompose A_2 into the disjoint union:

$$A_2 = \bigcup_{j=1}^{\infty} \mathcal{X}_j, \quad \mathcal{X}_j = \left\{\hat{p}_n \ge \lambda + \ell_n,\, \lambda - 2^j \ell_n^a \le p < \lambda - 2^{j-1}\ell_n^a\right\} \cap \mathcal{D}(r)\,.$$

Hence,

$$\mathbb{E}\int_{A_2} |p(x) - \lambda| \mathrm{d}Q(x) = \sum_{j=1}^{\infty} \mathbb{E}\int_{\mathcal{X}_j} |p(x) - \lambda| \mathrm{d}Q(x)\,. \tag{8.10}$$

Using the Fubini Theorem, the general term of the sum in the right-hand side of (8.10) can be bounded from above by

$$2^j \ell_n^a \int_{\mathcal{D}(r)} \mathbb{P}\left[|\hat{p}_n(x) - p(x)| > 2^{j-1}\ell_n^a\right] \mathbb{1}_{\{|p(x)-\lambda|<2^j\ell_n^a\}} \mathrm{d}Q(x)\,.$$

Using now (8.4) and the fact that p has γ-exponent at level λ, we get

$$\begin{aligned} \mathbb{E}\int_{A_2} |p(x)-\lambda|\mathrm{d}Q(x) &\le c_0 c_6 \sum_{j\ge 1} \exp\left(-c_7 n^a (2^{j-1}\ell_n^a)^2\right)(2^j\ell_n^a)^{1+\gamma} \\ &\le c_{12}(\ell_n^a)^{1+\gamma} = c_{12} n^{-\frac{(1+\gamma)a}{2}}\,. \end{aligned} \tag{8.11}$$

We now treat the integral over A_3 using the Fubini theorem and the assumption $Q(\hat{p}_n \ge \lambda) \le C$, a.s.

$$\begin{aligned} \mathbb{E}\int_{A_3} |p(x)-\lambda|\mathrm{d}Q(x) &\le \sup_{\substack{G\subseteq\mathcal{X}\\ Q(G)\le C}} \int_G |p(x)-\lambda| \mathbb{P}\left[|\hat{p}_n(x)-p(x)| > \ell_n^b\right] \mathrm{d}Q(x) \\ &\le (1+\lambda C) c_8 \exp\left(-c_9 n^a (\ell_n^b)^2\right) \le c_{13} n^{-\frac{(1+\gamma)a}{2}}\,, \end{aligned} \tag{8.12}$$

where in the last inequality, we used the fact that

$$\ell_n^b \ge \left(\frac{2c_4 n^a}{(1+\gamma)a\log n}\right)^{-1/2}\,.$$

In view of (8.8), if we combine (8.9), (8.11) and (8.12), we obtain

$$\mathbb{E}\int_{\tilde{\Gamma}_{\ell_n}\cap\Gamma^c} |p(x)-\lambda|\mathrm{d}Q(x) \le c_{14} n^{-\frac{(1+\gamma)a}{2}}\,,$$

where c_{14} depends on $c_6, c_7, c_8, c_9, C, a, b, \gamma$ and λ. In the same manner, using the fact that $\ell_n \le c_{15}\ell_n^a$, it can be shown that for $n \ge n_0$,

$$E\int_{\tilde{\Gamma}^c_{\ell_n}\cap\Gamma} |p(x)-\lambda|\mathrm{d}Q(x) \le c_{16} n^{-\frac{(1+\gamma)a}{2}}\,.$$

where c_{16} depends on $c_6, c_7, c_8, c_9, a, b, \gamma, \lambda$ but not on C. ∎

Remark 8.2. *If $\ell_n > 0$, i.e., we use penalized plug-in rules, we get*

$$\begin{aligned} \mathbb{E}Q\big(\tilde{\Gamma}_{\ell_n}\cap\Gamma^c\big) &= \mathbb{E}\int_{\tilde{\Gamma}_{\ell_n}\cap\Gamma^c} \mathrm{d}Q(x) \\ &\le \sup_{\substack{G\subseteq\mathcal{X}\\ Q(G)\le C}} \int_G \mathbb{P}\left[|\hat{p}_n(x)-p(x)| > \ell_n\right]\mathrm{d}Q(x) \\ &\le c_{6,8} C \exp\left(-c_{7,9} n^a \ell_n^2\right)\,, \end{aligned}$$

where $c_{6,8} \in \{c_6, c_8\}$ and $c_{7,9} \in \{c_7, c_9\}$. Thus the choice

$$\ell_n = \left(\frac{c_{7,9} n^a}{\alpha \log n}\right)^{-1/2}, \quad \alpha > 0\,,$$

yields

$$\mathbb{E}Q\big(\tilde{\Gamma}_{\ell_n}\cap\Gamma^c\big) = O(n^{-\alpha})\,,$$

whatever is $\gamma \ge 0$. This is of particular interest if γ is very small. Indeed, even in such cases we might be interested in the situation where the density level set estimator is

included in the density level set to be estimated with high probability. Note that for such a choice of ℓ_n, the resulting performance of the density level set estimator is only altered by a logarithmic factor. Indeed we have

$$\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_H(\Gamma_p(\lambda),\tilde{\Gamma}_{\ell_n})\right] \leq c_{17}\left(\frac{n^a}{\log n}\right)^{-(1+\gamma)/2},$$
$$\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda),\tilde{\Gamma}_{\ell_n})\right] \leq c_{18}\left(\frac{n^a}{\log n}\right)^{-\gamma/2}.$$

When the density p belongs to the class of locally Hölder densities $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$, a kernel density estimator defined in (8.2) can be used to obtain exponential inequalities as in (8.4) and (8.5). This choice is not the only possible one.

LEMMA 8.1. *Let P be a distribution on $\mathbb{R}^d$ having a density p w.r.t. the measure Q such that $\|p\|_\infty \leq L^*$ for some constant $L^* > 0$. Fix $\beta > 0$, $\beta^\star \geq \beta$, $L > 0$ and assume that $p \in \Sigma(\beta, L, x_0)$. Let $\hat{p}_n$ be a kernel density estimator with bandwidth $h > 0$ and $\beta^\star$-valid kernel K, given an i.i.d. sample $X_1, \ldots, X_n$ from P:*

$$\hat{p}_n(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right).$$

Set

$$\Delta = \frac{6L^*\|K\|^2}{\|K\|_\infty + L^* + L\int \|t\|_2^\beta K(t)\mathrm{d}t}.$$

Then, for all δ, h such that $\Delta > \delta > 2Lc_{19}h^\beta > 0$, we have,

$$\mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \geq \delta\} \leq 2\exp\left(-c_{20}nh^d\delta^2\right),$$

where $c_{19} = \int \|t\|^\beta K(t)\mathrm{d}t$ and $c_{20} = 1/(16L^*\|K\|^2)$.

PROOF. For any $x_0 \in \mathbb{R}^d$,

$$|\hat{p}_n(x_0) - p(x_0)| \leq \frac{1}{n}\Big|\sum_{i=1}^n Z_i(x_0)\Big|,$$

with

$$Z_i(x) = \frac{1}{h^d}K\left(\frac{X_i - x}{h}\right) - p(x).$$

The expectation of $Z_i(x_0)$ is the pointwise bias of a kernel density estimator with bandwidth h. Under the assumptions of the theorem, it is controlled in the following way

$$|\mathbb{E}Z_i(x_0)| \leq Lc_{19}h^\beta.$$

Indeed,

$$\begin{aligned}
|\mathbb{E}Z_i(x_0)| &= \Big|\int \frac{1}{h^d}K\Big(\frac{t}{h}\Big)\big[p(x_0+t) - p(x_0)\big]\mathrm{d}t\Big| \\
&= \Big|\int K(t)\big[p(x_0+ht) - p(x_0)\big]\mathrm{d}t\Big| \\
&= \Big|\int K(t)\big[p(x_0+ht) - p_{x_0}^{(\beta)}(x_0+ht)\big]\mathrm{d}t + \int K(t)\big[p_{x_0}^{(\beta)}(x_0+ht) - p(x_0)\big]\mathrm{d}t\Big| \\
&\leq Lh^\beta\int |K(t)|\|t\|_2^\beta\mathrm{d}t,
\end{aligned}$$

where the last inequality follows from the fact that p is in $\Sigma(\beta, L, x_0)$, and also from the fact that K is a kernel of order $\lfloor\beta\rfloor$ (cf. Proposition 8.2) and that $p_{x_0}^{(\beta)} - p(x_0)$ is a polynomial of degree at most $\lfloor\beta\rfloor$ with no constant term.

Now denote for simplicity $Z_i = Z_i(x_0)$ and let $\overline{Z_i}$ be the centered version of Z_i. Then, when $Lc_{19}h^\beta \le \delta/2$,

$$\begin{aligned}\mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \ge \delta\} &\le \mathbb{P}\left\{\frac{1}{n}\Big|\sum_{i=1}^n \overline{Z_i}\Big| \ge \delta - Lc_{19}h^\beta\right\} \\ &\le \mathbb{P}\left\{\frac{1}{n}\Big|\sum_{i=1}^n \overline{Z_i}\Big| \ge \frac{\delta}{2}\right\}.\end{aligned}$$

The right-hand side of the last inequality can be bounded applying Bernstein's inequality to $\overline{Z_i}$ and $-\overline{Z_i}$ successively. For $h \ge 1$, one has

$$\big|\overline{Z_i}\big| \le \|K\|_\infty h^{-d} + L^* + Lc_{19}h^\beta \le c_{21}h^{-d},$$

where $c_{21} = \|K\|_\infty + L^* + Lc_{19}$ and

$$\mathrm{Var}\{Z_i\} \le h^{-d}\int K(u)^2 p(hu)\mathrm{d}u \le c_{22}h^{-d},$$

where $c_{22} = L^*\|K\|^2$. Applying now Bernstein's inequality yields

$$\begin{aligned}\mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \ge \delta\} &\le 2\exp\left(-\frac{n(\delta/2)^2}{2\left(c_{22}h^{-d} + c_{21}h^{-d}\delta/6\right)}\right) \\ &\le 2\exp\left(-c_{20}nh^d\delta^2\right),\end{aligned}$$

for any $\delta \le \Delta$ and where $\Delta = 6c_{22}/c_{21}$ and $c_{20} = 1/(16c_{22})$. ∎

We can therefore apply Theorem 8.1. When the choice of h is optimal, i.e., $h = n^{-1/(2\beta+d)}$, it yields the following corollary.

COROLLARY 8.1. *Fix $\beta > 0$, $L > 0$, $\lambda > 0$, $\gamma > 0$ and consider the penalized plug-in estimator $\tilde{\Gamma}_{\ell_n}$, where $0 \le \ell_n \le c_{15}n^{-\beta/(2\beta+d)}$, $c_{15} > 0$. The nonparametric estimator $\hat{p}_n$ is the kernel density estimator defined in* (8.2) *with bandwidth parameter $h = n^{-1/(2\beta+d)}$ and $\beta^\star$-valid kernel K, where $\beta^\star = \max(\beta, \beta')$ and β' is the parameter from Definition 8.2. Then,*

$$\begin{aligned}\sup_{p\in\mathcal{P}_\Sigma(\beta,L,\lambda,\gamma)} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \tilde{\Gamma}_{\ell_n})\right] &\le c_{23}n^{-\frac{(1+\gamma)\beta}{2\beta+d}}, \\ \sup_{p\in\mathcal{P}_\Sigma(\beta,L,\lambda,\gamma)} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \tilde{\Gamma}_{\ell_n})\right] &\le c_{24}n^{-\frac{\gamma\beta}{2\beta+d}},\end{aligned}$$

where $c_{23} > 0$ and $c_{24} > 0$ depend on the constants c_{19} and c_{20} that appear in Lemma 8.1, on $c_0, \beta, \beta', \gamma, d$ and on λ.

PROOF. The results are direct consequences of Theorem 8.1 when $\hat{p}_n$ is chosen as in (8.2). We need to check that for such an estimator we have $Q(\hat{p}_n \ge \lambda) \le C$, almost surely for some $C > 0$. Note that since $K \in L_1(\mathbb{R}^d)$, we have

$$\infty > \int_{\mathbb{R}^d} |K(x)|\mathrm{d}Q(x) = \int_{\mathbb{R}^d} |\hat{p}_n(x)|\mathrm{d}Q(x) \ge \int_{\{\hat{p}_n\ge\lambda\}} |\hat{p}_n(x)|\mathrm{d}Q(x) \ge \lambda Q\{\hat{p}_n \ge \lambda\}.$$

Hence, the condition is satisfied with $C = \lambda^{-1} \int |K|$. All the other conditions of Theorem 8.1 are satisfied and we can apply it with $a = 2\beta/(2\beta + d)$ and $b = 2\beta'/(2\beta + d)$. ∎

4. Minimax lower bounds

The following theorem shows that the rate obtained in Corollary 8.1 is optimal in a minimax sense.

THEOREM 8.2. *Assume that Q is the Lebesgue measure on $\mathcal{X}$. Fix $\lambda > 0$ and let L, β, γ be positive constants such that $\gamma\beta \leq d$. Then, there exists constants $c_{25} > 0$ and $c_{26} > 0$ such that for any $n \geq 1$ and any estimator $\hat{G}_n$ of $\Gamma_p(\lambda)$ constructed from the sample $X_1, \ldots, X_n$, we have*

$$\sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \hat{G}_n)\right] \geq c_{25} n^{-\frac{(1+\gamma)\beta}{2\beta+d}} ,$$

$$\sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \hat{G}_n)\right] \geq c_{26} n^{-\frac{\gamma\beta}{2\beta+d}} . \tag{8.13}$$

PROOF. In view of Proposition 8.1, we only have to prove (8.13). To that end, we will use Lemma A.5 with $d = d_\triangle$, $\varepsilon = \varepsilon_n \sim n^{-\frac{\gamma\beta}{2\beta+d}}$ and $\mathcal{P} = \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$. Assume without loss of generality that $\lambda = 1$.

We now describe the construction of the family $\mathcal{N}$. Define the support density by

$$p_0(x) = \begin{cases} 1 & \text{if } 0 \leq x^{(j)} \leq 1, \text{ for any } j = 1, \ldots, d, \\ 0 & \text{else.} \end{cases}$$

We now define perturbations of the support density. Consider the integer $q = \lfloor c_{27} n^{\frac{1}{2\beta+d}} \rfloor$ where c_{27} is a positive constant chosen large enough to ensure that $q \geq 1$, and the regular grid $\mathcal{G}$ on $[0,1]^d$ defined as

$$\mathcal{G} = \left\{ \left(\frac{2k_1 + 1}{2q}, \ldots, \frac{2k_d + 1}{2q} \right), k_i \in \{0, \ldots, q-1\}, i = 1, \ldots, d \right\} .$$

Re-index the grid by $\mathcal{G} = \{g_j\}_{1 \leq j \leq q^d}$ and define the integer $m = \lfloor c_{28} q^{d-\gamma\beta} \rfloor$ for some positive constant c_{28}. The condition $\gamma\beta \leq d$ ensures that $m \geq 2$ if c_{28} is chosen large enough. Let $\mathcal{J} = \{1, 3, \ldots, 2m-1\}$ be the set of odd integers between 1 and $2m-1$ and for any $j = 1, \ldots, 2m$, define the disjoint balls $B_j = \mathcal{B}_{\tilde{\beta}}(g_j, (4q)^{-1})$, where $\tilde{\beta} = \beta$ if $\beta > 1$ and $\tilde{\beta} = 2$ if $\beta < 1$. Set $B_0 = [0,1]^d \setminus \bigcup_{j=1}^{2m} B_j$ and consider a partition of B_0 into two measurable sets $B_0 = C_1 \cup C_2$ such that $Q(C_1) = Q(C_2) = Q(B_0)/2$. For any $j \in \mathcal{J}$, define the function φ_j on $[0,1]^d$ by

$$\varphi_j(x) = \tilde{L} \|x - g_j\|_{\tilde{\beta}}^{\beta} \mathbb{1}_{\{x \in B_j\}} - \tilde{L} \|x - g_{j+1}\|_{\tilde{\beta}}^{\beta} \mathbb{1}_{\{x \in B_{j+1}\}} , \quad 0 < \tilde{L} < 1 .$$

For any $\omega = (\omega_1, \ldots, \omega_m) \in \{-1, 1\}^m$, define the density

$$p_\omega(x) = 1 + \sum_{j \in \mathcal{J}} \omega_j \varphi_j(x) + \kappa \left(\mathbb{1}_{\{x \in C_1\}} - \mathbb{1}_{\{x \in C_2\}} \right) ,$$

where $\kappa < 1$ is a tuning parameter. Consider a set $\Omega \subset \{-1, 1\}^m$ of cardinality s and define the family $\mathcal{N}$ as

$$\mathcal{N} = \{p_\omega, \omega \in \Omega\} .$$

The tuning parameters, $\tilde{L}$, κ and Ω, will be chosen in order to fulfill the conditions of Lemma A.5.

First condition: $\mathcal{N} \subset \mathcal{P}_\Sigma(\beta, L, 1, \gamma)$.

Remark first that for any $\omega \in \Omega$, p_ω is a density that satisfies $\|p_\omega\|_\infty \leq 2$. Next, fix $\eta > 0$ and $x_0 \in \mathcal{D}(\eta) = p_\omega^{-1}\{]1-\eta, 1+\eta[\}$ outside of the set of null Lebesgue measure where p_ω is not differentiable. If x_0 is in the interior of B_0 then p is constant in a neighborhood of x_0 and obviously belongs to $\Sigma(\beta, L, x_0)$. Fix $j \in \{1, \dots, 2m\}$ and assume that $x_0 \in B_j, x_0 \neq g_j$, for some $j = 1, \dots, 2m$.

We begin by treating the case $\beta > 1$. For any $x \in B_j$, we have

$$p_\omega(x) = 1 + \sigma_j \tilde{L} \sum_{l=1}^{d} \left| x^{(l)} - g_j^{(l)} \right|^\beta ,$$

where $\sigma_j \in \{-1, 1\}$. Fix $l \in \{1, 2, \dots, d\}$, and assume without loss of generality that $x_0^{(l)} > g_j^{(l)}$. Consider now a real number $x^{(l)}$ such that $|x^{(l)} - x_0^{(l)}| \leq (x_0^{(l)} - g_j^{(l)})/2$. A Laplace-Taylor expansion at point $x_0^{(l)}$ gives

$$\left| x^{(l)} - g_j^{(l)} \right|^\beta = \sum_{k=0}^{\lfloor \beta \rfloor} C(k, \beta) \big(x^{(l)} - x_0^{(l)} \big)^k \big(x_0^{(l)} - g_j^{(l)} \big)^{\beta - k} + R^{(l)} ,$$

where $C(k, \beta) \neq 0$ depends only on β, k and the sign of $x^{(l)} - g_j^{(l)}$. The residual term is given by

$$R^{(l)} = C'(\beta) \int_{x_0^{(l)}}^{x^{(l)}} \big(x^{(l)} - t \big)^{\lfloor \beta \rfloor} (t - g_j^{(l)})^{\beta - \lfloor \beta \rfloor - 1} \mathrm{d}t ,$$

where $C'(\beta)$ depends only on β and the sign of $x^{(l)} - g_j^{(l)}$. Note now that since $\beta - \lfloor \beta \rfloor - 1 \leq 0$, we have

$$|R^{(l)}| \leq |C'(\beta)| |x^{(l)} - x_0^{(l)}|^\beta .$$

Summing up over l allows us to bound from above the difference between p_ω and its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x_0 by the quantity

$$c_{29} \tilde{L} \|x - x_0\|_\beta^\beta \leq c_{30} \tilde{L} \|x - x_0\|_2^\beta ,$$

where c_{29} and c_{30} are positive constants that depend only on β and d.

When $\beta \leq 1$, $\tilde{\beta} = 2$ and the Taylor polynomial of p_ω at point x_0 has degree $\lfloor \beta \rfloor = 0$. Therefore, it is constant and equals $p_\omega(x_0)$. It yields

$$\begin{aligned} |p_\omega(x) - p_\omega(x_0)| &= \tilde{L} \Big| \|x - g_j\|_2^\beta - \|x_0 - g_j\|_2^\beta \Big| \\ &\leq \tilde{L} \big| \|x - g_j\|_2 - \|x_0 - g_j\|_2 \big|^\beta \\ &\leq \tilde{L} \|x - x_0\|_2^\beta . \end{aligned}$$

For any $\beta > 0$, taking $\tilde{L}$ such that $c_{30} \tilde{L} \leq L \wedge 1/2$, yields that $\mathcal{N} \subset \Sigma(\beta, L, x_0)$.

For large enough η, $\mathcal{D}(\eta) = [0, 1]^d$ and we can take $\beta' = \beta$ in condition (ii). Therefore we only have to check that p_ω has γ-exponent at level 1 with respect to the Lebesgue measure. The following decomposition holds:

(8.14)

$$\begin{aligned} Q(x : |p - 1| \leq \varepsilon) &= 2 \sum_{j \in \mathcal{J}} Q(x : |p(x) - 1| \leq \varepsilon, x \in B_j) + Q(x : |p(x) - 1| \leq \varepsilon, x \in B_0) \\ &= 2m Q(x : |p(x) - 1| \leq \varepsilon, x \in B_1) + Q(B_0) \mathbb{I}_{\{\varepsilon \geq \kappa\}} . \end{aligned}$$

We have on the one hand,

$$mQ(x : |p(x) - 1| \leq \varepsilon, x \in B_1) \leq c_{31} m q^{-d} \mathbb{1}_{\{\varepsilon \geq \tilde{L} q^{-\beta}\}} + c_{32} m \varepsilon^{d/\beta} \mathbb{1}_{\{\varepsilon < \tilde{L} q^{-\beta}\}}.$$

Since $m \leq c_{28} q^{d-\gamma\beta}$ we have $mq^{-d} \leq c_{28} q^{-\gamma\beta}$. It yields

$$mq^{-d} \mathbb{1}_{\{\varepsilon \geq \tilde{L} q^{-\beta}\}} \leq c_{28} q^{-\gamma\beta} \mathbb{1}_{\{\varepsilon^\gamma \geq \tilde{L} q^{-\gamma\beta}\}} \leq c_{33} \varepsilon^\gamma \,,$$

and

$$m\varepsilon^{d/\beta} \mathbb{1}_{\{\varepsilon < \tilde{L} q^{-\beta}\}} \leq c_{28} q^{d-\gamma\beta} \varepsilon^{d/\beta} \mathbb{1}_{\{q^{d-\gamma\beta} < (\varepsilon/\tilde{L})^{\gamma - d/\beta}\}} \leq c_{34} \varepsilon^\gamma.$$

Therefore

$$mQ(x : |p(x) - 1| \leq \varepsilon, x \in B_1) \leq c_{35} \varepsilon^\gamma. \tag{8.15}$$

On the other hand,

$$Q(B_0) \mathbb{1}_{\{\varepsilon \geq \kappa\}} = (1 - c_{36} m q^{-d}) \mathbb{1}_{\{\varepsilon \geq \kappa\}} \leq c_{37} \varepsilon^\gamma \,, \tag{8.16}$$

when $\kappa = c_{38}(1 - c_{36} m q^{-d})^{1/\gamma} < 1$. Equation (8.14) together with (8.15) and (8.16) yield

$$Q(x : |p - 1| \leq \varepsilon) \leq c_{39} \varepsilon^\gamma \,.$$

for $\kappa = c_{38}(1 - c_{36} m q^{-d})^{1/\gamma}$.

SECOND CONDITION (A.2): $d_\triangle(\Gamma_p, \Gamma_q) \geq \varepsilon_n, \forall\ p, q \in \mathcal{N}, p \neq q$.

By construction, for any $\omega, \omega' \in \{-1, 1\}^m$,

$$d_\triangle(\Gamma_{p_\omega}, \Gamma p_{\omega'}) = 2 \sum_{j=1}^m \mathbb{1}_{\{\omega_j \neq \omega'_j\}} q^{-d} \,.$$

We need to bound from below the Hamming distance $\rho(\omega, \omega') = \sum_{j=1}^m \mathbb{1}_{\{\omega_j \neq \omega'_j\}}$ between ω and ω' for any $\omega, \omega' \in \Omega$. To do so we use the Varshamov-Gilbert bound (cf. Lemma A.1) that guarantees the existence of Ω such that $\mathrm{card}(\Omega) \geq 2^{m/8}$ and $\rho(\omega, \omega') \geq m/8$ for any $\omega, \omega' \in \Omega$. For such Ω we have

$$d_\triangle(\Gamma_{p_\omega}, \Gamma_{p_{\omega'}}) \geq \frac{m q^{-d}}{4} \geq c_{40} q^{-\gamma\beta} \geq c_{41} n^{-\frac{\gamma\beta}{2\beta + d}} \,.$$

THIRD CONDITION: $\max_{p,q \in \mathcal{N}} K(p, q) \leq c_{42} \log(s)$.

Note that for the above choice of Ω, we have $s = \mathrm{card}(\mathcal{N}) = \mathrm{card}(\Omega) \geq 2^{m/8}$. Therefore $\log(s) \geq c_{43} m$ and we only have to prove that

$$\max_{p,q \in \mathcal{N}} K(p, q) \leq c_{44} m \,.$$

For any $p_\omega, p_{\omega'} \in \mathcal{N}$, we have,

$$\begin{aligned} K(p_\omega, p_{\omega'}) &= n \sum_{j \in \mathcal{J}} \int_{B_j \cup B_{j+1}} \log\left(\frac{1 + \omega_j \varphi_j(x)}{1 + \omega'_j \varphi_j(x)}\right) \big(1 + \omega_j \varphi_j(x)\big) \mathrm{d}x \\ &\leq n \sum_{j \in \mathcal{J}} \int_{B_j \cup B_{j+1}} \log\left(\frac{1 + \omega_j \varphi_j(x)}{1 - \omega_j \varphi_j(x)}\right) \big(1 + \omega_j \varphi_j(x)\big) \mathrm{d}x \,. \end{aligned}$$

Fix $j \in \mathcal{J}$ and assume without loss of generality that $\omega_j = 1$. For any $x \in B_j$ we have

$$\int_{B_j} \log\left(\frac{1+\varphi_j(x)}{1-\varphi_j(x)}\right)\big(1+\varphi_j(x)\big)\mathrm{d}x = \int_{B_j} \log\left(1+\frac{2\tilde{L}\|x-g_j\|_{\tilde{\beta}}^{\beta}}{1-\tilde{L}\|x-g_j\|_{\tilde{\beta}}^{\beta}}\right)\big(1+\tilde{L}\|x-g_j\|_{\tilde{\beta}}^{\beta}\big)\mathrm{d}x$$
$$= \int_{\mathcal{B}_{\tilde{\beta}}(0,(4q)^{-1})} \log\left(1+\frac{2z(x)}{1-z(x)}\right)\big(1+z(x)\big)\mathrm{d}x\,,$$

where $z(x) = \tilde{L}\|x\|_{\tilde{\beta}}^{\beta}$. In the same manner, we have

$$\int_{B_{j+1}} \log\left(\frac{1+\varphi_j(x)}{1-\varphi_j(x)}\right)\big(1+\varphi_j(x)\big)\mathrm{d}x = \int_{\mathcal{B}_{\tilde{\beta}}(0,(4q)^{-1})} \log\left(1-\frac{2z(x)}{1+z(x)}\right)\big(1-z(x)\big)\mathrm{d}x\,.$$

Consider the function $F : [0,1) \to \mathbb{R}$ defined by

$$F(z) = \log\left(1+\frac{2z}{1-z}\right)(1+z) + \log\left(1-\frac{2z}{1+z}\right)(1-z)\,.$$

For any $z \geq 0$ such that $z \leq \tilde{L}/4 \leq 1/4$, we have

$$F(z) \leq \frac{2z(1+z)}{1-z} - \frac{2z(1-z)}{1+z} = \frac{8z^2}{1-z^2} \leq 9z^2\,.$$

It yields

$$\int_{B_j \cup B_{j+1}} \log\left(\frac{1+\varphi_j(x)}{1-\varphi_j(x)}\right)\big(1+\varphi_j(x)\big)\mathrm{d}x \leq 9\int_{\mathcal{B}_{\tilde{\beta}}(0,(4q)^{-1})} \|x\|_{\tilde{\beta}}^{2\beta}\mathrm{d}x \leq c_{45}q^{(2\beta+d)}\,.$$

Hence

$$K(p_\omega, p_{\omega'}) \leq c_{45}nmq^{(2\beta+d)} \leq c_{46}m \leq c_{47}\log(s)\,.$$

We can therefore apply Lemma A.5 and the Theorem 8.2 is proved. ∎

5. Exponentially fast rates

We can strengthen the γ-exponent condition in a way corresponding to the case $\gamma \to \infty$. This condition imposes a jump of size at least $\eta_0 > 0$ on both sides of the level λ and is stated as follows.

Definition 8.4. *For any $\lambda, \eta_0 > 0$, a function $f : \mathcal{X} \to \mathbb{R}$ is said to have η_0-jump around level λ w.r.t measure Q if the set $\{x \in \mathcal{X} : |f(x) - \lambda| \leq \eta_0\}$ has null measure Q.*

When p has η_0-jump around level λ, no restriction such as $\gamma\beta \leq d$ is needed and the lower bounds of Theorem 8.2 have no more meaning in this context. Under such a condition it is even possible to attain exponential rates of convergence. In this context, the pseudo-distances d_H and $d_\triangle$ are equivalent as shown in the next proposition.

Proposition 8.4. *Fix $\lambda > 0$ and $\eta_0 > 0$. The two following statements are equivalent.*

(i) $Q\{x \in \mathcal{X} : |p(x) - \lambda| \leq \eta_0\} = 0$.

(ii) $\forall C \subseteq \mathcal{X}$, *such that* $Q(C) > 0$, *we have* $Q(C) < \eta_0^{-1}\int_C |p(x) - \lambda|\mathrm{d}Q(x)$.

In particular, taking $C = G_1 \triangle G_2$ with G_1, G_2 two closed subsets of $\mathcal{X}$, if the density p has η_0-jump around level λ w.r.t Q, we have

$$\eta_0 d_\triangle(G_1, G_2) \leq d_H(G_1, G_2)\,.$$

PROOF. Denote by $\mathcal{A} = \{x : |p(x) - \lambda| > \eta_0\}$,

$$\int_C |p(x) - \lambda| \mathrm{d}Q(x) = \int_{C \cap \mathcal{A}} |p(x) - \lambda| \mathrm{d}Q(x) > \eta_0 Q(C)\,,$$

and we have proved that (i) $\Longrightarrow$ (ii). We now prove (ii) $\Longrightarrow$ (i). Note first that $Q\{|p-\lambda| = \eta_0\} = 0$. Indeed, if $Q\{|p - \lambda| = \eta_0\} > 0$, from (ii), we have

$$Q\{|p - \lambda| = \eta_0\} < \eta_0^{-1} \int_{\{|p-\lambda|=\eta_0\}} |p(x) - \lambda| \mathrm{d}Q(x) = Q\{|p - \lambda| = \eta_0\}\,.$$

Therefore,

$$Q\left\{|p - \lambda| \leq \eta_0\right\} = Q\left\{|p - \lambda| < \eta_0\right\}\,,$$

If $Q\left\{|p - \lambda| \leq \eta_0\right\} > 0$, we get by (ii)

$$Q\left\{|p - \lambda| \leq \eta_0\right\} = \eta_0^{-1} \int_{\{|p-\lambda|<\eta_0\}} |p(x) - \lambda| \mathrm{d}Q(x) < Q\left\{|p - \lambda| \leq \eta_0\right\}\,.$$

Therefore $Q\left\{|p - \lambda| \leq \eta_0\right\} = 0$ and the equivalence is proved. ∎

The following theorem states that exponential rates of convergence are attainable under a jump condition.

THEOREM 8.3. *Fix $\lambda > 0$ and $\Delta > 0$. Let $\hat{p}_n$ be an estimator of the density p such that $Q(\hat{p}_n \geq \lambda) \leq C$ a.s. and let $\mathcal{P}$ be class of densities on $\mathcal{X}$. Assume that there exists constants $c_{48} > 0$ and $c_{49} > 0$ and a positive sequence $(a_n)_{n\geq 1}$ tending to infinity, such that for any $\eta_0 \leq \delta < \Delta$ and for Q-almost all $x \in \mathcal{X}$, we have*

$$\sup_{p\in\mathcal{P}} \mathbb{P}\left(|\hat{p}_n(x) - p(x)| \geq \delta\right) \leq c_{49} \mathrm{e}^{-c_{49} a_n \delta^2}\,,\ n \geq 1\,. \tag{8.17}$$

Then if p has η_0-jump around level λ for any $p \in cP$, the simple plug-in rule $\hat{\Gamma}$ satisfies

$$\begin{aligned}
\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \hat{\Gamma})\right] &\leq c_{50} \mathrm{e}^{-c_{49} a_n \eta_0^2}\,, \\
\sup_{p\in\mathcal{P}} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \hat{\Gamma})\right] &\leq c_{50} \eta_0 \mathrm{e}^{-c_{49} a_n \eta_0^2}\,,
\end{aligned} \tag{8.18}$$

for a positive constant c_{50} and all $n \geq 1$.

PROOF. In view of Proposition 8.4, we only have to prove (8.18). Recall that

$$\mathbb{E}\left[d_H\left(\Gamma, \hat{\Gamma}\right)\right] = \mathbb{E}\int_{\hat{\Gamma}\cap\Gamma^c} |p(x) - \lambda| \mathrm{d}Q(x) + \mathbb{E}\int_{\hat{\Gamma}^c\cap\Gamma} |p(x) - \lambda| \mathrm{d}Q(x)\,.$$

We only give details for the treatment of the first term. The second term is treated exactly in the same manner except that the assumption $Q(\hat{p}_n \geq \lambda) \leq C$ a.s. is not needed. Indeed, we need $Q(p \geq \lambda)$ which is trivially upper bounded by λ^{-1} since p is a probability density.

By the Fubini theorem and the jump assumption on p, we have

$$\begin{aligned}
\mathbb{E}\int_{\hat{\Gamma}\cap\Gamma^c} |p(x) - \lambda| \mathrm{d}Q(x) &\leq \sup_{\substack{G\subset\mathcal{X} \\ Q(G)\leq C}} \int_G |p(x) - \lambda| \mathbb{P}\left[|\hat{p}_n(x) - p(x)| > \eta_0\right] \mathrm{d}Q(x) \\
&\leq (1 + \lambda C) c_{48} \exp\left(-c_{49} a_n \eta_0^2\right)\,.
\end{aligned}$$

where in the last inequality, we use inequality (8.17) ∎

Define the class of densities $\widetilde{\mathcal{P}}_\Sigma(\beta, L, \mathcal{X}, \eta_0)$ as the class $\mathcal{P}_\Sigma(\beta, L, \mathcal{X}, \gamma)$ in Definition 8.2 with condition (iii) replaced by:

(iii') p has η_0-jump around level λ with respect to Q

If p is estimated by a kernel density estimator with a bandwidth that is sufficiently small but constant with respect to n, we get the following corollary.

COROLLARY 8.2. *Fix $\lambda > 0, \beta > 0$ and $\eta_0 > 0$. Consider the simple plug-in estimator $\hat{\Gamma}(\lambda) = \{\hat{p}_n \geq \lambda\}$ of the density level set $\Gamma_p(\lambda)$ where $\hat{p}_n$ is the kernel density estimator defined in* (8.2) *with constant bandwidth parameter $h \leq c_{51}$, where $c_{51} > 0$ depends only on β, L and η_0. Then,*

$$\begin{aligned} sup_{p\in\mathcal{P}} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \hat{\Gamma})\right] \leq & c_{52} \mathrm{e}^{-c_{53} n}, \\ sup_{p\in\mathcal{P}} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \hat{\Gamma})\right] \leq & c_{52} \eta_0 \mathrm{e}^{-c_{53} n}, \end{aligned}$$

where $c_{52} > 0$ and $c_{53} > 0$ depend on β, L and η_0.

PROOF. In view of Theorem 8.3, it is sufficient to prove that (8.17) holds. To that end, we apply Lemma 8.1. Taking $h < \min\left(1, (\eta_0/(2Lc_{19}))^{1/\beta}\right)$ yields inequality (8.17) for any δ satisfying $\eta_0 \leq \delta < \Delta$, where Δ and c_{19} are defined in Lemma 8.1. ∎

Note that the rate of convergence crucially depends on the value of the constant c_{53}. In this chapter, we do not address the problem of finding the best possible constant. The message of the corollary is to show that it is possible to estimate with exponential rates of convergence, the level sets of the density p that jumps around the level under consideration and has Hölder smoothness $\beta > 0$ arbitrary close to 0 elsewhere.

Part 4

Additional material and bibliography

APPENDIX A

Statistical background

Contents

1. Minimax lower bounds

This section contains useful tools to prove minimax lower bounds. For a recent survey on this topic, see Tsybakov (2004a)[Chap. 2]. We first give two lemmas linked to subset extraction.

1.1. Subset extraction. Fix an integer $m \geq 1$ and define

$$\Omega = \big\{\omega = (\omega_1, \ldots, \omega_m), \omega_i \in \{0,1\}\big\} = \{0,1\}^m$$

For any two $\omega = (\omega_1, \ldots, \omega_m)$ and $\omega' = (\omega'_1, \ldots, \omega'_m)$ in Ω define the *Hamming distance* between ω and ω' by

$$\rho(\omega, \omega') = \sum_{i=1}^{m} \mathbb{1}_{\{\omega_i \neq \omega'_i\}} .$$

The following lemma holds

LEMMA A.1 (Varshamov-Gilbert bound, 1962). *Fix $m \geq 8$. Then there exists a subset $\{\omega^{(0)}, \ldots, \omega^{(M)}\}$ of Ω such that $M \geq 2^{m/8}$ and*

$$\rho(\omega^{(j)}, \omega^{(k)}) \geq \frac{m}{8}, \quad \forall\ 0 \leq j < k \leq M .$$

Moreover, we can always take $\omega^{(0)} = (0, \ldots, 0)$.

For a proof of this lemma, see Tsybakov (2004a, Lemma 2.8, p. 89). Statements as the next lemma should be probably attributed to Gilbert (1952) but we present here a corollary of Lemma 4, p. 264 of Birgé and Massart (2001) which is presented in a form that is more appropriate to our purposes. Let m and ℓ be two integers such that $m \geq \ell \geq 1$ and define

$$\Omega_\ell = \Big\{\omega = (\omega_1, \ldots, \omega_m), \omega_i \in \{0,1\}, \sum_{i=1}^{m} \omega_i = \ell\Big\}$$

LEMMA A.2. *Let m and ℓ be two positive integers such that $m \geq 6\ell$. Then there exists a subset $\{\omega^{(1)}, \ldots, \omega^{(M)}\}$ of $\Omega_{2\ell}$ such that*

$$\log(M) \geq \ell \Big[\log\big(\frac{m}{\ell}\big) - \log(16) + 1 \Big]$$

and

$$\rho(\omega^{(j)}, \omega^{(k)}) \geq \ell + 1, \quad \forall\ 1 \leq j < k \leq M\,.$$

PROOF. Consider the bijection f between Ω and $\mathcal{P}\big(\{1, \ldots, m\}\big)$, the set of all subsets of $\{1, \ldots, m\}$, such that for any $\omega \in \Omega$, $f(\omega)$ is the set of integers that corresponding to the positions of the ones in ω. Clearly $\mathcal{M} = f(\Omega_{2\ell})$ is the set of all subsets of cardinality 2ℓ of $\{1, \ldots, m\}$. Applying now Birgé and Massart (2001, Lemma 4) with $N = m$ and $n = \ell$ yields the assertion of the lemma. ∎

1.2. Minimax lower bounds for density estimation. We now give to lemmas that can be used to obtain minimax lower bounds in the context of nonparametric density estimation with L_2 risk. Let P_p be a distribution on $\mathbb{R}^d$ that admits a probability density p with respect to the Lebesgue measure. Let $(X_1, \ldots, X_n)$ be i.i.d with distribution P_p^n. For any estimator estimator $\hat{p}_n$ based on $(X_1, \ldots, X_n)$, define its L_2 risk by

$$R_n(\hat{p}_n, p) = E_p^n \|\hat{p}_n - p\|^2$$

where E_p^n denotes the expectation with respect to the sample $(X_1, \ldots, X_n)$ and for any $g \in L_2(\mathbb{R}^d)$,

$$\|g\| = \Big(\int_{\mathbb{R}^d} g^2(x) \mathrm{d}x \Big)^{1/2}.$$

1.2.1. *Minimax lower bounds based on Assouad's lemma.* The next lemma is an adaptation of Birgé (1986, Corollary 4.1, p.281) which is a corollary of Assouad's lemma for a particular construction of Assouad's cube. It is particularly suitable when proving minimax lower bounds in the context of density estimation. It involves the *Hellinger distance* between two probability densities p on and q on $\mathbb{R}^d$

$$h(p, q) = \Big(\int_{\mathbb{R}^d} \big(\sqrt{p(x)} - \sqrt{q(x)}\big)^2 \mathrm{d}x \Big)^{1/2}.$$

LEMMA A.3. *Let $\mathcal{C}$ be a set of functions of the following type*

$$\mathcal{C} = \Big\{ f + \sum_{i=1}^{r} \delta_i g_i,\ \delta_i \in \{0, 1\},\ i = 1, \ldots, r \Big\},$$

where the g_i are functions on $\mathbb{R}^d$ with disjoint supports, such that $\int g_i(x)\mathrm{d}x = 0$, f is a probability density on $\mathbb{R}^d$ which is constant on the union of the supports of g_i's, and $f + g_i \geq 0$ for all i. Assume that

$$\min_{1 \leq i \leq r} \|g_i\|^2 \geq \alpha > 0 \quad \textit{and} \quad \max_{1 \leq i \leq r} h^2(f, f + g_i) \leq \beta < 1\,, \tag{A.1}$$

where $h^2(f, g) = (1/2) \int (\sqrt{f(x)} - \sqrt{g(x)})^2 \mathrm{d}x$ is the squared Hellinger distance between two probability densities f and g. Then

$$\inf_{T_n} \sup_{p \in \mathcal{C}} R_n(T_n, p) \geq \frac{r\alpha}{4} (1 - \sqrt{2n\beta})$$

where $\inf_{T_n}$ denotes the infimum over all estimators.

1.2.2. *Minimax lower bounds based on the general minimization scheme.* Tsybakov (2004a, Chap. 2) propose a general method to obtain minimax lower bounds. It is close to other methods using Fano's Lemma. The following lemma is quite standard and involves the Kullback-Leibler divergence between two probability densities p and q on $\mathbb{R}^d$

$$K(p,q) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} \log\Big(\frac{p(x)}{q(x)}\Big) p(x)\mathrm{d}x & \text{if } P_p \ll P_q\,, \\ +\infty & \text{else.} \end{cases}$$

LEMMA A.4. *Let $\mathcal{C} = \{q_0, \ldots, q_N\}, N \geq 1$, be a finite set of probability densities on $\mathbb{R}^d$. Assume that*

(i) $\|q_j - q_k\|^2 \geq 4\psi_n > 0, \quad \forall\ 0 \leq j < k \leq N,$

(ii) *$P_j \ll P_0$, $j = 1, \ldots, N$, where $P_j \triangleq P_{q_j}^n$, and the Kullback-Leibler divergences $K(q_j, q_0)$ between the densities q_j and q_0 satisfy*

$$\frac{1}{N}\sum_{j=1}^{N} K(q_j, q_0) \leq \frac{\log N}{16n}.$$

Then

$$\inf_{T_n} \sup_{p\in\mathcal{C}} R_n(T_n, p) \geq c_1\psi_n,$$

where $\inf_{T_n}$ denotes the infimum over all estimators of p and $c_1 > 0$ is an absolute constant.

The next lemma can be found in Tsybakov (1997, Lemma 4) and is stated here in a slightly weaker form. It allows to derive minimax lower bounds in the context of density level set estimation. Consider i.i.d random vectors $(X_1, \ldots, X_n)$ with values in $\mathcal{X}$ and distribution P, having an unknown probability density p with respect to the measure Leb_d. For a fixed $\lambda > 0$, define the λ-*level set* of the density p:

$$\Gamma_p(\lambda) \triangleq \{x \in \mathcal{X} \ :\ p(x) \geq \lambda\}\ .$$

LEMMA A.5. *Let d be a pseudo-metric between subsets of $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathcal{P}$ be a set of densities and assume that there exists a subset $\mathcal{N} \subset \mathcal{F}$ with cardinal $2 \leq \mathrm{card}(\mathcal{N}) = s < \infty$ and a constant $c_1 > 0$, such that*

$$d\big(\Gamma_p(\lambda), \Gamma_q(\lambda)\big) \geq \varepsilon, \quad \forall\ p, q \in \mathcal{N}, p \neq q\,, \tag{A.2}$$

and

$$\max_{p,q\in\mathcal{N}} K(p,q) \leq c_1 \log(s)\,, \tag{A.3}$$

where $K(\cdot,\cdot)$ denotes the Kullback-Leibler divergence. Then, there exists an absolute positive constant c_2 such that for any estimator $\hat{G}_n$ of $\Gamma_p(\lambda)$ constructed from the sample $X_1, \ldots, X_n$, we have

$$\sup_{p\in\mathcal{P}} \mathbb{E}\left[d(\Gamma_p(\lambda), \hat{G}_n)\right] \geq c_2\varepsilon\,.$$

2. Universal lower bound for kernel density estimators

The next lemma is in the spirit of Stone (1984) and gives a lower bound for kernel density estimators with a fixed kernel in a large class and for any positive bandwidths. Unlike the minimax setup where the lower bounds hold for any estimators and the density is supposed to belong to a certain class of densities, here the lower bound holds for any density and the estimators are supposed to belong to a particular family of estimators. We first recall the definition of a kernel density estimator. Consider independent identically

distributed (i.i.d.) random variables $X_1, \ldots, X_n$ having an unknown common probability density $p \in L_2(\mathbb{R})$. We measure the performance of $\hat{p}_n$ by its *mean integrated squared error* (MISE):

$$R_n(\hat{p}_n, p) = E_p \big\| \hat{p}_n - p \big\|^2 \tag{A.4}$$

where E_p denotes the expectation w.r.t. $(X_1, \ldots, X_n)$ and for any function $g \in L_2(\mathbb{R})$,

$$\|g\| = \Big(\int_{\mathbb{R}} g^2(x) \mathrm{d}x \Big)^{1/2}$$

Define the characteristic function

$$\varphi(\omega) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\omega x} p(x) \mathrm{d}x. \tag{A.5}$$

The empirical characteristic function (e.c.f.) is

$$\varphi_n(\omega) = \frac{1}{n} \sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}\omega X_k}. \tag{A.6}$$

Let $\hat{p}_{n,h}$ be a kernel density estimator of p defined by

$$\hat{p}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h} \Big), \tag{A.7}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an even integrable function such that $\int K(u) \mathrm{d}u = 1$ and $h > 0$ is the bandwidth parameter. Then, the Fourier transform of a kernel estimator defined in (A.7) presented in terms of its e.c.f. is given by

$$\mathcal{F}[\hat{p}_{n,h}](\omega) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\omega x}(x) \mathrm{d}[\hat{p}_{n,h} x = \varphi_n(\omega) \mathcal{F}[K](h\omega). \tag{A.8}$$

The next lemma is similar to Stone (1984).

LEMMA A.6. *Let p be any density in $L_2(\mathbb{R})$ and $K \in L_1(\mathbb{R})$ be a symmetric kernel satisfying $\int K(x) \mathrm{d}x = 1$ and one of the two following conditions*

(i) *K is non-negative,*

(ii) *K is a kernel of order $2s$ for a positive integer s (i.e. with moments $\alpha_k = \int t^k K(t) \mathrm{d}t = 0,\ \forall\, 1 \leq k < 2s$) such that $(-1)^{s+1} \alpha_{2s} > 0$ and $|K|$ has a finite absolute moment of order $2s + \delta,\ 0 < \delta \leq 1$, that is, $\beta_{2s+\delta} \triangleq \int |t|^{2s+\delta} |K(t)| \mathrm{d}t < \infty$.*

Then, there exists positive constants c, depending on p and K and $a < 1$, depending on K, such that

$$\inf_{h>0} R_n(\hat{p}_{n,h}, p) \geq c n^{-a}. \tag{A.9}$$

PROOF. The usual bias/variance decomposition is given by

$$R_n(\hat{p}_{n,h}, p) = E_p \big\| \hat{p}_{n,h} - E_p\left[\hat{p}_{n,h}\right] \big\|^2 + \big\| E_p\left[\hat{p}_{n,h}\right] - p \big\|^2.$$

To bound the bias from below, we begin as in the proof of Stone (1984). According to the Plancherel identity,

$$2\pi \big\| E_p\left[\hat{p}_{n,h}\right] - p \big\|^2 = \int_{\mathbb{R}} |\mathcal{F}[K](h\omega)\varphi(\omega) - \varphi(\omega)|^2 \mathrm{d}\omega = \int_{\mathbb{R}} \left(1 - \mathcal{F}[K](h\omega)\right)^2 |\varphi(\omega)|^2 \mathrm{d}\omega.$$

Since φ is continuous and $\varphi(0) = 1$, there exists $\eta > 0$ such that $|\varphi(\omega)|^2 \geq 1/2$ for $|\omega| \leq \eta$. Then,

$$\int_{\mathbb{R}} \left(1 - \mathcal{F}[K](h\omega)\right)^2 |\varphi(\omega)|^2 \mathrm{d}\omega \geq \frac{1}{2} \int_{-\eta}^{\eta} \left(1 - \mathcal{F}[K](h\omega)\right)^2 \mathrm{d}\omega.$$

• Suppose that K satisfies *(i)*. Then it is a probability density and $\mathcal{F}[K]$ is its characteristic function. From Lukacs (1970, Theorem 4.1.2), we obtain

$$\int_{-\eta}^{\eta} \left(1 - \mathcal{F}[K](h\omega)\right)^2 \mathrm{d}\omega \geq \frac{1}{2^{4m}} \int_{-\eta}^{\eta} \left(1 - \mathcal{F}[K](2^m h\omega)\right)^2 \mathrm{d}\omega, \quad \forall\, m \in \mathbb{N}.$$

Now choose m such that $2^{-(m+1)} \leq h < 2^{-m}$:

$$\int_{-\eta}^{\eta} \left(1 - \mathcal{F}[K](h\omega)\right)^2 \mathrm{d}\omega \geq h^4 \int_{-\eta/2}^{\eta/2} \left(1 - \mathcal{F}[K](\omega)\right)^2 \mathrm{d}\omega = c_1 h^4,$$

where $c_1 = \int_{-\eta/2}^{\eta/2} \left(1 - \mathcal{F}[K](\omega)\right)^2 \mathrm{d}\omega$ is a positive constant. Indeed if $c_1 = 0$ then, by continuity,

$$1 - \mathcal{F}[K](\omega) = 0, \quad \forall\, \omega \in (-\eta/2, \eta/2).$$

Thus, by Lukacs (1970, Theorem 4.1.1), $\mathcal{F}[K] \equiv 1$ which contradicts the conclusion of the Riemann-Lebesgue lemma. Therefore, there exists a positive constant c_1 such that

$$\left\| E_p\left[\hat{p}_{n,h}\right] - p \right\|^2 \geq \frac{c_1}{4\pi}(h^4 \wedge 1), \qquad \forall\, h \in \mathbb{R}_+. \tag{A.10}$$

• Suppose now that K satisfies *(ii)*. By Theorem 2.2.1 of Lukacs (1983), since K is symmetric and of order $2s$, $\mathcal{F}[K](t)$ admits an expansion of the form

$$\mathcal{F}[K](t) = 1 + \sum_{k=1}^{s} (-1)^k \frac{\alpha_{2k} t^{2k}}{(2k)!} + O\left(|t|^{2s+\delta}\right) = 1 + (-1)^s \frac{\alpha_{2s} t^{2s}}{(2s)!} + O\left(|t|^{2s+\delta}\right),$$

as $|t| \to 0$. Note that Lukacs' theorem is stated for characteristic functions but it can be easily extended to our case. It is enough to suppose that $h \leq h_0$ for some positive constant h_0.

If $h > h_0$, the bias is bounded from below by a positive constant and the lemma is trivially proved. For $h \leq h_0$ and $|\omega| \leq \eta$ with sufficiently small η, $1 - \mathcal{F}[K](h\omega) \geq (-1)^{s+1} \alpha_{2s} h^{2s} \omega^{2s} / [2(2s)!] > 0$. Therefore, for $h \leq h_0$,

$$\left\| E_p\left[\hat{p}_{n,h}\right] - p \right\|^2 \geq c_2 h^{2s}, \tag{A.11}$$

where c_2 is a positive constant depending on K.

If $h \leq h_0$ for sufficiently small h_0, the variance term can be written (see Tsybakov, 2004a, Proposition 1.7)

$$\begin{aligned} E_p \left\| \hat{p}_{n,h} - E_p\left[\hat{p}_{n,h}\right] \right\|^2 &= \frac{1}{nh^2} E_p \left[\int_{\mathbb{R}} K^2\left(\frac{X_1 - x}{h}\right) \mathrm{d}x - \left(\int_{\mathbb{R}} E_p\left[K\left(\frac{X_1 - x}{h}\right)\right] \mathrm{d}x \right)^2 \right] \\ &\geq \frac{c_3}{nh}, \quad c_3 > 0. \end{aligned}$$

Using (A.10) and (A.11), we find that in both cases *(i)* and *(ii)* of the lemma there exists a positive constant $\tilde{a}$ such that

$$\inf_{h>0} R_n(\hat{p}_{n,h}, p) \geq \inf_{h>0} \left[c_4 h^{\tilde{a}} + \frac{c_3}{nh} \right] \geq c n^{-a},$$

for positive constants c and $a < 1$. ∎

3. Technical lemma

The following technical result is proved in Nemirovskiĭ (1992, Appendix) (see also Juditsky and Nemirovski (2000, Lemma 2.1.)). For $q \geq 1$, denote by $\|\cdot\|_q$ the ℓ_q-norm on $\mathbb{R}^M$.

LEMMA A.7. *Let $M > 2$ and $q = 2\log M$. Then the function $W : \mathbb{R}^M \to \mathbb{R}$ defined by*

$$W(z) = \frac{1}{2}\|z\|_q^2$$

satisfies for every $z, d \in \mathbb{R}^M$ the inequality

$$W(z+d) \leq W(z) + d^\top \nabla W(z) + 4\mathrm{e}\log M \|d\|_\infty^2 ,$$

where ∇W denotes the gradient of W.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest, pp. 267–281.

Audibert, J.-Y. (2006). Model selection type aggregation with better variance control. Tech. rep., Laboratoire certis. Ecole Nationale des Ponts et Chaussées. Available at http://cermics.enpc.fr/∼audibert/.

Audibert, J.-Y. and Tsybakov, A. (2005). Fast learning rates for plug-in classifiers under the margin condition. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. To appear in the Annals of Statistics. Available at http://hal.ccsd.cnrs.fr/ccsd-00005882.

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. Probab. Lett.*, **65**(4), 411–417.

Baíllo, A., Cuesta-Albertos, J. A., and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.*, **53**(1), 27–35.

Balcan, M. F. and Blum, A. (2005). A PAC-style model for learning from labeled and unlabeled data. In P. Auer and R. Meir, eds., *COLT*, *Lecture Notes in Computer Science*, vol. 3559. Springer, pp. 111–126.

Barron, A. (1987). Are Bayes rules consistent in information? In T. Cover and B. Gopinath, eds., *Open Problems in Communication and Computation*. Springer-Verlag, New York, pp. 85–91.

Bartlett, P., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, **48**, 85–113.

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on riemannian manifolds. *Mach. Learn.*, **56**(1-3), 209–239.

Belomestny, D. and Spokoiny, V. (2004). Local likelihood modelling via stagewise aggregation. Tech. rep., WIAS-Berlin. Available at http://www.wias-berlin.de/publications/preprints/1000.

Ben-Tal, A. and Nemirovski, A. (1999). The conjugate barrier mirror descent method for non-smooth convex optimization. Tech. rep., MINERVA Optim. Center Report. Available at http://iew3.technion.ac.il/Labs/Opt/opt/Pap/CP_MD.pdf.

Bickel, P. J. and Ritov, Y. (2004). The golden chain. discussion of Boosting papers. *Ann. Statist.*, **32**(1), 91–96.

Bickel, P. J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *J. Mach. Learn. Res.*, **7**, 705–732.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, **71**(2), 271–291.

Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, *IMS Lecture Notes Monogr. Ser.*, vol. 36. Inst. Math. Statist., Beachwood, OH, pp. 113–133.

Birgé, L. (2003). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://www.proba.jussieu.fr/mathdoc/preprints.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, **3**(3), 203–268.

Blum, A. and Mitchell, T. M. (1998). Combining labeled and unlabeled data with co-training. In *COLT*. pp. 92–100.

Boĭko, L. L. and Golubev, G. K. (2000). How to improve the nonparametric density estimator in S-PLUS. *Problemy Peredachi Informatsii*, **36**(4), 80–88.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, **9**, 323–375 (electronic).

Bousquet, O., Chapelle, O., and Hein, M. (2004). Measure based regularization. In L. S. Thrun, S. and B. Schölkopf, eds., *NIPS*, vol. 16. MIT Press, Cambridge, MA USA.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**(2), 353–360.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *J. Amer. Statist. Assoc.*, **98**(462), 324–339.

Bunea, F. and Nobel, A. (2005). Sequential procedures for aggregating arbitrary estimators of a conditional mean. Tech. rep. Available at http://stat.fsu.edu/∼flori.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2004). Aggregation for Gaussian regression. Tech. rep. Available at http://stat.fsu.edu/∼flori.

Butucea, C. (2001). Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM Probab. Statist.*, **5**, 1–31 (electronic).

Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**(3), 898–924.

Castelli, V. and Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recogn. Lett.*, **16**(1), 105–111.

Castelli, V. and Cover, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Inform. Theory*, **42**(6, part 2), 2102–2117.

Catoni, O. (1997). A mixture approach to universal model selection. Tech. rep., Preprint LMENS-97-30, Ecole Normale Supérieure. Available at http://www.dma.ens.fr/edition/preprints/1997/titre97.html.

Catoni, O. (1999). "Universal" aggregation rules with exact bias bounds. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://www.proba.jussieu.fr/mathdoc/preprints.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, *Lecture Notes in Mathematics*, vol. 1851. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, **30**(3), 843–874. Dedicated to the memory of Lucien Le Cam.

Cavalier, L. and Tsybakov, A. B. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.*, **10**(3), 247–282. Meeting on Mathematical Statistics (Marseille, 2000).

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, Cambridge.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. In *NIPS*. pp. 57–64.

Cline, D. B. H. (1988). Admissible kernel estimators of a multivariate density. *Ann. Statist.*, **16**(4), 1421–1427.

Csörgő, S. (1985). Rates of uniform convergence for the empirical characteristic function. *Acta Sci. Math. (Szeged)*, **48**(1-4), 97–102.

Csörgő, S. and Totik, V. (1983). On how long interval is the empirical characteristic function uniformly consistent? *Acta Sci. Math. (Szeged)*, **45**(1-4), 141–149.

Cuevas, A., Febrero, M., and Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, **36**(4), 441–459.

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.*, **25**(6), 2300–2312.

Cuevas, A., González-Manteiga, W., and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Aust. N. Z. J. Stat.*, **48**(1), 7–19.

d'Alché Buc, F., Grandvalet, Y., and Ambroise, C. (2001). Semi-supervised marginboost. In T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., *NIPS*. MIT Press, pp. 553–560.

Dalelane, C. (2004). *Data driven kernel choice in non-parametric curve estimation*. Ph.D. thesis, Technische Universität Braunschweig.

Dalelane, C. (2005a). Exact minimax risk for density estimators in non-integer Sobolev classes. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://hal.ccsd.cnrs.fr/ccsd-00004754.

Dalelane, C. (2005b). Exact oracle inequality for a sharp adaptive kernel density estimator. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://hal.ccsd.cnrs.fr/ccsd-00004753.

Devroye, L. (1997). Universal smoothing factor selection in density estimation: theory and practice. *Test*, **6**(2), 223–320. With discussion and a rejoinder by the author.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, *Applications of Mathematics (New York)*, vol. 31. Springer-Verlag, New York.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York.

Devroye, L. and Penrod, C. S. (1984). Distribution-free lower bounds in density estimation. *Ann. Statist.*, **12**(4), 1250–1262.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, **38**(3), 480–488.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**(432), 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, **57**(2), 301–369. With discussion and a reply by the authors.

Efroĭmovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Teor. Veroyatnost. i Primenen.*, **30**(3), 524–534.

Efroĭmovich, S. Y. and Pinsker, M. S. (1982). Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii*, **18**(3), 19–38.

Efroĭmovich, S. Y. and Pinsker, M. S. (1984). A self-training algorithm for nonparametric filtering. *Automt. Remote Control*, (11), 1434–1440.

Efromovich, S. (2000). On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.*, **9**(2), 117–139.

Efromovich, S. (2004). Oracle inequalities for Efromovich-Pinsker blockwise estimates. *Methodol. Comput. Appl. Probab.*, **6**(3), 303–322.

Ekeland, I. and Temam, R. (1976). *Convex analysis and variational problems*. North-Holland Publishing Co., Amsterdam. Translated from the French, Studies in Mathematics and its Applications, Vol. 1.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Teor. Verojatnost. i Primenen.*, **14**, 156–162.

Gilbert, E. N. (1952). A comparison of signaling alphabets. *Bell System Technical Journal*, **31**, 504–522.

Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.*, **29**(6), 1601–1619.

Golubev, G. K. (1990). Quasilinear estimates for signals in L_2. *Problemy Peredachi Informatsii*, **26**(1), 19–24.

Golubev, G. K. (1991). Local asymptotic normality in problems of nonparametric estimation of functions, and lower bounds for quadratic risks. *Teor. Veroyatnost. i Primenen.*, **36**(1), 143–149.

Golubev, G. K. (1992). Nonparametric estimation of smooth densities of a distribution in L_2. *Problemy Peredachi Informatsii*, **28**(1), 52–62.

Golubev, G. K. and Levit, B. Y. (1996). Distribution function estimation: adaptive smoothing. *Math. Methods Statist.*, **5**(4), 383–403.

Golubev, G. K. and Nussbaum, M. (1992). Adaptive spline estimates in a nonparametric regression model. *Teor. Veroyatnost. i Primenen.*, **37**(3), 554–561.

Hall, P., Kerkyacharian, G., and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, **26**(3), 922–942.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*, *Lecture Notes in Statistics*, vol. 129. Springer-Verlag, New York.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, **82**(397), 267–270.

Hartigan, J. H. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA.

Herbei, R. and Wegkamp, M. (2006). Classification with rejection option. *Canad. J. Statist.* To appear.

Hertz, T., Bar-Hillel, A., and Weinshall, D. (2004). Boosting margin based distance functions for clustering. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. ACM Press, New York, NY, USA, p. 50.

Ibragimov, I. A. and Khas′minskiĭ, R. Z. (1981). *Statistical estimation*, *Applications of Mathematics*, vol. 16. Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

Ibragimov, I. A. and Khas′minskiĭ, R. Z. (1982). An estimate of the density of a distribution belonging to a class of entire functions. *Teor. Veroyatnost. i Primenen.*,

27(3), 514–524.

Ibragimov, I. A. and Khas′minskiĭ, R. Z. (1991). Asymptotically normal families of distributions and efficient estimation. *Ann. Statist.*, **19**(4), 1681–1724.

Iouditski, A., Nazin, A., Tsybakov, A., and Vayatis, N. (2005). Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Problems of Information Transmission*, **41**(4), 368–384.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, **28**(3), 681–712.

Juditsky, A., Rigollet, P., and Tsybakov, A. (2006). Learning by mirror averaging. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://hal.ccsd.cnrs.fr/ccsd-00014097.

Kivinen, J. and Warmuth, M. K. (1999). Averaging expert predictions. In *EuroCOLT '99: Proceedings of the 4th European Conference on Computational Learning Theory*. Springer-Verlag, London, UK, pp. 153–167.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.*, **22**(2), 835–866.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, **35**(3), 459–470.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, **36**(4), 645–659.

Lepskiĭ, O. V. (1992a). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, **37**(3), 468–481.

Lepskiĭ, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, *Adv. Soviet Math.*, vol. 12. Amer. Math. Soc., Providence, RI, pp. 87–106.

Leung, G. and Barron, A. (2004). Information theory and mixing least-squares regressions. Tech. rep. Available at http://people.qualcomm.com/gleung/mixLS/.

Li, J. and Barron, A. (1999). Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Müller, eds., *NIPS*. The MIT Press, pp. 279–285.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, **15**(3), 958–975.

Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.*, **32**(4), 1679–1697.

Lukacs, E. (1970). *Characteristic functions*. Hafner Publishing Co., New York. Second edition, revised and enlarged.

Lukacs, E. (1983). *Developments in characteristic function theory*. Macmillan Co., New York.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27**(6), 1808–1829.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**(2), 712–736.

Massart, P. (2006). *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Springer-Verlag, Berlin. Lecture notes from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003. To appear.

Molchanov, I. S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Statist.*, **25**(1), 235–242.

Müller, D. W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Tech. Rep. 398, SFB 123, Univ. Heidelberg.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, *Lecture Notes in Math.*, vol. 1738. Springer, Berlin, pp. 85–277.

Nemirovskiĭ, A. S. (1992). On nonparametric estimation of functions satisfying differential inequalities. In *Topics in nonparametric estimation*, *Adv. Soviet Math.*, vol. 12. Amer. Math. Soc., Providence, RI, pp. 7–43.

Parzen, E. (1958). On asymptotically efficient consistent estimates of the spectral density function of a stationary time series. *J. Roy. Statist. Soc. Ser. B*, **20**, 303–322.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

Petrov, V. V. (1995). *Limit theorems of probability theory*, *Oxford Studies in Probability*, vol. 4. The Clarendon Press Oxford University Press, New York. Sequences of independent random variables, Oxford Science Publications.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. of Inf. Transm.*, **16**, 52–68.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, **23**(3), 855–881.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, **69**, 1–24.

Polyak, B. T. and Tsybakov, A. B. (1990). Asymptotic optimality of the C_p-test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, **35**(2), 305–317.

Polyak, B. T. and Tsybakov, A. B. (1992). A family of asymptotically optimal methods for selecting the order of a projection estimator for a regression. *Teor. Veroyatnost. i Primenen.*, **37**(3), 502–512.

Rattray, M. (2000). A model-based distance for clustering. In *Proc. of the IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks*. IEEE Computer Society Press, pp. IV–13. ISBN 0769506216.

Rigollet, P. (2004). Adaptive density estimation using Stein's blockwise method. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2004.

Rigollet, P. (2006a). Adaptive density estimation using the blockwise Stein method. *Bernoulli*, **12**(2), 351–370.

Rigollet, P. (2006b). Generalization error bounds in semi-supervised classification under the cluster assumption. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://hal.ccsd.cnrs.fr/ccsd-00022528.

Rigollet, P. and Tsybakov, A. (2006). Linear and convex aggregation of density estimators. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://hal.ccsd.cnrs.fr/ccsd-00068216.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**(2), 65–78.

Samarov, A. and Tsybakov, A. (2005). Aggregation of density estimators and dimension reduction. In *Festschrift in honor of Kjell Doksum*. To appear. Available at http://hal.ccsd.cnrs.fr/ccsd-00014122.

Schipper, M. (1996). Optimal rates and constants in L_2-minimax estimation of probability density functions. *Math. Methods Statist.*, **5**(3), 253–274.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, **13**, 1443–1471.

Scott, D. W. (1992). *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. Theory, practice, and visualization, A Wiley-Interscience Publication.

Seeger, M. (2000). Learning with labeled and unlabeled data. Tech. rep., Institute for ANC, Edinburgh, UK. Available at http://www.dai.ed.ac.uk/∼seeger/papers.html.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, **53**(3), 683–690.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**(1), 45–54.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**(6), 1135–1151.

Steinwart, I., Hush, D., and Scovel, C. (2005). Density level detection is classification. In L. K. Saul, Y. Weiss, and L. Bottou, eds., *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**(4), 1285–1297.

Tarigan, B. and van de Geer, S. (2006). Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, **12**(6). To appear.

Tipping, M. (1999). Deriving cluster analytic distance functions from Gaussian mixture models. In *ICANN*.

Tsybakov, A. (2002). Discussion of 'Random rates in anisotropic regression' by M. Hoffmann and O. Lepski. *Ann. Statist.*, **30**(2), 379–385.

Tsybakov, A. (2003). Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, eds., *COLT*, *Lecture Notes in Computer Science*, vol. 2777. Springer, pp. 303–313.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, **25**(3), 948–969.

Tsybakov, A. B. (2004a). *Introduction à l'estimation non-paramétrique*, *Mathématiques & Applications (Berlin) [Mathematics & Applications]*, vol. 41. Springer-Verlag, Berlin.

Tsybakov, A. B. (2004b). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32**(1), 135–166.

Tsybakov, A. B. and van de Geer, S. A. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, **33**(3), 1203–1224.

van de Geer, S. A. (2000). *Applications of empirical process theory*, *Cambridge Series in Statistical and Probabilistic Mathematics*, vol. 6. Cambridge University Press, Cambridge.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New-York.

Vovk, V. (1990). Aggregating strategies. In M. A. Fulk and J. Case, eds., *COLT*. Morgan Kaufmann, pp. 371–383.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, *Monographs on Statistics and Applied Probability*, vol. 60. Chapman and Hall Ltd., London.

Watson, G. S. and Leadbetter, M. R. (1963). On the estimation of the probability density. I. *Ann. Math. Statist.*, **34**, 480–491.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.*, **31**(1), 252–273.

Wegkamp, M. H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.*, **27**(2), 409–420.

Yang, Y. (1999). Minimax nonparametric classification — part I: rates of convergence. *IEEE Trans. Inform. Theory*, **45**(7), 2271–2284.

Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28**(1), 75–87.

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10**(1), 25–47.

Zhang, T. (2006). From ε-entropy to kl-complexity: Analysis of minimum information complexity density estimation. *Ann. Statist.*, **34**(5). To appear.

Zhu, X. (2005). Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison. Http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

Inégalités d'oracle, agrégation et adaptation

Résumé : Historiquement, les inégalités d'oracle ont été développées comme des outils particulièrement efficaces pour l'adaptation à un paramètre inconnu en statistique mathématique. Initialement dédiées à la démonstration de propriétés statistiques de certains estimateurs, elles peuvent s'inscrire dans le cadre plus général du problème d'agrégation où elles sont au centre de la définition d'une *vitesse optimale d'agrégation*. Elles constituent alors d'une part des outils mathématiques et d'autre part des résultats précis et non asymptotiques. Les travaux faisant l'objet de cette thèse présentent différentes utilisations des inégalités d'oracle, d'abord dans un cadre général d'agrégation puis dans des modèles statistiques plus particuliers, comme l'estimation de densité et la classification. Les résultats obtenus sont une palette non exhaustive mais représentative de l'utilisation des inégalités d'oracle en statistique mathématique.

Mots-clés : Inégalités d'oracle, agrégation, apprentissage statistique, optimisation stochastique, estimation adaptative, bornes inférieures minimax.

Oracle inequalities, aggregation and adaptation

Abstract: Originally, oracle inequalities were developed as particularly efficient tools in mathematical statistics for deriving adaptation to an unknown parameter. Initially dedicated to the demonstration of statistical properties in the study of certain estimators, they can be extrapolated to the broader framework of aggregation in which they constitute the core of the definition of an *optimal rate of aggregation*. As such, they become not only mathematical tools but also precise finite sample results. In this thesis, several fields of application of oracle inequalities are presented. They are used first in a general aggregation framework and then in particular statistical models such as nonparametric density estimation, regression and classification. The obtained results map the variety of applications of oracle inequalities in mathematical statistics.

Keywords: Oracle inequalities, aggregation, statistical learning, stochastic optimization, adaptive estimation, minimax lower bounds.