Aggregation of estimators and classifiers : theory and methods

Benjamin Guedj

École Doctorale 386
Sciences Mathématiques de Paris Centre

THÈSE DE DOCTORAT

en vue de l'obtention du grade de

Docteur ès Sciences de l'Université Pierre et Marie Curie

Discipline : Mathématiques

Spécialité : Statistique

présentée par

Benjamin GUEDJ

AGRÉGATION D'ESTIMATEURS ET DE CLASSIFICATEURS : THÉORIE ET MÉTHODES

---

dirigée par MM. Gérard BIAU et Éric MOULINES

Au vu des rapports établis par
MM. Arnak DALALYAN et Jean-Michel MARIN

Soutenue publiquement le 4 décembre 2013 devant le jury composé de :

M.	Pierre	ALQUIER	University College Dublin	examinateur
M.	Gérard	BIAU	UPMC	directeur
M.	Arnak	DALALYAN	ENSAE ParisTech	rapporteur
M.	Jean-Michel	MARIN	Université Montpellier II	rapporteur
M.	Éric	MOULINES	Telecom ParisTech	directeur
M.	Olivier	WINTENBERGER	UPMC	président

Laboratoire de Statistique Théorique et Appliquée (LSTA)
Université Pierre & Marie Curie (UPMC)
Tour 25, couloirs 15-25 & 15-16
2ème étage – Boîte n°158
4, place Jussieu – 75252 Paris Cedex 05

Laboratoire Traitement et Communication de l'Information (LTCI)
UMR 5141 CNRS & Telecom ParisTech
46 rue Barrault – F-75634 Paris Cedex 13

École Doctorale 386 – Sciences Mathématiques de Paris Centre
Université Pierre & Marie Curie (UPMC)
Tour 25, couloir 15-25, bureau 115
1er étage – Boîte n°290
4, place Jussieu – 75252 Paris Cedex 05

À la mémoire
d'Odette Dubois,
Michel et Georgette Guedj

Remerciements

Si l'achèvement d'une thèse est un petit Everest personnel, l'écriture des remerciements en constitue sans hésitation la face nord. Permets-moi, lectrice, lecteur, de rendre ici un bref hommage aux rencontres qui m'ont amené jusqu'ici.

Gérard et Éric, pour avoir été d'extraordinaires directeurs tout au long de ce doctorat, c'est à vous que j'adresse mes premiers remerciements. Gérard, pour avoir toujours su, avec une rare habilité, doser encadrement et liberté de recherche. Pour ton soutien sans faille, fût-il mathématique ou plus personnel, dans les moments de doute. Pour la perfection de tes fichiers TEX, et ta réactivité surhumaine (dont j'espère avoir un peu hérité !). Pour avoir été celui par qui je suis arrivé à la statistique : c'est d'abord à toi que je dois d'être là. Éric, pour m'avoir fait partager ton incroyable culture scientifique. Pour m'avoir appris à prendre du recul sur mes propres recherches. Pour ton humour féroce et les heures passées à refaire le monde, fût-il académique ou politique. Pour la confiance, enfin, que vous m'avez tous deux témoignée, et pour tout ce qui n'est pas dans ces trop courtes lignes, merci.

À Arnak Dalalyan et Jean-Michel Marin, pour m'avoir fait l'honneur de rapporter ma thèse, j'exprime ma profonde reconnaissance. Vous savez tous deux combien vos recherches m'ont guidé : l'intérêt que vous avez porté à mes travaux est pour moi le plus bel encouragement à poursuivre dans cette voie.

Pierre, pour le merveilleux (et communicatif !) enthousiasme dont tu as su faire preuve, en me guidant dans cette jungle qu'était pour moi la théorie PAC-bayésienne au début de ma thèse.

Olivier, pour avoir sans hésiter accepté de faire partie de mon jury, et commencé à travailler avec moi alors que ton bureau n'était encore qu'un vaste champs de cartons.

Gilles, pour m'avoir, le premier, fait sentir combien la recherche, comme le sport de haut niveau, est une ascèse. De mon expérience danoise à tes côtés, je retiens la formation mathématique, et bien plus encore.

À Paul Deheuvels, je veux exprimer ma gratitude, pour m'avoir accueilli au sein de son laboratoire. Merci également pour avoir, de loin en loin, toujours eu des mots d'encouragement à mon égard.

D'une manière plus large, je veux remercier ici l'ensemble des membres du LSTA et du LPMA, qui ont contribué, de près ou de loin, à faire de ce doctorat une magnifique expérience. Mention toute particulière à Michel Broniatowski, John O'Quigley, et Daniel Pierre-Loti-Viaud pour les nombreuses discussions, mathématiques ou non, en partageant un café ou au détour d'un couloir. Pour tous les bons moments, les conseils avisés, et l'invraisemblable quantité de caféine partagée, merci à mes aînés, Agathe, Bertrand, Étienne, Fanny, Jean-Patrick, Olivier, Philippe, Stéphane. À l'ensemble des doctorants, passés et présents, pour ce savant mélange de procrastination et d'émulation, merci ! En première ligne, mes compagnons du bureau 208, et Cécile (je t'avais dit que la cohabitation avec un bayésien se passerait bien !). À mes frangins de thèse, Baptiste, Clément, Erwan, pour nos fines équi-

pées. Je veux dire ma reconnaissance et mon amitié à Corinne et Louise, à la gentillesse et l'efficacité sans limites.

J'ai eu la chance de trouver à Telecom ParisTech une seconde équipe pour m'accueillir : je veux ici en remercier les membres. Merci aux doctorants, et à Émilie et Sylvain, mes compagnons de bureau, et plus encore. À Joseph, prince de la bibliographie, une immense reconnaissance.

Une pensée toute particulière pour mes grands frères et sœur de thèse, Aurélie, Kevin et Julien, ainsi qu'à Christophe, Damien, Gaëlle, Laure, Lise-Marie, Olivier, Sarah, Sébastien et bien d'autres, qui se reconnaîtront dans ces lignes.

I would like to thank Jim Malley: Your enthusiasm has been a driving force in the COBRA adventure from the very beginning. To my many friends in Copenhagen, and former colleagues at DTU, I want to express my gratitude. *Jeg takker jer alle, se dig snart i København!*

Parmi mes nombreux professeurs, je tiens à distinguer Omer Adelman et Michel Vaugon : vous avez été les premiers à m'enseigner une forme d'esthétisme mathématique, et votre façon de faire m'a profondément marqué.

Plus largement, merci aux (très) nombreux membres de l'UPMC dont j'ai croisé la route. J'ai franchi pour la première fois les portes de cette grande université en septembre 2004. Près de dix ans et quelques diplômes plus tard, je crois avoir fait le bon choix !

J'ai eu la chance de participer à de nombreux séminaires et conférences : à toutes celles et ceux qui ont croisé mon chemin, je veux exprimer ma reconnaissance. Ces rencontres font de la communauté statistique une belle et dynamique famille, que j'ai toujours plaisir à retrouver. Mention toute particulière pour mes compagnons de Saint-Flour, où ce manuscrit a été rédigé en grande partie.

À mes ami-e-s : de lycée, du monde politique, de la fac, de voyages, d'ami-e-s, d'ailleurs. Si je suis heureux d'arriver au labo le matin, je le suis encore plus de le quitter le soir, et c'est grâce à vous. Et quand il ne doit plus en rester qu'un, c'est bien souvent à la coloc, avec Mehdi.

À Delphine, pour tout, tout simplement, et Faustine et Apolline.

À ma famille, et à la confiance qu'elle m'apporte. À mes parents, toujours présents dans les bons et moins bons moments. Votre fierté est la plus belle des récompenses.

À toi, enfin, lectrice, lecteur, qui parcours les premières pages de ce manuscrit. Ne t'y trompe pas : cette thèse, je crois, n'est que le début de l'aventure !

Table des matières

Avant-propos

Cette thèse a été financée par un contrat doctoral à l'Université Pierre & Marie Curie, du 1er février 2011 au 31 janvier 2014, et réalisée au Laboratoire de Statistique Théorique et Appliquée (LSTA), et au Laboratoire Traitement et Communication de l'Information (LTCI).

Plan du manuscrit

Le Chapitre 1 de ce manuscrit vise le double objectif de présenter le cadre mathématique des travaux de cette thèse et d'en synthétiser les principaux résultats. Nous présentons tout d'abord la problématique de l'apprentissage statistique et les approches considérées : théorie statistique de l'agrégation, approches oracle et minimax, contrainte de grande dimension, et les deux grandes familles d'estimateurs que sont les méthodes pénalisées et les agrégés à poids exponentiels. Une majorité des résultats contenus dans ce manuscrit s'inscrit dans l'approche PAC-bayésienne, qui fait ensuite l'objet d'un rappel, ainsi que les différentes implémentations présentes dans la littérature. Nous présentons enfin un panorama des résultats issus de la thèse : agrégation PAC-bayésienne pour les modèles additifs et logistiques en grande dimension, agrégation non linéaire d'estimateurs de la fonction de régression, modélisation bayésienne de l'hybridation de populations.

Le Chapitre 2 est la reproduction *in extenso* d'un article publié dans *Electronic Journal of Statistics* (Guedj and Alquier, 2013), écrit en collaboration avec Pierre Alquier (University College Dublin). Nous présentons une approche PAC-bayésienne pour l'estimation dans les modèles additifs de grande dimension. Les estimateurs introduits sont optimaux au sens minimax (à un terme logarithmique près), et sont implémentés dans le paquet *pacbpred* de R (Guedj, 2013b).

Le Chapitre 3 transpose les techniques du chapitre précédent au problème de l'estimation dans les modèles logistiques en grande dimension, à l'aide d'outils PAC-bayésiens. Les inégalités oracles qui y sont introduites justifient l'utilisation comme estimateur de l'agrégé *a posteriori* de Gibbs. L'algorithme permettant de calculer cet estimateur est également présenté.

Le Chapitre 4 est un travail réalisé en collaboration avec Gérard Biau (Université Pierre & Marie Curie, et Institut Universitaire de France), Aurélie Fischer (Université Denis Diderot) et James D. Malley (National Institutes of Health). Nous proposons une stratégie non linéaire d'agrégation d'estimateurs de la fonction de régression, soutenue par des résultats oracles et une vitesse de convergence explicite. La méthode est également présentée dans Biau et al. (2013) et implémentée dans le paquet *COBRA* de R (Guedj, 2013a).

Enfin, le Chapitre 5 reprend un article publié dans *Molecular Ecology Resources* (Guedj and Guillot, 2011), en collaboration avec Gilles Guillot (Danmarks Tekniske Universitet). Dans des travaux antérieurs au sujet de cette thèse, nous introduisons une modélisation

bayésienne spatiale de la présence d'hybridation génétique entre des populations. Cette approche est implémentée par des techniques de Monte-Carlo par chaînes de Markov (MCMC) dans le paquet *Geneland* de R.

Ce manuscrit se referme par une courte annexe contenant les principaux lemmes techniques utilisés.

Avertissement

Les Chapitres 2 à 5 se présentent avec leurs propres notations et peuvent être lus indépendamment les uns des autres.

Chapitre 1

État de l'art et contributions

Ce mémoire de doctorat est consacré à l'étude théorique et l'implémentation de procédures d'agrégation d'estimateurs dans un contexte de grande dimension.

Sommaire

1.1 État de l'art

Sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, considérons une variable aléatoire $(\mathbf{X}, Y)$ à valeurs dans $\mathcal{X} \times \mathcal{Y}$ et de loi inconnue $\mathcal{P}$. Nous nommerons *design* (respectivement *réponse*) la variable aléatoire $\mathbf{X}$ (respectivement Y), et nous nous limiterons au cas $\mathcal{X} \subseteq \mathbb{R}^d$. Les deux problèmes abordés dans ce manuscrit sont la régression ($\mathcal{Y} \subseteq \mathbb{R}$) et, dans une moindre mesure, la classification ($\mathcal{Y} \subseteq \mathbb{N}$).

L'une des principales tâches du statisticien consiste, à l'aide d'un n-échantillon $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de réplications indépendantes et identiquement distribuées de $(\mathbf{X}, Y)$, à inférer tout ou partie des propriétés de $\mathcal{P}$. Cette formulation, bien trop vaste, n'admet pas de solution générale : il est nécessaire de raffiner le contexte pour prétendre y répondre. Une première approche repose sur l'hypothèse que $\mathcal{P}$ appartient à un sous-ensemble restreint de l'ensemble des mesures de probabilités sur l'espace mesurable $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$, où $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$ désigne la σ-algèbre engendrée par les boréliens de $\mathcal{X} \times \mathcal{Y}$. Il est par exemple commode de supposer que $\mathcal{P}$ est une loi classique dont les paramètres sont à estimer ou que $\mathcal{P}$ admet certaines propriétés de régularité. La seconde approche — adoptée dans ce mémoire — est à l'origine de la théorie statistique de l'apprentissage, pendant théorique du *machine learning*, développée initialement par Cervonenkis and Vapnik (1968) et formalisée, entre

autres auteurs, par Vapnik (1998, 2000). Il ne s'agit plus de faire d'hypothèses sur $\mathcal{P}$, mais plutôt sur l'ensemble des procédures à la disposition du statisticien. Le jeu consiste alors non pas à approcher des quantités inconnues liées à $\mathcal{P}$, mais à imiter le meilleur élément de cet ensemble de procédures, au sens d'un certain critère de contraste. Cette approche, ouvrant la voie à la *minimisation structurelle du risque empirique*, est présentée dans de très nombreux travaux. Pour une introduction au sujet, nous renvoyons le lecteur aux articles Bousquet et al. (2004) et Bartlett et al. (2004) et aux ouvrages Devroye et al. (1996) et Hastie et al. (2009).

Les principaux objets d'intérêt pour la régression et la classification diffèrent : il s'agit respectivement de la *fonction de régression*

$$\begin{array}{rccl} f: & \mathcal{X} & \to & \mathbb{R} \\ & \mathbf{x} & \mapsto & \mathbb{E}[Y|\mathbf{X}=\mathbf{x}], \end{array}$$

et de l'ensemble des probabilités *a posteriori* :

$$\left\{\mathbb{P}(Y=k|\mathbf{X}=\mathbf{x}), \quad \mathbf{x}\in\mathcal{X}, \quad k\in\mathcal{Y}\right\}.$$

Si les méthodes d'apprentissage statistique cherchent globalement à "apprendre" le lien entre *design* et réponse, différents points de vue peuvent être envisagés. Considérons, à titre d'exemple, le modèle de régression linéaire $Y=\theta^\top\mathbf{X}+W$, où W désigne un bruit réel et $\theta\in\mathbb{R}^d$ le vecteur des cœfficients de la régression. Les trois principaux problèmes à résoudre sont les suivants :

- **Identification du support, sélection de variables :** estimer le support du vecteur θ (*i.e.*, les indices des composantes non nulles) pour sélectionner les covariables d'intérêt,
- **Estimation :** estimer le vecteur θ,
- **Prévision :** pour une réalisation $\mathbf{x}$ du *design* $\mathbf{X}$, estimer $f(\mathbf{x})=\theta^\top\mathbf{x}$.

Le contour de ces nuances est parfois flou dans la littérature, et nous les désignerons collectivement dans la suite par problèmes d'inférence.

Pour apprécier la pertinence d'une stratégie d'apprentissage, il est classique d'adopter le formalisme suivant, issu de la théorie de l'information. Considérons une *fonction de perte* $\ell:\mathcal{Y}\times\mathcal{Y}\to(0,\infty)$, le *risque* associé à un estimateur $\hat{f}$ de la fonction de régression f ou à un classifieur $\hat{f}$ est

$$R(\hat{f})=\mathbb{E}\ell(\hat{f}(\mathbf{X}),Y).$$

Comme ce risque dépend de la distribution inconnue $\mathcal{P}$, il est souvent avantageux de lui substituer sa version empirique construite sur l'échantillon $\mathcal{D}_n$, notée

$$R_n(\hat{f})=\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{f}(\mathbf{X}_i),Y_i).$$

De très nombreuses fonctions de perte ont été étudiées dans la littérature. Nous nous intéressons dans ce mémoire aux trois fonctions classiques suivantes :

- $\ell:(u,v)\mapsto(u-v)^2$ (perte quadratique),
- $\ell:(u,v)\mapsto\mathbb{1}_{\{u\neq v\}}$ (perte de classification),
- $\ell:(u,v)\mapsto\log[1+\exp(-uv)]$ (perte logistique).

Précisons enfin qu'il existe différents contrôles théoriques dans la littérature. Reprenons l'exemple précédent du modèle de régression linéaire : si $\hat{\theta}$ désigne un estimateur de θ, chercher à borner $|\hat{\theta}-\theta|_2^2$ correspond au problème de l'estimation, alors que l'on cherchera à contrôler $|\mathbf{X}(\hat{\theta}-\theta)|_2^2$ pour valider une stratégie de prévision ($|\cdot|_2$ désigne la norme euclidienne usuelle).

1.1.1 Agrégation, approche oracle, théorie minimax

Face à la très grande diversité de techniques d'inférence dans la littérature, a émergé au cœur des années 1980 l'idée de les *agréger* pour tirer le meilleur parti de leurs avantages respectifs (voir, par exemple, Vovk, 1990; Littlestone and Warmuth, 1994). Cette piste de recherche s'est arrimée à la communauté statistique après avoir été abondamment étudiée dans les communautés du *machine learning*, du traitement d'image, *etc*. Certaines des principales procédures nées au cours des années 1990 comme le *bagging* (Breiman, 1996), le *boosting* (Freund, 1990; Schapire, 1990) ou les forêts aléatoires (Amit and Geman, 1997; Breiman, 2001 ; voir également Biau et al., 2008, Biau and Devroye, 2010, Genuer, 2010 et Biau, 2012 pour un panorama récent) sont encore aujourd'hui largement plébiscitées pour leurs performances expérimentales. L'agrégation a été au cœur de nombreux travaux, portant notamment sur l'estimation de densité (Rigollet, 2006; Rigollet and Tsybakov, 2007) et la classification (Catoni, 2004; Audibert and Tsybakov, 2007; Lecué, 2007b,a). Nous nous intéressons dans la suite de ce chapitre aux techniques d'agrégation dans le cadre de la régression.

Le but est d'estimer la fonction de régression f par une combinaison d'éléments d'un ensemble connu appelé dictionnaire, composé de fonctions déterministes (par exemple une base d'un espace fonctionnel comme $L^2(\mathcal{X})$) ou d'estimateurs préliminaires. Cette approche est particulièrement pertinente lorsqu'on ne dispose d'aucune indication quant à la nature des données. En notant le dictionnaire $\mathbb{D} = \{\varphi_1, \dots, \varphi_M\}$ avec $\varphi_1, \dots, \varphi_M : \mathcal{X} \to \mathcal{Y}$, introduisons l'espace vectoriel engendré par les éléments de $\mathbb{D}$

$$\mathcal{F}_\Theta = \left\{ f_\theta = \sum_{j=1}^{M} \theta_j \varphi_j : \theta = (\theta_1, \dots, \theta_M) \in \Theta \right\}, \tag{1.1}$$

pour un certain ensemble de paramètres $\Theta \subseteq \mathbb{R}^M$. Le choix de Θ dans (1.1) caractérise le type du problème d'agrégation, de sorte que nous sommes ramenés à choisir convenablement un élément $\hat{\theta}$ à l'aide de $\mathcal{D}_n$ tel que $\hat{f} = f_{\hat{\theta}}$ soit un bon estimateur de f. La première formalisation de la théorie statistique de l'agrégation — cadre de ce mémoire — remonte à Nemirovski (2000), étendue depuis par de nombreux auteurs (voir par exemple la discussion dans Yang, 2003, et le panorama dressé par Tsybakov, 2008). Il est d'usage de distinguer cinq grands types d'agrégation.

(1) **Sélection de modèle (MS) :** construire $\hat{\theta}^{\mathrm{MS}}$ tel que $f_{\hat{\theta}^{\mathrm{MS}}}$ imite le meilleur des éléments φ_j de $\mathbb{D}$. Dans ce cas,

$$\Theta = \{e_1, \dots, e_M\},$$

où les e_j sont les vecteurs de la base canonique de $\mathbb{R}^M$, *i.e.*, les vecteurs dont la j-ème composante vaut 1 et toutes les autres 0. Ainsi $f_{e_j} = \varphi_j$ pour tout $j = 1, \dots, M$.

(2) **Agrégation convexe (C) :** construire $\hat{\theta}^{\mathrm{C}}$ tel que $f_{\hat{\theta}^{\mathrm{C}}}$ imite la meilleure combinaison convexe des φ_j, c'est-à-dire $\Theta = \Lambda^M$ où Λ^M désigne le simplexe de $\mathbb{R}^M$:

$$\Lambda^M = \left\{ \theta = (\theta_1, \dots, \theta_M) \in \mathbb{R}^M : \sum_{j=1}^{M} \theta_j = 1, \quad \theta_j \geq 0, \quad j = 1, \dots, M \right\}.$$

(3) **Agrégation linéaire (L) :** construire $\hat{\theta}^{\mathrm{L}}$ tel que $f_{\hat{\theta}^{\mathrm{L}}}$ imite la meilleure combinaison linéaire des φ_j, c'est-à-dire $\Theta = \mathbb{R}^M$.

(4) **Agrégation convexe *sparse* (CS) :** construire $\hat{\theta}^{\mathrm{CS}}$ tel que $f_{\hat{\theta}^{\mathrm{CS}}}$ imite la meilleure combinaison convexe *parcimonieuse* des φ_j, c'est-à-dire que $\hat{\theta}^{\mathrm{CS}}$ a exactement s composantes non nulles, avec $s \in \{1, \dots, M\}$ petit devant M. En notant $\mathcal{B}_0(s)$ la boule de

rayon $s \in \{1, \dots, M\}$ dans $\mathbb{R}^M$ muni de la norme ℓ_0 :

$$\mathcal{B}_0(s) = \left\{ \theta \in \mathbb{R}^M : |\theta|_0 = \sum_{j=1}^{M} \mathbb{1}_{\{\theta_j \neq 0\}} \leq s \right\},$$

on obtient $\Theta = \Lambda^M \cap \mathcal{B}_0(s)$.

(5) **Agrégation linéaire *sparse* (LS) :** construire $\hat{\theta}^{\mathrm{LS}}$ tel que $f_{\hat{\theta}^{\mathrm{LS}}}$ imite la meilleure combinaison linéaire *parcimonieuse* des φ_j, c'est-à-dire $\Theta = \mathbb{R}^M \cap \mathcal{B}_0(s)$ avec la notation précédente.

Les deux derniers types d'agrégation sont au cœur des travaux présentés dans ce mémoire. L'hypothèse de parcimonie (on rencontre également l'anglicisme sparsité) signifie que la dimension effective de la fonction de régression n'est pas d mais $d_0 \ll n$. Ce paradigme s'est imposé en apprentissage ces dernières années avec l'émergence de ce que l'on appelle le *big data*, qui se caractérise par le fait que les jeux de données sont gigantesques tant par leurs tailles (n grand) que par la dimension des données (d grand). Ce cadre — habituel en génétique, par exemple — se justifie à mesure que s'accumulent les indices empiriques de représentation parcimonieuse de nombreux phénomènes en grande dimension : compression d'images, puces à ADN, *etc*.

L'utilisation massive des techniques d'agrégation a rendu nécessaire l'introduction de nouveaux outils pour attester de la qualité des procédures proposées. Le contrôle de la performance des méthodes d'agrégation repose essentiellement sur la *théorie minimax* et les *inégalités oracles*. Les inégalités oracles ont été développées initialement comme des outils particulièrement efficaces pour l'adaptation à un paramètre inconnu, et dédiées à la démonstration de propriétés statistiques d'estimateurs. Dans le cadre de l'agrégation, elles sont au cœur de la notion de vitesse optimale d'agrégation ou vitesse *minimax*. On dira d'un élément $\theta^\star \in \Theta$ (ou, de façon indifférenciée, de $f_{\theta^\star}$) qu'il est un oracle pour le problème d'agrégation si

$$R(f_{\theta^\star}) = \inf_{\theta \in \Theta} R(f_\theta),$$

où Θ dépend du problème d'agrégation envisagé, comme exposé plus haut. Comme le risque fait intervenir la distribution $\mathcal{P}$ inconnue, l'oracle $\theta^\star$ n'est bien sûr pas accessible. En revanche, il est parfois possible de construire des estimateurs $\hat{\theta}$ qui imitent les performances et le comportement de l'oracle $\theta^\star$ en terme de risque sans pour autant chercher à l'approcher. Cette propriété est au cœur de la notion d'inégalité oracle :

$$R(f_{\hat{\theta}}) \leq \mathcal{C} \inf_{\theta \in \Theta} \left\{ R(f_\theta) + \Delta_{n,M}(\theta) \right\},$$

avec grande probabilité, ou

$$\mathbb{E} R(f_{\hat{\theta}}) \leq \mathcal{C} \inf_{\theta \in \Theta} \left\{ R(f_\theta) + \Delta_{n,M}(\theta) \right\}, \tag{1.2}$$

où, dans les deux cas, $\mathcal{C} \geq 1$ est une quantité déterministe bornée et $\Delta_{n,M}(\cdot)$ un terme résiduel indépendant de $\mathcal{P}$, qui décroît vers 0 avec n. Dans une telle situation, le risque de $f_{\hat{\theta}}$ est du même ordre que celui de l'oracle $f_{\theta^\star}$, à la constante multiplicative $\mathcal{C}$ près.

Le terme d'oracle et plus précisémment d'inégalité oracle a été introduit par Donoho and Johnston (1994). Les premiers exemples d'inégalités oracles ont ensuite été développés dans des articles proches (Donoho and Johnston, 1995; Donoho et al., 1995) dans lesquels le risque de l'oracle $R(f_{\theta^\star})$ était appelé risque idéal. Les inégalités oracles permettent d'obtenir des propriétés non asymptotiques d'adaptation à l'oracle : en effet, lorsque l'oracle possède

des propriétés statistiques intéressantes, l'inégalité permet de les transmettre à l'estimateur. Nous retrouvons ici le célèbre compromis biais-variance : plus Θ sera grand (au sens de l'inclusion), plus l'oracle sur Θ aura de bonnes propriétés (*i.e.*, la quantité $\inf_{\theta\in\Theta} R(f_\theta)$, semblable à un biais, sera petite), au prix cependant d'un terme résiduel $\Delta_{n,M}(\theta)$ d'autant plus important. Il est donc en général impossible de minimiser à la fois les deux termes. Pour les ensembles Θ donnés par les cinq types d'agrégation décrits précédemment, il est alors naturel de chercher à obtenir le terme résiduel $\Delta_{n,M}(\cdot)$ le plus petit possible. Cette approche, utilisée par Yang and Barron (1999) pour l'estimation de densité et formalisée par Tsybakov (2003) dans le cadre qui est celui de ce mémoire, a permis de définir les vitesses minimax pour l'agrégation. Supposons que la fonction de régression f appartienne à un espace fonctionnel $\mathcal{F}_{\beta,L}$, par exemple où la dérivée d'ordre β est bornée par la constante positive L. On dira d'une séquence $\phi_{n,M,\beta}$ qu'elle est une vitesse optimale ou minimax s'il existe deux constantes positives c et C, indépendantes de β et L et vérifiant :

$$\sup_{f\in\mathcal{F}_{\beta,L}} \{R(f_{\hat\theta})-R(f)\} \le C\phi_{n,M,\beta} \quad \text{et} \quad \inf_{f_{\tilde\theta}} \sup_{f\in\mathcal{F}_{\beta,L}} \{R(f_{\tilde\theta})-R(f)\} \ge c\phi_{n,M,\beta}, \tag{1.3}$$

où $\inf_{f_{\tilde\theta}}$ est l'infimum pris sur tous les estimateurs de f basés sur l'échantillon $\mathcal{D}_n$. Par extension, un estimateur $f_{\hat\theta}$ vérifiant (1.3) sera réputé optimal ou minimax sur $\mathcal{F}_{\beta,L}$. Le cas des inégalités oracles en espérance (les inégalités en probabilité sont plus difficiles à obtenir) du type (1.2) avec constante $\mathcal{C}=1$ (on parle alors d'inégalité *sharp* ou exacte) est particulièrement intéressant puisqu'il permet de borner l'excès de risque $R(f_{\hat\theta})-R(f)$ et d'évaluer les vitesses optimales correspondant aux cinq types d'agrégation (Tsybakov, 2003; Rigollet, 2006), qui peuvent ainsi se reformuler de la façon suivante :

(1) **Sélection de modèle (MS) :** construire $\hat\theta^{\mathrm{MS}}$ qui vérifie une inégalité oracle du type

$$\mathbb{E}R(f_{\hat\theta^{\mathrm{MS}}}) \le \inf_{\theta\in\{e_1,\dots,e_M\}} \left\{R(f_\theta)+\Delta^{\mathrm{MS}}_{n,M}(\theta)\right\}.$$

Notons que

$$\inf_{\theta\in\{e_1,\dots,e_M\}} R(f_\theta) = \min_{1\le j\le M} R(\varphi_j).$$

(2) **Agrégation convexe (C) :** construire $\hat\theta^{\mathrm{C}}$ qui vérifie une inégalité oracle du type

$$\mathbb{E}R(f_{\hat\theta^{\mathrm{C}}}) \le \inf_{\theta\in\Lambda^M} \left\{R(f_\theta)+\Delta^{\mathrm{C}}_{n,M}(\theta)\right\}. \tag{1.4}$$

(3) **Agrégation linéaire (L) :** construire $\hat\theta^{\mathrm{L}}$ qui vérifie une inégalité oracle du type

$$\mathbb{E}R(f_{\hat\theta^{\mathrm{L}}}) \le \inf_{\theta\in\mathbb{R}^M} \left\{R(f_\theta)+\Delta^{\mathrm{L}}_{n,M}(\theta)\right\}.$$

(4) **Agrégation convexe *sparse* (CS) :** construire $\hat\theta^{\mathrm{CS}}$ qui vérifie une inégalité oracle du type

$$\mathbb{E}R(f_{\hat\theta^{\mathrm{CS}}}) \le \inf_{\theta\in\Lambda^M\cap\mathcal{B}_0(s)} \left\{R(f_\theta)+\Delta^{\mathrm{CS}}_{n,M,s}(\theta)\right\}, \quad s\in\{1,\dots,M\}.$$

(5) **Agrégation linéaire *sparse* (LS) :** construire $\hat\theta^{\mathrm{LS}}$ qui vérifie une inégalité oracle du type

$$\mathbb{E}R(f_{\hat\theta^{\mathrm{LS}}}) \le \inf_{\theta\in\mathbb{R}^M\cap\mathcal{B}_0(s)} \left\{R(f_\theta)+\Delta^{\mathrm{LS}}_{n,M,s}(\theta)\right\}, \quad s\in\{1,\dots,M\}.$$

Les vitesses optimales $\Delta^{\mathrm{MS}}_{n,M}$, $\Delta^{\mathrm{C}}_{n,M}$, $\Delta^{\mathrm{L}}_{n,M}$, $\Delta^{\mathrm{CS}}_{n,M,s}$ et $\Delta^{\mathrm{LS}}_{n,M,s}$ ont été, à notre connaissance, explicitées pour un *design* déterministe ou aléatoire dans des modèles linéaires à bruit gaussien avec une fonction de perte quadratique, et peuvent être trouvées par exemple dans Lecué (2007a) (notamment les liens entre vitesses optimales et hypothèses de marge), Rigollet and Tsybakov (2011) et Tsybakov (2013). Il est clair que

$$\inf_{\theta\in\mathbb{R}^M} R(f_\theta) \le \inf_{\theta\in\Lambda^M} R(f_\theta) \le \inf_{\theta\in\{e_1,\dots,e_M\}} R(f_\theta),$$

alors que bien souvent,

$$\Delta^{\mathrm{MS}}_{n,M}(\theta) \le c_1\Delta^{\mathrm{C}}_{n,M}(\theta) \le c_2\Delta^{\mathrm{L}}_{n,M}(\theta),$$

où c_1 et c_2 sont des constantes positives : il n'y a en général pas de hiérarchie qui se dégage parmi les techniques d'agrégation.

1.1.2 Approches pénalisées

Le problème de la minimisation du risque empirique

$$\hat\theta \in \operatorname*{arg\,inf}_{\theta\in\Theta} R_n(\theta),$$

a été l'objet d'un nombre considérable de travaux (voir par exemple Vapnik, 1998). Cette approche, importante dans la littérature, se heurte à des difficultés pratiques : si Θ est trop grand, il peut s'avérer impossible de calculer son minimum en un temps raisonnable (problème d'identifiabilité, ou de surapprentissage). De nombreux travaux cherchant à pallier à cette difficulté ont donné naissance à une famille désormais classique d'estimateurs, solutions du problème de minimisation

$$\hat\theta \in \operatorname*{arg\,inf}_{\theta\in\Theta} \left\{R_n(f_\theta) + \mathrm{pen}_\lambda(\theta)\right\},$$

où $\mathrm{pen}_\lambda(\cdot)$ est un terme pénalisant la structure de Θ et dépendant d'un paramètre de régularisation $\lambda > 0$. Dans la suite, R_n désigne le risque empirique basé sur la fonction de perte quadratique $\ell : (u,v) \mapsto (u-v)^2$:

$$R_n(f_\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{j=1}^M \theta_j\varphi_j(X_i)\right)^2, \quad \theta\in\Theta. \tag{1.5}$$

Il existe de nombreux types de pénalité. Pour la sélection de modèle, mentionnons le cœfficient C_p (Mallows, 1973), le critère AIC (Akaike, 1974) et les travaux de Barron et al. (1999), Birgé and Massart (2001, 2007) ainsi que Arlot (2007), Massart (2007), Arlot and Massart (2009), Arlot and Celisse (2010) et Arlot and Bartlett (2011), parmi beaucoup d'autres. Une vaste majorité des méthodes pénalisées dans le cadre de l'agrégation a été introduite pour le modèle de régression linéaire $Y = \theta^\top \mathbf{X} + W$, où la variance du bruit W est supposée connue (à l'exception notable de Giraud (2008); Dalalyan et al. (2013) dans le cas hétéroscédastique, et Belloni et al. (2011); Dalalyan (2012); Giraud et al. (2012); Sun and Zhang (2012) pour le cas homoscédastique). Nous présentons ci-dessous certaines des méthodes les plus populaires :

(1) Pénalisation ℓ_0 ou *hard thresholding* (BIC, Schwarz, 1978) :

$$\hat\theta^{\mathrm{BIC}} \in \operatorname*{arg\,inf}_{\theta\in\Theta} \{R_n(f_\theta) + \lambda|\theta|_0\}, \quad \lambda > 0. \tag{1.6}$$

(2) Pénalisation ℓ_1 ou *soft thresholding* (Lasso, Tibshirani, 1996) :

$$\hat{\theta}^{\text{LASSO}} \in \operatorname*{arginf}_{\theta\in\Theta} \{R_n(f_\theta) + \lambda|\theta|_1\}, \quad \lambda > 0,$$

et l'une de ses nombreuses variantes, le *fused Lasso* pénalisant la variation totale de θ (Tibshirani et al., 2005; Rinaldo, 2009) :

$$\hat{\theta}^{\text{FLASSO}} \in \operatorname*{arginf}_{\theta\in\Theta} \left\{R_n(f_\theta) + \lambda_1|\theta|_1 + \lambda_2 \sum_{j=2}^{M} |\theta_j - \theta_{j-1}|\right\}, \quad \lambda = (\lambda_1, \lambda_2) \in \mathbb{R}_+ \times \mathbb{R}_+.$$

Citons également dans cette famille le sélecteur de Dantzig (Candès and Tao, 2007; Lounici, 2008; Bickel et al., 2009), le *Group Lasso* (Yuan and Lin, 2006), le *bootstraped Lasso* (Bach, 2008a; Meinshausen and Bühlmann, 2010), le *Smooth-Lasso* (Hebiri and van de Geer, 2010) dont la pénalité s'exprime par

$$\text{pen}_\lambda : \theta \mapsto \lambda_1|\theta|_1 + \lambda_2 \sum_{j=2}^{M} (\theta_j - \theta_{j-1})^2, \quad \lambda = (\lambda_1, \lambda_2) \in \mathbb{R}_+ \times \mathbb{R}_+.$$

(3) Pénalisation ℓ_2, également connue comme régression *ridge* ou régularisation de Tikhonov (Hastie et al., 2009) :

$$\hat{\theta}^{\text{ridge}} \in \operatorname*{arginf}_{\theta\in\Theta} \{R_n(f_\theta) + \lambda|\theta|_2^2\}, \quad \lambda > 0.$$

(4) Une combinaison des pénalisations $\ell_1 + \ell_2$ ou *elastic net* (Zou and Hastie, 2005) :

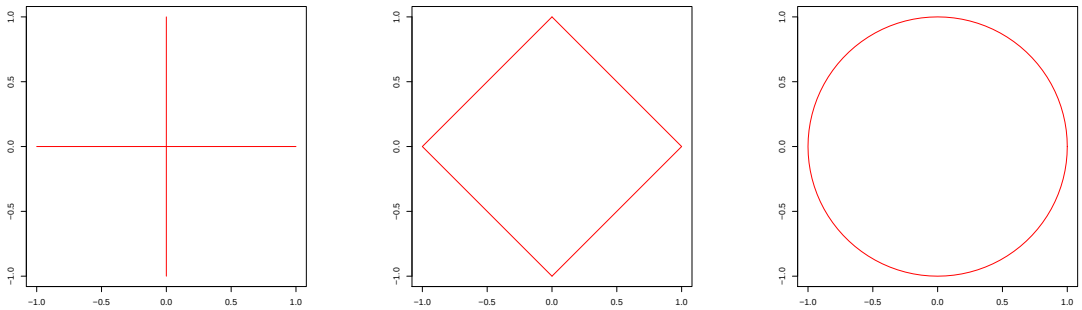
$$\hat{\theta}^{\text{EN}} \in \operatorname*{arginf}_{\theta\in\Theta} \{R_n(f_\theta) + \lambda_1|\theta|_1 + \lambda_2|\theta|_2^2\}, \quad \lambda = (\lambda_1, \lambda_2) \in \mathbb{R}_+ \times \mathbb{R}_+.$$

Ces approches sont judicieuses dans l'optique d'obtenir des inégalités oracles. En effet, elles font naturellement apparaître une approximation du terme de droite de l'inégalité oracle (1.4) où le risque R est remplacé par sa contrepartie empirique R_n, avec une pénalité pen_λ impactant le terme de vitesse (minimax dans plusieurs cas).

Historiquement, la pénalisation ℓ_0 a été parmi les premières à être étudiées. L'estimateur BIC présente l'avantage d'atteindre les vitesses minimax pour plusieurs problèmes d'agrégation (voir par exemple Bunea et al., 2007a). En revanche, son calcul est un problème NP-dur (il n'est pas résolvable en un temps polynomial) et nécessite une recherche exhaustive sur Θ, le rendant hors de portée dès que M dépasse quelques dizaines.

Pour sortir de l'ornière algorithmique, de nombreux auteurs ont proposé de convexifier le problème d'optimisation à résoudre. C'est l'approche qui prévaut dans la construction des estimateurs Lasso et variantes, *elastic net*, *etc*. Comme illustré par la Figure 1.1, la boule unité pour la norme ℓ_p est convexe dès que $p \geq 1$. Les autres approches présentées ci-dessus (Lasso, sélecteur de Dantzig, *elastic net*, *etc.*) proviennent de la relaxation convexe du problème des moindres carrés (1.6). Ces estimateurs atteignent, dans des cas spécifiques, les vitesses minimax pour certains des problèmes d'agrégation (par exemple, le Lasso permet de résoudre le problème (MS), voir Massart and Meynet, 2011). De plus, ils sont obtenus comme solutions d'un problème convexe ou linéaire et sont donc numériquement efficaces : parmi les algorithmes les plus populaires pour résoudre ce problème, citons *lars* (Efron et al., 2004) et *glmnet* (Friedman et al., 2010). En revanche, et contrairement à l'estimateur BIC, ces procédures n'ont des propriétés statistiques intéressantes (voir par exemple Bunea et al., 2007b; Bach, 2008b; Lounici, 2008; van de Geer, 2008; Koltchinskii, 2010, 2011)

que sous des conditions techniques bien souvent restrictives (hypothèses de cohérence mutuelle, condition *restricted isometry property*). Dans un modèle de régression linéaire, ces hypothèses impliquent typiquement que tous les sous-ensembles de régresseurs sont approximativement orthogonaux, ce qui est raisonnable lorsque l'objectif est d'estimer le support de $\theta^\star$, c'est-à-dire de déterminer les variables significatives dans un (potentiellement grand) ensemble de régresseurs. Ce contexte devient en revanche restrictif lorsque l'objectif est de construire des algorithmes de prédiction, bien que des travaux récents se soient attachés à affaiblir ces hypothèses (Bickel et al., 2009; van de Geer and Bühlmann, 2009).



(A) Boule ℓ_0 (B) Boule ℓ_1 (C) Boule ℓ_2

FIGURE 1.1 – Boule unité de $\mathbb{R}^2$ pour les normes ℓ_0, ℓ_1 et ℓ_2.

Les méthodes pénalisées se divisent donc en deux : d'une part, la pénalisation ℓ_0 permet, *sans hypothèses sur le design*, d'atteindre les vitesses minimax pour les différents types d'agrégation, mais l'estimateur BIC n'est pas calculable en pratique. D'autre part, les relaxations convexes permettent d'atteindre les vitesses minimax dans certains cas, et sont numériquement efficaces, mais ce compromis n'est accessible qu'au prix d'hypothèses restrictives. La recherche de méthodes moins contraignantes préservant cet équilibre a motivé l'ouverture d'une nouvelle ligne de recherche basée sur une approche d'agrégation d'estimateurs utilisant une loi a priori favorisant les solutions parcimonieuses.

1.1.3 Approche à poids exponentiels

Les méthodes d'agrégation exponentielles ont été introduites dans la communauté du *machine learning* (Vovk, 1990; Littlestone and Warmuth, 1994; Kivinen and Warmuth, 1997) et étudiées en apprentissage par renforcement, en particulier pour construire des prédicteurs dans des modèles de séquences individuelles (voir, parmi beaucoup d'autres références, Cesa-Bianchi and Lugosi, 1999, 2006, Stoltz, 2005, 2010, et Gerchinovitz, 2011). Dans ce cadre séquentiel où les points de *design* (déterministes) arrivent progressivement à la connaissance du statisticien, l'enjeu est de mélanger des prédicteurs ou *experts* pouvant reposer sur divers paramètres pour obtenir la meilleure prévision de la prochaine observation.

Dans le cadre stochastique qui nous intéresse, la méthode *progressive mixture* défendue par Yang (2000) et Catoni (2004) pour l'estimation de densité et la régression a été nommée "méthode de descente miroir avec moyennisation" (*mirror averaging*) par Juditsky et al. (2005). Le principe de cet agrégat repose sur la théorie de l'optimisation convexe, et renvoie plus spécifiquement aux algorithmes de descente de gradient dans l'espace conjugué (voir Nemirovski and Yudin, 1983), auquels furent ajoutés des termes permettant de régulari-

ser les poids successifs lorsque le nombre d'observations augmente. Cette approche a été appliquée par Bunea and Nobel (2008) au problème de la régression et par Lounici (2007, 2009) et Dalalyan and Tsybakov (2012a) sous contrainte de parcimonie. Les poids exponentiels apparaissent enfin en contraignant les mises à jour des poids successifs à être proches au sens de la divergence de Kullback-Leibler. Nous présentons ci-après l'estimateur *mirror averaging*, qui mène à l'agrégé à poids exponentiels EWA (*exponentially weighted aggregate*).

Désignons par π une distribution *a priori* sur Θ muni de sa tribu borélienne notée $\mathcal{T}$. Notons $r_{i,\theta}$, pour $i = 1,\dots,n$, le risque empirique du pré-estimateur f_θ mesuré sur les i premières observations :

$$r_{i,\theta} = \sum_{k=1}^{i} (Y_k - f_\theta(\mathbf{X}_k))^2, \quad \theta \in \Theta.$$

On définit ensuite les coefficients partiels

$$\hat{w}_{i,\theta} = \frac{\exp(-r_{i,\theta}/\beta)}{\int_\Theta \exp(-r_{i,\theta'}/\beta)\pi(\mathrm{d}\theta')}, \quad \theta \in \Theta, \tag{1.7}$$

pour un certain paramètre de température $\beta > 0$. On en tire l'expression des poids moyennés :

$$\hat{w}^{\mathrm{MA}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_{i,\theta}, \quad \theta \in \Theta,$$

et l'estimateur *mirror averaging* est donné par

$$\hat{f}^{\mathrm{MA}} = f_{\hat{\theta}^{\mathrm{MA}}},$$

où

$$\hat{\theta}^{\mathrm{MA}} = \int_\Theta \theta \hat{w}^{\mathrm{MA}}(\theta)\pi(\mathrm{d}\theta). \tag{1.8}$$

Cet estimateur possède de bonnes propriétés théoriques, sans conditions sur le *design* (voir par exemple Juditsky et al., 2008). Le calcul de l'intégrale dans (1.8) est cependant une gageure algorithmique dès que Θ est un espace compliqué, et son implémentation requiert des simplifications par randomisation ou recherche dichotomique (Catoni, 2004).

Les travaux de Leung and Barron (2006) ont représenté une avancée théorique majeure, permettant de s'affranchir de l'étape de moyennisation. Les poids EWA sont définis comme suit :

$$\hat{w}^{\mathrm{EWA}}(\theta) = \frac{\exp(-nR_n(\theta)/\beta)}{\int_\Theta \exp(-nR_n(\theta')/\beta)\pi(\mathrm{d}\theta')},$$

pour un certain paramètre de température $\beta > 0$ et où R_n désigne le risque empirique défini par (1.5), et l'on en tire l'estimateur EWA :

$$\hat{f}^{\mathrm{EWA}} = f_{\hat{\theta}^{\mathrm{EWA}}},$$

avec

$$\hat{\theta}^{\mathrm{EWA}} = \int_\Theta \theta \hat{w}^{\mathrm{EWA}}(\theta)\pi(\mathrm{d}\theta).$$

Le gain apporté par EWA est substantiel puisqu'il n'est plus nécessaire de construire les n agrégés intermédiaires de (1.7). Leung and Barron (2006), en considérant une famille d'indexation Θ finie (le cas d'espaces Θ continus et de grandes dimensions sera développé par Dalalyan and Tsybakov, 2007, 2008, 2012b), produisent une inégalité oracle pour le problème (MS). Ce résultat porte sur des pré-estimateurs qui sont des projecteurs en les

données, simplifiant considérablement l'implémentation puisqu'il s'agit de calculer des intégrales unidimensionnelles. De nombreux travaux ont depuis démontré les très bonnes garanties théoriques offertes par l'estimateur EWA, citons par exemple Alquier and Lounici (2011); Arias-Castro and Lounici (2012); Tsybakov (2013). Mentionnons également l'estimateur *Exponential Screening* développée par Rigollet and Tsybakov (2012), dont le principe est d'agréger des estimateurs de type moindres carrés dans tous les sous-ensembles de covariables, pénalisés par un *prior* favorisant les solutions parcimonieuses.

Les poids exponentiels peuvent être envisagés de deux façons. La première est "quasi"-bayésienne, et fait des poids exponentiels la densité de la distribution *a posteriori* par rapport au *prior* π. L'injection d'information via le *prior* va "tordre" les poids, en accord avec leur adéquation aux données via le risque empirique. Nous y reviendrons dans la suite de ce chapitre.

Une seconde interprétation, d'inspiration variationnelle, permet de faire le lien avec les approches pénalisées, et est notamment défendue dans Rigollet and Tsybakov (2012). En ne considérant plus des paramètres $\theta \in \Theta$ mais des distributions sur Θ muni de la tribu engendrée par ses boréliens, notée $\mathcal{T}$, et la divergence de Kullback-Leibler comme mesure naturelle de divergence entre les mesures, les auteurs considèrent le problème suivant :

$$\hat{\theta} \in \operatorname*{arg\,inf}_{\theta \in \Lambda^M} \{R_n(f_\theta) + \mathrm{pen}(\theta)\}. \tag{1.9}$$

La minimisation sur le simplexe Λ^M peut cependant être difficile. Dès que ℓ est une fonction convexe (ce qui est le cas de la perte quadratique), il vient

$$\sum_{j=1}^{M} \theta_j R_n(\varphi_j) \geq R_n(f_\theta), \quad \theta \in \Theta,$$

et Rigollet and Tsybakov (2012) propose de substituer au problème (1.9) la formulation suivante :

$$\hat{\theta} \in \operatorname*{arg\,inf}_{\theta \in \Lambda^M} \left\{ \sum_{j=1}^{M} \theta_j R_n(\varphi_j) + \mathrm{pen}(\theta) \right\}. \tag{1.10}$$

Notons que cette substitution apparaît également dans Catoni (2004). En assimilant les vecteurs du simplexe Λ^M à des probabilités sur $\{1,\dots,M\}$, on peut définir la divergence de Kullback-Leibler entre $\theta = (\theta_1,\dots,\theta_M)$ et $\pi = (\pi_1,\dots,\pi_M)$ par

$$\mathcal{KL}(\theta,\pi) = \sum_{j=1}^{M} \theta_j \log\left(\frac{\theta_j}{\pi_j}\right) \geq 0.$$

Fixons un paramètre de température $\beta > 0$, et une distribution *a priori* $\pi \in \Lambda^M$. Avec le choix de pénalité $\mathrm{pen}(\cdot) = \beta\mathcal{KL}(\cdot,\pi)/n$, le problème (1.10) devient

$$\hat{\theta} \in \operatorname*{arg\,inf}_{\theta \in \Lambda^M} \left\{ \sum_{j=1}^{M} \theta_j R_n(\varphi_j) + \frac{\beta}{n}\mathcal{KL}(\theta,\pi) \right\}.$$

Ce problème d'optimisation convexe sous contrainte admet une unique solution. En effet, il vient des conditions de Karush-Kuhn-Tucker (KKT) que les composantes $\hat{\theta}_j$ vérifient

$$nR_n(\varphi_j) + \beta\log\left(\frac{\hat{\theta}_j}{\pi_j}\right) + \mu - \delta_j = 0, \quad j = 1,\dots,M,$$

où μ, $\delta_1,\dots,\delta_M \geq 0$ sont les multiplicateurs de Lagrange et

$$\hat{\theta}_j \geq 0, \quad \delta_j\hat{\theta}_j = 0, \quad \sum_{j=1}^{M} \hat{\theta}_j = 1.$$

Ces conditions conduisent à l'unique solution suivante :

$$\hat{\theta}_j = \frac{\exp(-nR_n(\varphi_j)/\beta)\pi_j}{\sum_{k=1}^{M}\exp(-nR_n(\varphi_k)/\beta)\pi_k}, \quad j = 1,\dots,M,$$

i.e., les poids exponentiels. Rigollet and Tsybakov (2012) démontre, pour un *design* déterministe, le caractère optimal universel des poids exponentiels pour les cinq types d'agrégation.

1.1.4 Approche PAC-bayésienne

Comme nous l'avons vu, les approches pénalisées se concentrent pour une large part sur les modèles linéaires et à *design* déterministe (voir de Castro, 2011, pour un panorama récent). Les méthodes d'agrégation à poids exponentiels permettent d'envisager des modèles plus généraux, et ont permis l'émergence d'un nouveau paradigme, appelé *PAC-bayésien*. Cette approche constitue un nouvel arsenal de techniques de preuves, proposant de placer une structure sur les modèles à travers le prior π : à la pénalisation portant sur le vecteur θ des paramètres défendue par les approches pénalisées, se substitue ainsi une autre façon de mesurer la complexité des modèles, basée sur la divergence de Kullback-Leibler. Le rôle de ce *prior* équivaut à un certain choix de représentation de l'espace des paramètres.

L'idée centrale est d'obtenir des bornes sur le risque de règles de classification (on utilisera également l'anglicisme classificateurs) ou d'estimateurs bayésiens, et ceci de manière PAC (*Probably Approximately Correct*), c'est-à-dire en probabilité et avec un contrôle explicite sur la borne inférieure de cette probabilité. L'acronyme PAC s'applique à toute stratégie de choix d'un prédicteur au sein d'un ensemble, de sorte qu'avec grande probabilité, les prévisions soient approximativement aussi bonnes que possible. La couche bayésienne (voir Robert, 2007, pour une introduction aux méthodes bayésiennes) intervient dans l'introduction du *prior*, et l'estimation du volume de l'espace des paramètres qui soit consistant avec l'échantillon d'appprentissage : cette approche produit des estimateurs *a posteriori*, contrairement à l'analyse PAC "classique" qui développe des bornes *a priori* (voir par exemple Valiant, 1984).

Comme précédemment, désignons par π une distribution de probabilité *a priori* sur Θ. Pour un paramètre de température inverse $\beta > 0$, la distribution *a posteriori* de Gibbs construite sur l'échantillon $\mathcal{D}_n$ est définie par :

$$\hat{\rho}_\beta(\mathrm{d}\theta) \propto \exp(-\beta R_n(\theta))\pi(\mathrm{d}\theta).$$

La littérature sur le sujet s'intéresse essentiellement à deux estimateurs PAC-bayésiens : le premier est l'estimateur randomisé

$$\hat{\theta} \sim \hat{\rho}_\beta,$$

et le second l'estimateur agrégé ou moyenne *a posteriori*

$$\hat{\theta}^{\mathrm{a}} = \int_\Theta \theta\hat{\rho}_\beta(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\beta}\theta.$$

Illustrons ici cette approche, par un résultat prouvé dans le Chapitre 2. Désignons par $\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$ l'ensemble des mesures de probabilités sur $(\Theta,\mathcal{T})$ absolument continues par rapport à π, et considérons le modèle de régression non paramétrique $Y = f(\mathbf{X}) + W$, où W désigne une variable de bruit. Sous des hypothèses faibles (existence d'un moment exponentiel, pour l'utilisation d'inégalités de concentration et propriété de bornitude de la fonction

de régression et des éléments du dictionnaire), nous avons les inégalités oracles suivantes : pour tout $\eta \in (0,1)$, avec probabilité au moins $1-\eta$,

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(f) \\ R(f_{\hat{\theta}^{\mathrm{a}}}) - R(f) \end{array}\right\} \leq (1+\mathcal{C}) \inf_{\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) - R(f) + \frac{\mathcal{KL}(\rho,\pi) + \log\frac{2}{\eta}}{n} \right\}, \tag{1.11}$$

où $\mathcal{C} > 0$ est une constante. Ce résultat généralise une inégalité proche démontrée dans Alquier and Biau (2013), et est à rapprocher de l'inégalité oracle exacte en espérance apportée par Dalalyan and Tsybakov (2008) (les auteurs utilisent un schéma de preuve différent, avec un *design* déterministe).

Les inégalités (1.11) s'interprètent de la façon suivante. Pour un prior π et un paramètre de température inverse β fixés, avec grande probabilité, le risque des estimateurs PAC-bayésiens est borné à une constante multiplicative près par le meilleur des risques intégrés et un terme de complexité de la mesure par rapport au *prior*. Notons que ce terme permet de faire apparaître des vitesses proches des taux minimax pour un choix convenable de *prior*, comme présenté notamment dans le Chapitre 2.

La théorie PAC-bayésienne a été amorcée par les travaux de Shawe-Taylor and Williamson (1997) et McAllester (1999), qui ont introduit le principe d'un nouveau type de bornes en classification. Ce principe a ensuite été étendu et formalisé par Catoni (2004), pour la classification et la régression sous perte quadratique, où les vitesses atteintes sont minimax optimales dans certains cas. Mentionnons également les travaux de Seeger (2002, 2003) dans le cas de processus gaussiens. Les premiers résultats adaptatifs remontent à Audibert (2004a,b), étendus ensuite dans Catoni (2007). La régression sous perte générale a été traitée par Alquier (2006, 2008), et les techniques PAC-bayésiennes ont ensuite été étendues dans le cadre de la théorie des processus (Alquier et al., 2012; Alquier and Wintenberger, 2012), aux martingales (Seldin et al., 2012), au modèle semi-paramétrique *single-index* (Alquier and Biau, 2013) et à la régression sous contrainte de parcimonie (Alquier and Lounici, 2011). En outre, Suzuki (2012) démontre que les techniques PAC-bayésiennes obtiennent des vitesses rapides dans le cadre du *multiple kernel learning*.

Dans l'approche PAC-bayésienne, la marge de manœuvre du statisticien réside essentiellement dans le choix du paramètre β, et de façon plus importante dans la construction de la mesure *a priori* π. Cette mesure, dans le cadre de la grande dimension, portera exclusivement l'hypothèse de parcimonie et sa calibration est centrale pour faire apparaître les vitesses minimax, et pour la bonne tenue des versions implémentées. Une approche devenue classique dans les méthodes bayésiennes en grande dimension est de pénaliser exponentiellement la taille des modèles visités : pour un réel $\alpha \in (0,1)$ fixé,

$$\pi(\theta) \propto \alpha^{|\theta|_0} \binom{M}{|\theta|_0}^{-1}, \quad \theta \in \Theta.$$

Nous proposons, dans les chapitres 2 et 3, des mesures *a priori* inspirées de ce choix. Plusieurs travaux se sont attachés aux phénomènes liés à la très grande dimension et à l'hypothèse de parcimonie, nous renvoyons en particulier à Giraud (2008), Golubev (2010), Gaïffas and Lecué (2011), Giraud et al. (2012) et Verzelen (2012).

1.1.5 Implémentations de l'agrégation à poids exponentiels

Comme nous l'avons vu, les approches d'agrégation à poids exponentiels dans le cadre stochastique bénéficient depuis la fin des années 2000 de solides garanties théoriques et

se présentent ainsi comme de sérieux compétiteurs des méthodes pénalisées. En revanche, si leur implémentation ne pose pas de difficulté quand l'ensemble d'indexation Θ est de cardinalité petite, les poids exponentiels ont longtemps souffert de la concurrence algorithmique du Lasso, notamment quand Θ est de cardinalité très grande, ou continu. Dans ces contextes, il est hors de portée de calculer explicitement les intégrales potentiellement de grandes dimensions, et la meilleure stratégie consiste à les approximer à l'aide de méthodes dites de *Monte-Carlo*.

Ces méthodes jouissent d'une très grande popularité dans de larges pans de la statistique contemporaine. Nous renvoyons aux monographies Robert and Casella (2004), Marin and Robert (2007), Meyn and Tweedie (2009) et aux nombreuses références qu'elles contiennent, ainsi qu'à l'article Andrieu and Thoms (2008). Nous présentons ci-après trois implémentations récentes d'approches à poids exponentiels, visant à concurrencer les implémentations du Lasso et d'autres méthodes pénalisées.

La première approche, baptisée *Langevin Monte Carlo* et formalisée dans ce contexte par Dalalyan and Tsybakov (2012b), propose de remplacer la notion de chaîne de Markov par celle de diffusion. Comme l'estimateur cible peut être interprété comme l'espérance sous la distribution *a posteriori* $\hat{\rho}$, il s'agit d'approcher

$$\hat{\theta} = \mathbb{E}_{\hat{\rho}}\theta.$$

Supposons que l'on puisse écrire $\hat{\rho} \propto \exp \circ V$ pour une fonction V (appelée *potentiel*) suffisament régulière. L'équation différentielle stochastique

$$\mathrm{d}L_t = \nabla V(L_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad L_0 = \theta_0, \quad t \geq 0,$$

admet une unique solution, appelée diffusion de Langevin, avec $W = (W_t)_{t>0}$ un mouvement brownien M-dimensionnel et θ_0 un vecteur de $\mathbb{R}^M$, typiquement le vecteur nul dans ce qui suit. La diffusion $L = (L_t)_{t>0}$ est un processus Markovien homogène (Rogers and Williams, 1987). Sous certaines conditions précisées dans Dalalyan and Tsybakov (2012b), la distribution stationnaire de cette diffusion présente la propriété remarquable d'avoir pour densité $\exp \circ V$ par rapport à la mesure de Lebesgue, et il est alors raisonnable d'approcher $\mathbb{E}_{\hat{\rho}}\theta$ par

$$\bar{L}_T = \frac{1}{T}\int_0^T L_t \mathrm{d}t, \tag{1.12}$$

la vitesse de convergence étant de l'ordre de $1/\sqrt{T}$. Pour approximer cette intégrale, les auteurs proposent un schéma de discrétisation d'Euler, obtenant

$$L_{k+1} = \nabla V(L_k) + \sqrt{2h}W_k, \quad L_0 = \theta_0, \quad k = 0, \ldots, \lfloor T/h \rfloor - 1,$$

pour un pas de discrétisation h, et où $W_1, W_2, \ldots$ sont des vecteurs gaussiens normalisés de $\mathbb{R}^M$ et $\lfloor x \rfloor$ désigne la partie entière de x. L'intégrale dans (1.12) est alors approchée par

$$\bar{L}_{T,h} = \frac{1}{\lfloor T/h \rfloor}\sum_{k=0}^{\lfloor T/h \rfloor - 1} L_k,$$

qui est donc un estimateur de $\hat{\theta}$. Cette stratégie a notamment été adaptée au problème de traitement d'image par méthodes à patchs (Salmon and Le Pennec, 2009a,b; Salmon, 2010).

Les deux autres approches sont des variations de l'algorithme de *Metropolis-Hastings*, dont nous présentons dans ce chapitre une version générale (Algorithme 1.1) dans un souci

Algorithme 1.1 Algorithme de Metropolis-Hastings

1: **Input :** $\hat{\rho}$, θ_0, $T_{\min}$, $T_{\max}$.
2: **Output :** $\hat{\theta}$.
3: Initialiser $\theta^{(0)} = \theta_0$.
4: **for** $t = 1, \dots, T_{\max}$ **do**
5: Générer v suivant $k(\theta^{(t-1)}, \cdot)$.
6:

$$\theta^{(t)} = \begin{cases} v, & \text{avec probabilité } R(\theta^{(t)}, v) \\ \theta^{(t-1)}, & \text{avec probabilité } 1 - R(\theta^{(t)}, v) \end{cases}$$

où

$$R(x, y) = \min\left(1, \frac{\hat{\rho}(y)k(y,x)}{\hat{\rho}(x)k(x,y)}\right).$$

7: **end for**
8: Moyenner la chaîne après $T_{\min}$ itérations (*burnin*) : $\hat{\theta} = \frac{1}{T_{\max} - T_{\min}} \sum_{t=T_{\min}+1}^{T_{\max}} \theta^{(t)}$.

de lisibilité. Rigollet and Tsybakov (2011) proposent, pour l'estimateur *Exponential Screening*, une version adaptée aux *sparsity patterns*, où le noyau de transition $k(\cdot,\cdot)$ se réduit à la distribution uniforme sur les *patterns* voisins. Ce noyau, en particulier, est symétrique, simplifiant les calculs du ratio d'acceptation de Metropolis-Hastings. Cette méthode se compare très favorablement à différentes variantes du Lasso dans Rigollet and Tsybakov (2011), cependant pour des tailles d'échantillon et de dictionnaire modestes ($n \vee M \leq 500$).

Alquier and Biau (2013) adaptent l'algorithme *Reversible Jump Monte Carlo Markov Chain* (RJMCMC) introduit par Green (1995) au cas de la grande dimension pour le modèle semi-paramétrique single-index. Le noyau de transition favorise dans ce cas des sauts entre modèles de dimensions différentes, et tente à chaque pas de l'algorithme de rajouter ou d'enlever une covariable, dont le pouvoir explicatif influe sur le ratio d'acceptation. Notons que le choix du noyau k est crucial, et aura un impact déterminant sur la convergence de la chaîne et ce d'autant plus que la dimension de Θ sera importante. Un choix astucieux de noyau se concentrera en particulier sur des portions restreintes de l'espace Θ, en lien avec l'hypothèse de parcimonie.

1.2 Contributions

Les travaux présentés dans ce manuscrit sont regroupés en quatre chapitres. Les deux premiers étendent la méthodologie PAC-bayésienne au modèle de régression additive (Chapitre 2) et de régression logistique (Chapitre 3), dans un contexte de grande dimension. Le Chapitre 4 introduit une stratégie originale d'agrégation non linéaire d'estimateur de la fonction de régression. Enfin, le Chapitre 5 représente une incursion vers d'autres travaux, de modélisation bayésienne de l'hybridation de populations.

1.2.1 Approche PAC-bayésienne pour le modèle de régression additive sous contrainte de parcimonie

Le modèle de régression additive s'écrit comme suit : en conservant les notations introduites dans ce chapitre,

$$Y = \sum_{j=1}^{d} f_j(X_j) + W.$$

Cette formulation non paramétrique, étudiée notamment par Stone (1985), Hastie and Tibshirani (1986, 1990) et Härdle (1990), représente un bon compromis analytique entre une approche purement paramétrique et le modèle non paramétrique simple $Y = f(\mathbf{X}) + W$. De plus, la décomposition additive des covariables permet d'en expliquer les effets sur la réponse de façon bien plus intuitive. Sans perte de généralité, supposons $\mathcal{X} = (-1,1)^d$. Sur la base d'un échantillon $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ de réplications i.i.d. de $(\mathbf{X}, Y)$, nous proposons de construire une stratégie d'estimation de la fonction de régression $f(X_1, \dots, X_d) = \sum_{j=1}^d f_j(X_j)$. Pour cela, on se munit d'un dictionnaire $\mathbb{D} = \{\phi_1, \dots, \phi_K\}$, de fonctions connues $\phi_1, \dots, \phi_K : (-1,1) \to \mathbb{R}$. Notre ensemble d'estimateurs est

$$\mathcal{F}_\Theta = \left\{ f_\theta = \sum_{j=1}^d \sum_{k=1}^K \theta_{jk} \phi_k : \theta = (\theta_{11}, \dots, \theta_{1K}, \theta_{21}, \dots, \theta_{dK}) \in \Theta \right\},$$

où $\Theta \subseteq \mathbb{R}^{dK}$ est l'ensemble des paramètres. Notre stratégie ne fait pas d'hypothèse sur le *design* $\mathbf{X}$, et se démarque en ce sens d'autres travaux portant sur l'utilisation du Lasso et de variantes pour le modèle additif en grande dimension (Meier et al., 2009, Ravikumar et al., 2009, Koltchinskii and Yuan, 2010 et Suzuki and Sugiyama, 2013 essentiellement). En l'espèce, nous n'avons besoin que des deux hypothèses suivantes ($|\cdot|_\infty$ désigne la norme du supremum) :

Hypothèse 1.1. *Pour tout entier $k \geq 1$, $\mathbb{E}[|W|^k] < \infty$, $\mathbb{E}[W|\mathbf{X}] = 0$ et il existe deux constantes positives L et σ^2 telles que pour tout entier $k \geq 2$,*

$$\mathbb{E}[|W|^k | \mathbf{X}] \leq \frac{k!}{2} \sigma^2 L^{k-2}.$$

Hypothèse 1.2. *Il existe une constante $C > \max(1, \sigma)$ telle que $|f|_\infty \leq C$.*

Ces hypothèses sont relativement faibles : en particulier, il est important de souligner que l'hypothèse 1.1 est vérifiée si W est une variable gaussienne. L'hypothèse de bornitude de la fonction de régression joue un rôle central dans les approches PAC-bayésienne, et permet d'utiliser une inégalité de concentration de type Bernstein. Notons également que cette hypothèse est plus qu'un simple prérequis technique : si la dimension effective de la fonction de régression est encore importante devant n, la bornitude de f et des fonctions de $\mathcal{F}_\Theta$ permet d'obtenir des vitesses bien plus rapides, comme expliqué dans les travaux de Raskutti et al. (2012).

Dans le cadre de la grande dimension ($d \gg n$), l'estimation effective n'est généralement possible qu'au prix d'une hypothèse de parcimonie. En l'espèce, cela revient à supposer que l'ensemble $\{f_j \neq 0 : j = 1, \dots, d\}$ est de cardinalité faible devant n. Comme précédemment, nous considérons, sous la fonction de perte quadratique, pour une mesure de probabilité *a priori* π sur Θ muni de sa σ-algèbre borélienne, et pour un paramètre de température inverse $\beta > 0$, la distribution *a posteriori* de Gibbs :

$$\hat{\rho}_\beta(\mathrm{d}\theta) = \frac{\exp(-\beta R_n(f_\theta))\pi(\mathrm{d}\theta)}{\int_\Theta \exp(-\beta R_n(f_{\theta'}))\pi(\mathrm{d}\theta')},$$

et les deux estimateurs considérés sont l'estimateur randomisé

$$\hat{\theta} \sim \hat{\rho}_\beta,$$

et l'estimateur agrégré ou moyenne *a posteriori*

$$\hat{\theta}^{\mathrm{a}} = \int_\Theta \theta \hat{\rho}_\beta(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\beta} \theta.$$

Nous considérons deux types de *prior*. Le premier consiste à pénaliser uniquement le nombre de régresseurs f_j, et non leur expansion sur le dictionnaire $\mathbb{D}$. Si l'on note $\mathbf{m} = (m_1, \dots, m_d)$ le vecteur codant le développement de chaque estimateur de f_j (c'est-à-dire $m_j \in \{0, \dots, K\}$ pour tout $j = 1, \dots, d$), le *prior* prend la forme

$$\pi^1(\mathrm{d}\theta) = \sum_{\mathbf{m} \in \mathcal{M}} \frac{1 - \frac{\alpha}{1-\alpha}}{1 - \left(\frac{\alpha}{1-\alpha}\right)^{d+1}} \binom{d}{|S(\mathbf{m})|}^{-1} \alpha^{\sum_{j=1}^d m_j} \frac{1}{\mathrm{Vol}(\mathcal{B}^1_{\mathbf{m}}(C))} \mathbb{1}_{\{\theta \in \mathcal{B}^1_{\mathbf{m}}(C)\}} \lambda_{\mathbf{m}}(\mathrm{d}\theta),$$

où $\mathcal{M}$ est l'ensemble des modèles $\{\mathbf{m} \in \{0, \dots, K\}^d\}$, $\mathcal{B}^1_{\mathbf{m}}(C)$ la boule ℓ_1 dans $\mathbb{R}^{\mathbf{m}}$ de rayon C, $S(\mathbf{m}) = \{j : m_j \neq 0\}$, $\alpha \in (0, 1/2)$ et $\lambda_{\mathbf{m}}$ la mesure de Lebesgue dans $\mathbb{R}^{\mathbf{m}}$. Ce choix revient à considérer des modèles emboîtés : l'estimateur de f_j est construit sur les m_j premières fonctions du dictionnaire $\mathbb{D}$. En d'autres termes, π^1 pénalise les modèles ayant trop de régresseurs.

Une seconde approche consiste à chercher des développements parcimonieux sur le nombre de régresseurs et sur le développement des estimateurs dans le dictionnaire $\mathbb{D}$. Le *prior* prend alors la forme suivante :

$$\pi^2(\mathrm{d}\theta) = \sum_{\mathbf{m} \in \mathcal{M}} \frac{1 - \alpha \frac{1-\alpha^{K+1}}{1-\alpha}}{1 - \left(\alpha \frac{1-\alpha^{K+1}}{1-\alpha}\right)^{d+1}} \binom{d}{|S(\mathbf{m})|}^{-1} \prod_{j \in S(\mathbf{m})} \binom{K}{|S(\mathbf{m}_j)|}^{-1} \alpha^{|S(\mathbf{m}_j)|} \frac{1}{\mathrm{Vol}(\mathcal{B}^1_{\mathbf{m}}(C))} \times \mathbb{1}_{\{\theta \in \mathcal{B}^1_{\mathbf{m}}(C)\}} \lambda_{\mathbf{m}}(\mathrm{d}\theta),$$

avec les notations $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_d) \in \{0,1\}^{dK}$ et $\mathbf{m}_j = (m_{j1}, \dots, m_{jK}) \in \{0,1\}^K$, et

$$S(\mathbf{m}) = \{\mathbf{m}_j \neq \mathbf{0}, \quad j \in \{1, \dots, d\}\}, \qquad S(\mathbf{m}_j) = \{m_{jk} \neq 0, \quad k \in \{1, \dots, K\}\}.$$

Ce second *prior* pénalise donc le choix d'un certain ensemble de variables parmi dK (qui est supposé immense devant n). Une fois les régresseurs et leurs expansions choisies, le vecteur θ est tiré uniformément sur la boule ℓ_1 de rayon C dans le modèle courant.

Sous les hypothèses 1.1 et 1.2, nous prouvons dans le Chapitre 2 les résultats suivants :

Théorème 1.1. *Pour le choix $\pi = \pi^1$, et tout $\varepsilon \in (0,1)$, nous avons avec probabilité au moins $1 - \varepsilon$,*

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(f) \\ R(f_{\hat{\theta}^{\mathrm{a}}}) - R(f) \end{array}\right\} \leq (1 + \mathcal{D}) \inf_{\mathbf{m} \in \mathcal{M}} \inf_{\theta \in \mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(f_\theta) - R(f) + \frac{|S(\mathbf{m})|}{n} \log\left(\frac{p}{|S(\mathbf{m})|}\right) + \frac{\log(n)}{n} \sum_{j \in S(\mathbf{m})} m_j + \frac{\log(2/\varepsilon)}{n} \right\},$$

où $\mathcal{D} > 0$ est une constante.

Théorème 1.2. *Pour le choix $\pi = \pi^2$, et tout $\varepsilon \in (0,1)$, nous avons avec probabilité au moins $1 - \varepsilon$,*

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(f) \\ R(f_{\hat{\theta}^{\mathrm{a}}}) - R(f) \end{array}\right\} \leq (1 + \mathcal{D}) \inf_{\mathbf{m} \in \mathcal{M}} \inf_{\theta \in \mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(f_\theta) - R(f) + \frac{|S(\mathbf{m})|}{n} \log\left(\frac{p}{|S(\mathbf{m})|}\right) + \frac{\log(nK)}{n} \sum_{j \in S(\mathbf{m})} |S(\mathbf{m}_j)| + \frac{\log(2/\varepsilon)}{n} \right\},$$

où $\mathcal{D} > 0$ est une constante.

Ces résultats prouvent que s'il existe au moins un modèle "petit" dans la collection $\mathcal{M}$, c'est-à-dire un modèle $\mathbf{m}$ tel que, d'une part, $\sum_{j\in S(\mathbf{m})} m_j$ et $|S(\mathbf{m})|$ sont petits devant dK, et d'autre part, pour $\theta \in \mathcal{B}^1_{\mathbf{m}}(C)$, le risque de l'oracle f_θ est proche du risque $R(f)$, alors les risques des estimateurs PAC-bayésiens $f_{\hat{\theta}}$ et $f_{\hat{\theta}^{\mathrm{a}}}$ sont également proches de celui de f, à des termes en $\log(n)/n$ et $\log(d/|S(\mathbf{m})|)/n$ près. Par suite, si f est effectivement parcimonieuse, notre stratégie offre de bonnes garanties d'estimation.

Si de plus, nous supposons que les f_j appartiennent à des espaces fonctionnels de régularité suffisante, nous pouvons montrer l'optimalité au sens minimax de nos estimateurs. $\mathbb{D}$ désigne désormais le système trigonométrique, et $S^\star$ est l'ensemble des régresseurs non nuls. Supposons que pour tout $j = 1,\dots,d$, f_j appartienne à l'ellipsoïde de Sobolev de paramètres r_j et d_j, définie par

$$\mathcal{W}(r_j,d_j) = \left\{ f \in \mathrm{L}^2([-1,1]) \colon f = \sum_{k=1}^{\infty} \theta_k \varphi_k \quad \text{et} \quad \sum_{i=1}^{\infty} i^{2r_j}\theta_i^2 \le d_j \right\},$$

où les d_j sont choisis de sorte que $\sum_{j\in S^\star} \sqrt{d_j} \le C\sqrt{6}/\pi$, et les $r_1,\dots,r_{|S^\star|} \ge 1$ sont inconnus : le résultat suivant fait des deux estimateurs PAC-bayésiens des estimateurs adaptatifs quasi-optimaux au sens minimax (à un terme $\log(n)$ près).

Théorème 1.3. *Pour le choix $\pi = \pi^1$, et tout $\varepsilon \in (0,1)$, nous avons avec probabilité au moins $1-\varepsilon$,*

$$\left.\begin{array}{l} R(f_{\hat{\theta}}) - R(f) \\ R(f_{\hat{\theta}^{\mathrm{a}}}) - R(f) \end{array}\right\} \le (1+\mathcal{D}) \left\{ \sum_{j\in S^\star} d_j^{\frac{1}{2r_j+1}} \left(\frac{\log(n)}{2nr_j}\right)^{\frac{2r_j}{2r_j+1}} + \frac{|S^\star|\log(d/|S^\star|)}{n} + \frac{\log(2/\varepsilon)}{n} \right\},$$

où $\mathcal{D} > 0$ est une constante.

L'implémentation des deux estimateurs exige d'échantillonner selon la distribution $\hat{\rho}_\beta$: nous proposons dans le Chapitre 2 une approche de type Metropolis-Hastings, où le noyau de transition favorise les mouvements dans un voisinage du modèle courant, au sens de l'ajout ou de la suppression d'un régresseur ou de termes dans son expansion sur $\mathbb{D}$. Cette approche, inspirée par les travaux de Carlin and Chib (1995), Hans et al. (2007), Petralias (2010) et Petralias and Dellaportas (2012), est présentée dans l'algorithme 2.1. La Figure 1.2 présente deux examples d'estimation de la fonction de régression. Le paquet *pacbpred* (pour ***PAC-B**ayesian **Pred**iction*, Guedj, 2013b) implémente [1] notre stratégie.

1.2.2 Approche PAC-bayésienne pour le modèle de régression logistique sous contrainte de parcimonie

Le modèle de régression logistique bénéficie depuis longtemps, et dans des domaines d'applications variés (médecine, génomique, sociologie, parmi de nombreux autres), d'une grande popularité. Ce modèle s'écrit comme suit : Y est une variable aléatoire à valeurs dans $\{\pm 1\}$, et en notant

$$p(\mathbf{x}) = \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}), \quad \text{pour tout } x \in \mathbb{R}^d,$$

le modèle "classique" de régression logistique suppose que

$$\operatorname{logit} p(\mathbf{x}) = \sum_{j=1}^{d} \theta_j x_j, \quad \mathbf{x} \in \mathbb{R}^d, \tag{1.13}$$

1. http://cran.r-project.org/web/packages/pacbpred/index.html

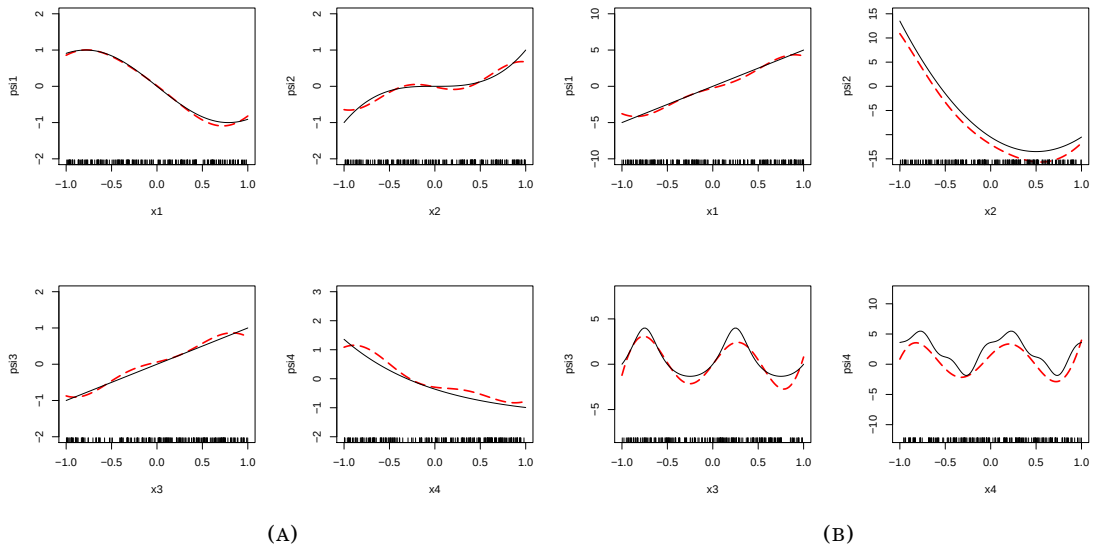


FIGURE 1.2 – Deux exemples d'estimation (rouge) de la fonction de régression (noir), sur les quatre premières covariables.

où logit $(a) = \log \frac{a}{1-a}$ pour tout $a \in (0,1)$. Dans le cas $n \geq d$, l'estimation par maximum de vraisemblance est implémentée dans de nombreux logiciels de statistique et ne présente pas de difficultés. En revanche, dans le contexte de grande dimension $n \ll d$, cette approche n'est plus accessible. Des travaux récents ont montré que l'utilisation de techniques de régularisation de type ℓ_1 (van de Geer, 2008; Meier et al., 2008; Kwemou, 2012) permettait d'obtenir de bonnes garanties théoriques des estimateurs, qui sont de plus calculables en temps raisonnable. Hélas, cette approche nécessite, comme dans le cas de la régression linéaire, de drastiques conditions de régularité sur la matrice de Gram (cohérence mutuelle des covariables, *restricted isometry property*).

Nous démontrons dans le Chapitre 3 que l'utilisation des techniques PAC-bayésiennes permet, sans conditions sur le *design*, d'obtenir également de solides résultats théoriques. Les preuves présentées dans le Chapitre 3 utilisent une inégalité de concentration dans l'esprit du Lemme A.2, adaptée au cas de la perte logistique et présentée dans le Lemme A.3.

La formulation du modèle logistique présentée en (1.13) est adaptée à l'utilisation de versions du Lasso. Nous adoptons ici une formulation plus générale :

$$\text{logit } p(\mathbf{x}) = \sum_{j=1}^{d} f_j(x_j), \quad \mathbf{x} \in \mathbb{R}^d, \tag{1.14}$$

où les $f_1, \ldots, f_d$ sont des fonctions $\mathbb{R} \to \mathbb{R}$. Nous cherchons donc à estimer la fonction de lien $\sum_{j=1}^{d} f_j$, et l'approche présentée dans le Chapitre 2 y est bien adaptée. Considérant donc un dictionnaire $\mathbb{D} = \{\phi_1, \ldots, \phi_M\}$ de fonctions $\mathbb{R} \to \mathbb{R}$ connues, l'ensemble des estimateurs que nous étudions est

$$\mathcal{F}_\Theta = \left\{ f_\theta = \sum_{j=1}^{d} \sum_{k=1}^{M} \theta_{jk} \phi_k : \theta = (\theta_{11}, \ldots, \theta_{1M}, \theta_{21}, \ldots, \theta_{dM}) \in \Theta \right\},$$

où $\Theta \subseteq \mathbb{R}^{dM}$ est l'ensemble des paramètres.

Dans la suite de l'approche présentée dans le Chapitre 2, considérons une mesure *a priori* π sur l'espace Θ muni de la tribu engendrée par ses boréliens. Pour un paramètre de température inverse $\beta > 0$, la mesure *a posteriori* de Gibbs est définie par

$$\hat{\rho}_\beta(\mathrm{d}\theta) \propto \exp(-\beta n R_n(f_\theta))\pi(\mathrm{d}\theta).$$

Sous la seule hypothèse que l'on sache borner (au sens du supremum) les éléments de $\mathcal{F}_\Theta$, nous obtenons le premier résultat suivant.

Théorème 1.4. *Pour tout* $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[R(f_{\hat{\theta}^a}) \leq \frac{c_2(\beta)}{c_1(\beta)} \inf_{\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{B}(\Theta))} \left\{\int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) + \frac{2\mathcal{KL}(\rho,\pi)}{\beta n c_2(\beta)} + \frac{2\log(2/\varepsilon)}{\beta n c_2(\beta)}\right\}\right] \geq 1-\varepsilon,$$

où $c_1(\beta)$ *et* $c_2(\beta)$ *sont des constantes.*

Il s'agit de l'analogue du résultat classique rappelé en (1.11), pour le modèle de régression logistique.

De plus, définissons le prior π comme suit :

$$\pi(\mathrm{d}\theta) = \sum_{m \in \mathcal{M}} \frac{1-\alpha^{d+1}}{1-\alpha} \binom{d}{|m|_0}^{-1} \alpha^{|m|_0} \frac{\mathbb{1}_{\mathcal{B}_{M|m|_0}(r)}(\theta)}{\mathcal{V}(r)}, \tag{1.15}$$

avec $\alpha \in (0,1)$, $\mathcal{B}_{M|m|_0}(r)$ désigne la boule ℓ_2 de rayon r dans $\mathbb{R}^{M|m|_0}$ et $\mathcal{V}(r)$ son volume, et $\mathcal{M}$ est l'ensemble des modèles, c'est-à-dire $\mathcal{M} = \{0,1\}^d$. Sous l'hypothèse technique (et classique dans la littérature PAC-bayésienne) que pour un certain modèle m, θ appartient à la boule $\mathcal{B}_{M|m|_0}(r)$, nous prouvons le résultat central du Chapitre 3 :

Théorème 1.5. *Pour tout* $\varepsilon \in (0,1)$,

$$\mathbb{P}\Bigg[R(f_{\hat{\theta}^a}) \leq \frac{c_2(\beta)}{c_1(\beta)} \inf_{m \in \mathcal{M}} \inf_{\theta \in \mathcal{B}_{M|m|_0}(r)} \Bigg\{R(f_\theta) + \frac{2}{\beta n c_2(\beta)}\Bigg[M|m|_0\left(1+\log\left(\frac{\beta n r c_2(\beta) C'}{2M|m|_0}\right)\right)$$
$$+|m|_0 \log\left(\frac{de}{|m|_0}\right) + |m|_0 \log(1/\alpha) + \log\left(\frac{1}{1-\alpha}\right) + \log(2/\varepsilon)\Bigg]\Bigg\}\Bigg] \geq 1-\varepsilon.$$

Notons que bien que nous soyons dans un contexte profondément différent, nous parvenons à faire apparaître les vitesses optimales de convergence, de l'ordre de $|m|_0/n$.

Le Chapitre 3 se referme par une section dédiée à l'implémentation effective de notre méthode (voir Algorithme 3.1). Il s'agit d'une adaptation de l'approche présentée dans le Chapitre 2.

1.2.3 Agrégation non linéaire d'estimateur (COBRA)

Comme on l'a vu, une part importante des travaux issus de cette thèse a consisté à étendre la méthodologie PAC-bayésienne. Nous abordons ici une stratégie (inspirée par les travaux de Mojirsheibani, 1999, en classification supervisée) qui se veut une alternative aux techniques d'agrégation existantes. Plutôt que de construire une combinaison linéaire ou convexe d'estimateurs de la fonction de régression (dont nous changeons la notation, dans la suite du chapitre, en $r^\star$), nous les utilisons comme indicateurs de distance entre échantillon d'apprentissage et nouveaux points de *design*. Pour un nouveau point de *design* $\mathbf{X}$, nous considérons qu'un point de l'échantillon d'apprentissage en est proche si l'ensemble des estimateurs prédisent pour ces deux points à distance au plus ε, où ε est un seuil fixé,

(A) Comment prédire la réponse du point de *design* en pointillés ?

(B) Prédictions des deux estimateurs.

(C) Quels sont les points "proches" du nouveau selon les deux estimateurs ?

(D) Prédiction finale.

FIGURE 1.3 – Exemple d'agrégation non linéaire de deux estimateurs.

correspondant à un paramètre de lissage. La prédiction faite pour le point $\mathbf{X}$ est alors la moyenne des réponses des points d'apprentissage proches. Cette définition informelle est illustrée par la Figure 1.3.

Présentons maintenant la méthode plus en détails. Supposons que l'on dispose d'un échantillon d'apprentissage $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de réplication i.i.d. de $(\mathbf{X}, Y)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$. Nous coupons cet échantillon en deux sous-échantillons, notés $\mathcal{D}_k$ et $\mathcal{D}_\ell$: sur $\mathcal{D}_k$, nous construisons M estimateurs de la fonction de régression $r^\star$, notés $r_{k,1}, \ldots, r_{k,M}$. Insistons ici sur un point saillant de notre approche : ces estimateurs peuvent être paramétriques, semi-paramétriques ou non paramétriques. N'importe quelle stratégie d'estimation (suggérée par le contexte expérimental ou le statisticien) est acceptable, à l'unique condition

qu'elle soit capable, sur la base de $\mathcal{D}_k$, de fournir une estimation de $r^\star$. Ces estimateurs "primitifs" — appelés également *machines* dans le Chapitre 4 — sont fixés par le statisticien. Nous définissons notre estimateur final T_n, nommé collectif de la régression, par

$$T_n(\mathbf{r}_k(\mathbf{x})) = \sum_{i=1}^{\ell} W_{n,i}(\mathbf{x}) Y_i, \quad \mathbf{x} \in \mathbb{R}^d, \tag{1.16}$$

avec la notation $\mathbf{r}_k = (r_{k,1}, \ldots, r_{k,M})$ et où les poids aléatoires $W_{n,i}(\mathbf{x})$ sont définis par

$$W_{n,i}(\mathbf{x}) = \frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \le \varepsilon_\ell\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \le \varepsilon_\ell\}}}, \tag{1.17}$$

avec $\varepsilon_\ell > 0$. L'originalité de cette approche réside dans le fait que l'estimateur final de la fonction de régression dépend de façon non linéaire des estimateurs primitifs. À notre connaissance, il n'y a pas de technique comparable dans la littérature portant sur l'agrégation.

Avec la fonction de perte quadratique, et sous des hypothèses techniques (vérifiée par exemple dès que les machines sont bornées), nous obtenons en particulier le résultat suivant.

Théorème 1.6 (Biau et al., 2013)**.** *Si $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$, alors*

$$\mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})\right|^2 \le \min_{m=1,\ldots,M} \mathbb{E}\left|r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X})\right|^2 + C\ell^{-\frac{2}{M+2}},$$

pour une certaine constante positive C indépendante de k.

Ce résultat porte plusieurs enseignements. Tout d'abord, l'estimateur combiné T_n est asymptotiquement aussi bon que le meilleur estimateur initial, au sens de l'erreur quadratique moyenne. De plus, cette inégalité oracle exacte fait apparaître un terme résiduel de vitesse satisfaisant dans notre cadre : comme le nombre de machines M est fixé, le terme $\ell^{-2/(M+2)}$ est négligeable devant la vitesse non paramétrique classique $\ell^{-2/(d+2)}$, par exemple. Bien sûr, des vitesses plus rapides peuvent être atteintes si la distribution de $(\mathbf{X}, Y)$ peut être approchée de façon paramétrique, cependant notre stratégie est conçue pour des problèmes de régression plus périlleux. Insistons ici sur le fait que contrairement à d'autres approches, comme les méthodes pénalisées en norme ℓ_1, il n'est fait ici aucune hypothèse sur le *design*, et les conditions portant sur les estimateurs initiaux sont faibles.

En particulier, il est important de noter que si l'un des estimateurs initiaux est consistant, c'est-à-dire que pour un certain m_0,

$$\mathbb{E}\left|r_{k,m_0}(\mathbf{X}) - r^\star(\mathbf{X})\right|^2 \to 0 \quad \text{quand } k \to \infty,$$

alors l'estimateur combiné T_n hérite de cette consistance :

$$\lim_{k,\ell\to\infty} \mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})\right|^2 = 0.$$

Notre procédure d'agrégation non linéaire offre ainsi une alternative aux techniques existantes, permettant au praticien de mélanger des estimateurs très différents quand il se retrouve confronté à un problème difficile.

Cette approche est implémentée dans le paquet pour le logiciel R *COBRA* (acronyme de ***CO**m**B**ined **R**egression **A**lternative*, Guedj, 2013a), téléchargeable gratuitement [2]. Nous

2. http://cran.r-project.org/web/packages/COBRA/index.html

présentons dans le Chapitre 4 un protocole de validation expérimentale conséquent, démontrant les excellentes performances de *COBRA*, en terme d'erreur de prédiction et de temps de calcul (le paquet sait nativement tirer parti des architectures multi-cœur). Nous menons ensuite une étude comparative avec l'algorithme "Super Learner" (van der Laan et al., 2007; Polley and van der Laan, 2010), ainsi qu'avec l'agrégation à poids exponentiels. Les figures 1.4 et 1.5 représentent des exemples d'exécution du paquet. Notons que la version implémentée fait intervenir une définition plus générale des poids que celle introduite en (1.17). En effet, il était demandé à l'ensemble des machines de tomber d'accord sur l'importance des points Y_i intervenant dans la moyenne (1.16). Nous relâchons cette contrainte d'unanimité, en modifiant les poids de la façon suivante : pour une fraction $\alpha \in (0,1)$,

$$W_{n,i}(\mathbf{x}) = \frac{\mathbb{1}_{\{\sum_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \leq \varepsilon_\ell\}} \geq M\alpha\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\sum_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \leq \varepsilon_\ell\}} \geq M\alpha\}}}.$$

Autrement dit, il suffit qu'une proportion au moins α des machines place $\mathbf{x}$ à distance (au sens de la métrique sur les machines) au plus ε_ℓ de $\mathbf{X}_i$ pour que Y_i contribue à la prédiction de la réponse de $\mathbf{x}$. Cette proportion α peut être interprétée comme une mesure de l'homogénéité de l'ensemble des machines initiales.

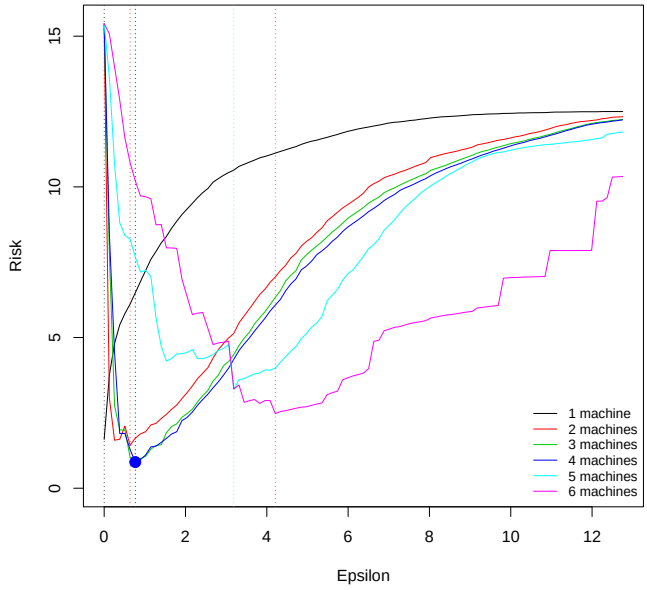


FIGURE 1.4 – Exemple de calibration des paramètres ε_ℓ et α. Le point en gras détermine le couple $(\varepsilon_\ell, \alpha)$.

1.2.4 Modélisation bayésienne de l'hybridation de populations

Nous présentons enfin dans le Chapitre 5 des travaux antérieurs à cette thèse, effectués sous la direction de Gilles Guillot (*Associate Professor*, *Danmarks Tekniske Universitet*). Ces travaux visaient à modéliser l'hybridation de populations par des techniques de type MCMC (Guedj and Guillot, 2011).

L'apparition de méthodes d'inférences adaptées aux volumes gigantesque des données collectés dans des problèmes de génétique des populations est relativement récente. Une

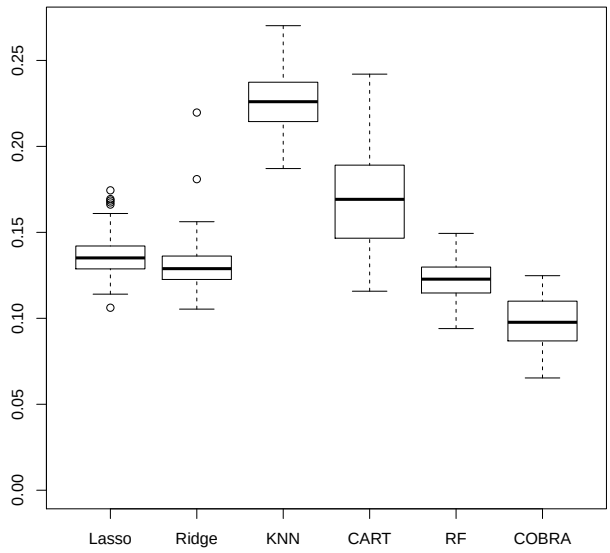


FIGURE 1.5 – Exemple d'erreurs quadratiques. De gauche à droite : Lasso, régression ridge, k-plus proches voisins, algorithme CART, forêts aléatoires, COBRA.

partie des travaux s'est concentrée ces dernières années sur la modélisation du phénomène d'hybridation : le génome d'un individu donné n'est pas nécessairement issu d'un seul bassin (*pool*) génétique, mais de plusieurs suffisamment différenciés, dans des proportions à estimer. Rappelons dans les lignes qui suivent la modélisation classique de l'hybridation (*admixture model*) introduite dans les logiciels *Structure* (Pritchard et al., 2000) et *Geneland* (Guillot et al., 2005, 2009).

Supposons que les individus de l'échantillon soient porteurs d'allèles provenant de K populations distinctes, caractérisé par des fréquences alléliques différentes. Désignons par $z = (z_{i\ell})$ la matrice des génotypes, où $z_{i\ell} = (z_{i\ell 1}, z_{i\ell 2})$ désigne le génotype de l'invidivu i au locus ℓ, et $f_{k\ell a}$ est la fréquence de l'allèle a au locus ℓ au sein de la population k. Introduisons la matrice $q = (q_{ik})$, où q_{ik} est la proportion du génome de l'individu i provenant de la population k (ainsi $\sum_{k=1}^{K} q_{ik} = 1$). Il est important de noter que bien que chaque individu se verra *in fine* affecter un label parmi $\{1, \ldots, K\}$ dans notre approche inférentielle, la réalité est plus nuancée, et ce label désigne le *cluster* dont la proportion est majoritaire dans son génome. En supposant que les deux allèles portés par le même locus de chromosomes homologues sont indépendants, la vraisemblance pour un individu diploïde s'écrit

$$\mathcal{L}(z_{i\ell}|f,q) = \sum_{k=1}^{K} q_{ik} f_{k\ell z_{i\ell 1}} f_{k\ell z_{i\ell 2}} (2 - \mathbb{1}[z_{i\ell 1} = z_{i\ell 2}]).$$

Pour un individu haploïde,

$$\mathcal{L}(z_{i\ell}|f,q) = \sum_{k=1}^{K} q_{ik} f_{k\ell z_{i\ell}}.$$

De plus, en supposant que les différents loci sont indépendants,

$$\mathcal{L}(z|f,q) = \prod_{i=1}^{n} \prod_{\ell=1}^{L} \mathcal{L}(z_{i\ell}|f,q).$$

L'approche défendue par *Geneland*, contrairement à *Structure*, intègre une dimension spatiale. Nous faisons l'hypothèse que les différentes populations sont relativement homogènes spatialement. Un exemple du modèle spatial reposant sur la tesselation de Poisson-Voronoi est présenté dans la Figure 1.6.

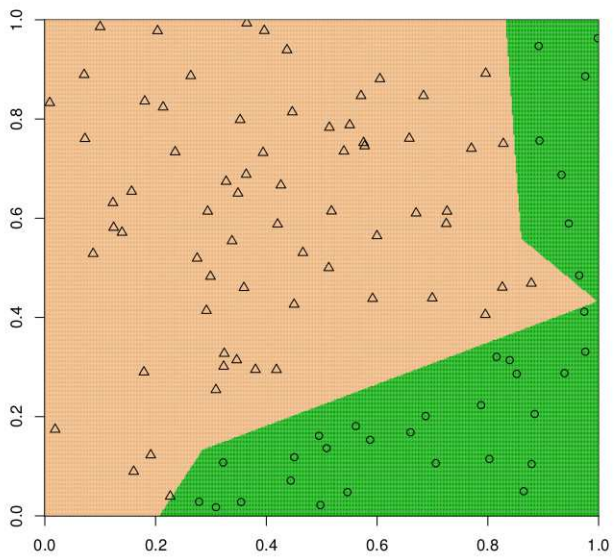


FIGURE 1.6 – Exemple de $K = 2$ populations simulées à partir d'un modèle *a priori* de Poisson-Voronoi. Les points représentent les lieux d'échantillonage des individus (les symboles repèrent les populations). Dans un souci de lisibilité, les réalisations du processus poissonnien qui gouverne la tesselation sont masquées.

Le travail présenté dans ce chapitre intègre l'hybridation dans *Geneland*. Notre approche est la suivante : supposons que chaque vecteur (*i.e.*, chaque individu) des proportions d'hybridation $q_{i.} = (q_{ik})_{k=1}^{K}$ suit une distribution de Dirichlet $\mathcal{D}(\alpha_{i1}, \ldots, \alpha_{iK})$. Désignons par d_{ik} la distance (euclidienne) de l'individu i au *cluster* k (en particulier, la distance d'un individu à son *cluster* est nulle). Nous supposons la relation déterministe

$$\alpha_{ik} = a \exp(-d_{ik}/b).$$

Un exemple de simulation suivant ce modèle est donné dans la Figure 1.7. Le modèle hiérarchique complet est quant à lui présenté dans la Figure 1.8.

Ce modèle est proche de celui présenté par Durand et al. (2009), cependant notre stratégie d'estimation est radicalement différente. Les quantités inconnues dans la Figure 1.8 (notamment le nombre de *clusters* K) sont estimées de façon bayésienne et nous proposons une implémentation par MCMC.

Par rapport à la version implémentées dans les versions de *Geneland* antérieures à la 3.3.0, trois nouveaux paramètres sont à estimer : q, a et b. L'idée la plus naturelle consiste à les estimer de façon jointe, avec les autres paramètres (nombre de *clusters*, paramètres de la tesselation, fréquences alléliques notamment). La très grande dimension des objets en question, et les difficultés pratiques d'estimer conjointement K et $q = (q_{ik})$ nous ont cependant fait préférer une approximation en deux étapes : la chaîne de Markov que nous lançons a pour distribution stationnaire la distribution de (q, a, b) conditionnellement aux

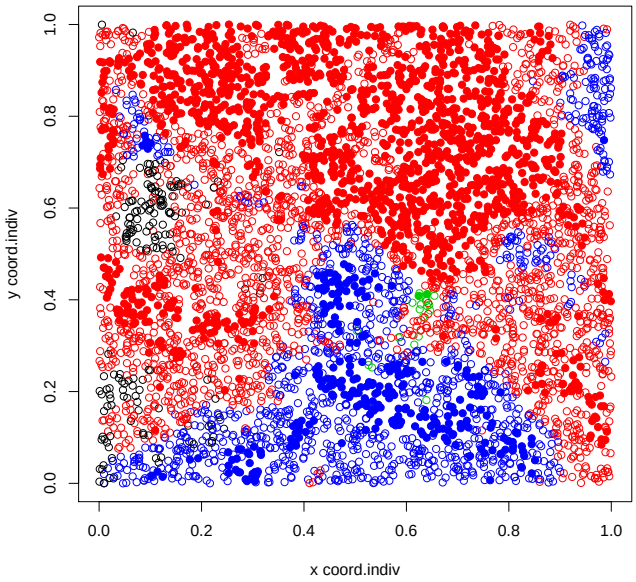


FIGURE 1.7 – Les points repèrent la position des individus échantillonnés sur un domaine. Quatre populations sont présentes : les cercles pleins désignent les individus dont le génome est "pur", et les cercles creux les individus dont le génome est hybride : dans ce cas la couleur désigne la population majoritaire dans le génome.

données *et* aux autres paramètres du modèle obtenus avec un premier appel à *Geneland* sous le modèle sans hybridation.

Un exemple d'estimation de q est présenté dans la Figure 1.9. Des simulations étudiant l'impact des paramètres a, b et L notamment sont présentées dans le Chapitre 5. En particulier, notre méthode se montre efficace même pour de faibles valeurs de L, c'est-à-dire dans la situation où peu de matériel génétique est disponible pour attester de la différenciation génétique. Bien sûr, les performances déclinent lorsque b devient grand : la dépendance spatiale est alors lâche et l'hypothèse d'homogénéité spatiale n'est plus vérifiée. C'est, en d'autres mots, le prix à payer pour l'utilisation d'un modèle spatial.

Notre approche présente donc l'avantage de modéliser explicitement la présence d'une zone d'hybridation (c'est-à-dire d'une singularité spatiale de la variabilité génétique), et d'estimer ainsi l'intensité et la portée du phénomène.

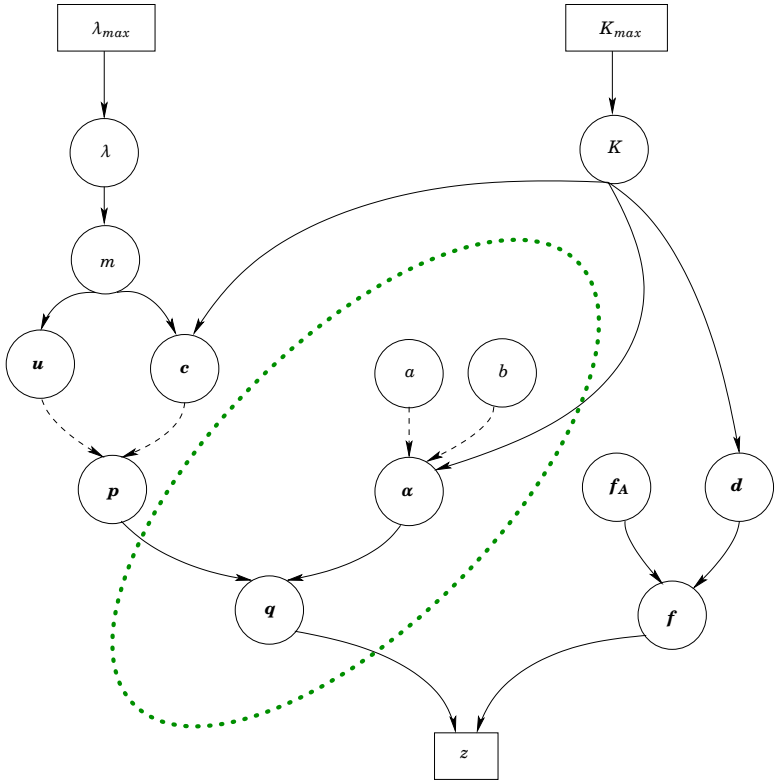


FIGURE 1.8 – Graphe acyclique dirigé du modèle proposé. Les lignes continues (respectivement pointillées) symbolisent des dépendances stochastiques (respectivement déterministes). Les données et les hyperparamètres fixés sont placés dans des carrés, et les paramètres estimés sont dans des cercles. La ligne pointillée verte encadre nos contributions. Les autres éléments sont empruntés aux modélisations de *Structure* ou *Geneland*.

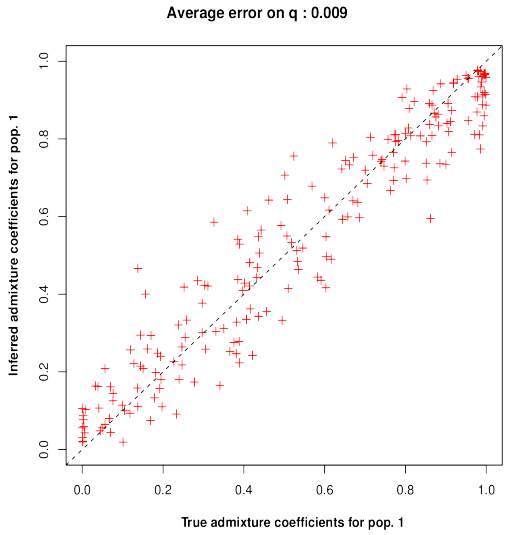


FIGURE 1.9 – Exemple en présence de deux populations : proportion estimée d'hybridation.

Chapitre 2

PAC-Bayesian Estimation and Prediction in Sparse Additive Models

Abstract. The present chapter is about estimation and prediction in high-dimensional additive models under a sparsity assumption ($p \gg n$ paradigm). A PAC-Bayesian strategy is investigated, delivering oracle inequalities in probability. The implementation is performed through recent outcomes in high-dimensional MCMC algorithms, and the performance of our method is assessed on simulated data.

The authors are grateful to Gérard Biau and Éric Moulines for their constant implication, and to Christophe Giraud and Taiji Suzuki for valuable insights and comments. They also thank an anonymous referee and an associate editor for providing constructive and helpful remarks.

Contents

This chapter has been published in *Electronic Journal of Statistics* (Guedj and Alquier, 2013).

2.1 Introduction

Substantial progress has been achieved over the last years in estimating very high-dimensional regression models. A thorough introduction to this dynamic field of contemporary statistics is provided by the recent monographs Hastie et al. (2009) and Bühlmann and van de Geer (2011). In the popular framework of linear and generalized linear models, the Lasso estimator introduced by Tibshirani (1996) immediately proved successful. Its theoretical properties have been extensively studied and its popularity has never wavered

since then, see for example Bunea et al. (2006), van de Geer (2008), Bickel et al. (2009) and Meinshausen and Yu (2009). However, even though numerous phenomena are well captured within this linear context, restraining high-dimensional statistics to this setting is unsatisfactory. To relax the strong assumptions required in the linear framework, one idea is to investigate a more general class of models, such as nonparametric regression models of the form $Y = f(X) + W$, where Y denotes the response, X the predictor and W a zero-mean noise. A good compromise between complexity and effectiveness is the additive model. It has been extensively studied and formalized for thirty years now. Amongst many other references, the reader is invited to refer to Stone (1985), Hastie and Tibshirani (1986, 1990) and Härdle (1990). The core of this model is that the regression function is written as a sum of univariate functions $f = \sum_{i=1}^{p} f_i$, easing its interpretation. Indeed, each covariate's effect is assessed by a unique function. This class of nonparametric models is a popular setting in statistics, despite the fact that classical estimation procedures are known to perform poorly as soon as the number of covariates p exceeds the number of observations n in that setting.

In the present chapter, our goal is to investigate a PAC-Bayesian-based prediction strategy in the high-dimensional additive framework ($p \gg n$ paradigm). In that context, estimation is essentially possible at the price of a sparsity assumption, *i.e.*, most of the f_i functions are zero. More precisely, our setting is nonasymptotic. As empirical evidence of sparse representations accumulates, high-dimensional statistics are more and more coupled with a sparsity assumption, namely that the intrinsic dimension p_0 of the data is much smaller than p and n, see e.g. Giraud et al. (2012). Additive modelling under a sparsity constraint has been essentially studied under the scope of the Lasso in Meier et al. (2009), Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2013) or of a combination of functional grouped Lasso and backfitting algorithm in Ravikumar et al. (2009). Those papers inaugurated the study of this problem and contain essential theoretical results consisting in asymptotic (Meier et al., 2009; Ravikumar et al., 2009) and nonasymptotic (Koltchinskii and Yuan, 2010; Suzuki and Sugiyama, 2013) oracle inequalities. The present chapter should be seen as a constructive contribution towards a deeper understanding of prediction problems in the additive framework. It should also be stressed that our work is to be seen as an attempt to relax as much as possible assumptions made on the model, such as restrictive conditions on the regressors' matrix. We consider them too much of a non-realistic burden when it comes to prediction problems.

Our *modus operandi* will be based on PAC-Bayesian results, which is original in that context to our knowledge. The PAC-Bayesian theory originates in the two seminal papers Shawe-Taylor and Williamson (1997) and McAllester (1999) and has been extensively formalized in the context of classification by Catoni (2004, 2007) and regression by Audibert (2004a,b), Alquier (2006, 2008) and Audibert and Catoni (2010, 2011). However, the methods presented in these references are not explicitly designed to cover the high-dimensional setting under the sparsity assumption. Thus, the PAC-Bayesian theory has been worked out in the sparsity perspective lately, by Dalalyan and Tsybakov (2008, 2012b), Alquier and Lounici (2011) and Rigollet and Tsybakov (2012). The main message of these studies is that aggregation with a properly chosen prior is able to deal effectively with the sparsity issue. Interesting additional references addressing the aggregation outcomes would be Rigollet (2006) and Audibert (2009). The former aggregation procedures rely on an exponential weights approach, achieving good statistical properties. Our method should be seen as an extension of these techniques, and is particularly focused on additive modelling specificities. Contrary to procedures such as the Lasso, the Dantzig selector and other penalized methods which are provably consistent under restrictive assumptions on the Gram matrix associated to the predictors, PAC-Bayesian aggregation requires only minimal assumptions

on the model. Our method is supported by oracle inequalities in probability, that are valid in both asymptotic and nonasymptotic settings. We also show that our estimators achieve the optimal rate of convergence over traditional smoothing classes such as Sobolev ellipsoids. It should be stressed that our work is inspired by Alquier and Biau (2013), which addresses the celebrated single-index model with similar tools and philosophy. Let us also mention that although the use of PAC-Bayesian techniques are original in this context, parallel work has been conducted in the deterministic design case by Suzuki (2012).

A major difficulty when considering high-dimensional problems is to achieve a favorable compromise between statistical and computational performances. The recent and thorough monograph Bühlmann and van de Geer (2011) shall provide the reader with valuable insights that address this drawback. As a consequence, the explicit implementation of PAC-Bayesian techniques remains unsatisfactory as existing routines are only put to test with small values of p (typically $p < 100$), contradicting with the high-dimensional framework. In the meantime, as a solution of a convex problem the Lasso proves computable for large values of p in reasonable amounts of time. We therefore focused on improving the computational aspect of our PAC-Bayesian strategy. Monte Carlo Markov Chains (MCMC) techniques proved increasingly popular in the Bayesian community, for they probably are the best way of sampling from potentially complex probability distributions. The reader willing to find a thorough introduction to such techniques is invited to refer to the comprehensive monographs Marin and Robert (2007) and Meyn and Tweedie (2009). While Alquier and Lounici (2011) and Alquier and Biau (2013) explore versions of the reversible jump MCMC method (RJMCMC) introduced by Green (1995), Dalalyan and Tsybakov (2008, 2012b) investigate a Langevin-Monte Carlo-based method, however only a deterministic design is considered. We shall try to overcome those limitations by considering adaptations of a recent procedure whose comprehensive description is to be found in Petralias (2010) and Petralias and Dellaportas (2012). This procedure called Subspace Carlin and Chib algorithm originates in the seminal paper by Carlin and Chib (1995), and has a close philosophy of Hans et al. (2007), as it favors local moves for the Markov chain. We provide numerical evidence that our method is computationally efficient, on simulated data.

The chapter is organized as follows. Section 2.2 presents our PAC-Bayesian prediction strategy in additive models. In particular, it contains the main theoretical results of this chapter which consist in oracle inequalities. Section 2.3 is devoted to the implementation of our procedure, along with numerical experiments on simulated data, presented in Section 2.4. Finally, and for the sake of clarity, proofs have been postponed to Section 2.5.

2.2 PAC-Bayesian prediction

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space on which we denote by $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ a sample of n independent and identically distributed (i.i.d.) random vectors in $(-1,1)^p \times \mathbb{R}$, with $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, satisfying

$$Y_i = \psi^\star(\mathbf{X}_i) + \xi_i = \sum_{j=1}^p \psi_j^\star(X_{ij}) + \xi_i, \quad i \in \{1, \ldots, p\},$$

where $\psi_1^\star, \ldots, \psi_p^\star$ are p continuous functions $(-1,1) \to \mathbb{R}$ and $\{\xi_i\}_{i=1}^n$ is a set of i.i.d. (conditionaly to $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$) real-valued random variables. Let $\mathcal{P}$ denote the distribution of the sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Denote by $\mathbb{E}$ the expectation computed with respect to $\mathbb{P}$ and let $\|\cdot\|_\infty$ be the supremum norm. We make the two following assumptions.

Assumption 2.1. *For any integer k, $\mathbb{E}[|\xi_1|^k] < \infty$, $\mathbb{E}[\xi_1|\mathbf{X}_1] = 0$ and there exist two positive constants L and σ^2 such that for any integer $k \geq 2$,*

$$\mathbb{E}[|\xi_1|^k|\mathbf{X}_1] \leq \frac{k!}{2}\sigma^2 L^{k-2}.$$

Assumption 2.2. *There exists a constant $C > \max(1,\sigma)$ such that $\|\psi^\star\|_\infty \leq C$.*

Note that Assumption 2.1 implies that $\mathbb{E}\xi_1 = 0$ and that the distribution of ξ_1 may depend on $\mathbf{X}_1$. In particular, Assumption 2.1 holds if ξ_1 is a zero-mean gaussian with variance $\gamma^2(\mathbf{X}_1)$ where $x \mapsto \gamma^2(x)$ is bounded.

Further, note that the boundedness assumption Assumption 2.2 plays a key role in our approach, as it allows to use a version of Bernstein's inequality which is one of the two main technical tools we use to state our results. This assumption is not only a technical prerequisite since it proved crucial for critical regimes: Indeed, if the intrinsic dimension p_0 of the regression function $\psi^\star$ is still large, the boundedness of the function class allows much faster estimation rates. This point is profusely discussed in Raskutti et al. (2012).

We are mostly interested in sparse additive models, in which only a few $\{\psi_j^\star\}_{j=1}^p$ are not identically zero. Let $\{\varphi_k\}_{k=1}^\infty$ be a known countable set of continuous functions $\mathbb{R} \to (-1,1)$ called the dictionary. In the sequel, $|\mathcal{H}|$ stands for the cardinality of a set $\mathcal{H}$. For any p-th tuple $\mathbf{m} = (m_1,\dots,m_p) \in \mathbb{N}^p$, denote by $S(\mathbf{m}) \subset \{1,\dots,p\}$ the set of indices of nonzero elements of $\mathbf{m}$, *i.e.*,

$$|S(\mathbf{m})| = \sum_{j=1}^p \mathbb{1}[m_j > 0],$$

and define

$$\Theta_{\mathbf{m}} = \left\{\theta \in \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_p}\right\},$$

with the convention $\mathbb{R}^0 = \emptyset$. The set $\Theta_{\mathbf{m}}$ is embedded with its canonical Borel field $\mathcal{B}(\Theta_m) = \mathcal{B}(\mathbb{R}^{m_1}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R}^{m_p})$. Denote by

$$\Theta \stackrel{\text{def}}{=} \bigcup_{\mathbf{m}\in\mathcal{M}} \Theta_{\mathbf{m}},$$

which is equipped with the σ-algebra $\mathcal{T} = \sigma(\bigvee_{\mathbf{m}\in\mathcal{M}} \mathcal{B}(\Theta_{\mathbf{m}}))$, where $\mathcal{M}$ is the collection of models $\mathcal{M} = \{\mathbf{m} = (m_1,\dots,m_p) \in \mathbb{N}^p\}$. Consider the span of the set $\{\varphi_k\}_{k=1}^\infty$, *i.e.*, the set of functions

$$\mathbb{F} = \left\{\psi_\theta = \sum_{j\in S(\mathbf{m})} \psi_j = \sum_{j\in S(\mathbf{m})} \sum_{k=1}^{m_j} \theta_{jk}\varphi_k : \theta \in \Theta_{\mathbf{m}}, \mathbf{m} \in \mathcal{M}\right\},$$

equipped with a countable generated σ-algebra denoted by $\mathcal{F}$. The risk and empirical risk associated to any $\psi_\theta \in \mathbb{F}$ are defined respectively as

$$R(\psi_\theta) = \mathbb{E}\left[Y_1 - \psi_\theta(\mathbf{X}_1)\right]^2 \quad \text{and} \quad R_n(\psi_\theta) = r_n(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \psi_\theta),$$

where

$$r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \psi_\theta(\mathbf{x}_i)\right)^2.$$

Consider the probability η_α on the set $\mathcal{M}$ defined by

$$\eta_\alpha : \mathbf{m} \mapsto \frac{1 - \frac{\alpha}{1-\alpha}}{1 - \left(\frac{\alpha}{1-\alpha}\right)^{p+1}} \binom{p}{|S(\mathbf{m})|}^{-1} \alpha^{\sum_{j=1}^p m_j},$$

for some $\alpha \in (0, 1/2)$. Let us stress the fact that the probability η_α acts as a penalization term over a model $\mathbf{m}$ through $\alpha^{\sum_{j=1}^p m_j}$ (and to a lesser extent, through the combinatorial term $\binom{p}{|S(\mathbf{m})|}^{-1}$).

Our procedure relies on the following construction of the probability π, referred to as the prior, in order to promote the sparsity properties of the target regression function $\psi^\star$. For any $\mathbf{m} \in \mathcal{M}$, $\zeta > 0$ and $\mathbf{x} \in \Theta_{\mathbf{m}}$, denote by $\mathcal{B}^1_{\mathbf{m}}(\mathbf{x}, \zeta)$ the ℓ^1-ball centered in $\mathbf{x}$ with radius ζ. For any $\mathbf{m} \in \mathcal{M}$, denote by $\pi_{\mathbf{m}}$ the uniform distribution on $\mathcal{B}^1_{\mathbf{m}}(0, C)$. Define the probability π on $(\Theta, \mathcal{T})$,

$$\pi(A) = \sum_{\mathbf{m} \in \mathcal{M}} \eta_\alpha(\mathbf{m}) \pi_{\mathbf{m}}(A), \quad A \in \mathcal{T}.$$

Note that the volume $V_{\mathbf{m}}(C)$ of $\mathcal{B}^1_{\mathbf{m}}(0, C)$ is given by

$$V_{\mathbf{m}}(C) = \frac{(2C)^{\sum_{j \in S(\mathbf{m})} m_j}}{\Gamma\left(\sum_{j \in S(\mathbf{m})} m_j + 1\right)} = \frac{(2C)^{\sum_{j \in S(\mathbf{m})} m_j}}{\left(\sum_{j \in S(\mathbf{m})} m_j\right)!}.$$

Finally, set $\delta > 0$ (which may be interpreted as an inverse temperature parameter) and the posterior Gibbs transition density is

$$\hat{\rho}_\delta(\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \theta) = \sum_{\mathbf{m} \in \mathcal{M}} \frac{\eta_\alpha(\mathbf{m})}{V_{\mathbf{m}}(C)} \mathbb{1}_{\mathcal{B}^1_{\mathbf{m}}(0,C)}(\theta) \frac{\exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta)]}{\int \exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta)] \pi(\mathrm{d}\theta)}. \tag{2.1}$$

We then consider two competing estimators. The first one is the randomized Gibbs estimator $\hat{\psi}$, constructed with parameters $\hat{\theta}$ sampled from the posterior Gibbs density, *i.e.*, for any $A \in \mathcal{F}$,

$$\mathbb{P}(\hat{\psi} \in A | \{\mathbf{X}_i, Y_i\}_{i=1}^n) = \int_A \hat{\rho}_\delta(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \theta) \pi(\mathrm{d}\theta), \tag{2.2}$$

while the second one is the aggregated Gibbs estimator $\hat{\psi}^{\mathrm{A}}$ defined as the posterior mean

$$\hat{\psi}^{\mathrm{A}} = \int \psi_\theta \hat{\rho}_\delta(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \theta) \pi(\mathrm{d}\theta) = \mathbb{E}[\hat{\psi} | \{\mathbf{X}_i, Y_i\}_{i=1}^n]. \tag{2.3}$$

These estimators have been introduced in Catoni (2004, 2007) and investigated in further work by Audibert (2004a), Alquier (2006, 2008) and Dalalyan and Tsybakov (2008, 2012b).

For the sake of clarity, denote by $\mathcal{D}$ a generic numerical constant in the sequel. We are now in a position to write a PAC-Bayesian oracle inequality.

Theorem 2.2.1. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be the Gibbs estimators defined by (2.2)–(2.3), respectively. Let Assumption 2.1 and Assumption 2.2 hold. Set $w = 8C \max(L, C)$ and $\delta = n\ell/[w + 4(\sigma^2 + C^2)]$, for $\ell \in (0,1)$, and let $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability at least $1 - 2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \le \mathcal{D} \inf_{\mathbf{m} \in \mathcal{M}} \inf_{\theta \in \mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(\psi_\theta) - R(\psi^\star) \right.$$
$$\left. + |S(\mathbf{m})| \frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(n)}{n} \sum_{j \in S(\mathbf{m})} m_j + \frac{\log(1/\varepsilon)}{n} \right\}, \tag{2.4}$$

where $\mathcal{D}$ depends upon w, σ, C, ℓ and α defined above.

Under mild assumptions, Theorem 2.2.1 provides inequalities which admit the following interpretation. If there exists a "small" model in the collection $\mathcal{M}$, *i.e.*, a model $\mathbf{m}$ such that $\sum_{j \in S(\mathbf{m})} m_j$ and $|S(\mathbf{m})|$ are small, such that ψ_θ (with $\theta \in \Theta_{\mathbf{m}}$) is close to $\psi^\star$, then $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ are also close to $\psi^\star$ up to $\log(n)/n$ and $\log(p)/n$ terms. However, if no such model exists, at least one of the terms $\sum_{j \in S(\mathbf{m})} m_j/n$ and $|S(\mathbf{m})|/n$ starts to emerge, thereby deteriorating the global quality of the bound. A satisfying estimation of $\psi^\star$ is typically possible when $\psi^\star$ admits a sparse representation.

To go further, we derive from Theorem 2.2.1 an inequality on Sobolev ellipsoids. We show that our procedure achieves the optimal rate of convergence in this setting. For the sake of shortness, we consider Sobolev spaces, however one can easily derive the following results in other functional spaces such as Besov spaces. See Tsybakov (2009) and the references therein.

The notation $\{\varphi_k\}_{k=1}^{\infty}$ now refers to the (non-normalized) trigonometric system, defined as

$$\varphi_1 : t \mapsto 1, \quad \varphi_{2j} : t \mapsto \cos(\pi j t), \quad \varphi_{2j+1} : t \mapsto \sin(\pi j t),$$

with $j \in \mathbb{N}^*$ and $t \in (-1,1)$. Let us denote by $S^\star$ the set of indices of non-identically zero regressors. That is, the regression function $\psi^\star$ is

$$\psi^\star = \sum_{j \in S^\star} \psi_j^\star.$$

Assume that for any $j \in S^\star$, $\psi_j^\star$ belongs to the Sobolev ellipsoid $\mathcal{W}(r_j, d_j)$ defined as

$$\mathcal{W}(r_j, d_j) = \left\{ f \in \mathrm{L}^2([-1,1]) : f = \sum_{k=1}^{\infty} \theta_k \varphi_k \quad \text{and} \quad \sum_{i=1}^{\infty} i^{2r_j} \theta_i^2 \le d_j \right\}.$$

with d_j chosen such that $\sum_{j \in S^\star} \sqrt{d_j} \le C\sqrt{6}/\pi$, and $r_1, \ldots, r_{|S^\star|} \ge 1$ are unknown regularity parameters. Let us stress the fact that this assumption casts our results onto the adaptive setting. It also implies that $\psi^\star$ belongs to the Sobolev ellipsoid $\mathcal{W}(r,d)$, with $r = \min_{j \in S^\star} r_j$ and $d = \sum_{j \in S^\star} d_j$, *i.e.*,

$$\psi^\star = \sum_{j \in S^\star} \sum_{k=1}^{\infty} \theta_{jk}^\star \varphi_k. \tag{2.5}$$

It is worth pointing out that in that setting, the Sobolev ellipsoid is better approximated by the ℓ^1-ball $\mathcal{B}^1_{\mathbf{m}}(0,C)$ as the dimension of $\mathbf{m}$ grows. Further, make the following assumption.

Assumption 2.3. *The distribution of the data $\mathcal{P}$ has a probability density with respect to the corresponding Lebesgue measure, bounded from above by a constant $B > 0$.*

Theorem 2.2.2. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be the Gibbs estimators defined by (2.2)–(2.3), respectively. Let Assumption 2.1, Assumption 2.2 and Assumption 2.3 hold. Set $w = 8C\max(L,C)$ and $\delta = n\ell/[w + 4(\sigma^2 + C^2)]$, for $\ell \in (0,1)$, and let $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability at least $1 - 2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \le \mathcal{D} \left\{ \sum_{j \in S^\star} d_j^{\frac{1}{2r_j+1}} \left(\frac{\log(n)}{2nr_j} \right)^{\frac{2r_j}{2r_j+1}} + \frac{|S^\star| \log(p/|S^\star|)}{n} + \frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ is a constant depending only on w, σ, C, ℓ, α and B.

Theorem 2.2.2 illustrates that we obtain the minimax rate of convergence over Sobolev classes up to a $\log(n)$ term. Indeed, the minimax rate to estimate a single function with regularity r is $n^{\frac{2r}{2r+1}}$, see for example Tsybakov (2009, Chapter 2). Theorem 2.2.1 and Theorem 2.2.2 thus validate our method.

A salient fact about Theorem 2.2.2 is its links with existing work: Assume that all the $\psi_j^\star$ belong to the same Sobolev ellipsoid $\mathcal{W}(r,d)$. The convergence rate is now $\log(n) n^{-\frac{2r}{2r+1}} + \log(p)/n$. This rate (down to a $\log(n)$ term) is the same as the one exhibited by Koltchinskii and Yuan (2010) in the context of multiple kernel learning ($n^{-\frac{2r}{2r+1}} + \log(p)/n$). Suzuki and Sugiyama (2013) even obtain faster rates which correspond to smaller functional spaces. However, the results presented by both Koltchinskii and Yuan (2010) and Suzuki and

Sugiyama (2013) are obtained under stringent conditions on the design, which are not necessary to prove Theorem 2.2.2.

A natural extension is to consider sparsity on both regressors and their expansion, instead of sparse regressors and nested expansion as before. That is, we no longer consider the first m_j dictionary functions for the expansion of regressor j. To this aim, we slightly extend the previous notation. Let $K \in \mathbb{N}^*$ be the length of the dictionary. A model is now denoted by $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_p)$ and for any $j \in \{1, \dots, p\}$, $\mathbf{m}_j = (m_{j1}, \dots, m_{jK})$ is a K-sized vector whose entries are 1 whenever the corresponding dictionary function is present in the model and 0 otherwise. Introduce the notation

$$S(\mathbf{m}) = \{\mathbf{m}_j \neq \mathbf{0}, j \in \{1, \dots, p\}\}, \quad S(\mathbf{m}_j) = \{m_{jk} \neq 0, k \in \{1, \dots, K\}\}.$$

The prior distribution on the models space $\mathcal{M}$ is now

$$\eta_\alpha \colon \mathbf{m} \mapsto \frac{1 - \alpha \frac{1-\alpha^{K+1}}{1-\alpha}}{1 - \left(\alpha \frac{1-\alpha^{K+1}}{1-\alpha}\right)^{p+1}} \binom{p}{|S(\mathbf{m})|}^{-1} \prod_{j \in S(\mathbf{m})} \binom{K}{|S(\mathbf{m}_j)|}^{-1} \alpha^{|S(\mathbf{m}_j)|},$$

for any $\alpha \in (0, 1/2)$.

Theorem 2.2.3. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be the Gibbs estimators defined by (2.2)–(2.3), respectively. Let Assumption 2.1 and Assumption 2.2 hold. Set $w = 8C \max(L, C)$ and $\delta = n\ell/[w + 4(\sigma^2 + C^2)]$, for $\ell \in (0, 1)$, and let $\varepsilon \in (0, 1)$.Then with $\mathbb{P}$-probability at least $1 - 2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \leq \mathcal{D} \inf_{\mathbf{m} \in \mathcal{M}} \inf_{\theta \in \mathcal{B}^1_{\mathbf{m}}(0, C)} \left\{ R(\psi_\theta) - R(\psi^\star) \right.$$
$$\left. + |S(\mathbf{m})| \frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(nK)}{n} \sum_{j \in S(\mathbf{m})} |S(\mathbf{m}_j)| + \frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ depends upon w, σ, C, ℓ and α defined above.

2.3 MCMC implementation

In this section, we describe an implementation of the method outlined in the previous section. Our goal is to sample from the Gibbs posterior distribution $\hat{\rho}_\delta$. We use a version of the so-called Subspace Carlin and Chib (SCC) developed by Petralias (2010) and Petralias and Dellaportas (2012) which originates in the Shotgun Stochastic Search algorithm (Hans et al., 2007). The key idea of the algorithm lies in a stochastic search heuristic that restricts moves in neighborhoods of the visited models. Let $T \in \mathbb{N}^*$ and denote by $\{\theta(t), \mathbf{m}(t)\}_{t=0}^T$ the Markov chain of interest, with $\theta(t) \in \Theta_{\mathbf{m}(t)}$. Define $i \colon t \mapsto \{+, -, =\}$, the three possible moves performed by the algorithm: An addition, a deletion or an adjustment of a regressor. Let $\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ be the canonical base of $\mathbb{R}^p$. For any model $\mathbf{m}(t) = (m_1(t), \dots, m_p(t)) \in \mathcal{M}$, define its neighborhood $\{\mathcal{V}^+[\mathbf{m}(t)], \mathcal{V}^-[\mathbf{m}(t)], \mathcal{V}^=[\mathbf{m}(t)]\}$, where

$$\mathcal{V}^+[\mathbf{m}(t)] = \{\mathbf{k} \in \mathcal{M} \colon \mathbf{k} = \mathbf{m}(t) + x\mathbf{e}_j, x \in \mathbb{N}^*, j \in \{1, \dots, p\} \setminus S[\mathbf{m}(t)]\},$$

$$\mathcal{V}^-[\mathbf{m}(t)] = \{\mathbf{k} \in \mathcal{M} \colon \mathbf{k} = \mathbf{m}(t) - m_j(t)\mathbf{e}_j, j \in S[\mathbf{m}(t)]\},$$

and

$$\mathcal{V}^=[\mathbf{m}(t)] = \{\mathbf{k} \in \mathcal{M} \colon S(\mathbf{k}) = S[\mathbf{m}(t)]\}.$$

A move $i(t)$ is chosen with probability $q[i(t)]$. By convention, if $|S[\mathbf{m}(t)]| = \min(pK, n)$ (respectively $|S[\mathbf{m}(t)]| = 1$) the probability of performing an addition move (respectively a deletion move) is zero. Note $\xi\colon \{+,-\} \mapsto \{-,+\}$ and let $D_{\mathbf{m}}$ be the design matrix in model $\mathbf{m} \in \mathcal{M}$. Denote by $\mathrm{LSE}_{\mathbf{m}}$ the least square estimate $\mathrm{LSE}_{\mathbf{m}} = (D'_{\mathbf{m}} D_{\mathbf{m}})^{-1} D'_{\mathbf{m}} \mathbf{Y}$ (with $\mathbf{Y} = (Y_1, \ldots, Y_n)$) in model $\mathbf{m} \in \mathcal{M}$. This is possible since we prevent the chain from visiting models with more covariates than the sample size n. For ease of notation, let $\mathcal{I}$ denote the identity matrix. Finally, denote by $\phi(\cdot; \mu, \Gamma)$ the density of a Gaussian distribution $\mathcal{N}(\mu, \Gamma)$ with mean μ and covariance matrix Γ. A description of the full algorithm is presented in Algorithm 2.1.

The estimates $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ are obtained as

$$\hat{\psi} = \sum_{j=1}^{p} \sum_{k=1}^{K} \theta_{jk}(T) \varphi_k,$$

and for some burnin $b \in \{1, \ldots, T-1\}$,

$$\hat{\psi}^{\mathrm{A}} = \sum_{j=1}^{p} \sum_{k=1}^{K} \left(\frac{1}{T-b} \sum_{\ell=b+1}^{T} \theta_{jk}(\ell) \right) \varphi_k.$$

The transition kernel of the chain defined above is reversible with respect to $\hat{\rho}_\delta \otimes \eta_\alpha$, hence this procedure ensures that $\{\theta(t)\}_{t=1}^{T}$ is a Markov Chain with stationary distribution $\hat{\rho}_\delta$.

2.4 Numerical studies

In this section we validate the effectiveness of our method on simulated data. All our numerical studies have been performed with the software R (R Core Team, 2013). The method is available on the CRAN website (`http://www.cran.r-project.org/web/packages/pacbpred/index.html`), under the name `pacbpred` (Guedj, 2013b).

Some comments are in order here about how to calibrate the constants C, K, σ^2, δ and α. Clearly, a too small value for C will stuck the algorithm, preventing the chain to escape from the initial model. Indeed, most proposed models will be discarded since the acceptance ratio will frequently take the value 0. Conversely, a large value for C deteriorates the quality of the bound in Theorem 2.2.1, Theorem 2.2.2, Theorem 2.2.3 and Theorem 2.5.1. However, this only influences the theoretical bound, as its contribution to the acceptance ratio is limited to $\log(2C)$. We thereby proceeded with typically large values of C (such as $C = 10^6$). The size K of the dictionary determines the quality of the approximation of the functions $\psi_j^\star$: A good compromise between approximation and computational complexity is achieved for K ranging approximately from 6 to 12 (the default value in the package `pacbpred` is $K = 8$). As the parameter σ^2 is the variance of the proposal distribution ϕ, the practioner should tune it in accordance with the noise level of the data. The parameter requiring the finest calibration is δ: The convergence of the algorithm is sensitive to its choice. Dalalyan and Tsybakov (2008, 2012b) exhibit the theoretical value $\delta = n/4\sigma^2$. This value leads to very good numerical performances, as it has been also noticed by Dalalyan and Tsybakov (2008, 2012b) and Alquier and Biau (2013). The choice for α is guided by a similar reasoning to the one for C. Its contribution to the acceptance ratio is limited to a $\log(1/\alpha)$ term. The value $\alpha = 0.25$ was used in the simulations for its apparent good properties. Although it would be computationally challenging, a finer calibration through methods such as cross-validation or bayesian integration is possible.

Finally and as a general rule, we strongly encourage practitioners to run several chains of inequal lengths and to adjust the number of iterations needed by observing if the empirical risk is stabilized.

Algorithm 2.1 A Subspace Carlin and Chib-based algorithm

Initialize $(\theta(0), \mathbf{m}(0))$.
for $t = 1$ to T **do**
  Choose a move $i(t)$ with probability $q[i(t)]$.
  For any $\mathbf{k} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]$, generate $\theta_{\mathbf{k}}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{k}}, \sigma^2 \mathcal{I})$.
  Propose a model $\mathbf{k} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]$ with probability

$$\gamma(\mathbf{m}(t-1), \mathbf{k}) = \frac{A_{\mathbf{k}}}{\sum_{\mathbf{j} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]} A_{\mathbf{j}}},$$

  where

$$A_{\mathbf{j}} = \frac{\hat{\rho}_{\delta}(\theta_{\mathbf{j}})}{\phi(\theta_{\mathbf{j}}; \mathrm{LSE}_{\mathbf{j}}, \sigma^2 \mathcal{I})}.$$

  if $i(t) \in \{+, -\}$ **then**
    For any $\mathbf{h} \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]$, generate $\theta_{\mathbf{h}}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{h}}, \sigma^2 \mathcal{I})$. Note that $\mathbf{m}(t-1) \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]$.
    Accept model $\mathbf{k}$, *i.e.*, set $\mathbf{m}(t) = \mathbf{k}$ and $\theta(t) = \theta_{\mathbf{k}}$, with probability

$$\alpha = \min\left(1, \frac{A_{\mathbf{k}} q[i(t)] \gamma(\mathbf{k}, \mathbf{m}(t-1))}{A_{\mathbf{m}(t-1)} q[\xi(i(t))] \gamma(\mathbf{m}(t-1), \mathbf{k})}\right) = \min\left(1, \frac{q[i(t)] \sum_{\mathbf{h} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]} A_{\mathbf{h}}}{q[\xi(i(t))] \sum_{\mathbf{h} \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]} A_{\mathbf{h}}}\right).$$

    Otherwise, set $\mathbf{m}(t) = \mathbf{m}(t-1)$ and $\theta(t) = \theta_{\mathbf{m}(t-1)}$.
  else
    Generate $\theta_{\mathbf{m}(t-1)}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{m}(t-1)}, \sigma^2 \mathcal{I})$.
    Accept model $\mathbf{k}$, *i.e.*, set $\mathbf{m}(t) = \mathbf{k}$ and $\theta(t) = \theta_{\mathbf{k}}$, with probability

$$\alpha = \min\left(1, \frac{A_{\mathbf{k}} \gamma(\mathbf{k}, \mathbf{m}(t-1))}{A_{\mathbf{m}(t-1)} \gamma(\mathbf{m}(t-1), \mathbf{k})}\right).$$

    Otherwise, set $\mathbf{m}(t) = \mathbf{m}(t-1)$ and $\theta(t) = \theta_{\mathbf{m}(t-1)}$.
  end if
end for

TABLE 2.1 – Each number is the mean (standard deviation) of the RSS over 10 independent runs

MCMC	$p = 50$ 3000 it.	$p = 200$ 10000 it.	$p = 400$ 20000 it.
Model 2.1	0.0318 (0.0047)	0.0320 (0.0029)	0.0335 (0.0056)
Model 2.2	0.0411 (0.0061)	0.1746 (0.0639)	0.2201 (0.0992)
Model 2.3	0.0665 (0.0421)	0.1151 (0.0399)	0.1597 (0.0579)

Model 2.1. *$n = 200$ and $S^\star = \{1,2,3,4\}$. This model is similar to Meier et al. (2009, Section 3, Example 1) and is given by*

$$Y_i = \psi_1^\star(X_{i1}) + \psi_2^\star(X_{i2}) + \psi_3^\star(X_{i3}) + \psi_4^\star(X_{i4}) + \xi_i,$$

with

$$\psi_1^\star : x \mapsto -\sin(2x), \quad \psi_2^\star : x \mapsto x^3, \quad \psi_3^\star : x \mapsto x, \quad \psi_4^\star : x \mapsto e^{-x} - e/2, \quad \xi_i \sim \mathcal{N}(0, 0.1),$$

with $i \in \{1,\dots,n\}$. The covariates are sampled from independent uniform distributions over $(-1,1)$.

Model 2.2. *$n = 200$ and $S^\star = \{1,2,3,4\}$. As above but correlated. The covariates are sampled from a multivariate gaussian distribution with covariance matrix $\Sigma_{ij} = 2^{-|i-j|-2}$, i, $j \in \{1,\dots,p\}$.*

Model 2.3. *$n = 200$ and $S^\star = \{1,2,3,4\}$. This model is similar to Meier et al. (2009, Section 3, Example 3) and is given by*

$$Y_i = 5\psi_1^\star(X_{i1}) + 3\psi_2^\star(X_{i2}) + 4\psi_3^\star(X_{i3}) + 6\psi_4^\star(X_{i4}) + \xi_i,$$

with

$$\psi_1^\star : x \mapsto x, \quad \psi_2^\star : x \mapsto 4(x^2 - x - 1), \quad \psi_3^\star : x \mapsto \frac{\sin(2\pi x)}{2 - \sin(2\pi x)},$$
$$\psi_4^\star : x \mapsto 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x),$$
$$\xi_i \sim \mathcal{N}(0, 0.5), \quad i \in \{1,\dots,n\}.$$

The covariates are sampled from independent uniform distributions over $(-1,1)$.

The results of the simulations are summarized in Table 2.1 and illustrated by Figure 2.1 and Figure 2.2. The reconstruction of the true regression function $\psi^\star$ is achieved even in very high-dimensional situations, pulling up our method at the level of the gold standard Lasso.

2.5 Proofs

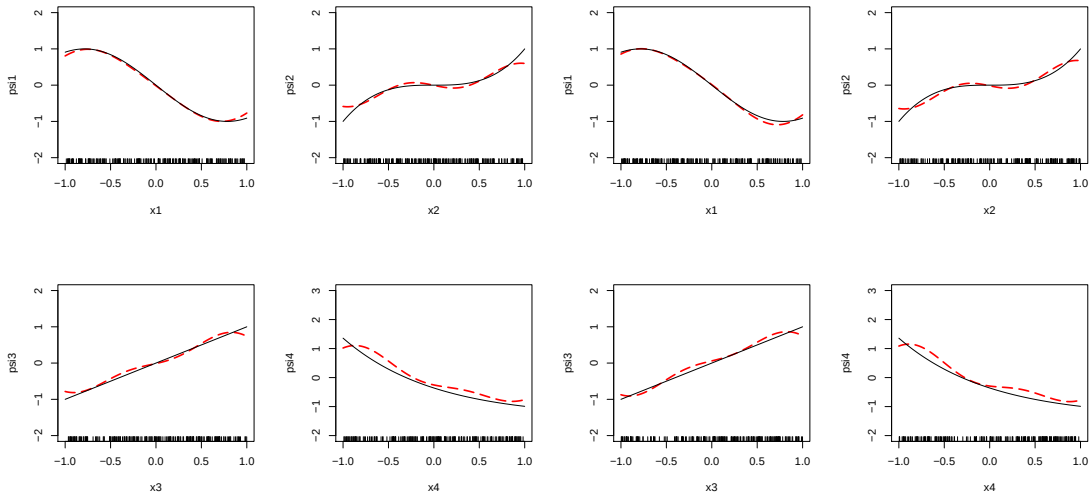
To start the chain of proofs leading to Theorem 2.2.1, Theorem 2.2.2 and Theorem 2.2.3, we recall and prove some lemmas to establish Theorem 2.5.1 which consists in a general PAC-Bayesian inequality in the spirit of Catoni (2004, Theorem 5.5.1) for classification or Catoni (2004, Lemma 5.8.2) for regression. Note also that Dalalyan and Tsybakov (2012b, Theorem 1) provides a similar inequality in the deterministic design case. A salient fact on Theorem 2.5.1 is that the validity of the oracle inequalities only involves the distribution of the noise variable ξ_1, and that distribution is independent of the sample size n.

The proofs of the following two classical results are omitted. Lemma 2.5.1 is a version of Bernstein's inequality which originates in Massart (2007, Proposition 2.9), whereas Lemma 2.5.2 appears in Catoni (2004, Equation 5.2.1).

For $x \in \mathbb{R}$, denote $(x)_+ = \max(x, 0)$. Let μ_1, μ_2 be two probabilities. The Kullback-Leibler divergence of μ_1 with respect to μ_2 is denoted $\mathcal{KL}(\mu_1, \mu_2)$ and is

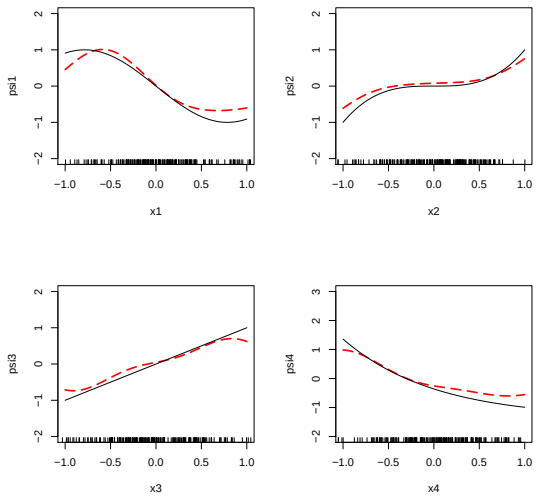
$$\mathcal{KL}(\mu_1, \mu_2) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_1 & \text{if } \mu_1 \ll \mu_2, \\ \infty & \text{otherwise.} \end{cases}$$

Finally, for any measurable space $(A, \mathcal{A})$ and any probability π on $(A, \mathcal{A})$, let us denote by $\mathcal{M}^1_{+,\pi}(A, \mathcal{A})$ the set of probabilities on $(A, \mathcal{A})$ absolutely continuous with respect to π.

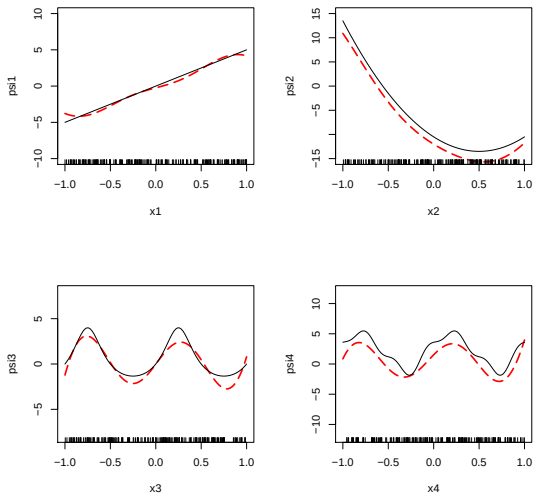


(A) Model 2.1, $p = 200$.

(B) Model 2.1, $p = 400$.

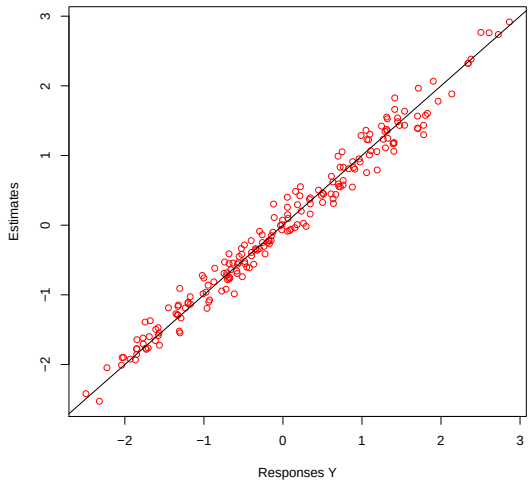


(C) Model 2.2, $p = 50$.

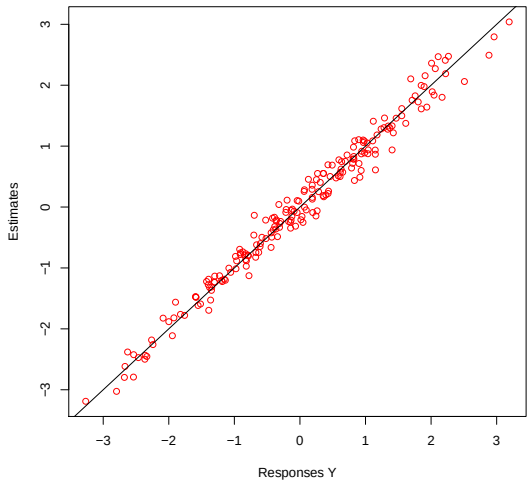


(D) Model 2.3, $p = 50$.

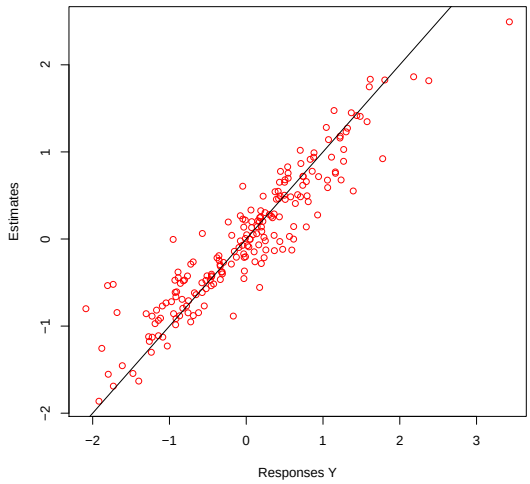
FIGURE 2.1 – Estimates (red dashed lines) for $\psi_1^\star$, $\psi_2^\star$, $\psi_3^\star$ and $\psi_4^\star$ (solid black lines). Other estimates (for $\psi_j^\star$, $j \notin \{1,2,3,4\}$) are mostly zero.



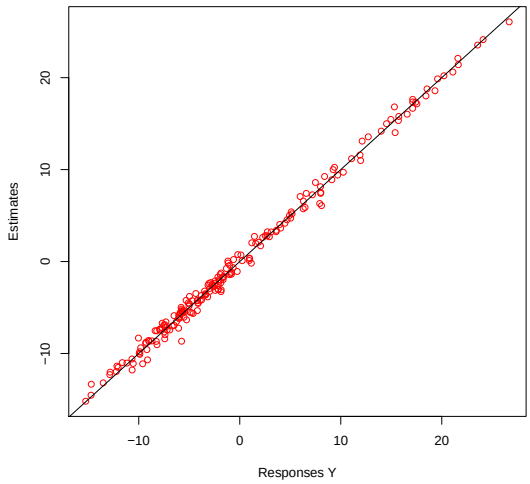
(A) Model 2.1, $p = 200$.



(B) Model 2.1, $p = 400$.



(C) Model 2.2, $p = 50$.



(D) Model 2.3, $p = 50$.

FIGURE 2.2 – Plot of the responses $Y_1, \ldots, Y_n$ against their estimates. The more points on the first bisectrix (solid black line), the better the estimation.

Lemma 2.5.1. *Let $(T_i)_{i=1}^n$ be independent real-valued variables. Assume that there exist two positive constants v and w such that*

$$\sum_{i=1}^n \mathbb{E}T_i^2 \le v,$$

and for any integer $k \ge 3$,

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \le \frac{k!}{2} v w^{k-2}.$$

Then for any $\gamma \in \left(0, \frac{1}{w}\right)$,

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^n (T_i - \mathbb{E}T_i)\right)\right] \le \exp\left(\frac{v\gamma^2}{2(1-w\gamma)}\right).$$

Lemma 2.5.2. *Let $(A, \mathcal{A})$ be a measurable space. For any probability μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int (\exp \circ h) \mathrm{d}\mu < \infty$,*

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}_{+,\pi}^1(A,\mathcal{A})} \int h \mathrm{d}m - \mathcal{KL}(m, \mu),$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

Note that Theorem 2.5.1 is valid in the general regression framework. In the proofs of Lemma 2.5.3, Lemma 2.5.4, Lemma 2.5.5 and Theorem 2.5.1, we consider a general regression function $\psi^\star$. Denote by $(\Theta, \mathcal{T})$ a space of functions equipped with a countable generated σ-algebra, and let π be a probability on $(\Theta, \mathcal{T})$, referred to as the prior. Lemma 2.5.3, Lemma 2.5.4, Lemma 2.5.5 and Theorem 2.5.1 follow from the work of Catoni (2004), Alquier (2008), Dalalyan and Tsybakov (2008, 2012b) and Alquier and Biau (2013). Let $\delta > 0$ and consider the so-called *posterior* Gibbs transition density $\hat{\rho}_\delta$ with respect to π, defined as

$$\hat{\rho}_\delta(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi) = \frac{\exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi)]}{\int \exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi)] \pi(\mathrm{d}\psi)}. \tag{2.6}$$

In the following three lemmas, denote by ρ a *so-called* posterior probability absolutely continuous with respect to π. Let ψ be a random variable sampled from ρ.

Lemma 2.5.3. *Let Assumption 2.1 and Assumption 2.2 hold. Set $w = 8C\max(L, C)$, $\delta \in (0, n/[w + 4(\sigma^2 + C^2)])$ and $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability at least $1 - \varepsilon$*

$$R(\psi) - R(\psi^\star) \le \frac{1}{1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta}} \left(R_n(\psi) - R_n(\psi^\star) + \frac{\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) + \log \frac{1}{\varepsilon}}{\delta} \right).$$

Proof. Apply Lemma 2.5.1 to the variables T_i defined as follow: For any $\psi \in \mathbb{F}$,

$$T_i = -(Y_i - \psi(\mathbf{X}_i))^2 + (Y_i - \psi^\star(\mathbf{X}_i))^2, \quad i \in \{1, \dots, n\}. \tag{2.7}$$

First, let us note that

$$\begin{aligned}
R(\psi) - R(\psi^\star) &= \mathbb{E}[(Y_1 - \psi(\mathbf{X}_1))^2] - \mathbb{E}[(Y_1 - \psi^\star(\mathbf{X}_1))^2] \\
&= \mathbb{E}[(2Y_1 - \psi(\mathbf{X}_1) - \psi^\star(\mathbf{X}_1))(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))] \\
&= \mathbb{E}\left[(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))\mathbb{E}[(2W_1 + \psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1)) | \mathbf{X}_1]\right] \\
&= 2\mathbb{E}[(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))\mathbb{E}[\xi_1 | \mathbf{X}_1]] + \mathbb{E}[\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1)]^2.
\end{aligned}$$

As $\mathbb{E}[\xi_1|\mathbf{X}_1]=0$,

$$R(\psi)-R(\psi^\star)=\mathbb{E}[\psi^\star(\mathbf{X})-\psi(\mathbf{X})]^2. \tag{2.8}$$

By (2.7), the random variables $(T_i)_{i=1}^n$ are independent. Using Lemma 2.5.1, we get

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}T_i^2 &= \sum_{i=1}^n \mathbb{E}\left[(2Y_i-\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i))^2(\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i))^2\right]\\ &= \sum_{i=1}^n \mathbb{E}\mathbb{E}\left[(2W_i+\psi^\star(\mathbf{X}_i)-\psi(\mathbf{X}_i))^2(\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i))^2|\mathbf{X}_i\right].\end{aligned}$$

Next, using that $|a+b|^k\le 2^{k-1}(|a|^k+|b|^k)$ for any $a,b\in\mathbb{R}$ and $k\in\mathbb{N}^*$, we get

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}T_i^2 &\le 2\sum_{i=1}^n \mathbb{E}\left[(\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i))^2\mathbb{E}\left[(4W_i^2+4C^2)|\mathbf{X}_i\right]\right]\\ &\le 8\left(\sigma^2+C^2\right)\sum_{i=1}^n \mathbb{E}\left[(\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i))^2\right]\\ &= 8n\left(\sigma^2+C^2\right)\left(R(\psi)-R(\psi^\star)\right)\stackrel{\text{def}}{=} v,\end{aligned} \tag{2.9}$$

where we have used (2.8) in the last equation. It follows that for any integer $k\ge 3$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] &= \sum_{i=1}^n \mathbb{E}\mathbb{E}[(T_i)_+^k|\mathbf{X}_i]\\ &\le \sum_{i=1}^n \mathbb{E}\mathbb{E}\left[|2Y_i-\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i)|^k|\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i)|^k|\mathbf{X}_i\right]\\ &= \sum_{i=1}^n \mathbb{E}\mathbb{E}\left[|2W_i+\psi^\star(\mathbf{X}_i)-\psi(\mathbf{X}_i)|^k|\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i)|^k|\mathbf{X}_i\right]\\ &\le 2^{k-1}\sum_{i=1}^n \mathbb{E}\mathbb{E}\left[\left(2^k|\xi_i|^k+|\psi^\star(\mathbf{X}_i)-\psi(\mathbf{X}_i)|^k\right)|\psi(\mathbf{X}_i)-\psi^\star(\mathbf{X}_i)|^k|\mathbf{X}_i\right].\end{aligned}$$

Using that $|\psi(\mathbf{x}_i)-\psi^\star(\mathbf{x}_i)|^k\le(2C)^{k-2}|\psi(\mathbf{x}_i)-\psi^\star(\mathbf{x}_i)|^2$ and (2.8), we get

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] &\le 2^{k-1}\sum_{i=1}^n\left(2^{k-1}k!\sigma^2L^{k-2}+(2C)^k\right)(2C)^{k-2}[R(\psi)-R(\psi^\star)]\\ &= \frac{k!}{2}v(2C)^{k-2}\left(\frac{2^{2k-4}\sigma^2L^{k-2}+\frac{2}{k!}2^{2k-4}C^k}{\sigma^2+C^2}\right).\end{aligned}$$

Recalling that $C>\max(1,\sigma)$ gives

$$\begin{aligned}\frac{2^{2k-4}\sigma^2L^{k-2}+\frac{2}{k!}2^{2k-4}C^k}{\sigma^2+C^2} &\le \frac{4^{k-2}\sigma^2L^{k-2}}{2\sigma^2}+\frac{\frac{2}{k!}4^{k-2}C^k}{C^2}\\ &\le \frac{1}{2}(4L)^{k-2}+\frac{1}{2}(4C)^{k-2}\\ &= [4\max(L,C)]^{k-2}.\end{aligned}$$

Hence

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k]\le\frac{k!}{2}vw^{k-2},\quad\text{with}\quad w\stackrel{\text{def}}{=}8C\max(L,C). \tag{2.10}$$

Applying Lemma 2.5.1, we obtain, for any real $\delta\in\left(0,\frac{n}{w}\right)$, with $\gamma=\frac{\delta}{n}$,

$$\mathbb{E}\exp[\delta(R_n(\psi^\star)-R_n(\psi)+R(\psi)-R(\psi^\star))]\le\exp\left(\frac{v\delta^2}{2n^2\left(1-\frac{w\delta}{n}\right)}\right),$$

that is, that for any real number $\varepsilon \in (0,1)$,

$$\mathbb{E}\exp\left[\delta[R_n(\psi^\star)-R_n(\psi)]+\delta[R(\psi)-R(\psi^\star)]\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)-\log\frac{1}{\varepsilon}\right]\le\varepsilon. \tag{2.11}$$

Next, we use a standard PAC-Bayesian approach as developed in Audibert (2004a), Catoni (2004, 2007) and Alquier (2008). For any prior probability π on $(\Theta,\mathcal{T})$,

$$\int\mathbb{E}\exp\left[\delta[R(\psi)-R(\psi^\star)]\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi^\star)-R_n(\psi)]-\log\frac{1}{\varepsilon}\right]\pi(\mathrm{d}\psi)\le\varepsilon.$$

By the Fubini-Tonelli theorem

$$\mathbb{E}\int\exp\left[\delta[R(\psi)-R(\psi^\star)]\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi^\star)-R_n(\psi)]-\log\frac{1}{\varepsilon}\right]\pi(\mathrm{d}\psi)\le\varepsilon.$$

Therefore, for any data-dependent posterior probability measure ρ absolutely continuous with respect to π, adopting the convention $\infty\times 0=0$,

$$\mathbb{E}\int\exp\left[\delta[R(\psi)-R(\psi^\star)]\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi^\star)-R_n(\psi)]-\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi)\right.$$
$$\left.-\log\frac{1}{\varepsilon}\right]\rho(\mathrm{d}\psi)\le\varepsilon. \tag{2.12}$$

Recalling that $\mathbb{E}$ stands for the expectation computed with respect to $\mathbb{P}$, the integration symbol may be omitted and we get

$$\mathbb{E}\exp\left[\delta[R(\psi)-R(\psi^\star)]\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi^\star)-R_n(\psi)]-\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi)-\log\frac{1}{\varepsilon}\right]\le\varepsilon.$$

Using the elementary inequality $\exp(\delta x)\ge\mathbb{1}_{\mathbb{R}_+}(x)$, we get, with $\mathbb{P}$-probability at most ε

$$\left(1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)[R(\psi)-R(\psi^\star)]\ge R_n(\psi)-R_n(\psi^\star)+\frac{\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi)+\log\frac{1}{\varepsilon}}{\delta}.$$

Taking $\delta<n/[w+4(\sigma^2+C^2)]$ implies

$$1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}>0,$$

and with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\psi)-R(\psi^\star)\le\frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}\left(R_n(\psi)-R_n(\psi^\star)+\frac{\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi)+\log\frac{1}{\varepsilon}}{\delta}\right).$$

□

Lemma 2.5.4. *Let Assumption 2.1 and Assumption 2.2 hold. Set $w=8C\max(L,C)$, $\delta\in(0,n/[w+4(\sigma^2+C^2)])$ and $\varepsilon\in(0,1)$. Then with $\mathbb{P}$-probability at least $1-\varepsilon$*

$$\int R_n(\psi)\rho(\mathrm{d}\psi)-R_n(\psi^\star)\le\left[1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right]\left[\int R(\psi)\rho(\mathrm{d}\psi)-R(\psi^\star)\right]$$
$$+\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta}. \tag{2.13}$$

Proof. Set $\psi \in \mathbb{F}$ and $Z_i = (Y_i - \psi(\mathbf{X}_i))^2 - (Y_i - \psi^\star(\mathbf{X}_i))^2$, $i \in \{1,\dots,n\}$. Since $Z_i = -T_i$ where T_i is defined in (2.7), using the same arguments that lead to (2.11), we get that for any $\delta \in (0, n/w)$ and $\varepsilon \in (0,1)$

$$\mathbb{E}\int \exp\left[-\delta[R(\psi)-R(\psi^\star)]\left(1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi)-R_n(\psi^\star)]-\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) \right.$$
$$\left. -\log\frac{1}{\varepsilon}\right]\rho(\mathrm{d}\psi) \le \varepsilon.$$

Using Jensen's inequality, we get

$$\mathbb{E}\exp\left[-\int\left\{\delta[R(\psi)-R(\psi^\star)]\left(1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\delta[R_n(\psi)-R_n(\psi^\star)]-\log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) \right.\right.$$
$$\left.\left. -\log\frac{1}{\varepsilon}\right\}\rho(\mathrm{d}\psi)\right] \le \varepsilon.$$

Since $\exp(\delta x) \ge \mathbb{1}_{\mathbb{R}_+}(x)$, we obtain with $\mathbb{P}$-probability at most ε

$$\left[-\int R(\psi)\rho(\mathrm{d}\psi)+R(\psi^\star)\right]\left(1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\int R_n(\psi)\rho(\mathrm{d}\psi)-R_n(\psi^\star)$$
$$-\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \ge 0.$$

Taking $\delta < n/[w+4(\sigma^2+C^2)]$ yields (2.13). □

Lemma 2.5.5. *Let Assumption 2.1 and Assumption 2.2 hold. Set $w = 8C\max(L,C)$, $\delta \in (0, n/[w+4(\sigma^2+C^2)])$ and $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability at least $1-\varepsilon$*

$$\int R(\psi)\rho(\mathrm{d}\psi)-R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}\left(\int R_n(\psi)\rho(\mathrm{d}\psi)-R_n(\psi^\star)+\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta}\right).$$

Proof. Recall (2.12). By Jensen's inequality,

$$\mathbb{E}\exp\left[\delta\left(\int R(\psi)\rho(\mathrm{d}\psi)-R(\psi^\star)\right)\left[1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right]+\delta\left(R_n(\psi^\star)-\int R_n(\psi)\rho(\mathrm{d}\psi)\right)\right.$$
$$\left. -\mathcal{KL}(\rho,\pi)-\log\frac{1}{\varepsilon}\right] \le \varepsilon.$$

Using $\exp(\delta x) \ge \mathbb{1}_{\mathbb{R}_+}(x)$ yields the expected result. □

Theorem 2.5.1. *Let $\hat\psi$ and $\hat\psi^{\mathrm{A}}$ be the Gibbs estimators defined by (2.2)–(2.3), respectively. Let Assumption 2.1 and Assumption 2.2 hold. Set $w = 8C\max(L,C)$ and $\delta = n\ell/[w+4(\sigma^2+C^2)]$, for $\ell \in (0,1)$, and let $\varepsilon \in (0,1)$. Then with probability at least $1-2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat\psi)-R(\psi^\star) \\ R(\hat\psi^{\mathrm{A}})-R(\psi^\star) \end{array}\right\} \le \mathcal{D} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})}\left\{\int R(\psi)\rho(\mathrm{d}\psi)-R(\psi^\star)+\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{n}\right\}, \tag{2.14}$$

where $\mathcal{D}$ is a constant depending only upon w, σ, C and ℓ.

Proof. Recall that the randomized Gibbs estimator $\hat\psi$ is sampled from $\hat\rho_\delta$. By Lemma 2.5.3, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat\psi)-R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}\left(R_n(\hat\psi)-R_n(\psi^\star)+\frac{\log\frac{\mathrm{d}\hat\rho_\delta}{\mathrm{d}\pi}(\hat\psi)+\log\frac{1}{\varepsilon}}{\delta}\right).$$

Note that

$$\log \frac{\mathrm{d}\hat{\rho}_\delta}{\mathrm{d}\pi}(\hat{\psi}) = \log \frac{\exp[-\delta R_n(\hat{\psi})]}{\int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi)}$$
$$= -\delta R_n(\hat{\psi}) - \log \int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi).$$

Thus, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \left(-R_n(\psi^\star) - \frac{1}{\delta}\log \int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi) \frac{1}{\delta}\log\frac{1}{\varepsilon} \right).$$

By Lemma 2.5.2, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left(\int R_n(\psi)\rho(\mathrm{d}\psi) - R_n(\psi^\star) + \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right).$$

Finally, by Lemma 2.5.4, with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \le \frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) \right.$$
$$\left. + \frac{2}{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

Apply Lemma 2.5.5 with the Gibbs posterior probability defined by (2.6). With $\mathbb{P}$-probability at least $1-\varepsilon$,

$$\int R(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \left(\int R_n(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) - R_n(\psi^\star) + \frac{\mathcal{KL}(\hat{\rho}_\delta,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right).$$

Note that

$$\mathcal{KL}(\hat{\rho}_\delta,\pi) = \int \log \frac{\exp[-\delta R_n(\psi)]}{\int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi)} \hat{\rho}_\delta(\mathrm{d}\psi)$$
$$= -\delta \int R_n(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) - \log\left(\int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi) \right).$$

By Lemma 2.5.2, with $\mathbb{P}$-probability at least $1-\varepsilon$

$$\int R(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R_n(\psi)\rho(\mathrm{d}\psi) - R_n(\psi^\star) \right.$$
$$\left. + \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

By Lemma 2.5.4, with $\mathbb{P}$-probability at least $1-2\varepsilon$

$$\int R(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) \right.$$
$$\left. + \frac{2}{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

As R is a convex function, applying Jensen's inequality gives

$$\int R(\psi)\hat{\rho}_\delta(\mathrm{d}\psi) \geq R(\hat{\psi}^{\mathrm{A}}).$$

Finally, note that

$$\frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} = 1 + \frac{8\ell(\sigma^2+C^2)}{(1-\ell)(w+4\sigma^2+4C^2)}.$$

□

Proof of Theorem 2.2.1. Let $\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$. First, for any $A \in \mathcal{T}$, note that $\rho(A) = \sum_{\mathbf{m}\in\mathcal{M}} \rho_{\mathbf{m}}(A)$ where $\rho_{\mathbf{m}}(\cdot) = \rho(\cdot \cap \Theta_{\mathbf{m}})$, the trace of ρ on $\Theta_{\mathbf{m}}$. By Theorem 2.5.1, with $\mathbb{P}$-probability at least $1-2\varepsilon$

$$R(\hat{\psi}) - R(\psi^\star) \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho_{\mathbf{m}}(\mathrm{d}\psi) - R(\psi^\star) + \frac{\mathcal{KL}(\rho_{\mathbf{m}},\pi) + \log\frac{1}{\varepsilon}}{n} \right\}. \tag{2.15}$$

Note that for any $\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$ and any $\mathbf{m} \in \mathcal{M}$,

$$\begin{aligned}\mathcal{KL}(\rho_{\mathbf{m}},\pi) &= \int \log\left(\frac{\mathrm{d}\rho_{\mathbf{m}}}{\mathrm{d}\pi_{\mathbf{m}}}\right)\mathrm{d}\rho_{\mathbf{m}} + \int \log\left(\frac{\mathrm{d}\pi_{\mathbf{m}}}{\mathrm{d}\pi}\right)\mathrm{d}\rho_{\mathbf{m}} \\ &= \mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + \log(1/\alpha)\sum_{j\in S(\mathbf{m})} m_j + \log\left(\binom{p}{|S(\mathbf{m})|}\right) + \log\left(\frac{1-\left(\frac{\alpha}{1-\alpha}\right)^{p+1}}{1-\frac{\alpha}{1-\alpha}}\right).\end{aligned}$$

Next, using the elementary inequality $\log\binom{n}{k} \leq k\log(ne/k)$ and that $\frac{\alpha}{1-\alpha} < 1$,

$$\mathcal{KL}(\rho_{\mathbf{m}},\pi) \leq \mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + \log(1/\alpha)\sum_{j\in S(\mathbf{m})} m_j + |S(\mathbf{m})|\log\left(\frac{pe}{|S(\mathbf{m})|}\right) + \log\left(\frac{1-\alpha}{1-2\alpha}\right).$$

We restrict the set of all probabilities absolutely continuous with respect to $\pi_{\mathbf{m}}$ to uniform probabilities on the ball $\mathcal{B}^1_{\mathbf{m}}(\mathbf{x},\zeta)$, with $\mathbf{x} \in \mathcal{B}^1_{\mathbf{m}}(0,C)$ and $0 < \zeta \leq C - \|\theta\|_1$. Such a probability is denoted by $\mu_{\mathbf{x},\zeta}$. With $\mathbb{P}$-probability at least $1-2\varepsilon$, it yields that

$$\begin{aligned}R(\hat{\psi}) - R(\psi^\star) \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta},0<\zeta\leq C-\|\theta\|_1} \Bigg\{ &\int R(\psi_{\bar{\theta}})\mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}) - R(\psi^\star) \\ &+ \frac{1}{n}\left[\mathcal{KL}(\mu_{\theta,\zeta},\pi_{\mathbf{m}}) + \log\frac{1}{\varepsilon} + |S(\mathbf{m})|\log\left(\frac{p}{|S(\mathbf{m})|}\right) + \sum_{j\in S(\mathbf{m})} m_j\right]\Bigg\}.\end{aligned}$$

Next, note that

$$\mathcal{KL}(\mu_{\theta,\zeta},\pi_{\mathbf{m}}) = \log\left(\frac{V_{\mathbf{m}}(C)}{V_{\mathbf{m}}(\zeta)}\right) = \log\left(\frac{C}{\zeta}\right)\sum_{j\in S(\mathbf{m})} m_j.$$

Note also that

$$\begin{aligned}\int R(\psi_{\bar{\theta}})\mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}) &= \int \mathbb{E}\left[Y_1 - \psi_{\bar{\theta}}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}) \\ &= \int \mathbb{E}\left[Y_1 - \psi_\theta(\mathbf{X}_1) + \psi_\theta(\mathbf{X}_1) - \psi_{\bar{\theta}}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}),\end{aligned}$$

and

$$\int \mathbb{E}\left[Y_1 - \psi_\theta(\mathbf{X}_1) + \psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$= \int R(\psi_\theta)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) + \int \mathbb{E}\left[\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$+ 2\int \mathbb{E}\{[Y_1 - \psi_\theta(\mathbf{X}_1)][\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\}\mu_{\theta,\zeta}(\mathrm{d}\bar\theta).$$

Since $\bar\theta \in \mathcal{B}^1_{\mathbf{m}}(\theta,\zeta)$,

$$\int \mathbb{E}\left[\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta) = \int \mathbb{E}\left[\sum_{j\in S(\mathbf{m})}\sum_{k=1}^{m_j}(\theta_{jk} - \bar\theta_{jk})\varphi_k(X_{1j})\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$\le \|\theta - \bar\theta\|_1^2 \max_k \|\varphi_k\|_\infty^2$$
$$\le \zeta^2,$$

and by the Fubini-Tonelli theorem,

$$2\int \mathbb{E}\{[Y_1 - \psi_\theta(\mathbf{X}_1)][\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\}\mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$= 2\mathbb{E}\left[[Y_1 - \psi_\theta(\mathbf{X}_1)]\int [\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\mu_{\theta,\zeta}(\mathrm{d}\bar\theta)\right] = 0,$$

since $\int \psi_{\bar\theta}(\mathbf{X}_1)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) = \psi_\theta(\mathbf{X}_1)$. Consequently, as

$$\int R(\psi_\theta)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) = R(\psi_\theta),$$

we get

$$\int R(\psi_{\bar\theta})\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) \le R(\psi_\theta) + \zeta^2.$$

So with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat\psi) - R(\psi^\star) \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta},0<\zeta\le C-\|\theta\|_1} \Bigg\{R(\psi_\theta) + \zeta^2 - R(\psi^\star)$$
$$+ \frac{1}{n}\left[\log(C/\zeta)\sum_{j\in S(\mathbf{m})} m_j + \log\frac{1}{\varepsilon} + |S(\mathbf{m})|\log\left(\frac{p}{|S(\mathbf{m})|}\right) + \sum_{j\in S(\mathbf{m})} m_j\right]\Bigg\}.$$

The function $t \mapsto t^2 + \log(C/t)\sum_{j\in S(\mathbf{m})} m_j/n$ is convex. Its minimum is unique and is reached for $t = [\sum_{j\in S(\mathbf{m})} m_j/(2n)]^{1/2}$. With $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat\psi) - R(\psi^\star) \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \Bigg\{R(\psi_\theta) - R(\psi^\star) + |S(\mathbf{m})|\frac{\log(p/|S(\mathbf{m})|)}{n}$$
$$+ \frac{\log(n)}{n}\sum_{j\in S(\mathbf{m})} m_j + \frac{\log(1/\varepsilon)}{n}\Bigg\},$$

where $\mathcal{D}$ is a constant depending only on w, σ, C, ℓ and α. As the same inequality holds for $\hat\psi^{\mathrm{A}}$, this concludes the proof. □

Proof of Theorem 2.2.2. Recall Theorem 2.2.1. Assumption 2.3 gives

$$R(\psi_\theta)-R(\psi^\star)=\int(\psi_\theta(\mathbf{x})-\psi^\star(\mathbf{x}))^2\mathrm{d}\mathcal{P}(\mathbf{x})\le B\int(\psi_\theta(\mathbf{x})-\psi^\star(\mathbf{x}))^2\mathrm{d}\mathbf{x}.$$

For any $\mathbf{m}\in\mathcal{M}$, define

$$\psi^\star_{\mathbf{m}}=\sum_{j\in S^\star}\sum_{k=1}^{m_j}\theta^\star_{jk}\varphi_k.$$

To proceed, we need to check that the projection of $\theta^\star$ onto model $\mathbf{m}$ lies in $\mathcal{B}^1_{\mathbf{m}}(0,C)$, *i.e.*,

$$\sum_{j\in S^\star}\sum_{k=1}^{m_j}|\theta^\star_{jk}|\le C.$$

Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\sum_{j\in S^\star}\sum_{k=1}^{m_j}|\theta^\star_{jk}|&=\sum_{j\in S^\star}\sum_{k=1}^{m_j}k^{r_j}|\theta^\star_{jk}|k^{-r_j}\\&\le\sum_{j\in S^\star}\sqrt{\sum_{k=1}^{m_j}k^{2r_j}(\theta^\star_{jk})^2}\sqrt{\sum_{k=1}^{m_j}k^{-2r_j}}.\end{aligned}$$

Since for any $t\ge1$, $\sum_{k=1}^{m_j}k^{-2t}\le\sum_{k=1}^{\infty}k^{-2t}=\pi^2/6$, the previous inequality yields

$$\sum_{j\in S^\star}\sum_{k=1}^{m_j}|\theta^\star_{jk}|\le\frac{\pi}{\sqrt{6}}\sum_{j\in S^\star}\sqrt{d_j}\le C.$$

Recalling (2.8) and Assumption 2.3, for a $\mathbf{m}\in\mathcal{M}$ we may now write that

$$\begin{aligned}\inf_{\theta\in\Theta_{\mathbf{m}}}R(\psi_\theta)-R(\psi^\star)&\le R(\psi^\star_{\mathbf{m}})-R(\psi^\star)\\&\le B\int(\psi^\star(\mathbf{x})-\psi^\star_{\mathbf{m}}(\mathbf{x}))^2\mathrm{d}\mathbf{x}\\&=B\int\left(\sum_{j\in S^\star}\sum_{k=1+m_j}^{\infty}\theta^\star_{jk}\varphi_k(\mathbf{x})\right)^2\mathrm{d}\mathbf{x}.\end{aligned}$$

As $\{\varphi_k\}_{k=1}^{\infty}$ forms an orthogonal basis,

$$B\int\left(\sum_{j\in S^\star}\sum_{k=1+m_j}^{\infty}\theta^\star_{jk}\varphi_k(\mathbf{x})\right)^2\mathrm{d}\mathbf{x}=B\sum_{j\in S^\star}\sum_{k=1+m_j}^{\infty}(\theta^\star_{jk})^2\le B\sum_{j\in S^\star}d_j(1+m_j)^{-2r_j},$$

where the normalizing numerical factors are included in the now generic constant B. As a consequence, with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat\psi)-R(\psi^\star)\le\mathcal{D}\inf_{\mathbf{m}\in\mathcal{M}}\left\{B\sum_{j\in S^\star}\left\{d_j(1+m_j)^{-2r_j}+\frac{m_j}{n}\log(n)\right\}+|S^\star|\frac{\log(p/|S^\star|)}{n}+\frac{\log(1/\varepsilon)}{n}\right\},$$

where $\mathcal{D}$ is the same constant as in Theorem 2.2.1.

For any $r\ge2$, the function $t\mapsto d_j(1+t)^{-2r_j}+\frac{\log(n)}{n}t$ is convex and admits a minimum in $\left(\frac{\log(n)}{2r_jd_jn}\right)^{-\frac{1}{2r_j+1}}-1$. Accordingly, choosing $m_j\sim\left(\frac{\log(n)}{2r_jd_jn}\right)^{-\frac{1}{2r_j+1}}-1$ yields that with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat\psi)-R(\psi^\star)\le\mathcal{D}\left\{\sum_{j\in S^\star}d_j^{\frac{1}{2r_j+1}}\left(\frac{\log(n)}{2nr_j}\right)^{\frac{2r_j}{2r_j+1}}+|S^\star|\frac{\log\left(\frac{p}{|S^\star|}\right)}{n}+\frac{\log(1/\varepsilon)}{n}\right\},$$

where $\mathcal{D}$ is a constant depending only on α, w, σ, C, ℓ and B, and that ends the proof. □

Proof of Theorem 2.2.3. The proof is similar to the proof of Theorem 2.2.1. From (2.15) and for any $\rho \in \mathcal{M}^1_{+,\pi}(\Theta, \mathcal{T})$ and any $\mathbf{m} \in \mathcal{M}$,

$$\mathcal{KL}(\rho_{\mathbf{m}}, \pi) = \mathcal{KL}(\rho_{\mathbf{m}}, \pi_{\mathbf{m}}) + \log(1/\alpha)|S(\mathbf{m})| + \log\binom{p}{|S(\mathbf{m})|} + \log\left(\frac{1-\left(\alpha\frac{1-\alpha^{K+1}}{1-\alpha}\right)^{p+1}}{1-\alpha\frac{1-\alpha^{K+1}}{1-\alpha}}\right) + \sum_{j\in S(\mathbf{m})} \log\binom{K}{|S(\mathbf{m}_j)|}.$$

Using the elementary inequality $\log\binom{n}{k} \le k\log(ne/k)$ and that $\alpha\frac{1-\alpha^{K+1}}{1-\alpha} \in (0,1)$ since $\alpha < 1/2$,

$$\mathcal{KL}(\rho_{\mathbf{m}}, \pi) \le \mathcal{KL}(\rho_{\mathbf{m}}, \pi_{\mathbf{m}}) + |S(\mathbf{m})|\left[\log(1/\alpha) + \log\left(\frac{pe}{|S(\mathbf{m})|}\right)\right] + \sum_{j\in S(\mathbf{m})} |S(\mathbf{m}_j)|\log\left(\frac{Ke}{|S(\mathbf{m}_j)|}\right) + \log\left(\frac{1-\alpha}{1-2\alpha}\right).$$

Thus with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta}, 0<\zeta\le C-\|\theta\|_1} \Big\{ R(\psi_\theta) + \zeta^2 - R(\psi^\star) + \frac{1}{n}\Big[[\log(C/\zeta) + \log(K)] \sum_{j\in S(\mathbf{m})} |S(\mathbf{m}_j)| + \log\frac{1}{\varepsilon} + |S(\mathbf{m})|\log\left(\frac{p}{|S(\mathbf{m})|}\right)\Big]\Big\}.$$

Hence with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \Big\{ R(\psi_\theta) - R(\psi^\star) + |S(\mathbf{m})|\frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(nK)}{n}\sum_{j\in S(\mathbf{m})} |S(\mathbf{m}_j)| + \frac{\log(1/\varepsilon)}{n}\Big\},$$

where $\mathcal{D}$ is a numerical constant depending upon w, σ, C, ℓ and α. □

Chapitre 3

PAC-Bayesian Estimation and Prediction in Sparse Logistic Models

Abstract. The present chapter extends the results of Chapter 2 to the high-dimensional logistic regression model. A PAC-Bayesian approach is adopted, leading to nonasymptotic oracle inequalities in probability. We also present an implementation relying on a metropolized Carlin & Chib algorithm.

Contents

3.1 Introduction

A tremendous amount of work has focused on the logistic regression model in the past decades. Its flexibility is a valuable asset in many applications (see Cramer, 2003). Whenever the sample size n is larger than the number of covariates d, classical estimation techniques such as Maximum Likelihood exhibit nice theoretical properties and are computationally feasible. However, the adversarial high-dimensional situation $n \ll d$ arises in a widespread spectrum of scientific studies. This big data paradigm is more and more standard in studies that attempt to identify risk factors for disease and clinical outcomes (see for example Wu et al., 2009, in the case of genomics studies). Bayesian techniques have been used in that context (Zhou et al., 2004; Cawley and Talbot, 2006; Genkin et al., 2007), yet they face tough computational issues when confronted with very high values of d (see Gelman et al., 2004, for a survey of the topic). In parallel works, results have been obtained from a classification perspective (Ng and Jordan, 2002; Tsybakov, 2004; Zhang, 2004; Boucheron et al., 2005) (note that classification and logistic regression are close yet distinct problems, as pointed out by Audibert and Tsybakov, 2007), on the quality of the boosting procedure (Friedman et al., 2000; Blanchard et al., 2003; Lugosi and Vayatis, 2004), and on the use of a convex surrogate of the classification loss (Bartlett et al., 2006). Yet all those references do not explicitly address the high-dimensional issue, as they are not worked out

under the sparsity assumption. This typically expresses the belief that among $d \gg n$ covariates, only d_0 are of interest and this dimension reduction paradigm allows for proper estimation in numerous settings. We refer to the monograph Bühlmann and van de Geer (2011) for a thorough presentation of the sparsity approach.

In the linear regression framework, the Lasso estimator introduced by Tibshirani (1996) proved highly effective to deal with the sparsity approach (see Meinshausen and Yu, 2009, among many other references). Many variations such as the Group Lasso (Yuan and Lin, 2006), fused Lasso (Tibshirani et al., 2005; Rinaldo, 2009), or the smooth Lasso (Hebiri and van de Geer, 2010) have proven successful in numerous settings, exhibiting both oracle inequalities and efficient implementation. For the logistic model, the use of ℓ_1-regularized methods has mainly been studied by van de Geer (2008), Meier et al. (2008), Bach (2010) and Kwemou (2012). Here and after, we will refer to the following model as the "classic" logistic model:

$$\text{logit}\, \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \theta^\top \mathbf{x}, \quad \theta, \mathbf{x} \in \mathbb{R}^d, \quad Y \in \{\pm 1\}. \tag{3.1}$$

This is the model assumed in Meier et al. (2008) and Bach (2010), along with a deterministic design. Meier et al. (2008) provides consistency results for the Group Lasso, whereas Bach (2010) derives non asymptotic bounds for the excess risk, using tools from the convex optimisation theory (namely, self-concordant functions). In a setting which is not solely dedicated to logistic regression, van de Geer (2008) proves non asymptotic results for the Lasso in high dimensional generalized linear models. The work presented in Kwemou (2012) is closer to the approach we adopt in this chapter: The author studies the non asymptotic performance of the Group Lasso in the following more general high dimensional logistic regression model:

$$\text{logit}\, \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \eta(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad Y \in \{\pm 1\}. \tag{3.2}$$

To the best of our knowledge, only van de Geer (2008), Bach (2010) and Kwemou (2012) have provided non asymptotic theoretical results in the high dimensional logistic regression context. However, all these methods suffer from technical restrictions which hamper their predictive performance. Indeed, they all assume restrictive conditions on the Gram matrix (mutual coherence, restricted isometry property). These conditions are somewhat natural when one's purpose is to identify relevant covariates, yet we consider them too much of a non realistic burden when it comes to fashioning prediction-oriented strategies. It is to be noted that these conditions may be slighly toned down in some settings (see Bickel et al., 2009; van de Geer and Bühlmann, 2009; Arias-Castro and Lounici, 2012).

In this chapter, we design a strategy which significantly differs from the previously cited works. We rely on a PAC-Bayesian strategy, carrying no assumption whatsoever. The PAC-Bayesian theory originates in the two seminal papers Shawe-Taylor and Williamson (1997) and McAllester (1999) and has been extensively formalized in the context of classification by Catoni (2004, 2007) and regression by Audibert (2004a,b), Alquier (2006, 2008) and Audibert and Catoni (2010, 2011). However, these methods are not explicitly designed to cover the high-dimensional setting under the sparsity assumption. Thus, the PAC-Bayesian theory has been worked out in the sparsity perspective more recently, by Dalalyan and Tsybakov (2008, 2012b), Alquier and Lounici (2011), Dalalyan and Salmon (2012), Suzuki (2012), Alquier and Biau (2013) and Guedj and Alquier (2013). The main message of these studies is that PAC-Bayesian aggregation with a properly chosen prior is able to deal effectively with the sparsity issue in a regression setting under the squared loss. The purpose of this chapter is to extend the use of such techniques to the case of the logistic loss: As far as we are aware, this approach is original. Let us mention however the parallel work of Rigollet (2012), studying Kullback-Leibler aggregation in misspecified generalized linear models.

From a methodological perspective, we also feel that our procedure may represent a nice alternative to Lasso-based algorithms, whenever the Gram matrix is known to blithely violate the restricted eigenvalues condition. We rely on MCMC to compute our estimators, and we adapt the point of view presented in Guedj and Alquier (2013) and implemented in the R package pacbpred (Guedj, 2013b), which is inspired by the seminal work of Carlin and Chib (1995) and its later extensions Hans et al. (2007) and Petralias and Dellaportas (2012). The key idea is to define a neighborhood relationship between visited models, promoting local moves of the Markov chain.

The chapter is organized as follows. In Section 3.2, we present our PAC-Bayesian estimation strategy and introduce our main theoretical results. Section 3.3 is devoted to the implementation of our estimator via MCMC, along with some simulations. Finally, and for the sake of clarity, proofs are gathered in Section 3.4.

3.2 PAC-Bayesian Logistic Regression

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Consider a random variable $(\mathbf{X}, Y) \in \mathcal{X} \times \{\pm 1\}$ where $\mathcal{X} \subseteq \mathbb{R}^d$, whose distribution is denoted by $\mathcal{P}$. Our goal is to infer the posterior probabilities

$$\mathbb{P}(Y = \pm 1 | \mathbf{X} = \mathbf{x}), \quad \mathbf{x} \in \mathcal{X}.$$

Let

$$p : \mathbf{x} \mapsto \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}).$$

To accomplish this task, we rely on a n-sample of i.i.d. replications of $(\mathbf{X}, Y)$, denoted by $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. We assume that $(\mathbf{X}, Y)$ is drawn according to the logistic model, *i.e.*, for all $\mathbf{x} \in \mathcal{X}$,

$$\text{logit } p(\mathbf{x}) = \eta(\mathbf{x}),$$

where

$$\text{logit } p(\mathbf{x}) = \log \frac{p(\mathbf{x})}{1 - p(\mathbf{x})},$$

and we assume further that the link function η has the additive form

$$\eta(\mathbf{x}) = \sum_{j=1}^{d} \eta_j(x_j).$$

Consequently, our goal is to estimate the unknown link function η. To this aim, we benefit from a known dictionary $\mathbb{D} = \{\phi_1, \ldots, \phi_M\}$ and each of the $\phi_1, \ldots, \phi_M$ is an $\mathbb{R} \to \mathbb{R}$ function. For some parameter space $\Theta \subseteq \mathbb{R}^{dM}$, we consider the functional space

$$\mathcal{F}_\Theta = \left\{ f_\theta = \sum_{j=1}^{d} \sum_{k=1}^{M} \theta_{jk} \phi_k, \quad \theta = (\theta_{11}, \ldots, \theta_{1M}, \theta_{21}, \ldots, \theta_{dM}) \in \mathbb{R}^{dM} \right\},$$

so the problem becomes to construct some $\hat{\theta} \in \Theta$ based on the sample $\mathcal{D}_n$ such that $f_{\hat{\theta}}$ is "close" in some sense of η.

To this aim, we define the logistic loss of an estimator f_θ as

$$\ell(Y, f_\theta(\mathbf{x})) = \log[1 + \exp(-Y f_\theta(\mathbf{x}))], \quad \mathbf{x} \in \mathcal{X},$$

and the risk function is

$$R(f_\theta) = \mathbb{E}\ell(Y, f_\theta(\mathbf{X})),$$

with its empirical counterpart

$$R_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(\mathbf{X}_i)).$$

Our procedure goes as follows. Consider a prior measure π on Θ embedded with its canonical Borel field (denoted by $\mathcal{B}(\Theta)$ in the sequel). We will discuss the choice of π further in the chapter. Fix some inverse temperature parameter β, and define the Gibbs posterior distribution as

$$\hat{\rho}_\beta(\mathrm{d}\theta) = \frac{\exp(-\beta n R_n(f_\theta))\pi(\mathrm{d}\theta)}{\int_\Theta \exp(-\beta n R_n(f_{\theta'}))\pi(\mathrm{d}\theta')}.$$

In this work, we state oracle results for the following PAC-Bayesian aggregated estimator (which is the posterior mean with respect to the Gibbs posterior distribution):

$$\hat{\theta}^{\mathrm{a}} = \int_\Theta \theta \hat{\rho}_\beta(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\beta}\theta. \tag{3.3}$$

The resulting estimator for η is thus

$$f_{\hat{\theta}^{\mathrm{a}}} = \sum_{j=1}^{d}\sum_{k=1}^{M} \hat{\theta}^{\mathrm{a}}_{jk}\phi_k,$$

which yields the estimate

$$\hat{p}(\mathbf{x}) = \frac{1}{1+\exp(-f_{\hat{\theta}^{\mathrm{a}}}(\mathbf{x}))}, \quad \mathbf{x} \in \mathcal{X}.$$

It should be noted that we obtain analogous oracle inequalities to the ones which follow, for the randomized estimator

$$\hat{\theta} \sim \hat{\rho}_\beta, \tag{3.4}$$

with similar proofs (see Chapter 2). However we focus on the aggregated estimator in this chapter since it offers much more numerical stability. Finally, we introduce the two following classifiers:

$$\mathcal{C}^1_n(\mathbf{x}) = \begin{cases} 1 & \text{with probability } \hat{p}(\mathbf{x}), \\ -1 & \text{with probability } 1-\hat{p}(\mathbf{x}), \end{cases} \quad \mathbf{x} \in \mathbb{R}^d, \tag{3.5}$$

the randomized classifier, and

$$\mathcal{C}^2_n : \mathbf{x} \mapsto \mathbb{1}_{\{\hat{p}(\mathbf{x})>1/2\}} - \mathbb{1}_{\{\hat{p}(\mathbf{x})\le 1/2\}}, \tag{3.6}$$

the deterministic classifier.

In the following, we assume that there exists a constant $C < \infty$ such that for any $\theta \in \Theta$, $|f_\theta|_\infty \le C$ with $|\cdot|_\infty$ denoting the supremum norm.

Theorem 3.2.1. *For any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\left[R(f_{\hat{\theta}^a}) \le \frac{c_2(\beta)}{c_1(\beta)} \inf_{\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{B}(\Theta))} \left\{\int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) + \frac{2\mathcal{KL}(\rho,\pi)}{\beta n c_2(\beta)} + \frac{2\log(2/\varepsilon)}{\beta n c_2(\beta)}\right\}\right] \ge 1-\varepsilon,$$

where $c_1(\beta)$ and $c_2(\beta)$ are constants depending only on β and C.

This preliminary result is conclusive: It is similar to the PAC-Bayesian inequalities in expectation presented in Dalalyan and Tsybakov (2008) or in probability in Alquier and Lounici (2011) and Guedj and Alquier (2013), however the risk is computed with respect to the logistic loss function. The main message is that with high probability, the risk of the estimator $f_{\hat{\theta}^{\mathrm{a}}}$ is comparable to the best integrated risk on the class $\mathcal{M}^1_{+,\pi}(\Theta, \mathcal{B}(\Theta))$, up to the term $\mathcal{KL}(\rho,\pi)/n$ which measures the complexity of the class. Finally, it is to be noted that this result relies on the sole assumption that the estimators we consider are bounded in the sense of the supremum norm. This mild condition is common in the PAC-Bayesian literature, and allows to use Bernstein-like concentration inequalities.

Next, let us cast Theorem 3.2.1 onto the sparsity perspective. We define the prior π as

$$\pi(\mathrm{d}\theta) = \sum_{m\in\mathcal{M}} \frac{1-\alpha^{d+1}}{1-\alpha} \binom{d}{|m|_0}^{-1} \alpha^{|m|_0} \frac{\mathbb{1}_{\mathcal{B}_{M|m|_0}(r)}(\theta)}{\mathcal{V}(r)}, \tag{3.7}$$

where $\alpha \in (0,1)$, $\mathcal{B}_{M|m|_0}(r)$ is the ℓ_2 ball of radius r in $\mathbb{R}^{M|m|_0}$ and $\mathcal{V}(r)$ is its volume, and $\mathcal{M}$ is the set of models, *i.e.*, $\mathcal{M} = \{0,1\}^d$. The prior π favors sparse vectors θ, as it exponentially penalizes how many regressors are kept. Once the regressors are chosen, we only draw from a uniform distribution over the ℓ_2 ball of radius r. We now make the following mild assumption on the dictionary $\mathbb{D}$.

Assumption 3.1. *There exists a constant $C' < \infty$ such that*

$$\sup_{\mathbf{x}\in\mathcal{X}} |\mathbb{D}(\mathbf{x})|_2 = C',$$

with the slight abuse of notation

$$\mathbb{D}(\mathbf{x}) = \big(\phi_1(x_1), \phi_2(x_1), \ldots, \phi_{M-1}(x_d), \phi_M(x_d)\big) \in \mathbb{R}^{dM}.$$

In particular, this condition is met if the functions $\phi_1, \ldots, \phi_M$ are bounded, consider for example the trigonometric system. This is the case in many studies of the aggregation literature.

We may now derive the main theorem of this chapter.

Theorem 3.2.2. *Under Assumption 3.1, for any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\Bigg[R(f_{\hat{\theta}^a}) \le \frac{c_2(\beta)}{c_1(\beta)} \inf_{m\in\mathcal{M}} \inf_{\theta\in\mathcal{B}_{M|m|_0}(r)} \Bigg\{R(f_\theta) + \frac{2}{\beta n c_2(\beta)} \Bigg[M|m|_0 \left(1 + \log\left(\frac{\beta n r c_2(\beta) C'}{2M|m|_0}\right)\right) + |m|_0 \log\left(\frac{de}{|m|_0}\right) + |m|_0 \log(1/\alpha) + \log\left(\frac{1}{1-\alpha}\right) + \log(2/\varepsilon)\Bigg]\Bigg\}\Bigg] \ge 1-\varepsilon.$$

This result is remarkable as it exhibits the optimal rates of aggregation under the squared loss, in the logistic regression context. If the model is well-specified, that is η may be well approximated by a sparse estimator f_θ (*i.e.*, $R(f_\theta)$ is small), then our PAC-Bayesian aggregate $f_{\hat{\theta}^{\mathrm{a}}}$ also has a small risk.

3.3 Implementation

In this section, we present our implementation strategy, which is an adaptation of Algorithm 2.1 in Chapter 2. It is inspired by the seminal paper Carlin and Chib (1995) and Hans et al. (2007). This "Metropolized" Carlin and Chib algorithm has been worked out

recently by Petralias (2010) and Petralias and Dellaportas (2012). The key idea is to favor local moves of the Markov Chain by defining a neighborhood relationship between models, via three elementary steps: Addition, deletion, or resampling of a regressor. Let $m = (m_1, \dots, m_d)$ be a model, where $m_j \in \{0,1\}^M$.

- $\mathcal{V}^+$ is the set of models which have all the regressors of m plus one.
- $\mathcal{V}^-$ is the set of models which have all the regressors of m but one.
- In the case of a resampling move, the neighborhood is limited to m.

When adding a regressor, the maximal development on the dictionary $\mathbb{D}$ is considered. This is an improvement over Algorithm 2.1 as it fastens convergence (at each MCMC step, fewer models are evaluated) and improves flexibility (for example, if $\mathbb{D}$ is the trigonometric system, it is unlikely that only the two or three first functions will be enough). Having chosen a move with a certain probability (typically, one whould choose numbers of the order of 1/4 for addition and deletion moves, and 1/2 for a resampling move), a candidate vector is drawn from a Gaussian distribution, centered in the maximum likelihood estimator in the current model m and with covariance matrix $\sigma^2 \times \mathrm{Id}$. Maximum likelihood estimators (MLE) are efficiently computed in most statistical programs for models with reasonable dimension, which is the case here since the prior (3.7) prevents visiting models containing more covariates than observations. Further, σ^2 is the variance of the proposal distribution and is chosen by the user: The MCMC literature abounds with heuristic rules. One may try different values for σ^2 and check acceptance rates, and whether the empirical risk has stabilized.

One should try to find a balance between the two parameters α and β. The sparsity-inducing parameter α is to be tuned according to the level of sparsity one expects. The choice $\alpha = 1$ bears no sparsity and large models will be visited by the chain, so the Metropolis-Hastings acceptance ratio will only depend on how the candidate model fits the data. In the same spirit, no matter what the value for α is, a very large value for β will discard most candidate models as soon as the risk of the associated MLE is stricty larger than the risk of the current model of the chain. On the contrary, if β is chosen very close to 0, the chain will ignore how the candidate models fit the data, and is reduced to a random walk on the models space. In a nutshell, we leave the choice of α to the user since its calibration is subject to some sparsity belief; however, we strongly advise to launch several chains with different values for β, and check wether the chain has converged or not. A good starting value is $\beta \propto 1/4\sigma^2$, as exhibited for example in Dalalyan and Tsybakov (2012b) and Alquier and Biau (2013).

Finally, note that the bounding constant r plays a minor role in the implementation, since its influence is very limited in the acceptation ratio: One may take rather large values when testing the method. The full algorithm is presented in Algorithm 3.1.

As a first step to validate our method from an empirical perspective, we tested our method in the three following synthetic models (let us remind here that $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+\exp(-\eta(\mathbf{x}))}$):

Model 3.1. $n = 1000$, $d = 50$, $\alpha = 0.1$, $\sigma^2 = 10^{-2}$, $\beta = 0.25$, $\eta(\mathbf{X}) = X_1^2 + X_2 + X_3 - X_4^3 - 5X_5$.

Model 3.2. $n = 500$, $d = 500$, $\alpha = 10^{-10}$, $\sigma^2 = 10^{-2}$, $\beta = 250$, $\eta(\mathbf{X}) = X_1^2 + X_2 + X_3 - X_4^3 - 5X_5$.

Model 3.3. $n = 300$, $d = 1000$, $\alpha = 10^{-20}$, $\sigma^2 = 10^{-2}$, $\beta = 250$, $\eta(\mathbf{X}) = X_1^2 + X_2 + X_3 - X_4^3 - 5X_5$.

TABLE 3.1 – Quadratic estimation error $\frac{1}{200}\sum_{i=1}^{200}(p(\mathbf{X}_i)-\hat{p}(\mathbf{X}_i))^2$ on a testing sample $(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_{200},Y_{200})$, over 100 replications.

	Model 3.1	Model 3.2	Model 3.3
Mean	0.013	0.151	0.133
S.d.	0.032	0.004	0.018

TABLE 3.2 – Missclassification rates $\frac{1}{200}\sum_{i=1}^{200}\mathbb{1}_{\{Y_i\neq\mathcal{C}_n^j(\mathbf{X}_i)\}}$, $j=1,2$, on a testing sample $(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_{200},Y_{200})$, over 100 replications.

		Model 3.1	Model 3.2	Model 3.3
$\mathcal{C}_n^1$	Mean	0.200	0.224	0.211
	S.d.	0.064	0.005	0.016
$\mathcal{C}_n^2$	Mean	0.144	0.224	0.212
	S.d.	0.073	0.004	0.014

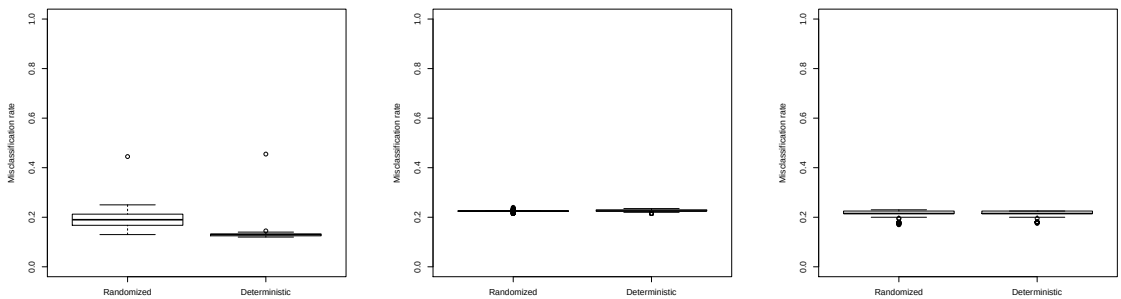
In our simulations, we let $M=7$ and as a dictionary, we used the Legendre polynomials, defined as follows:

$$\phi_1: x\mapsto x,\qquad \phi_2: x\mapsto \frac{1}{2}(3x^2-1),\qquad \phi_3: x\mapsto \frac{1}{2}(5x^3-3x),\qquad \phi_4: x\mapsto \frac{1}{8}(35x^4-30x^2+3),$$
$$\phi_5: x\mapsto \frac{1}{8}(63x^5-70x^3+15x),\qquad \phi_6: x\mapsto \frac{1}{16}(231x^6-315x^4+105x^2-5),$$
$$\phi_7: x\mapsto \frac{1}{16}(429x^7-693x^5+315x^3-35x).$$

In Table 3.1, we report the quadratic error which is committed when substituting $\hat{p}$ to p, over a hundred independent replications. What is remarkable is that we achieve a fair reconstruction of the link function. We also show, in Table 3.2 and Figure 3.1, that both classifiers (defined in (3.5) and (3.6)) exhibit very satisfactory misclassification rates, pulling up our method to high standards.

FIGURE 3.1 – Boxplots of misclassification rates, for both classifiers (randomized $\mathcal{C}_n^1$ and deterministic $\mathcal{C}_n^2$) over 100 replications.

(A) Model 3.1. (B) Model 3.2. (C) Model 3.3.



Algorithm 3.1 A Subspace Carlin and Chib-based algorithm

Input: $\hat{\rho}_\beta$, $T_{\min}$, $T_{\max}$, r and σ^2.
Output: $\hat{\theta}^{\mathrm{a}}$.
Initialize $\Upsilon^{(t)} = \mathbf{0} \in \mathbb{R}^{dM}$ for $t = 0, \dots, T_{\max}$.
for $t = 1$ to $T_{\max}$ **do**
  Denote the current model by $m = \operatorname{supp}\left(\Upsilon^{(t-1)}\right)$.
  Pick a movement $i \in \{+, -, =\}$ with probability $q(i)$.
  if $i \in \{+, -\}$ **then**
    Build up the set of candidate models $\mathcal{V}^i$, and for each $k \in \mathcal{V}^i$, draw a parameter vector θ_k from $\mathcal{N}_{|k|_0 M}\left(\mathrm{MLE}, \sigma^2 \mathrm{Id}\right)$, where MLE denotes the maximum likelihood estimator in that model, and Id is the identity matrix.
    Select candidate θ_k with probability

$$\frac{\hat{\rho}_\beta(\theta_k)/\phi_k(\theta_k)}{\sum_{j \in \mathcal{V}^i} \hat{\rho}_\beta(\theta_j)/\phi_j(\theta_j)},$$

    where ϕ_k denotes the density of the distribution $\mathcal{N}_{|k|_0 M}\left(\mathrm{MLE}, \sigma^2 \mathrm{Id}\right)$.
    Set

$$\Upsilon^{(t)}_{|m} = \begin{cases} \theta_k, & \text{with probability } \alpha, \\ \Upsilon^{(t-1)}_{|m}, & \text{with probability } 1-\alpha, \end{cases}$$

    where

$$\alpha = \min\left(1, \frac{\hat{\rho}_\beta(\theta_k)\phi_m\left(\Upsilon^{(t-1)}_{|m}\right)}{\hat{\rho}_\beta\left(\Upsilon^{(t-1)}_{|m}\right)\phi_k(\theta_k)}\right),$$

    and $\Upsilon^{(t)}_{|m}$ is the restriction of $\Upsilon^{(t)}$ to its entries corresponding to model m.
  else
    Draw a parameter vector θ from $\mathcal{N}_{|m|_0 M}\left(\Upsilon^{(t-1)}_{|m}, \sigma^2 \mathrm{Id}\right)$.
    Set

$$\Upsilon^{(t)}_{|m} = \begin{cases} \theta, & \text{with probability } \alpha, \\ \Upsilon^{(t-1)}_{|m}, & \text{with probability } 1-\alpha, \end{cases}$$

    where

$$\alpha = \min\left(1, \frac{\hat{\rho}_\beta(\theta)\varphi\left(\Upsilon^{(t-1)}_{|m}\right)}{\hat{\rho}_\beta\left(\Upsilon^{(t-1)}_{|m}\right)\varphi(\theta_m)}\right),$$

    where φ denotes the density of the distribution $\mathcal{N}_{|m|_0 M}\left(\Upsilon^{(t-1)}_{|m}, \sigma^2 \mathrm{Id}\right)$.
  end if
end for
Compute

$$\hat{\theta}^{\mathrm{a}} = \frac{1}{T_{\max} - T_{\min}} \sum_{t=T_{\min}+1}^{T_{\max}} \Upsilon^{(t)}.$$

3.4 Proofs

The scheme of proofs goes as follows. We introduce the two following technical lemmas: Lemma 3.4.1 (Legendre transform of the Kullback-Leibler divergence) may be found in Catoni (2004, Equation 5.2.1). Lemma 3.4.2 is a version of Bernstein's inequality, which

is adapted from Massart (2007, Proposition 2.9) to the logistic context.

Lemma 3.4.1. *Let $(A,\mathcal{A})$ be a measurable space. For any probability μ on $(A,\mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int(\exp\circ h)\mathrm{d}\mu < \infty$,*

$$\log\int(\exp\circ h)\mathrm{d}\mu = \sup_{m\in\mathcal{M}^1_{+,\pi}(A,\mathcal{A})}\left\{\int h\mathrm{d}m - \mathcal{KL}(m,\mu)\right\},$$

with the convention $\infty-\infty=-\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int(\exp\circ h)\mathrm{d}\mu}, \quad a\in A.$$

Lemma 3.4.2. *Denote by ϕ the function $\phi : u \mapsto \exp(u)-u-1$. Let $\{T_i\}_{i=1}^n$ be a collection of independent real-valued variables, and let us assume that there exist two positive deterministic constants v and w such that*

$$\sum_{i=1}^n \mathbb{E}[T_i^2] \le v,$$

and for any integer $k\ge 3$,

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \le vw^{k-2}.$$

Then for any $\gamma>0$,

$$\mathbb{E}\left[\exp\left(\gamma\sum_{i=1}^n(T_i-\mathbb{E}T_i)\right)\right] \le \exp\left(\frac{v\phi(\gamma w)}{w^2}\right).$$

Proof. Note that for all $u\le 0$, $\phi(u)\le\frac{u^2}{2}$. Hence, for any $\gamma>0$ and any $i\in\{1,\dots,n\}$,

$$\phi(\gamma T_i) \le \frac{\gamma^2 T_i^2}{2} + \sum_{k=3}^\infty \frac{\gamma^k (T_i)_+^k}{k!},$$

which yields

$$\sum_{i=1}^n \mathbb{E}\phi(\gamma T_i) \le \frac{\gamma^2}{2}\sum_{i=1}^n \mathbb{E}T_i^2 + \sum_{k=3}^\infty \frac{\gamma^k}{k!}\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \le v\sum_{k=2}^\infty \frac{\gamma^k w^{k-2}}{k!} = \frac{v}{w^2}\phi(\gamma w).$$

Using the elementary inequality $\log u \le u-1$ when $u>0$, we obtain

$$\begin{aligned}\sum_{k=1}^n \mathbb{E}\phi(\gamma T_i) = \sum_{k=1}^n \left[\mathbb{E}\exp(\gamma T_i) - \gamma\mathbb{E}T_i - 1\right] &\ge \sum_{k=1}^n \left[\log\mathbb{E}\exp(\gamma T_i) - \gamma\mathbb{E}T_i\right] \\ &= \sum_{k=1}^n \log\mathbb{E}\exp\left[\gamma(T_i-\mathbb{E}T_i)\right].\end{aligned}$$

Putting all the pieces together, we finally get

$$\sum_{k=1}^n \log\mathbb{E}\exp\left[\gamma(T_i-\mathbb{E}T_i)\right] \le \frac{v\phi(\gamma w)}{w^2}.$$

As $\{T_i\}_{i=1}^n$ is a collection of independent variables, taking exponential on both sides of the latter inequality yields the expected result. □

Proof of Theorem 3.2.1. For some $\theta \in \Theta$, set $T_i(\theta) = \ell(Y_i, f_\theta(X_i))$ for all $i = 1, \dots, n$. First, note that since the variables $T_i(\theta)$ are i.i.d., we have

$$\sum_{i=1}^n \mathbb{E}T_i^2(\theta) = \sum_{i=1}^n \mathbb{E}\ell^2(Y_i, f_\theta(\mathbf{X}_i)) \le \sum_{i=1}^n \log\left(1+e^C\right)\mathbb{E}\ell(Y_i, f_\theta(\mathbf{X}_i)) = n\log\left(1+e^C\right)R(f_\theta) \stackrel{\text{def}}{=} v(\theta),$$

and

$$\sum_{i=1}^n \mathbb{E}\left[(T_i^k(\theta))_+\right] \le \sum_{i=1}^n \mathbb{E}\left[(T_i^{k-2}(\theta))_+ T_i^2(\theta)\right] \le \log^{k-2}\left(1+e^C\right)v(\theta),$$

so we set $w \stackrel{\text{def}}{=} \log^{k-2}\left(1+e^C\right)$. Applying Lemma 3.4.2, we get for any $\beta > 0$,

$$\mathbb{E}\exp\left[\beta n(R_n(f_\theta) - R(f_\theta))\right] \le \exp\left[\frac{n\log\left(1+e^C\right)R(f_\theta)\phi(\beta w)}{w^2}\right],$$

i.e., for any $\varepsilon \in (0,1)$,

$$\mathbb{E}\exp\left[\beta n(R_n(f_\theta) - R(f_\theta)) - \frac{n\phi(\beta w)}{w}R(f_\theta) - \log\frac{1}{\varepsilon}\right] \le \varepsilon.$$

For any prior π probability, we may write

$$\int_\Theta \mathbb{E}\exp\left[\beta n(R_n(f_\theta) - R(f_\theta)) - \frac{n\phi(\beta w)}{w}R(f_\theta) - \log\frac{1}{\varepsilon}\right]\pi(\mathrm{d}\theta) \le \varepsilon,$$

and for any posterior $\rho \in \mathcal{M}^1_{+,\pi}(\Theta, \mathcal{B}(\Theta))$,

$$\int_\Theta \mathbb{E}\exp\left[\beta n(R_n(f_\theta) - R(f_\theta)) - \frac{n\phi(\beta w)}{w}R(f_\theta) - \log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \log\frac{1}{\varepsilon}\right]\rho(\mathrm{d}\theta) \le \varepsilon.$$

Using the Fubini-Tonelli theorem, we obtain

$$\mathbb{E}\left[\int_\Theta \exp\left[\beta n(R_n(f_\theta) - R(f_\theta)) - \frac{n\phi(\beta w)}{w}R(f_\theta) - \log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \log\frac{1}{\varepsilon}\right]\rho(\mathrm{d}\theta)\right] \le \varepsilon,$$

and by Jensen, the previous inequality yields

$$\mathbb{E}\exp\left[\int_\Theta\left(\beta n(R_n(f_\theta) - R(f_\theta)) - \frac{n\phi(\beta w)}{w}R(f_\theta)\right)\rho(\mathrm{d}\theta) - \mathcal{KL}(\rho,\pi) - \log\frac{1}{\varepsilon}\right] \le \varepsilon.$$

Next, we use the elementary inequality $\exp(\beta x) \ge \mathbb{1}_{(0,\infty)}(x)$, hence

$$\mathbb{P}\left[\int_\Theta\left(R_n(f_\theta) - R(f_\theta) - \frac{\phi(\beta w)}{\beta w}R(f_\theta)\right)\rho(\mathrm{d}\theta) - \frac{\mathcal{KL}(\rho,\pi)}{\beta n} - \frac{\log(1/\varepsilon)}{\beta n} \ge 0\right] \le \varepsilon,$$

i.e.,

$$\mathbb{P}\left[\int_\Theta R_n(f_\theta)\rho(\mathrm{d}\theta) \le \left(1+\frac{\phi(\beta w)}{\beta w}\right)\int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{KL}(\rho,\pi)}{\beta n} + \frac{\log(1/\varepsilon)}{\beta n}\right] \ge 1-\varepsilon. \tag{3.8}$$

We can now proceed analogously with the variables $Z_i(\theta) = -T_i(\theta)$, and we obtain

$$\mathbb{P}\left[\int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) \le \frac{1}{1-\phi(\beta w)/\beta w}\left(\int_\Theta R_n(f_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{KL}(\rho,\pi)}{\beta n} + \frac{\log(1/\varepsilon)}{\beta n}\right)\right] \ge 1-\varepsilon, \tag{3.9}$$

whenever β is such that $\phi(\beta w) \le \beta w$. Since (3.9) holds for any posterior ρ absolutely continuous with respect to π, in particular, for the choice $\rho = \hat{\rho}_\beta$, we get

$$\mathbb{P}\left[\int_\Theta R(f_\theta)\hat{\rho}_\beta(\mathrm{d}\theta) \le \frac{1}{1-\phi(\beta w)/\beta w}\left(\int_\Theta R_n(f_\theta)\hat{\rho}_\beta(\mathrm{d}\theta) + \frac{\mathcal{KL}(\hat{\rho}_\beta,\pi)}{\beta n} + \frac{\log(1/\varepsilon)}{\beta n}\right)\right] \ge 1-\varepsilon.$$

First, note that by Jensen,

$$\int_\Theta R(f_\theta)\hat{\rho}_\beta(\mathrm{d}\theta) \geq R\left(\int_\Theta f_\theta \hat{\rho}_\beta(\mathrm{d}\theta)\right) = R(f_{\hat{\theta}^{\mathrm{a}}}).$$

Further,

$$\begin{aligned}\mathcal{KL}(\hat{\rho}_\beta,\pi) &= \int_\Theta \log \frac{\exp(-\beta n R_n(f_\theta))}{\int_\Theta \exp\left(-\beta n R_n(f_{\theta'})\right)\pi(\mathrm{d}\theta')}\hat{\rho}_\beta(\mathrm{d}\theta) \\ &= -\beta n \int_\Theta R_n(f_\theta)\hat{\rho}_\beta(\mathrm{d}\theta) - \log \int_\Theta \exp\left(-\beta n R_n(f_{\theta'})\right)\pi(\mathrm{d}\theta'),\end{aligned}$$

which yields

$$\mathbb{P}\left[R(f_{\hat{\theta}^{\mathrm{a}}}) \leq \frac{1}{1-\phi(\beta w)/\beta w}\left(-\frac{1}{\beta n}\log\int_\Theta \exp\left(-\beta n R_n(f_{\theta'})\right)\pi(\mathrm{d}\theta') + \frac{\log(1/\varepsilon)}{\beta n}\right)\right] \geq 1-\varepsilon.$$

We may now use Lemma 3.4.1:

$$\mathbb{P}\left[R(f_{\hat{\theta}^{\mathrm{a}}}) \leq \frac{1}{1-\phi(\beta w)/\beta w}\left(\inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{B}(\Theta))}\left\{\int_\Theta R_n(f_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{KL}(\rho,\pi)}{\beta n}\right\} + \frac{\log(1/\varepsilon)}{\beta n}\right)\right] \geq 1-\varepsilon.$$

Plugging (3.8) in the previous inequality, we obtain

$$\begin{aligned}\mathbb{P}\Bigg[R(f_{\hat{\theta}^{\mathrm{a}}}) \leq \frac{1}{1-\phi(\beta w)/\beta w}\Bigg(\inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{B}(\Theta))}\Bigg\{\left(1+\frac{\phi(\beta w)}{\beta w}\right)\int_\Theta R(f_\theta)\rho(\mathrm{d}\theta) + \frac{2\mathcal{KL}(\rho,\pi)}{\beta n}\Bigg\} \\ + \frac{2\log(1/\varepsilon)}{\beta n}\Bigg)\Bigg] \geq 1-2\varepsilon,\end{aligned}$$

which is the desired result, using the notation

$$c_1(\beta) = 1 - \frac{\phi(\beta w)}{\beta w}, \quad \text{and} \quad c_2(\beta) = 1 + \frac{\phi(\beta w)}{\beta w}.$$

□

Proof of Theorem 3.2.2. First, note that under Assumption 3.1 and from (3.7), we get that $C = rC'$ by the Cauchy-Schwarz inequality.

Now, observe that Theorem 3.2.1 yields

$$\mathbb{P}\left[R(f_{\hat{\theta}^{\mathrm{a}}}) \leq \frac{c_2(\beta)}{c_1(\beta)}\inf_{m\in\mathcal{M}}\inf_{\rho_m}\left\{\int_\Theta R(f_\theta)\rho_m(\mathrm{d}\theta) + \frac{2\mathcal{KL}(\rho_m,\pi)}{\beta n c_2(\beta)} + \frac{2\log(2/\varepsilon)}{\beta n c_2(\beta)}\right\}\right] \geq 1-\varepsilon,$$

where $\inf_{\rho_m}$ is a shortcut writing denoting the infimum taken on any distribution on $\mathbb{R}^{M|m|_0}$ absolutely continuous with respect to π. Next, adopting the notation $\pi = \sum_{m\in\mathcal{M}}\pi_m$, we proceed with

$$\begin{aligned}\mathcal{KL}(\rho_m,\pi) &= \int \log\left(\frac{\mathrm{d}\rho_m}{\mathrm{d}\pi_m}(\theta)\cdot\frac{\mathrm{d}\pi_m}{\mathrm{d}\pi}(\theta)\right)\rho_m(\mathrm{d}\theta) \\ &= \mathcal{KL}(\rho_m,\pi_m) + \log\binom{d}{|m|_0} + |m|_0\log(1/\alpha) + \log\left(\frac{1-\alpha^{d+1}}{1-\alpha}\right) \\ &\leq \mathcal{KL}(\rho_m,\pi_m) + |m|_0\log\left(\frac{de}{|m|_0}\right) + |m|_0\log(1/\alpha) + \log\left(\frac{1}{1-\alpha}\right),\end{aligned}$$

where we used the elementary inequality $\log\binom{n}{k} \le k\log(ne/k)$. Now, we replace $\inf_{\rho_m}$ by the infimum $\inf_{\theta\in\mathcal{B}_{M|m|_0}(r)}$ and we consider the uniform distribution μ_θ on the ball $\theta+\mathcal{B}_{M|m|_0}(t)$, *i.e.*, the ℓ_2 ball of radius t which is centered in θ (with $t \le r-|\theta|_2$). We obtain

$$\mathbb{P}\Bigg[R(f_{\hat\theta^{\mathrm{a}}}) \le \frac{c_2(\beta)}{c_1(\beta)} \inf_{m\in\mathcal{M}} \inf_{\theta\in\mathcal{B}_{M|m|_0}(r)} \inf_{0\le t\le r-|\theta|_2} \Bigg\{\int_\Theta R(f_{\theta'})\mu_\theta(\mathrm{d}\theta') + \frac{2}{\beta n c_2(\beta)}\left(\mathcal{KL}(\mu_\theta,\pi_m) + |m|_0\log\left(\frac{de}{|m|_0}\right) + |m|_0\log(1/\alpha) + \log\left(\frac{1}{1-\alpha}\right) + \log(2/\varepsilon)\right)\Bigg\}\Bigg] \ge 1-\varepsilon.$$

Note that

$$\mathcal{KL}(\mu_\theta,\pi_m) = \log\left(\frac{\mathcal{V}(r)}{\mathcal{V}(t)}\right) = M|m|_0\log\left(\frac{r}{t}\right),$$

and since R is Lipschitz,

$$\begin{aligned}\int_\Theta R(f_{\theta'})\mu_\theta(\mathrm{d}\theta') &= \int_\Theta [R(f_{\theta'}) - R(f_\theta) + R(f_\theta)]\,\mu_\theta(\mathrm{d}\theta') \\ &\le R(f_\theta) + \int_\Theta |R(f_{\theta'}) - R(f_\theta)|\mu_\theta(\mathrm{d}\theta') \\ &\le R(f_\theta) + \int_\Theta |f_{\theta'} - f_\theta|\mu_\theta(\mathrm{d}\theta') \\ &\le R(f_\theta) + tC',\end{aligned}$$

using again the Cauchy-Schwarz inequality. To conclude, observe that

$$t \mapsto tC' + \frac{2M|m|_0}{\beta n c_2(\beta)}\log\left(\frac{r}{t}\right)$$

admits a global minimum in

$$t = \frac{2M|m|_0}{\beta n c_2(\beta)C'}.$$

Therefore, under the condition

$$|\theta|_2 \le r - \frac{2M|m|_0}{\beta n c_2(\beta)C'},$$

we obtain

$$\mathbb{P}\Bigg[R(f_{\hat\theta^{\mathrm{a}}}) \le \frac{c_2(\beta)}{c_1(\beta)} \inf_{m\in\mathcal{M}} \inf_{\theta\in\mathcal{B}_{M|m|_0}(r)} \Bigg\{R(f_\theta) + \frac{2}{\beta n c_2(\beta)}\Bigg[M|m|_0\left(1+\log\left(\frac{\beta n r c_2(\beta)C'}{2M|m|_0}\right)\right) + |m|_0\log\left(\frac{de}{|m|_0}\right) + |m|_0\log(1/\alpha) + \log\left(\frac{1}{1-\alpha}\right) + \log(2/\varepsilon)\Bigg]\Bigg\}\Bigg] \ge 1-\varepsilon,$$

which is the desired result.

□

Chapitre 4

COBRA: A Nonlinear Aggregation Strategy

Abstract. A new method for combining several initial estimators of the regression function is introduced. Instead of building a linear or convex optimized combination over a collection of basic estimators $r_1, \ldots, r_M$, we use them as a collective indicator of the proximity between the training data and a test observation. This local distance approach is model-free and very fast. More specifically, the resulting collective estimator is shown to perform asymptotically at least as well in the L^2 sense as the best basic estimator in the collective. Moreover, it does so without having to declare which might be the best basic estimator for the given data set. A companion R package called `COBRA` (standing for COmBined Regression Alternative) is presented (downloadable on http://cran.r-project.org/web/packages/COBRA/index.html). Substantial numerical evidence is provided on both synthetic and real data sets to assess the excellent performance and velocity of our method in a large variety of prediction problems.

Contents

4.1 Introduction

Recent years have witnessed a growing interest in aggregated statistical procedures, supported by a considerable research and extensive empirical evidence. Indeed, the increasing number of available estimation and prediction methods (hereafter denoted *machines*) in a wide range of modern statistical problems naturally suggests using some efficient strategy

for combining procedures and estimators. If the combined strategy is known to be optimal in some sense and relatively free of assumptions that are hard to evaluate, then such a model-free strategy is a valuable and practical research tool.

In this regard, numerous contributions have enriched the aggregation literature with various approaches, such as model selection (select the optimal single estimator from a list of models), convex aggregation (searching for the optimal convex combination of given estimators, such as exponentially weighted aggregates) and linear aggregation (selecting the optimal linear combination).

Model selection, linear-type aggregation strategies and related problems have been studied by Catoni (2004), Juditsky and Nemirovski (2000), Nemirovski (2000), Yang (2000, 2001, 2004), Györfi et al. (2002), and Wegkamp (2003). Minimax results have been derived by Nemirovski (2000) and Tsybakov (2003), leading to the notion of optimal rates of aggregation. Similar results can be found in Bunea et al. (2007a). Further, upper bounds for the risk in model selection and convex aggregation have been established, for instance by Audibert (2004a), Birgé (2006), and Dalalyan and Tsybakov (2008). An interesting feature is that such aggregation problems may be treated within the scope of L^1-penalized least squares, as performed in Bunea et al. (2006, 2007a,b). This kind of framework is also considered by van de Geer (2008) and Koltchinskii (2009), with the L^2 loss replaced by another convex loss. In the agregation literature, let us also mention the work of Juditsky et al. (2005), Bunea and Nobel (2008), and Baraud et al. (2013). More recently, specific models such as single-index in Alquier and Biau (2013) and additive models in Guedj and Alquier (2013) have been studied in the context of aggregation under a sparsity assumption.

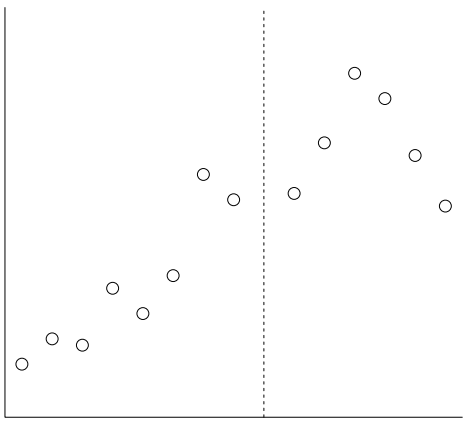
This chapter investigates a distinctly different point of view, motivated by the sense that nonlinear, data-dependent techniques are a source of analytic flexibility and might improve upon current aggregation procedures. Specifically, consider the following example of a classification problem: If the ensemble of machines happens to include a strong one, lurking but unnamed in the collection of which many might be very weak machines, it might make sense to consider a more sophisticated method than the previously cited ones for pooling the information across the machines. Choosing to set aside some of the machines, on some data-dependent criteria, seems only weakly motivated, since the performance of the collective, retaining those suspect machines, might be quite good on a nearby data set. Similarly, searching for some phantom strong machine in the collective could also be ruinous when presented with new and different data.

Instead of choosing either of these options—selecting out weak performers, searching for a hidden, universally strong performer—we propose an original nonlinear method for combining the outcomes over some list of plausibly good procedures. We call this combined scheme a regression collective over the given basic machines. More specifically, we consider the problem of building a new estimator by combining M estimators of the regression function, thereby exploiting an idea proposed in the context of classification by Mojirsheibani (1999). Given a set of preliminary estimators $r_1,\dots,r_M$, the idea behind this aggregation method is a "unanimity" concept, in that it is based on the values predicted by $r_1,\dots,r_M$ for the data and for a new observation $\mathbf{x}$. In a nutshell, a data point is considered to be "close" to $\mathbf{x}$, and consequently, reliable for contributing to the estimation of this new observation, if all estimators predict values which are close to each other for $\mathbf{x}$ and this data item, *i.e.*, not more distant than a prespecified threshold ε. The predicted value corresponding to this query point $\mathbf{x}$ is then set to the average of the responses of the selected observations. More precisely, the average is over the original outcome values of the selected observations, and *not* over the estimates provided by the several machines for these observations.

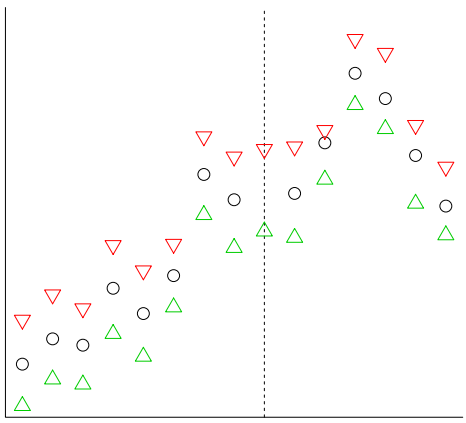
To make the concept clear, consider the following toy example illustrated by Figure 4.1. Assume we are given the observations plotted in circles, and the values predicted by two known machines f_1 and f_2 (triangles pointing up and down, respectively). The goal is to predict the response for the new point $\mathbf{x}$ (along the dotted line). Set a threshold ε, the black solid circles are the data points $(\mathbf{x}_i, y_i)$ within the two dotted intervals, *i.e.*, such that for $m = 1,2$, $|f_m(\mathbf{x}_i) - f_m(\mathbf{x}_0)| \leq \varepsilon$. Averaging the corresponding y_i's yields the prediction for $\mathbf{x}$ (diamond).

FIGURE 4.1 – A toy example: Nonlinear aggregation of two primal estimators.

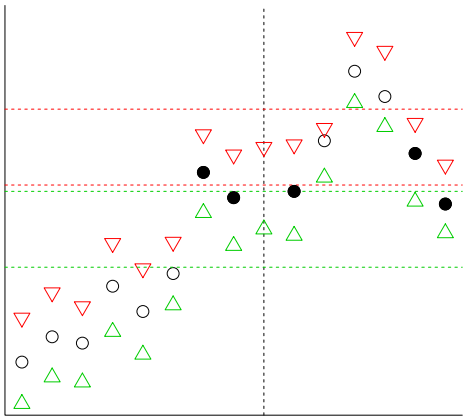
(A) How should we predict the query point's response (dotted line)?



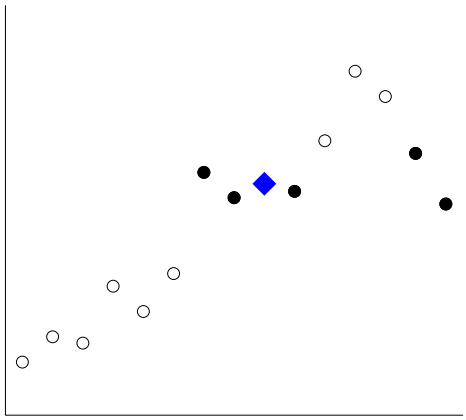
(B) The two primal estimators.



(C) The collective operates.



(D) Predicted value for the query point.



We stress that the central and original idea behind our approach is that the resulting regression predictor is a nonlinear, data-dependent function of the basic predictors $r_1, \ldots, r_M$, but where the predictors are used to determine a local distance between a new, test instance

and the original training data. To the best of our knowledge there exists no formalized procedure in the machine learning and aggregation literature that operates as does ours.

Along with this chapter, we release the software COBRA (Guedj, 2013a) which implements the method as an additional package to the statistical software R (see R Core Team, 2013). COBRA is freely downloadable on the CRAN website [1]. As detailed in Section 4.3, we undertook a lengthy series of numerical experiments, over which COBRA proved extremely successful. These stunning results lead us to believe that regression collectives can provide valuable insights on a wide range of prediction problems. Further, these same results demonstrate that COBRA has remarkable speed in terms of CPU timings. In the context of high-dimensional (such as genomic) data, such velocity is critical, and in fact COBRA can natively take advantage of multi-core parallel environments.

The chapter is organized as follows. In Section 4.2, we describe the combined estimator—the regression collective—and derive a nonasymptotic risk bound. Next we present the main result, that is, the collective is asymptotically at least as good as any of the basic estimators. We also provide a rate of convergence for our procedure that is faster than the usual nonparametric rate. Section 4.3 is devoted to the companion R package COBRA and presents benchmarks of its excellent performance on both simulated and real data sets, including high-dimensional models. We also show that COBRA compares favorably with two competitors, Super Learner (see the seminal paper van der Laan et al., 2007) and exponentially weighted aggregation (among many other references, see Dalalyan and Tsybakov, 2008), in that it performs similarly in most situations, much better in some, while it is consistently faster in every case (for the Super Learner). Finally, for ease of exposition, proofs are collected in Section 4.4.

4.2 The combined estimator

4.2.1 Notation

Throughout the chapter, we assume we are given a training sample denoted by $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. $\mathcal{D}_n$ is composed of i.i.d. random variables taking their values in $\mathbb{R}^d \times \mathbb{R}$, and distributed as an independent prototype pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}Y^2 < \infty$ (with the notation $\mathbf{X} = (X_1, \ldots, X_d)$). The space $\mathbb{R}^d$ is equipped with the standard Euclidean metric. Our goal is to consistently estimate the regression function $r^\star(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, $\mathbf{x} \in \mathbb{R}^d$, using the data $\mathcal{D}_n$.

Firstly, the original data set $\mathcal{D}_n$ is split into two sequences $\mathcal{D}_k = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_k, Y_k)\}$ and $\mathcal{D}_\ell = \{(\mathbf{X}_{k+1}, Y_{k+1}), \ldots, (\mathbf{X}_n, Y_n)\}$, with $\ell = n - k \geq 1$. For ease of notation, the elements of $\mathcal{D}_\ell$ are renamed $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_\ell, Y_\ell)\}$. There is a slight abuse of notation here, as the same letter is used for both subsets $\mathcal{D}_k$ and $\mathcal{D}_\ell$—however, this should not cause any trouble since the context is clear.

Now, suppose that we are given a collection of $M \geq 1$ competing candidates $r_{k,1}, \ldots, r_{k,M}$ to estimate $r^\star$. These basic estimators—basic machines—are assumed to be generated using only the first subsample $\mathcal{D}_k$. These machines can be any among the researcher's favorite toolkit, such as linear regression, kernel smoother, SVM, Lasso, neural networks, naive Bayes, or random forests. They could equally well be any ad hoc regression rules suggested by the experimental context. The essential idea is that these basic machines can be parametric or nonparametric, or indeed semi-parametric, with possible tuning rules. All

1. http://cran.r-project.org/web/packages/COBRA/index.html

that is asked for is that each of the $r_{k,m}(\mathbf{x})$, $m = 1, \ldots, M$, is able to provide an estimation of $r^\star(\mathbf{x})$ on the basis of $\mathcal{D}_k$ alone. Thus, any collection of model-based or model-free machines are allowed, and the collection is here called the regression collective. Let us emphasize here that the number of basic machines M is considered as fixed throughout this chapter. Hence, the number of machines is not expected to grow and is typically of a reasonable size (M is chosen on the order of 10 in Section 4.3).

Given the collection of basic machines $\mathbf{r}_k = (r_{k,1}, \ldots, r_{k,M})$, we define the collective estimator T_n to be

$$T_n(\mathbf{r}_k(\mathbf{x})) = \sum_{i=1}^{\ell} W_{n,i}(\mathbf{x}) Y_i, \quad \mathbf{x} \in \mathbb{R}^d,$$

where the random weights $W_{n,i}(\mathbf{x})$ take the form

$$W_{n,i}(\mathbf{x}) = \frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \le \varepsilon_\ell\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \le \varepsilon_\ell\}}}. \tag{4.1}$$

In this definition, ε_ℓ is some positive parameter and, by convention, $0/0 = 0$.

The weighting scheme used in our regression collective is distinctive but not obvious. Starting from Devroye et al. (1996) and Györfi et al. (2002), we see that T_n is a local averaging estimator in the following sense: The value for $r^\star(\mathbf{x})$, that is, the estimated outcome at the query point $\mathbf{x}$, is the unweighted average over those Y_i's such that $\mathbf{X}_i$ is "close" to the query point. More precisely, for each $\mathbf{X}_i$ in the sample $\mathcal{D}_\ell$, "close" means that the output at the query point, generated from each basic machine, is within an ε_ℓ distance of the output generated by the same basic machines at $\mathbf{X}_i$. If a basic machine evaluated at $\mathbf{X}_i$ is close to the basic machine evaluated at the query point $\mathbf{x}$, then the corresponding outcome Y_i is included in the average, and not otherwise. Also, as a further note of clarification: "Closeness" of the $\mathbf{X}_i$ is not here to be understood in the Euclidean sense. It refers to closeness of the basic machine outputs at the query point with basic machine outputs over all points in the training data. Training points $\mathbf{X}_i$ that are close, in the basic machine sense, to the corresponding basic machine output at the query point contribute to the indicator function for the corresponding outcome Y_i. This alternative approach is motivated by the fact that a major issue in learning problems consists of devising a metric that is suited to the data (see, *e.g.*, the monograph by Pekalska and Duin, 2005).

In this context, ε_ℓ plays the role of a smoothing parameter: Put differently, in order to retain Y_i, all basic estimators $r_{k,1}, \ldots, r_{k,M}$ have to deliver predictions for the query point $\mathbf{x}$ which are in a ε_ℓ-neighborhood of the predictions $r_{k,1}(\mathbf{X}_i), \ldots, r_{k,M}(\mathbf{X}_i)$. Note that the greater ε_ℓ, the more tolerant the process. It turns out that the practical performance of T_n strongly relies on an appropriate choice of ε_ℓ. This important question will thoroughly be discussed in Section 4.3, where we devise an automatic (*i.e.*, data-dependent) selection strategy of ε_ℓ.

Next, we note that the subscript n in T_n may be a little confusing, since T_n is a weighted average of the Y_i's in $\mathcal{D}_\ell$ only. However, T_n depends on the entire data set $\mathcal{D}_n$, as the rest of the data is used to set up the original machines $r_{k,1}, \ldots, r_{k,M}$. Finally, and most importantly, it should be noticed that the combined estimator T_n is nonlinear with respect to the basic estimators $r_{k,m}$'s. This makes it very different from techniques derived from model selection or convex and linear aggregation literature. As such, it is inspired by the preliminary work of Mojirsheibani (1999) in the supervised classification context.

In addition, let us mention that, in the definition of the weights (4.1), all original estimators are invited to have the same, equally valued opinion on the importance of the

observation $\mathbf{X}_i$ (within the range of ε_ℓ) for the corresponding Y_i to be integrated in the combination T_n. However, this unanimity constraint may be relaxed by imposing, for example, that a fixed fraction $\alpha \in \{1/M, 2/M, \ldots, 1\}$ of the machines agree on the importance of $\mathbf{X}_i$. In that case, the weights take the more sophisticated form

$$W_{n,i}(\mathbf{x}) = \frac{\mathbb{1}_{\{\sum_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \le \varepsilon_\ell\}} \ge M\alpha\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\{\sum_{m=1}^{M} \mathbb{1}_{\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \le \varepsilon_\ell\}} \ge M\alpha\}}}.$$

It turns out that adding the parameter α does not change the asymptotic properties of T_n, provided $\alpha \to 1$. Thus, to keep a sufficient degree of clarity in the mathematical statements and subsequent proofs, we have decided to consider only the case $\alpha = 1$ (i.e., unanimity). We leave as an exercise extension of the results to more general values of α. On the other hand, as highligthed by Section 4.3, α has a nonnegligible impact on the performance of the combined estimator. Accordingly, we will discuss in Section 4.3 an automatic procedure to select this extra parameter.

4.2.2 Theoretical performance

This section is devoted to the study of some asymptotic and nonasymptotic properties of the combined estimator T_n, whose quality will be assessed by the quadratic risk

$$\mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})\right|^2.$$

Here and later, $\mathbb{E}$ denotes the expectation with respect to both $\mathbf{X}$ and the sample $\mathcal{D}_n$. Everywhere in the document, it is assumed that $\mathbb{E}|r_{k,m}(\mathbf{X})|^2 < \infty$ for all $m = 1, \ldots, M$. Moreover, we shall need the following technical requirement: For any $m = 1, \ldots, M$,

$$r_{k,m}^{-1}((t, +\infty)) \underset{t \uparrow +\infty}{\searrow} \emptyset \quad \text{and} \quad r_{k,m}^{-1}((-\infty, t)) \underset{t \downarrow -\infty}{\searrow} \emptyset. \tag{4.2}$$

It is stressed that this is a mild assumption which is met, for example, whenever the machines are bounded. Throughout, we let

$$T(\mathbf{r}_k(\mathbf{X})) = \mathbb{E}[Y|\mathbf{r}_k(\mathbf{X})]$$

and note that, by the very definition of the L^2 conditional expectation,

$$\mathbb{E}|T(\mathbf{r}_k(\mathbf{X})) - Y|^2 \le \inf_f \mathbb{E}|f(\mathbf{r}_k(\mathbf{X})) - Y|^2, \tag{4.3}$$

where the infimum is taken over all square integrable functions of $\mathbf{r}_k(\mathbf{X})$.

Our first result is a nonasymptotic inequality, which states that the combined estimator behaves as well as the best one in the original list, within a term measuring how far T_n is from T.

Proposition 4.2.1. *Let $\mathbf{r}_k = (r_{k,1}, \ldots, r_{k,M})$ be the collection of basic estimators, and let $T_n(\mathbf{r}_k(\mathbf{x}))$ be the combined estimator. Then, for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$,*

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \le \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 + \inf_f \mathbb{E}|f(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2,$$

where the infimum is taken over all square integrable functions of $\mathbf{r}_k(\mathbf{X})$. In particular,

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \le \min_{m=1,\ldots,M} \mathbb{E}|r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X})|^2 + \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2.$$

Note that since, for example, Proposition 4.2.1 holds for any square integrable function of $\mathbf{r}_k(\mathbf{X})$, this result allows to derive inequalities linked to any existing aggregation procedure: One may consider linear or convex aggregation as well.

Proposition 4.2.1 reassures us on the performance of T_n with respect to the basic machines, whatever the distribution of $(\mathbf{X},Y)$ is and regardless of which initial estimator is actually the best. The term $\min_{m=1,\dots,M}\mathbb{E}|r_{k,m}(\mathbf{X})-r^\star(\mathbf{X})|^2$ may be regarded as a bias term, whereas the term $\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X}))-T(\mathbf{r}_k(\mathbf{X}))|^2$ is a variance-type term, which can be asymptotically neglected.

Proposition 4.2.2. *Assume that $\varepsilon_\ell \to 0$ and $\ell\varepsilon_\ell^M \to \infty$ as $\ell \to \infty$. Then*

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X}))-T(\mathbf{r}_k(\mathbf{X}))|^2 \to 0 \quad \text{as } \ell \to \infty,$$

for all distributions of $(\mathbf{X},Y)$ with $\mathbb{E}Y^2<\infty$. Thus,

$$\limsup_{\ell\to\infty}\mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X}))-r^\star(\mathbf{X})\right|^2 \le \min_{m=1,\dots,M}\mathbb{E}\left|r_{k,m}(\mathbf{X})-r^\star(\mathbf{X})\right|^2.$$

This result is remarkable, for these two reasons. Firstly, it shows that, in terms of predictive quadratic risk, the combined estimator does asymptotically at least as well as the best primitive machine. Secondly, the result is universal, in the sense that it is true for all distributions of $(\mathbf{X},Y)$, without exceptions.

This is especially interesting because the performance of any estimation procedure eventually depends upon some model and smoothness assumptions on the observations. For example, a linear regression fit performs well if the distribution is truly linear, but may behave poorly otherwise. Similarly, the Lasso procedure is known to do a good job for non-correlated designs (see van de Geer, 2008), with no clear guarantee however in adversarial situations. Likewise, rates of convergence of nonparametric procedures such as the k-nearest neighbor method, kernel estimators and random forests dramatically deteriorate as the ambient dimension increases, but may be significantly improved if the true underlying dimension is reasonable. This phenomenon is thoroughly analyzed for the random forests algorithm in Biau (2012).

The universal result exhibited in Proposition 4.2.2 does not require any regularity assumption on the basic machines. However, this universality comes at a price since we have no guarantee on the rate of convergence of the variance term. Nevertheless, assuming some light additional smoothness conditions, one has the following result.

Theorem 4.2.1. *Assume that Y and the basic machines $\mathbf{r}_k$ are bounded by some constant R. Assume moreover that there exists a constant $L \ge 0$ such that, for every $k \ge 1$,*

$$|T(\mathbf{r}_k(\mathbf{x}))-T(\mathbf{r}_k(\mathbf{y}))| \le L|\mathbf{r}_k(\mathbf{x})-\mathbf{r}_k(\mathbf{y})|, \quad \mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

Then, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$, one has

$$\mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X}))-r^\star(\mathbf{X})\right|^2 \le \min_{m=1,\dots,M}\mathbb{E}\left|r_{k,m}(\mathbf{X})-r^\star(\mathbf{X})\right|^2 + C\ell^{-\frac{2}{M+2}},$$

for some positive constant $C=C(R,L)$, independent of k.

Theorem 4.2.1 offers an oracle-type inequality with leading constant 1, stating that the risk of the regression collective is bounded by the lowest risk amongst those of the basic machines, *i.e.*, our procedure mimics the performance of the oracle over the set $\{r_{k,m} : m =$

$1,\dots,M\}$, plus a remainder term of the order of $\ell^{-2/(M+2)}$ which is the price to pay for aggregating. In our setting, it is important to observe that this term has a limited impact. As a matter of fact, since the number of basic machines M is assumed to be fixed and not too large (the implementation presented in Section 4.3 considers M at most 6), the remainder term is negligible compared to the standard nonparametric rate $\ell^{-2/(d+2)}$ in dimension d. While the rate $\ell^{-2/(d+2)}$ is affected by the curse of dimensionality when d is large, this is not the case for the term $\ell^{-2/(M+2)}$. Obviously, under the assumption that the distribution of $(\mathbf{X},Y)$ might be described parametrically and that one of the initial estimators is adapted to this distribution, faster rates of the order of $1/\ell$ could emerge in the bias term. Nonetheless, the regression collective is designed for much more adversarial regression problems, hence the rate exhibited in Theorem 4.2.1 appears satisfactory. As a final comment to this result, we stress that our approach carries no assumption on the random design and mild ones over the primal estimators, whereas stringent conditions over the deterministic design and/or the primal estimators are necessary to prove similar results in other aggregation procedures such as the Lasso (van de Geer, 2008; Bunea et al., 2007b).

The central motivation for our method is that model and smoothness assumptions are usually unverifiable, especially in modern high-dimensional and large scale data sets. To circumvent this difficulty, researchers often try many different methods and retain the one exhibiting the best empirical (*e.g.*, cross-validated) results. Our aggregation strategy offers a nice alternative, in the sense that if one of the initial estimators is consistent for a given class $\mathcal{M}$ of distributions, then, under light smoothness assumptions, T_n inherits the same property. To be more precise, assume that the aggregation problem is well-specified, *i.e.*, that one of the original estimators, say r_{k,m_0}, satisfies

$$\mathbb{E}\left|r_{k,m_0}(\mathbf{X})-r^\star(\mathbf{X})\right|^2 \to 0 \quad \text{as } k\to\infty,$$

for all distribution of $(\mathbf{X},Y)$ in some class $\mathcal{M}$. Then, under the assumptions of Theorem 4.2.1, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$, one has

$$\lim_{k,\ell\to\infty} \mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X}))-r^\star(\mathbf{X})\right|^2 = 0.$$

4.3 Implementation and numerical studies

This section is devoted to the implementation of the described method. Its excellent performance is then assessed in a series of experiments. The companion R package `COBRA` (standing for COmBined Regression Alternative) is available on the CRAN website `http://cran.r-project.org/web/packages/COBRA/index.html`, for Linux, Mac and Windows platforms, see Guedj (2013a). `COBRA` includes a `parallel` option, allowing for improved performance on multi-core computers (see Knaus, 2010).

As raised in the previous section, a precise calibration of the smoothing parameter ε_ℓ is crucial. Clearly, a value that is too small will discard many machines and most weights will be zero. Conversely, a large value sets all weights to $1/\Sigma$ with

$$\Sigma = \sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\le\varepsilon_\ell\}},$$

giving the naive predictor that does not account for any new data point and predicts the mean over the sample $\mathcal{D}_\ell$. We also consider a relaxed version of the unanimity constraint: Instead of requiring global agreement over the implemented machines, consider some $\alpha \in$

$(0,1]$ and keep observation Y_i in the construction of T_n if and only if at least a proportion α of the machines agree on the importance of $\mathbf{X}_i$. This parameter requires some calibration. To understand this better, consider the following toy example: On some data set, assume most machines but one have nice predictive performance. For any new data point, requiring global agreement will fail since the pool of machines is heterogeneous. In this regard, α should be seen as a measure of homogeneity: If a small value is selected, it should be seen as an indicator that some machines perform (possibly much) better than some others. Conversely, a large value indicates that the predictive abilities of the machines are close.

A natural measure of the risk in the prediction context is the empirical quadratic loss, namely

$$\hat{R}(\hat{\mathbf{Y}}) = \frac{1}{p}\sum_{j=1}^{p}(\hat{Y}_j - Y_j)^2,$$

where $\hat{\mathbf{Y}} = (\hat{Y}_1,\ldots,\hat{Y}_p)$ is the vector of predicted values for the responses $Y_1,\ldots,Y_p$ and $\{(\mathbf{X}_j,Y_j)\}_{j=1}^{p}$ is a testing sample. We adopted the following protocol: Using a simple data-splitting device, ε_ℓ and α are chosen by minimizing the empirical risk $\hat{R}$ over the set $\{\varepsilon_{\ell,\min},\ldots,\varepsilon_{\ell,\max}\}\times\{1/M,\ldots,1\}$, where $\varepsilon_{\ell,\min} = 10^{-300}$ and $\varepsilon_{\ell,\max}$ is proportional to the largest absolute difference between two predictions of the pool of machines. In the package, the discretization $\#\{\varepsilon_{\ell,\min},\ldots,\varepsilon_{\ell,\max}\}$ may be modified by the user, otherwise the default value 200 is chosen. It is also possible to choose either a linear or a logistic scale. Figure 4.2 illustrates the discussion about the choice of ε_ℓ and α.

By default, `COBRA` includes the following classical packages dealing with regression estimation and prediction. However, note that the user has the choice to modify this list to her/his own convenience:

— Lasso (R package `lars`, see Hastie and Efron, 2012).
— Ridge regression (R package `ridge`, see Cule, 2012).
— k-nearest neighbors (R package `FNN`, see Li, 2013).
— CART algorithm (R package `tree`, see Ripley, 2012).
— Random Forests algorithm (R package `randomForest`, see Liaw and Wiener, 2002).

First, `COBRA` is benchmarked on synthetic data. For each of the following eight models, two designs are considered: Uniform over $(-1,1)^d$ (referred to as "Uncorrelated" in Table 4.1, Table 4.2 and Table 4.3), and Gaussian with mean 0 and covariance matrix Σ with $\Sigma_{ij} = 2^{-|i-j|}$ ("Correlated"). Models considered cover a wide spectrum of contemporary regression problems. Indeed, Model 4.1 is a toy example, Model 4.2 comes from van der Laan et al. (2007), Model 4.3 and Model 4.4 appear in Meier et al. (2009). Model 4.5 is somewhat a classic setting. Model 4.6 is about predicting labels, Model 4.7 is inspired by high-dimensional sparse regression problems. Finally, Model 4.8 deals with probability estimation, forming a link with nonparametric model-free approaches such as in Malley et al. (2012). In the sequel, we let $\mathcal{N}(\mu,\sigma^2)$ denote a Gaussian random variable with mean μ and variance σ^2. In the simulations, the training data set was usually set to 80% of the whole sample, then split into two equal parts corresponding to $\mathcal{D}_k$ and $\mathcal{D}_\ell$.

Model 4.1. $n = 800$, $d = 50$, $Y = X_1^2 + \exp(-X_2^2)$.

Model 4.2. $n = 600$, $d = 100$, $Y = X_1X_2 + X_3^2 - X_4X_7 + X_8X_{10} - X_6^2 + \mathcal{N}(0,0.5)$.

Model 4.3. $n = 600$, $d = 100$, $Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + \mathcal{N}(0,0.5)$.

Model 4.4. $n = 600$, $d = 100$, $Y = X_1 + (2X_2-1)^2 + \sin(2\pi X_3)/(2-\sin(2\pi X_3)) + \sin(2\pi X_4) + 2\cos(2\pi X_4) + 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4) + \mathcal{N}(0,0.5)$.

Model 4.5. $n = 700$, $d = 20$, $Y = \mathbb{1}_{\{X_1>0\}} + X_2^3 + \mathbb{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0,0.5)$.

Model 4.6. $n = 500$, $d = 30$, $Y = \sum_{k=1}^{10} \mathbb{1}_{\{X_k^3<0\}} - \mathbb{1}_{\{\mathcal{N}(0,1)>1.25\}}$.

Model 4.7. $n = 600$, $d = 300$, $Y = X_1^2 + X_2^2 X_3 \exp(-|X_4|) + X_6 - X_8 + \mathcal{N}(0,0.5)$.

Model 4.8. $n = 600$, $d = 50$, $Y = \mathbb{1}_{\{X_1+X_4^3+X_9+\sin(X_{12}X_{18})+\mathcal{N}(0,0.1)>0.38\}}$.

Table 4.1 presents the mean quadratic error and standard deviation over 100 independent replications, for each model and design. Bold numbers identify the lowest error, *i.e.*, the apparent best competitor. Boxplots of errors are presented in Figure 4.3 and Figure 4.4. Further, Figure 4.5 and Figure 4.6 shows the predictive capacities of COBRA, and Figure 4.7 depicts its ability to reconstruct the functional dependence over the covariates in the context of additive regression, assessing the striking performance of our approach in a wide spectrum of statistical settings. A persistent and notable fact is that COBRA performs at least as well as the best machine, significantly so in Model 4.3, Model 4.5 and Model 4.6.

Next, since an increasing number of problems in contemporary statistics involve high-dimensional data, we have tested the abilities of COBRA in that context. As highlighted by Table 4.4 and Figure 4.8, the main message is that COBRA is perfectly able to deal with high-dimensional data, provided that it is generated over machines, at least some of which are known to perform well in such situations (possibly at the price of a sparsity assumption). In that context, we conducted 200 independent replications for the three following models:

Model 4.9. $n = 500$, $d = 1000$, $Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$. *Uncorrelated design.*

Model 4.10. $n = 500$, $d = 1000$, $Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$. *Correlated design*

Model 4.11. $n = 500$, $d = 1500$, $Y = \exp(-X_1) + \exp(X_1) + \sum_{j=2}^{d} X_j^{j/100}$. *Uncorrelated design.*

A legitimate question that arises is where one should cut the initial sample $\mathcal{D}_n$? In other words, for a given data set of size n, what is the optimal value for k? A naive approach is to cut the initial sample in two halfs (i.e., $k = n/2$): This appears to be satisfactory provided that n is large enough, which may be too much of an unrealistic assumption in numerous experimental settings. A more involved choice is to adopt a random cut scheme, where k is chosen uniformly in $\{1,\ldots,n\}$. Figure 4.9 presents the boxplot of errors of the five default machines and COBRA with that random cutting strategy, and also shows the risk of COBRA with respect to k: To illustrate this phenomenon, we tested a thousand random cuts on the following Model 4.12. As showed in Figure 4.9, for that particular model, the best value seems to be near $3n/4$.

Model 4.12. $n = 1200$, $d = 10$, $Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$. *Uncorrelated design.*

The average risk of COBRA on a thousand replications of Model 4.12 is 0.3124. Since this delivered a thousand prediction vectors, a natural idea is to take their mean or median. The risk of the mean is 0.2306, and the median has an even better risk (0.2184). Since a random cut scheme may generate some unstability, we advise practitioners to compute a few COBRA estimators, then compute the mean or median vector of their predictions.

Next, we compare COBRA to the Super Learner algorithm (Polley and van der Laan, 2012). This widely used algorithm was first described in van der Laan et al. (2007) and extended in Polley and van der Laan (2010). Super Learner is used in this section as the key competitor to our method. In a nutshell, the Super Learner trains basic machines

$r_1,\dots,r_M$ on the whole sample $\mathcal{D}_n$. Then, following a V-fold cross-validation procedure, Super Learner adopts a V-blocks partition of the set $\{1,\dots,n\}$ and computes the matrix

$$H = (H_{ij})_{1\le i\le n}^{1\le j\le M},$$

where H_{ij} is the prediction for the query point $\mathbf{X}_i$ made by machine j trained on all remaining $V-1$ blocks, *i.e.*, excluding the block containing $\mathbf{X}_i$. The Super Learner estimator is then

$$SL = \sum_{j=1}^{M} \hat{\alpha}_j r_j,$$

where

$$\hat{\alpha} \in \operatorname*{arg\,inf}_{\alpha\in\Lambda^M} \sum_{i=1}^{n} |Y_i - (H\alpha)_i|^2,$$

with Λ^M denoting the simplex

$$\Lambda^M = \left\{ \alpha \in \mathbb{R}^M : \sum_{j=1}^{M} \alpha_j = 1,\ \alpha_j \ge 0 \text{ for any } j = 1,\dots,M \right\}.$$

Although this convex aggregation scheme is significantly different from our regression collective scheme, it is similar to the approach used in the `SuperLearner` package, in that both allow the user to aggregate as many machines as desired, then blending them to deliver predictive outcomes. For that reason, it is reasonable to deploy Super Learner as a benchmark in our study of regression collectives.

Table 4.2 summarizes the performance of `COBRA` and `SuperLearner` (used with the functions `SL.randomForest`, `SL.ridge` and `SL.glmnet`, for the fairness of the comparison) through the described protocol. Both methods compete on similar terms in most models, although `COBRA` proves much more efficient on correlated design in Model 4.2 and Model 4.4. This already remarkable result is to be stressed by the flexibility and velocity showed by `COBRA`. Indeed, as emphasized in Table 4.3, without even using the `parallel` option, `COBRA` obtains similar or better results than `SuperLearner` roughly five times faster. Note also that `COBRA` suffers from a disadvantage: `SuperLearner` is built on the whole sample $\mathcal{D}_n$ whereas `COBRA` only uses $\ell < n$ data points. Finally, observe that the algorithmic cost of computing the random weights on n_{test} query points is $\ell \times M \times n_{\text{test}}$ operations. In the package, those calculations are handled in C language for optimal speed performance.

Super Learner is a natural competitor on the implementation side. However, on the theoretical side, we do not assume that it should be the only benchmark: Thus, we compared `COBRA` to the popular exponentially weighted aggregation method (EWA). We implemented the following version of the EWA: For all preliminary estimators $r_{k,1},\dots,r_{k,M}$, their empirical risks $\hat{R}_1,\dots,\hat{R}_M$ are computed on a subsample of $\mathcal{D}_\ell$ and the EWA is

$$\text{EWA}_\beta : \mathbf{x} \mapsto \sum_{j=1}^{M} \hat{w}_j r_{k,j}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

where

$$\hat{w}_j = \frac{\exp(-\beta \hat{R}_j)}{\sum_{i=1}^{M} \exp(-\beta \hat{R}_i)}, \quad j = 1,\dots,M.$$

The temperature parameter $\beta > 0$ is selected by minimizing the empirical risk of EWA_β over a data-based grid, in the same spirit as the selection of ε_ℓ and α. We conducted 200 independent replications, on Models 4.9 to 4.12. The conclusion is that `COBRA` outperforms

the EWA estimator in some models, and delivers similar performance in others, as shown in Figure 4.10 and Table 4.5.

Finally, COBRA is used to process the following real-life data sets:

— Concrete Slump Test [2] (see Yeh, 2007).
— Concrete Compressive Strength [3] (see Yeh, 1998).
— Wine Quality [4] (see Cortez et al., 2009). We point out that the Wine Quality data set involves supervised classification and leads naturally to a line of future research using COBRA as a regression collective over probability machines (see Malley et al., 2012).

The good predictive performance of COBRA is summarized in Figure 4.11 and errors are presented in Figure 4.12. For every data set, the sample is divided into a training set (90%) and a testing set (10%) on which the predictive performance is evaluated. Boxplots are obtained by randomly shuffling the data points a hundred times.

As a conclusion to this thorough experimental protocol, it is our belief that COBRA sets a new high standard of reference, a benchmark procedure, both in terms of performance and velocity, for prediction-oriented problems in the context of regression.

2. http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test.
3. http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.
4. http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

TABLE 4.1 – Quadratic errors of the implemented machines and COBRA. Means and standard deviations over 100 independent replications.

Uncorr.		lars	ridge	fnn	tree	rf	COBRA
Model 4.1	m.	0.1561	0.1324	0.1585	0.0281	0.0330	**0.0259**
	sd.	0.0123	0.0094	0.0123	0.0043	0.0033	0.0036
Model 4.2	m.	0.4880	0.2462	0.3070	0.1746	**0.1366**	0.1645
	sd.	0.0676	0.0233	0.0303	0.0270	0.0161	0.0207
Model 4.3	m.	0.2536	0.5347	1.1603	0.4954	0.4027	**0.2332**
	sd.	0.0271	0.4469	0.1227	0.0772	0.0558	0.0272
Model 4.4	m.	7.6056	6.3271	10.5890	3.7358	3.5262	**3.3640**
	sd.	0.9419	1.0800	0.9404	0.8067	0.3223	0.5178
Model 4.5	m.	0.2943	0.3311	0.5169	0.2918	0.2234	**0.2060**
	sd.	0.0214	0.1012	0.0439	0.0279	0.0216	0.0210
Model 4.6	m.	0.8438	1.0303	2.0702	2.3476	1.3354	**0.8345**
	sd.	0.0916	0.4840	0.2240	0.2814	0.1590	0.1004
Model 4.7	m.	1.0920	0.5452	0.9459	0.3638	0.3110	**0.3052**
	sd.	0.2265	0.0920	0.0833	0.0456	0.0325	0.0298
Model 4.8	m.	0.1308	0.1279	0.2243	0.1715	0.1236	**0.1021**
	sd.	0.0120	0.0161	0.0189	0.0270	0.0100	0.0155

Corr.		lars	ridge	fnn	tree	rf	COBRA
Model 4.1	m.	2.3736	1.9785	2.0958	0.3312	0.5766	**0.3301**
	sd.	0.4108	0.3538	0.3414	0.1285	0.1914	0.1239
Model 4.2	m.	8.1710	4.0071	4.3892	**1.3609**	1.4768	1.3612
	sd.	1.5532	0.6840	0.7190	0.4647	0.4415	0.4654
Model 4.3	m.	6.1448	6.0185	8.2154	4.3175	4.0177	**3.7917**
	sd.	11.9450	12.0861	13.3121	11.7386	12.4160	11.1806
Model 4.4	m.	60.5795	42.2117	51.7293	**9.6810**	14.7731	9.6906
	sd.	11.1303	9.8207	10.9351	3.9807	5.9508	3.9872
Model 4.5	m.	6.2325	7.1762	10.1254	3.1525	4.2289	**2.1743**
	sd.	2.4320	3.5448	3.1190	2.1468	2.4826	1.6640
Model 4.6	m.	1.2765	1.5307	2.5230	2.6185	1.2027	**0.9925**
	sd.	0.1381	0.9593	0.2762	0.3445	0.1600	0.1210
Model 4.7	m.	20.8575	4.4367	5.8893	3.6865	**2.7318**	2.9127
	sd.	7.1821	1.0770	1.2226	1.0139	0.8945	0.9072
Model 4.8	m.	0.1366	0.1308	0.2267	0.1701	0.1226	**0.0984**
	sd.	0.0127	0.0143	0.0179	0.0302	0.0102	0.0144

TABLE 4.2 – Quadratic errors of `SuperLearner` and `COBRA`. Means and standard deviations over 100 independent replications.

Uncorr.		SL	COBRA
Model 4.1	m.	0.0541	**0.0320**
	sd.	0.0053	0.0104
Model 4.2	m.	**0.1765**	0.3569
	sd.	0.0167	0.8797
Model 4.3	m.	**0.2081**	0.2573
	sd.	0.0282	0.0699
Model 4.4	m.	4.3114	**3.7464**
	sd.	0.4138	0.8746
Model 4.5	m.	**0.2119**	0.2187
	sd.	0.0317	0.0427
Model 4.6	m.	**0.7627**	1.0220
	sd.	0.1023	0.3347
Model 4.7	m.	**0.1705**	0.3103
	sd.	0.0260	0.0490
Model 4.8	m.	0.1081	**0.1075**
	sd.	0.0121	0.0235

Corr.		SL	COBRA
Model 4.1	m.	0.8733	**0.3262**
	sd.	0.2740	0.1242
Model 4.2	m.	2.3391	**1.3984**
	sd.	0.4958	0.3804
Model 4.3	m.	**3.1885**	3.3201
	sd.	1.5101	1.8056
Model 4.4	m.	25.1073	**9.3964**
	sd.	7.3179	2.8953
Model 4.5	m.	5.6478	**4.9990**
	sd.	7.7271	9.3103
Model 4.6	m.	**0.8967**	1.1988
	sd.	0.1197	0.4573
Model 4.7	m.	**3.0367**	3.1401
	sd.	1.6225	1.6097
Model 4.8	m.	0.1116	**0.1045**
	sd.	0.0111	0.0216

TABLE 4.3 – Average CPU-times in seconds. No parallelization. Means and standard deviations over 10 independent replications.

Uncorr.		SL	COBRA
Model 4.1	m.	53.92	**10.92**
	sd.	1.42	0.29
Model 4.2	m.	57.96	**11.90**
	sd.	0.95	0.31
Model 4.3	m.	53.70	**10.66**
	sd.	0.55	0.11
Model 4.4	m.	55.00	**11.15**
	sd.	0.74	0.18
Model 4.5	m.	28.46	**5.01**
	sd.	0.73	0.06
Model 4.6	m.	22.97	**3.99**
	sd.	0.27	0.05
Model 4.7	m.	127.80	**35.67**
	sd.	5.69	1.91
Model 4.8	m.	32.98	**6.46**
	sd.	1.33	0.33

Corr.		SL	COBRA
Model 4.1	m.	61.92	**11.96**
	sd.	1.85	0.27
Model 4.2	m.	70.90	**14.16**
	sd.	2.47	0.57
Model 4.3	m.	59.91	**11.92**
	sd.	2.06	0.41
Model 4.4	m.	63.58	**13.11**
	sd.	1.21	0.34
Model 4.5	m.	31.24	**5.02**
	sd.	0.86	0.07
Model 4.6	m.	24.29	**4.12**
	sd.	0.82	0.15
Model 4.7	m.	145.18	**41.28**
	sd.	8.97	2.84
Model 4.8	m.	31.31	**6.24**
	sd.	0.73	0.11

TABLE 4.4 – Quadratic errors of the implemented machines and COBRA in high-dimensional situations. Means and standard deviations over 200 independent replications.

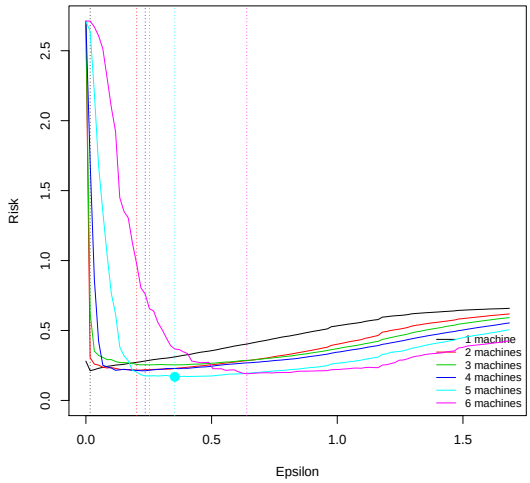
		lars	ridge	fnn	tree	rf	COBRA
Model 4.9	m.	1.5698	2.9752	3.9285	1.8646	1.5001	**0.9996**
	sd.	0.2357	0.4171	0.5356	0.3751	0.2491	0.1733
Model 4.10	m.	5.2356	5.1748	6.1395	6.1585	4.8667	**2.7076**
	sd.	0.6885	0.7139	0.9192	0.9298	0.6634	0.3810
Model 4.11	m.	0.1584	0.1055	0.1363	0.0058	0.0327	**0.0049**
	sd.	0.0199	0.0119	0.0176	0.0010	0.0052	0.0009

TABLE 4.5 – Quadratic errors of exponentially weighted aggregation (EWA) and COBRA. 200 independent replications.

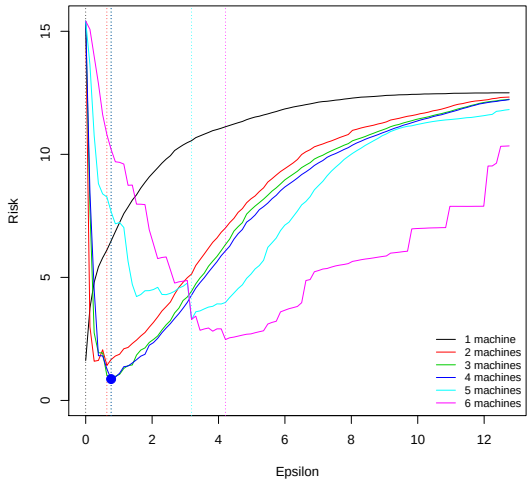
		EWA	COBRA
Model 4.9	m.	1.1712	**1.1360**
	sd.	0.2090	0.2468
Model 4.10	m.	**9.4789**	12.4353
	sd.	5.6275	9.1267
Model 4.11	m.	0.0244	**0.0128**
	sd.	0.0042	0.0237
Model 4.12	m.	0.4175	**0.3124**
	sd.	0.0513	0.0884

FIGURE 4.2 – Examples of calibration of parameters ε_ℓ and α. The bold point is the minimum.

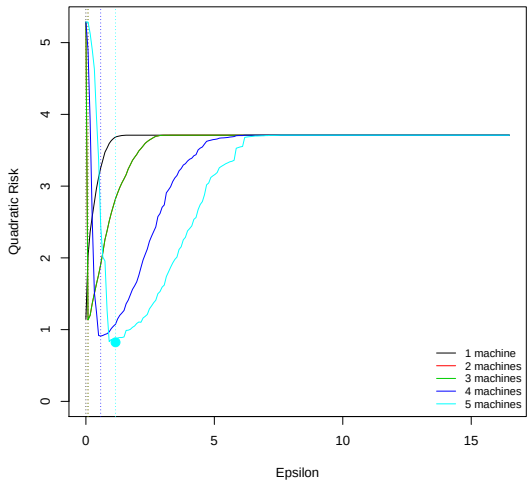
(A) Model 4.5, uncorrelated design.



(B) Model 4.5, correlated design.



(C) Model 4.9.



(D) Model 4.12.

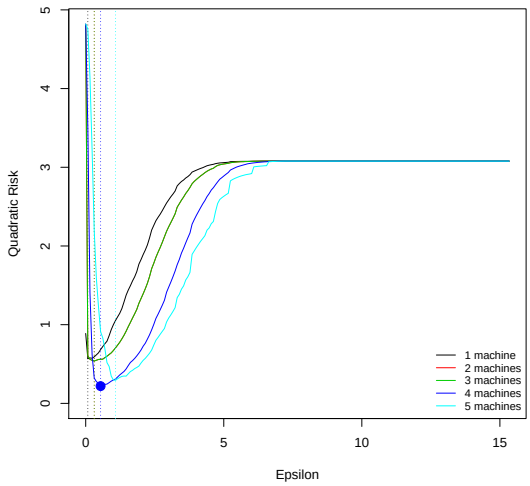
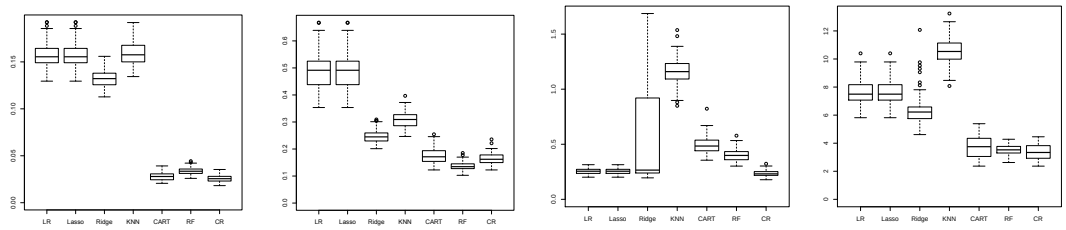


FIGURE 4.3 – Boxplots of quadratic errors, uncorrelated design. From left to right: `lars`, `ridge`, `fnn`, `tree`, `randomForest`, `COBRA`.

(A) Model 4.1. (B) Model 4.2. (C) Model 4.3. (D) Model 4.4.



(E) Model 4.5. (F) Model 4.6. (G) Model 4.7. (H) Model 4.8.

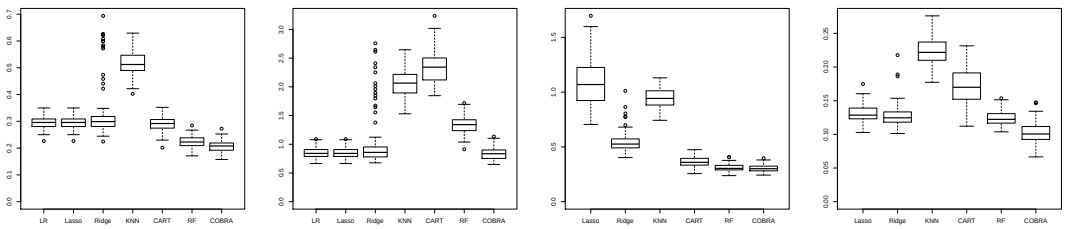
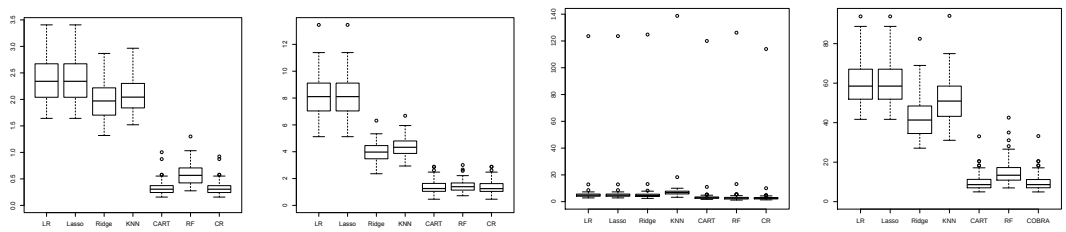


FIGURE 4.4 – Boxplots of quadratic errors, correlated design. From left to right: `lars`, `ridge`, `fnn`, `tree`, `randomForest`, `COBRA`.

(A) Model 4.1. (B) Model 4.2. (C) Model 4.3. (D) Model 4.4.



(E) Model 4.5. (F) Model 4.6. (G) Model 4.7. (H) Model 4.8.

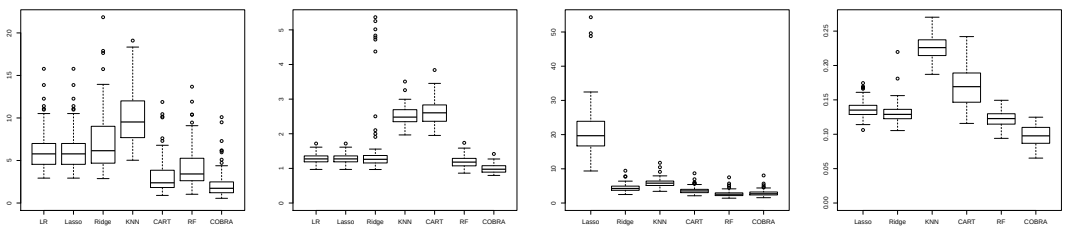
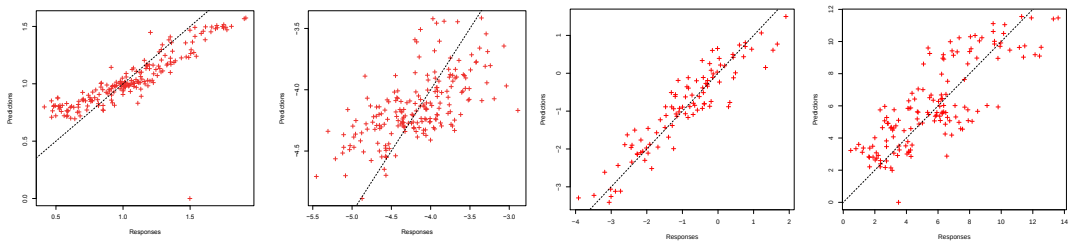


FIGURE 4.5 – Prediction over the testing set, uncorrelated design. The more points on the first bissectrix, the better the prediction.

(A) Model 4.1. (B) Model 4.2. (C) Model 4.3. (D) Model 4.4.



(E) Model 4.5. (F) Model 4.6. (G) Model 4.7. (H) Model 4.8.

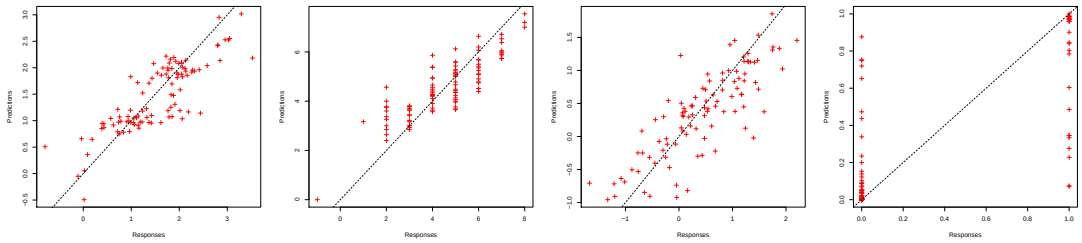
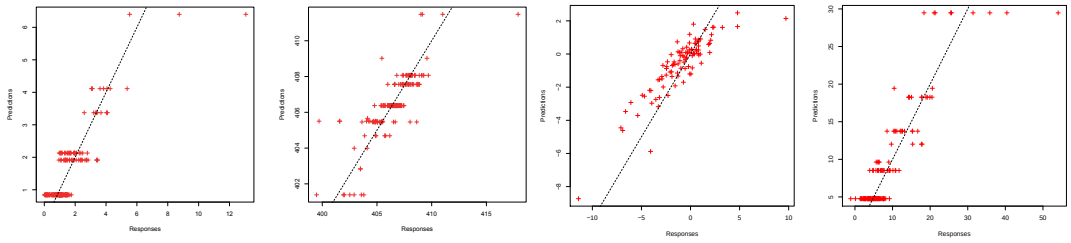


FIGURE 4.6 – Prediction over the testing set, correlated design. The more points on the first bissectrix, the better the prediction.

(A) Model 4.1. (B) Model 4.2. (C) Model 4.3. (D) Model 4.4.



(E) Model 4.5. (F) Model 4.6. (G) Model 4.7. (H) Model 4.8.

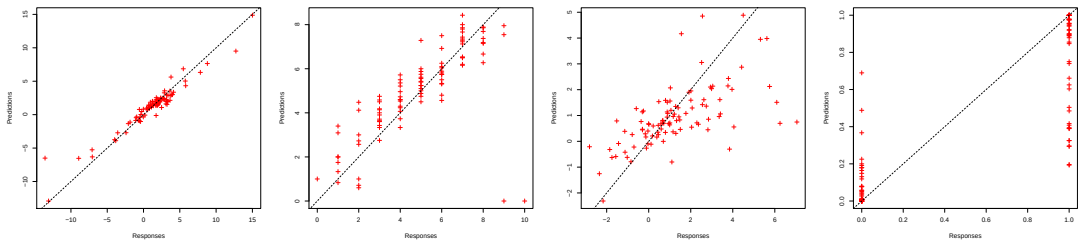
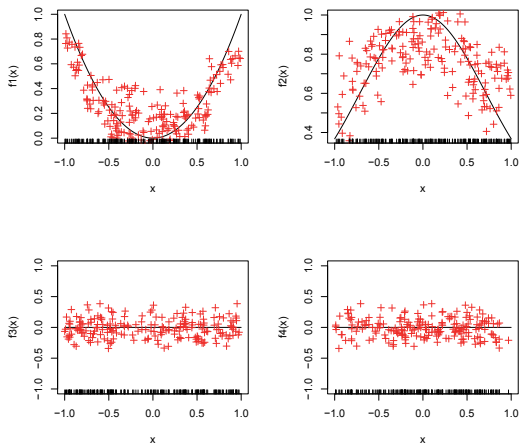
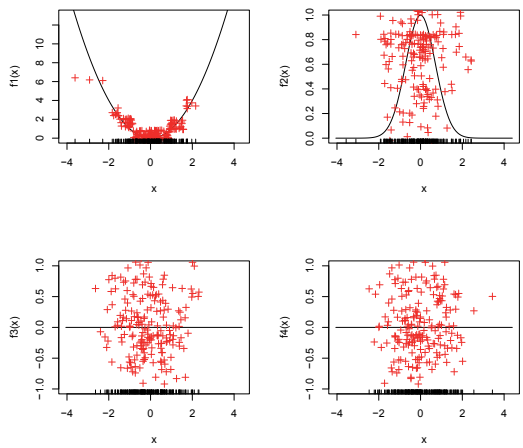


FIGURE 4.7 – Examples of reconstruction of the functional dependencies, for covariates 1 to 4.

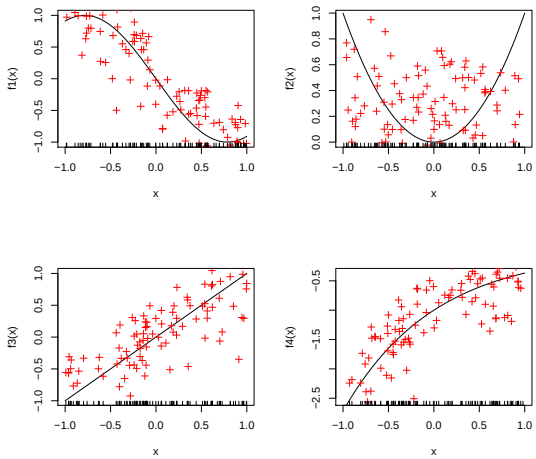
(A) Model 4.1, uncorrelated design.



(B) Model 4.1, correlated design.



(C) Model 4.3, uncorrelated design.



(D) Model 4.3, correlated design.

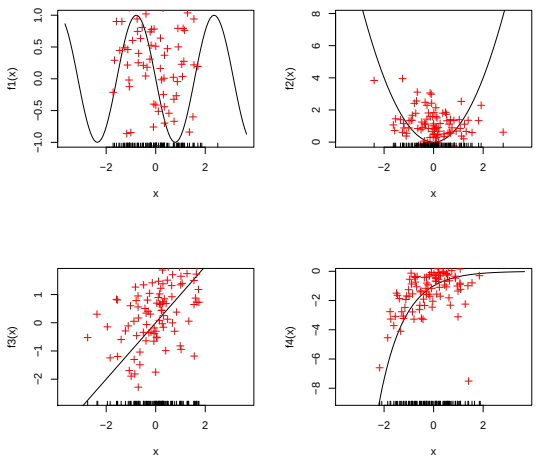


FIGURE 4.8 – Boxplot of errors, high-dimensional models.

(A) Model 4.9 (B) Model 4.10 (C) Model 4.11

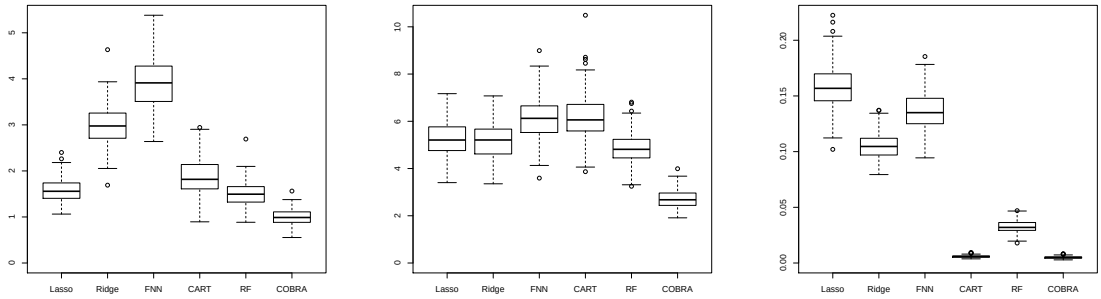


FIGURE 4.9 – How stable is COBRA?

(A) Boxplot of errors: Initial sample is randomly cut (1000 replications of Model 4.12).

(B) Empirical risk with respect to the size of subsample $\mathcal{D}_k$, in Model 4.12.

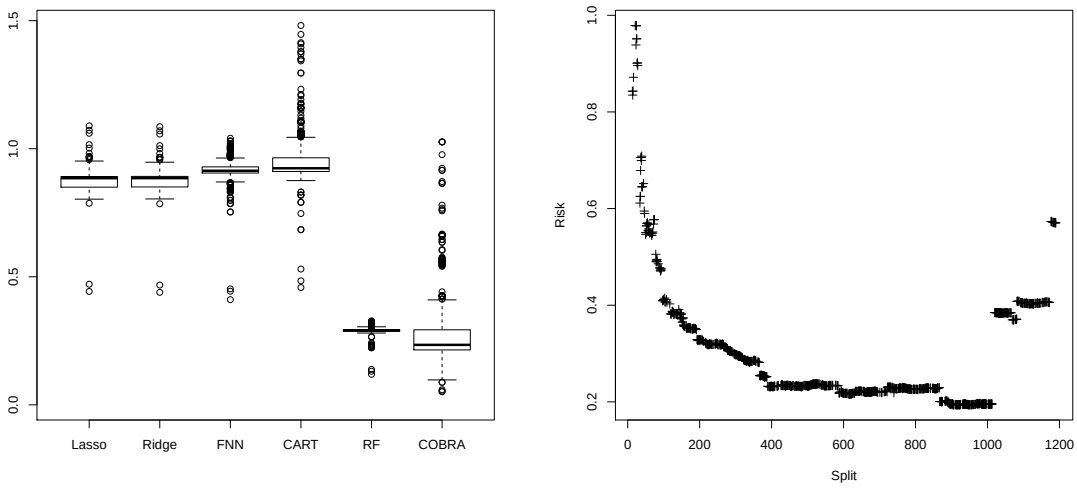


FIGURE 4.10 – Boxplot of errors: EWA vs COBRA

(A) Model 4.9. (B) Model 4.10. (C) Model 4.11. (D) Model 4.12.

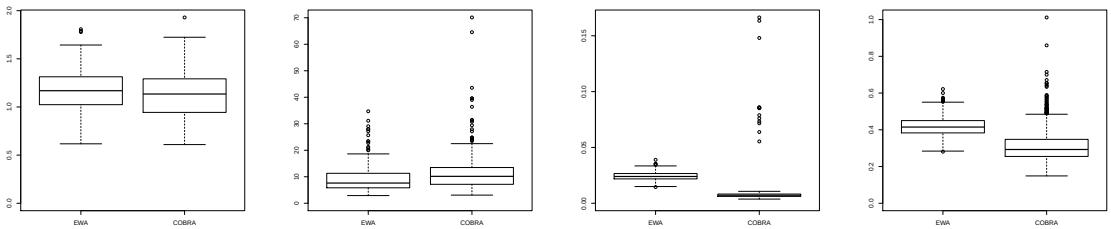
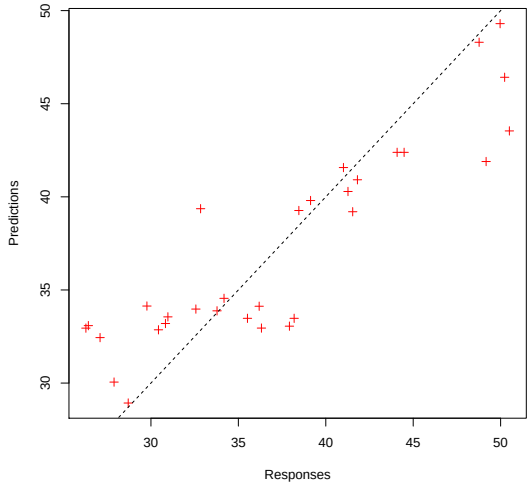
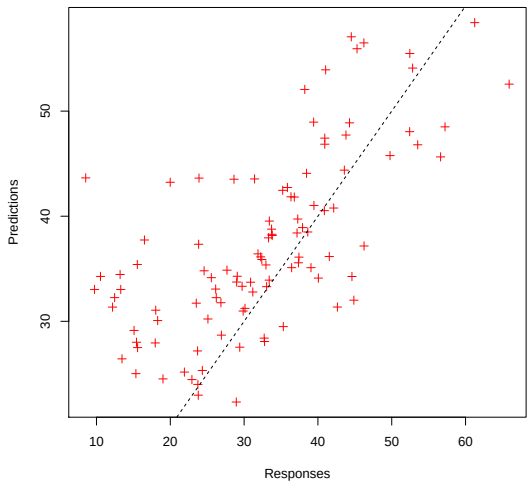


FIGURE 4.11 – Prediction over the testing set, real-life data sets.

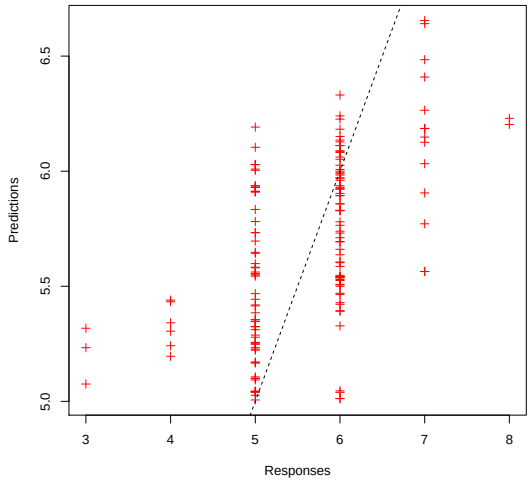
(A) Concrete Slump Test.



(B) Concrete Compressive Strength.



(C) Wine Quality, red wine.



(D) Wine Quality, white wine.

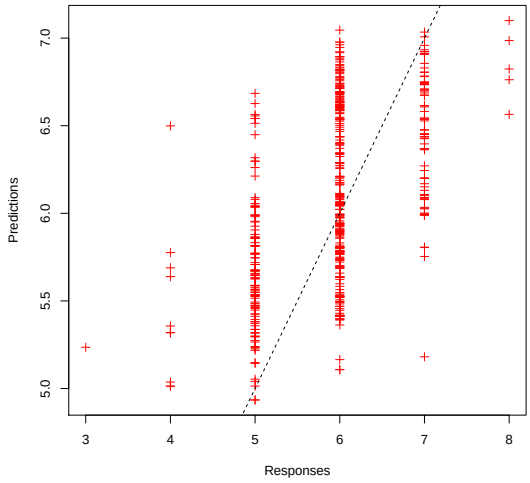
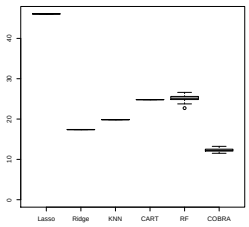
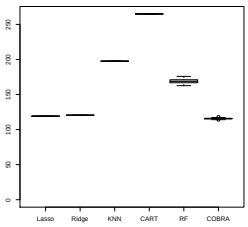


FIGURE 4.12 – Boxplot of quadratic errors, real-life data sets.

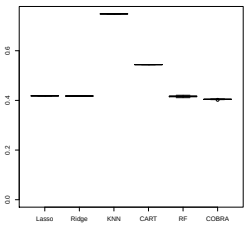
(A) Concrete Slump Test.



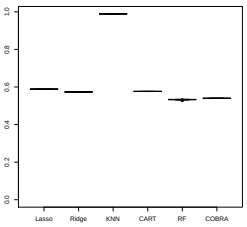
(B) Concrete Compressive Strength.



(C) Wine Quality, red wine.



(D) Wine Quality, white wine.



4.4 Proofs

4.4.1 Proof of Proposition 4.2.1

We have

$$\begin{aligned}\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 &= \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 + \mathbb{E}|T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \\ &\quad - 2\mathbb{E}[(T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X})))(T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X}))].\end{aligned}$$

As for the double product, notice that

$$\begin{aligned}&\mathbb{E}[(T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X})))(T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X}))] \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[(T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X})))(T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X}))|\mathbf{r}_k(\mathbf{X}), \mathcal{D}_n\right]\right] \\ &\quad = \mathbb{E}\left[(T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X})))\mathbb{E}\left[T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|\mathbf{r}_k(\mathbf{X}), \mathcal{D}_n\right]\right].\end{aligned}$$

But

$$\begin{aligned}\mathbb{E}[r^\star(\mathbf{X})|\mathbf{r}_k(\mathbf{X}), \mathcal{D}_n] &= \mathbb{E}[r^\star(\mathbf{X})|\mathbf{r}_k(\mathbf{X})] \\ &\quad \text{(by independence of } \mathbf{X} \text{ and } \mathcal{D}_n) \\ &= \mathbb{E}[\mathbb{E}[Y|\mathbf{X}]|\mathbf{r}_k(\mathbf{X})] \\ &= \mathbb{E}[Y|\mathbf{r}_k(\mathbf{X})] \\ &\quad \text{(since } \sigma(\mathbf{r}_k(\mathbf{X})) \subset \sigma(\mathbf{X})) \\ &= T(\mathbf{r}_k(\mathbf{X})).\end{aligned}$$

Consequently,

$$\mathbb{E}[(T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X})))(T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X}))] = 0$$

and

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 = \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 + \mathbb{E}|T(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2.$$

Thus, by the very definition of the conditional expectation, and using the fact that $T(\mathbf{r}_k(\mathbf{X})) = \mathbb{E}[r^\star(\mathbf{X})|\mathbf{r}_k(\mathbf{X})]$,

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \le \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 + \inf_f \mathbb{E}|f(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2,$$

where the infimum is taken over all square integrable functions of $\mathbf{r}_k(\mathbf{X})$. In particular,

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \quad \le \min_{m=1,\dots,M} \mathbb{E}|r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X})|^2 + \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2,$$

as desired.

4.4.2 Proof of Proposition 4.2.2

Note that the second statement is an immediate consequence of the first statement and Proposition 4.2.1, therefore we only have to prove that

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \to 0 \quad \text{as } \ell \to \infty.$$

We start with a technical lemma, whose proof can be found in the monograph by Györfi et al. (2002).

Lemma 4.4.1. *Let $B(n,p)$ be a binomial random variable with parameters $n \geq 1$ and $p > 0$. Then*

$$\mathbb{E}\left[\frac{1}{1+B(n,p)}\right] \leq \frac{1}{p(n+1)}$$

and

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{B(n,p)>0\}}}{B(n,p)}\right] \leq \frac{2}{p(n+1)}.$$

For all distribution of $(\mathbf{X},Y)$, using the elementary inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, note that

$$\begin{aligned}
&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \\
&= \mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(Y_i - T(\mathbf{r}_k(\mathbf{X}_i)) + T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X})) + T(\mathbf{r}_k(\mathbf{X}))) - T(\mathbf{r}_k(\mathbf{X}))\right|^2 \\
&\leq 3\mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X})))\right|^2 && (4.4)\\
&\quad + 3\mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(Y_i - T(\mathbf{r}_k(\mathbf{X}_i)))\right|^2 && (4.5)\\
&\quad + 3\mathbb{E}\left|\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) - 1\right) T(\mathbf{r}_k(\mathbf{X}))\right|^2. && (4.6)
\end{aligned}$$

Consequently, to prove the proposition, it suffices to establish that (4.4), (4.5) and (4.6) tend to 0 as ℓ tends to infinity. This is done, respectively, in Proposition 4.4.1, Proposition 4.4.2 and Proposition 4.4.3 below.

Proposition 4.4.1. *Under the assumptions of Proposition 4.2.2,*

$$\lim_{\ell\to\infty} \mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X})))\right|^2 = 0.$$

Proof of Proposition 4.4.1. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X})))\right|^2 \\
&= \mathbb{E}\left|\sum_{i=1}^{\ell} \sqrt{W_{n,i}(\mathbf{X})}\sqrt{W_{n,i}(\mathbf{X})}(T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X})))\right|^2 \\
&\leq \mathbb{E}\left[\sum_{j=1}^{\ell} W_{n,j}(\mathbf{X}) \sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&:= A_n.
\end{aligned}$$

The function T is such that $\mathbb{E}[T^2(\mathbf{r}_k(\mathbf{X}))] < \infty$. Therefore, it can be approximated in an L^2 sense by a continuous function with compact support, say $\tilde{T}$. This result may be found in many references, amongst them Györfi et al. (2002, Theorem A.1). More precisely, for any $\eta > 0$, there exists a function $\tilde{T}$ such that

$$\mathbb{E}\left|T(\mathbf{r}_k(\mathbf{X})) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))\right|^2 < \eta.$$

Consequently, we obtain

$$\begin{aligned}
A_n &= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&\leq 3\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|T(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{T}(\mathbf{r}_k(\mathbf{X}_i))|^2\right] \\
&\quad + 3\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|\tilde{T}(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&\quad + 3\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|\tilde{T}(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&:= 3A_{n1} + 3A_{n2} + 3A_{n3}.
\end{aligned}$$

Computation of A_{n3}. Thanks to the approximation of T by $\tilde{T}$,

$$\begin{aligned}
A_{n3} &= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|T(\mathbf{r}_k(\mathbf{X})) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&\leq \mathbb{E}\left|T(\mathbf{r}_k(\mathbf{X})) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))\right|^2 < \eta.
\end{aligned}$$

Computation of A_{n1}. Denote by μ the distribution of $\mathbf{X}$. Then,

$$\begin{aligned}
A_{n1} &= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|\tilde{T}(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{X}_i))|^2\right] \\
&= \ell\mathbb{E}\left[\frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{X})-r_{k,m}(\mathbf{X}_1)|\leq\varepsilon_\ell\}}}{\sum_{j=1}^{\ell}\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{X})-r_{k,m}(\mathbf{X}_j)|\leq\varepsilon_\ell\}}}|\tilde{T}(\mathbf{r}_k(\mathbf{X}_1)) - T(\mathbf{r}_k(\mathbf{X}_1))|^2\right]. \\
&= \ell\mathbb{E}\Bigg\{\int_{\mathbb{R}^d}|\tilde{T}(\mathbf{r}_k(\mathbf{u})) - T(\mathbf{r}_k(\mathbf{u}))|^2 \\
&\qquad \times \mathbb{E}\left[\int_{\mathbb{R}^d}\frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{u})|\leq\varepsilon_\ell\}}}{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{u})|\leq\varepsilon_\ell\}} + \sum_{j=2}^{\ell}\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\leq\varepsilon_\ell\}}}\mu(\mathrm{d}\mathbf{x})\middle|\mathcal{D}_k\right] \\
&\qquad \mu(\mathrm{d}\mathbf{u})\Bigg\}.
\end{aligned}$$

Let us prove that

$$\begin{aligned}
A'_{n1} &= \mathbb{E}\left[\int_{\mathbb{R}^d}\frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{u})|\leq\varepsilon_\ell\}}}{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{u})|\leq\varepsilon_\ell\}} + \sum_{j=2}^{\ell}\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\leq\varepsilon_\ell\}}}\mu(\mathrm{d}\mathbf{x})\middle|\mathcal{D}_k\right] \\
&\leq \frac{2^M}{\ell}.
\end{aligned}$$

To this aim, observe that

$$\begin{aligned}
A'_{n1} &= \mathbb{E}\left[\int_{\mathbb{R}^d}\frac{\mathbb{1}_{\{\mathbf{x}\in\bigcap_{m=1}^{M} r_{k,m}^{-1}([r_{k,m}(\mathbf{u})-\varepsilon_\ell, r_{k,m}(\mathbf{u})+\varepsilon_\ell])\}}}{1+\sum_{j=2}^{\ell}\mathbb{1}_{\{\mathbf{X}_j\in\bigcap_{m=1}^{M} r_{k,m}^{-1}([r_{k,m}(\mathbf{x})-\varepsilon_\ell, r_{k,m}(\mathbf{x})+\varepsilon_\ell])\}}}\mu(\mathrm{d}\mathbf{x})\middle|\mathcal{D}_k\right] \\
&= \mathbb{E}\left[\int_{\mathbb{R}^d}\frac{\mathbb{1}_{\{\mathbf{x}\in\bigcup_{(a_1,\dots,a_M)\in\{1,2\}^M} r_{k,1}^{-1}(I_{n,1}^{a_1}(\mathbf{u}))\cap\dots\cap r_{k,M}^{-1}(I_{n,M}^{a_M}(\mathbf{u}))\}}}{1+\sum_{j=2}^{\ell}\mathbb{1}_{\{\mathbf{X}_j\in\bigcap_{m=1}^{M} r_{k,m}^{-1}([r_{k,m}(\mathbf{x})-\varepsilon_\ell, r_{k,m}(\mathbf{x})+\varepsilon_\ell])\}}}\mu(\mathrm{d}\mathbf{x})\middle|\mathcal{D}_k\right] \\
&\leq \sum_{p=1}^{2^M}\mathbb{E}\left[\int_{\mathbb{R}^d}\frac{\mathbb{1}_{\{\mathbf{x}\in R_n^p(\mathbf{u})\}}}{1+\sum_{j=2}^{\ell}\mathbb{1}_{\{\mathbf{X}_j\in\bigcap_{m=1}^{M} r_{k,m}^{-1}([r_{k,m}(\mathbf{x})-\varepsilon_\ell, r_{k,m}(\mathbf{x})+\varepsilon_\ell])\}}}\mu(\mathrm{d}\mathbf{x})\middle|\mathcal{D}_k\right].
\end{aligned}$$

Here, $I^1_{n,m}(\mathbf{u}) = [r_{k,m}(\mathbf{u}) - \varepsilon_\ell, r_{k,m}(\mathbf{u})]$, $I^2_{n,m}(\mathbf{u}) = [r_{k,m}(\mathbf{u}), r_{k,m}(\mathbf{u}) + \varepsilon_\ell]$, and $R^p_n(\mathbf{u})$ is the p-th set of the form $r^{-1}_{k,1}(I^{a_1}_{n,1}(\mathbf{u})) \cap \cdots \cap r^{-1}_{k,M}(I^{a_M}_{n,M}(\mathbf{u}))$ assuming that they have been ordered using the lexicographic order of $(a_1, \ldots, a_M)$.

Next, note that

$$\mathbf{x} \in R^p_n(\mathbf{u}) \Rightarrow R^p_n(\mathbf{u}) \subset \bigcap_{m=1}^{M} r^{-1}_{k,m}([r_{k,m}(\mathbf{x}) - \varepsilon_\ell, r_{k,m}(\mathbf{x}) + \varepsilon_\ell]).$$

To see this, just observe that, for all $m = 1, \ldots, M$, if $r_{k,m}(\mathbf{z}) \in [r_{k,m}(\mathbf{u}) - \varepsilon_\ell, r_{k,m}(\mathbf{u})]$, i.e., $r_{k,m}(\mathbf{u}) - \varepsilon_\ell \le r_{k,m}(\mathbf{z}) \le r_{k,m}(\mathbf{u})$, then, as $r_{k,m}(\mathbf{u}) - \varepsilon_\ell \le r_{k,m}(\mathbf{x}) \le r_{k,m}(\mathbf{u})$, one has $r_{k,m}(\mathbf{x}) - \varepsilon_\ell \le r_{k,m}(\mathbf{z}) \le r_{k,m}(\mathbf{x}) + \varepsilon_\ell$. Similarly, if $r_{k,m}(\mathbf{u}) \le r_{k,m}(\mathbf{z}) \le r_{k,m}(\mathbf{u}) + \varepsilon_\ell$, then $r_{k,m}(\mathbf{u}) \le r_{k,m}(\mathbf{x}) \le r_{k,m}(\mathbf{u}) + \varepsilon_\ell$ implies $r_{k,m}(\mathbf{x}) - \varepsilon_\ell \le r_{k,m}(\mathbf{z}) \le r_{k,m}(\mathbf{x}) + \varepsilon_\ell$. Consequently,

$$\begin{aligned}
A'_{n1} &\le \sum_{p=1}^{2^M} \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{\mathbb{1}_{\{\mathbf{x} \in R^p_n(\mathbf{u})\}}}{1 + \sum_{j=2}^{\ell} \mathbb{1}_{\{\mathbf{X}_j \in R^p_n(\mathbf{u})\}}} \mu(\mathrm{d}\mathbf{x}) \middle| \mathcal{D}_k\right] \\
&= \sum_{p=1}^{2^M} \mathbb{E}\left[\frac{\mu\{R^p_n(\mathbf{u})\}}{1 + \sum_{j=2}^{\ell} \mathbb{1}_{\{\mathbf{X}_j \in R^p_n(\mathbf{u})\}}} \middle| \mathcal{D}_k\right] \\
&\le \sum_{p=1}^{2^M} \mathbb{E}\left[\frac{\mu\{R^p_n(\mathbf{u})\}}{\ell \mu\{R^p_n(\mathbf{u})\}} \middle| \mathcal{D}_k\right] \\
&\le \frac{2^M}{\ell}
\end{aligned}$$

(by the first statement of Lemma 4.4.1). Thus, returning to A_{n1}, we obtain

$$A_{n1} \le 2^M \mathbb{E}\left|\tilde{T}(\mathbf{r}_k(\mathbf{X}) - T(\mathbf{r}_k(\mathbf{X})))\right|^2 < 2^M \eta.$$

Computation of A_{n2}. For any $\delta > 0$, write

$$\begin{aligned}
A_{n2} &= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) |\tilde{T}(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))|^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) |\tilde{T}(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))|^2 \mathbb{1}_{\bigcup_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| > \delta\}}\right] \\
&\quad + \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) |\tilde{T}(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{T}(\mathbf{r}_k(\mathbf{X}))|^2 \mathbb{1}_{\bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| \le \delta\}}\right] \\
&\le 4 \sup_{\mathbf{u} \in \mathbb{R}^d} |\tilde{T}(\mathbf{r}_k(\mathbf{u}))|^2 \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) \mathbb{1}_{\bigcup_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| > \delta\}}\right] && (4.7) \\
&\quad + \Bigg(\sup_{\mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{u}) - r_{k,m}(\mathbf{v})| \le \delta\}} |\tilde{T}(\mathbf{r}_k(\mathbf{v})) - \tilde{T}(\mathbf{r}_k(\mathbf{u}))|\Bigg)^2. && (4.8)
\end{aligned}$$

With respect to the term (4.7), if $\delta > \varepsilon_\ell$, then

$$\begin{aligned}
&\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) \mathbb{1}_{\bigcup_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| > \delta\}} \\
&\quad = \sum_{i=1}^{\ell} \frac{\mathbb{1}_{\bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| \le \varepsilon_\ell\}} \mathbb{1}_{\bigcup_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)| > \delta\}}}{\sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M} \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_j)| \le \varepsilon_\ell\}}} \\
&\quad = 0.
\end{aligned}$$

It follows that, for all $\delta > 0$, this term converges to 0 as ℓ tends to infinity. On the other hand, letting $\delta \to 0$, we see that the term (4.8) tends to 0 as well, by uniform continuity of $\tilde{T}$. Hence, A_{n2} tends to 0 as ℓ tends to infinity. Letting finally η go to 0, we conclude that A_n vanishes as ℓ tends to infinity. □

Proposition 4.4.2. *Under the assumptions of Proposition 4.2.2,*

$$\lim_{\ell\to\infty} \mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(Y_i - T(\mathbf{r}_k(\mathbf{X}_i)))\right|^2 = 0.$$

Proof of Proposition 4.4.2.

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})(Y_i - T(\mathbf{r}_k(\mathbf{X}_i)))\right|^2 &= \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \mathbb{E}[W_{n,i}(\mathbf{X})W_{n,j}(\mathbf{X})(Y_i - T(\mathbf{r}_k(\mathbf{X}_i)))(Y_j - T(\mathbf{r}_k(\mathbf{X}_j)))] \\
&= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})|Y_i - T(\mathbf{r}_k(\mathbf{X}_i))|^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})\sigma^2(\mathbf{r}_k(\mathbf{X}_i))\right],
\end{aligned}$$

where

$$\sigma^2(\mathbf{r}_k(\mathbf{x})) = \mathbb{E}[|Y - T(\mathbf{r}_k(\mathbf{X}))|^2|\mathbf{r}_k(\mathbf{x})].$$

For any $\eta > 0$, using again Györfi et al. (2002, Theorem A.1), σ^2 can be approximated in an L^1 sense by a continuous function with compact support $\tilde{\sigma}^2$, i.e.,

$$\mathbb{E}|\tilde{\sigma}^2(\mathbf{r}_k(\mathbf{X})) - \sigma^2(\mathbf{r}_k(\mathbf{X}))| < \eta.$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})\sigma^2(\mathbf{r}_k(\mathbf{X}_i))\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})\tilde{\sigma}^2(\mathbf{r}_k(\mathbf{X}_i))\right] + \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})|\sigma^2(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{\sigma}^2(\mathbf{r}_k(\mathbf{X}_i))|\right] \\
&\le \sup_{\mathbf{u}\in\mathbb{R}^d} |\tilde{\sigma}^2(\mathbf{r}_k(\mathbf{u}))|\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})\right] + \mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|\sigma^2(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{\sigma}^2(\mathbf{r}_k(\mathbf{X}_i))|\right].
\end{aligned}$$

With the same argument as for A_{n1}, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X})|\sigma^2(\mathbf{r}_k(\mathbf{X}_i)) - \tilde{\sigma}^2(\mathbf{r}_k(\mathbf{X}_i))|\right] \le 2^M\eta.$$

Therefore, it remains to prove that $\mathbb{E}\left[\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X})\right] \to 0$ as $\ell \to \infty$. To this aim, fix $\delta > 0$, and note that

$$\begin{aligned}
\sum_{i=1}^{\ell} W_{n,i}^2(\mathbf{X}) &= \frac{\sum_{i=1}^{\ell} \mathbb{1}_{\cap_{m=1}^{M}\{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}}{\left(\sum_{j=1}^{\ell} \mathbb{1}_{\cap_{m=1}^{M}\{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_j)|\le\varepsilon_\ell\}}\right)^2} \\
&\le \min\left\{\delta, \frac{1}{\sum_{i=1}^{\ell} \mathbb{1}_{\cap_{m=1}^{M}\{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}}\right\} \\
&\le \delta + \frac{\mathbb{1}_{\left\{\sum_{i=1}^{\ell} \mathbb{1}_{\cap_{m=1}^{M}\{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}} > 0\right\}}}{\sum_{i=1}^{\ell} \mathbb{1}_{\cap_{m=1}^{M}\{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}}.
\end{aligned}$$

To complete the proof, we have to establish that the expectation of the right-hand term tends to 0. Denoting by I a bounded interval on the real line, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{1}_{\left\{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}>0\right\}}}{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}}\right]\\
&\le\mathbb{E}\left[\frac{\mathbb{1}_{\left\{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}>0\right\}}\mathbb{1}_{\left\{\mathbf{X}\in\bigcap_{m=1}^{M}r_{k,m}^{-1}(I)\right\}}}{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}}\right]+\mu\left(\bigcup_{m=1}^{M}r_{k,m}^{-1}(I^c)\right)\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}_{\left\{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}>0\right\}}\mathbb{1}_{\left\{\mathbf{X}\in\bigcap_{m=1}^{M}r_{k,m}^{-1}(I)\right\}}}{\sum_{i=1}^{\ell}\mathbb{1}_{\left\{\mathbf{X}_i\in\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell])\right\}}}\middle|\mathcal{D}_k,\mathbf{X}\right]\right]+\mu\left(\bigcup_{m=1}^{M}r_{k,m}^{-1}(I^c)\right)\\
&\le\frac{2}{(\ell+1)}\mathbb{E}\left[\frac{\mathbb{1}_{\left\{\mathbf{X}\in\bigcap_{m=1}^{M}r_{k,m}^{-1}(I)\right\}}}{\mu(\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell]))}\right]+\mu\left(\bigcup_{m=1}^{M}r_{k,m}^{-1}(I^c)\right).
\end{aligned}
$$

The last inequality arises from the second statement of Lemma 4.4.1. By an appropriate choice of I, according to the technical statement (4.2), the second term on the right-hand side can be made as small as desired. Regarding the first term, there exists a finite number N_ℓ of points $\mathbf{z}_1,\dots,\mathbf{z}_{N_\ell}$ such that

$$
\bigcap_{m=1}^{M}r_{k,m}^{-1}(I)\subset\bigcup_{(j_1,\dots,j_M)\in\{1,\dots,N_\ell\}^M}r_{k,1}^{-1}(I_{n,1}(\mathbf{z}_{j_1}))\cap\cdots\cap r_{k,M}^{-1}(I_{n,M}(\mathbf{z}_{j_M})),
$$

where $I_{n,m}(\mathbf{z}_j)=[\mathbf{z}_j-\varepsilon_\ell/2,\mathbf{z}_j+\varepsilon_\ell/2]$. Suppose, without loss of generality, that the sets

$$
r_{k,1}^{-1}(I_{n,1}(\mathbf{z}_{j_1}))\cap\cdots\cap r_{k,M}^{-1}(I_{n,M}(\mathbf{z}_{j_M}))
$$

are ordered, and denote by R_n^p the p-th among the $N_\ell^M=(\lceil|I|/\varepsilon_\ell\rceil)^M$ sets. Here $|I|$ denotes the length of the interval I and $\lceil x\rceil$ denotes the smallest integer greater than x. For all p,

$$
\mathbf{x}\in R_n^p\Rightarrow R_n^p\subset\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{x})-\varepsilon_\ell,r_{k,m}(\mathbf{x})+\varepsilon_\ell]).
$$

Indeed, if $\mathbf{v}\in R_n^p$, then, for all $m=1,\dots,M$, there exists $j\in\{1,\dots,N_\ell\}$ such that $r_{k,m}(\mathbf{v})\in[\mathbf{z}_j-\varepsilon_\ell/2,\mathbf{z}_j+\varepsilon_\ell/2]$, that is $\mathbf{z}_j-\varepsilon_\ell/2\le r_{k,m}(\mathbf{v})\le\mathbf{z}_j+\varepsilon_\ell/2$. Since we also have $\mathbf{z}_j-\varepsilon_\ell/2\le r_{k,m}(\mathbf{X})\le\mathbf{z}_j+\varepsilon_\ell/2$, we obtain $r_{k,m}(\mathbf{X})-\varepsilon_\ell\le r_{k,m}(\mathbf{v})\le r_{k,m}(\mathbf{X})+\varepsilon_\ell$. In conclusion,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{1}_{\left\{\mathbf{X}\in\bigcap_{m=1}^{M}r_{k,m}^{-1}(I)\right\}}}{\mu(\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell]))}\right]\\
&\le\sum_{p=1}^{N_\ell^M}\mathbb{E}\left[\frac{\mathbb{1}_{\{\mathbf{X}\in R_n^p\}}}{\mu(\bigcap_{m=1}^{M}r_{k,m}^{-1}([r_{k,m}(\mathbf{X})-\varepsilon_\ell,r_{k,m}(\mathbf{X})+\varepsilon_\ell]))}\right]\\
&\le\sum_{p=1}^{N_\ell^M}\mathbb{E}\left[\frac{\mathbb{1}_{\{\mathbf{X}\in R_n^p\}}}{\mu(R_n^p)}\right]\\
&=N_\ell^M\\
&=\left\lceil\frac{|I|}{\varepsilon_\ell}\right\rceil^M.
\end{aligned}
$$

The result follows from the assumption $\lim_{\ell\to\infty} \ell\varepsilon_\ell^M = \infty$. □

Proposition 4.4.3. *Under the assumptions of Proposition 4.2.2,*

$$\lim_{\ell\to\infty} \mathbb{E}\left|\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) - 1\right) T(\mathbf{r}_k(\mathbf{X}))\right|^2 = 0.$$

Proof of Proposition 4.4.3. Since $|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) - 1| \le 1$, one has

$$\left|\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) - 1\right) T(\mathbf{r}_k(\mathbf{X}))\right|^2 \le T^2(\mathbf{r}_k(\mathbf{X})).$$

Consequently, by Lebesgue's dominated convergence theorem, to prove the proposition, it suffices to show that $W_{n,i}(\mathbf{X})$ tends to 1 almost surely. Now,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) \neq 1\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{\ell} \mathbb{1}_{\cap_{m=1}^M \{|r_{k,m}(\mathbf{X}) - r_{k,m}(\mathbf{X}_i))| \le \varepsilon_\ell\}} = 0\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{\ell} \mathbb{1}_{\left\{\mathbf{X}_i \in \cap_{m=1}^M r_{k,m}^{-1}([r_{k,m}(\mathbf{X}) - \varepsilon_\ell, r_{k,m}(\mathbf{X}) + \varepsilon_\ell])\right\}} = 0\right) \\
&= \int_{\mathbb{R}^d} \mathbb{P}\left(\forall i = 1,\dots,\ell, \mathbb{1}_{\left\{\mathbf{X}_i \in \cap_{m=1}^M r_{k,m}^{-1}([r_{k,m}(\mathbf{x}) - \varepsilon_\ell, r_{k,m}(\mathbf{x}) + \varepsilon_\ell])\right\}} = 0\right) \mu(\mathrm{d}\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \left[1 - \mu(\cap_{m=1}^M r_{k,m}^{-1}([r_{k,m}(\mathbf{x}) - \varepsilon_\ell, r_{k,m}(\mathbf{x}) + \varepsilon_\ell]))\right]^\ell \mu(\mathrm{d}\mathbf{x}).
\end{aligned}$$

Denote by I a bounded interval. Then,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) \neq 1\right) \\
&\le \int_{\mathbb{R}^d} \exp\left(-\ell\mu(\cap_{m=1}^M r_{k,m}^{-1}([r_{k,m}(\mathbf{x}) - \varepsilon_\ell, r_{k,m}(\mathbf{x}) + \varepsilon_\ell]))\right) \\
&\quad \times \mathbb{1}_{\{\mathbf{x} \in \cap_{m=1}^M r_{k,m}^{-1}(I)\}} \mu(\mathrm{d}\mathbf{x}) + \mu\left(\bigcup_{m=1}^M r_{k,m}^{-1}(I^c)\right) \\
&\le \max_{\mathbf{u}} \mathbf{u} e^{-\mathbf{u}} \int_{\mathbb{R}^d} \frac{\mathbb{1}_{\{\mathbf{x} \in \cap_{m=1}^M r_{k,m}^{-1}(I)\}}}{\ell\mu(\cap_{m=1}^M r_{k,m}^{-1}([r_{k,m}(\mathbf{x}) - \varepsilon_\ell, r_{k,m}(\mathbf{x}) + \varepsilon_\ell]))} \mu(\mathrm{d}\mathbf{x}) + \mu\left(\bigcup_{m=1}^M r_{k,m}^{-1}(I^c)\right).
\end{aligned}$$

Using the same arguments as in the proof of Proposition 4.4.2, we may bound the probability $\mathbb{P}\left(\sum_{i=1}^{\ell} W_{n,i}(\mathbf{X}) \neq 1\right)$ by $\frac{e^{-1}}{\ell}\left\lceil \frac{|I|}{\varepsilon_\ell}\right\rceil^M$. This bound vanishes as n tends to infinity since, by assumption, $\lim_{\ell\to\infty} \ell\varepsilon_\ell^M = \infty$. □

4.4.3 Proof of Theorem 4.2.1

Choose $\mathbf{x} \in \mathbb{R}^d$. An easy calculation yields that

$$\begin{aligned}
&\mathbb{E}[|T_n(\mathbf{r}_k(\mathbf{x})) - T(\mathbf{r}_k(\mathbf{x}))|^2 | \mathbf{r}_k(\mathbf{X}_1), \dots, \mathbf{r}_k(\mathbf{X}_\ell), \mathcal{D}_k] \\
&= \mathbb{E}\left[\left|T_n(\mathbf{r}_k(\mathbf{x})) - \mathbb{E}[T_n(\mathbf{r}_k(\mathbf{x})) | \mathbf{r}_k(\mathbf{X}_1), \dots, \mathbf{r}_k(\mathbf{X}_\ell), \mathcal{D}_k]\right|^2 | \mathbf{r}_k(\mathbf{X}_1), \dots, \mathbf{r}_k(\mathbf{X}_\ell), \mathcal{D}_k\right] \\
&\quad + \left|\mathbb{E}[T_n(\mathbf{r}_k(\mathbf{x})) | \mathbf{r}_k(\mathbf{X}_1), \dots, \mathbf{r}_k(\mathbf{X}_\ell), \mathcal{D}_k] - T(\mathbf{r}_k(\mathbf{x}))\right|^2 \\
&:= E_1 + E_2. \qquad (4.9)
\end{aligned}$$

On the one hand, we have

$$
\begin{aligned}
E_1 &= \mathbb{E}\Big[\big|T_n(\mathbf{r}_k(\mathbf{x})) - \mathbb{E}[T_n(\mathbf{r}_k(\mathbf{x}))|\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k]\big|^2|\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k\Big] \\
&= \mathbb{E}\left[\left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{x})(Y_i - \mathbb{E}[Y_i|\mathbf{r}_k(\mathbf{X}_i)])\right|^2 |\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k\right].
\end{aligned}
$$

Developing the square and noticing that $\mathbb{E}\big[Y_j|Y_i,\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k\big] = \mathbb{E}[Y_j|\mathbf{r}_k(\mathbf{X}_j)]$, since Y_j is independent of Y_i and of the X_j's with $j \neq i$, we have

$$
\begin{aligned}
E_1 &= \mathbb{E}\left[\frac{\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}|Y_i - \mathbb{E}[Y_i|\mathbf{r}_k(\mathbf{X}_i)]|^2}{\left|\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}\right|^2}\middle|\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k\right] \\
&= \sum_{i=1}^{\ell} \mathbb{V}(Y_i|\mathbf{r}_k(\mathbf{X}_i))\frac{\mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}}{\left|\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}\right|^2}.
\end{aligned}
$$

Thus,

$$
E_1 \le 4R^2 \frac{\mathbb{1}_{\left\{\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}>0\right\}}}{\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}}, \tag{4.10}
$$

where $\mathbb{V}(Z)$ denotes the variance of a random variable Z. On the other hand, and recalling the notation Σ introduced in Section 4.3, we obtain for the second term E_2:

$$
\begin{aligned}
E_2 &= \big|\mathbb{E}[T_n(\mathbf{r}_k(\mathbf{x}))|\mathbf{r}_k(\mathbf{X}_1),\dots,\mathbf{r}_k(\mathbf{X}_\ell),\mathcal{D}_k] - T(\mathbf{r}_k(\mathbf{x}))\big|^2 \\
&= \left|\sum_{i=1}^{\ell} W_{n,i}(\mathbf{x})\mathbb{E}[Y_i|\mathbf{r}_k(\mathbf{X}_i)] - T(\mathbf{r}_k(\mathbf{x}))\right|^2 \mathbb{1}_{\{\Sigma>0\}} + T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}} \\
&\le \frac{\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}|\mathbb{E}[Y_i|\mathbf{r}_k(\mathbf{X}_i)] - T(\mathbf{r}_k(\mathbf{x}))|^2}{\sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\le\varepsilon_\ell\}}}\mathbb{1}_{\{\Sigma>0\}} \qquad (4.11) \\
&\quad + T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}} \\
&\quad \text{(by Jensen's inequality)} \\
&= \frac{\sum_{i=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\le\varepsilon_\ell\}}|T(\mathbf{r}_k(\mathbf{X}_i)) - T(\mathbf{r}_k(\mathbf{x}))|^2}{\sum_{j=1}^{\ell} \mathbb{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\le\varepsilon_\ell\}}}\mathbb{1}_{\{\Sigma>0\}} \qquad (4.12) \\
&\quad + T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}} \\
&\le L^2\varepsilon_\ell^2 + T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}}. \qquad (4.13)
\end{aligned}
$$

Now,

$$
\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \le \int_{\mathbb{R}^d} \mathbb{E}|(T_n(\mathbf{r}_k(\mathbf{x})) - T(\mathbf{r}_k(\mathbf{x}))|^2\mu(\mathrm{d}\mathbf{x}).
$$

Then, using the decomposition (4.9) and the upper bounds (4.10) and (4.13),

$$
\begin{aligned}
&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \\
&\le \int_{\mathbb{R}^d} \mathbb{E}\left[\frac{4R^2\mathbb{1}_{\{\Sigma>0\}}}{B}\right]\mu(\mathrm{d}\mathbf{x}) + L^2\varepsilon_\ell^2 + \int_{\mathbb{R}^d} \mathbb{E}\left[T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}}\right]\mu(\mathrm{d}\mathbf{x}) \\
&\le \int_{\mathbb{R}^d} \mathbb{E}\left\{\mathbb{E}\left[\frac{4R^2\mathbb{1}_{\{\Sigma>0\}}}{B}\middle|\mathcal{D}_k\right]\right\}\mu(\mathrm{d}\mathbf{x}) + L^2\varepsilon_\ell^2 + \int_{\mathbb{R}^d} \mathbb{E}\left\{\mathbb{E}\left[T^2(\mathbf{r}_k(\mathbf{x}))\mathbb{1}_{\{\Sigma=0\}}|\mathcal{D}_k\right]\right\}\mu(\mathrm{d}\mathbf{x}).
\end{aligned}
$$

Thus, thanks to Lemma 4.4.1,

$$\begin{aligned}
&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \\
&\quad \le \frac{8R^2}{(\ell+1)} \int_{\mathbb{R}^d} \frac{1}{\mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})} \mu(\mathrm{d}\mathbf{x}) + L^2\varepsilon_\ell^2 \\
&\qquad + \int_{\mathbb{R}^d} T^2(\mathbf{r}_k(\mathbf{x})) \left(1 - \mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})\right)^\ell \mu(\mathrm{d}\mathbf{x}).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \\
&\quad \le \frac{8R^2}{(\ell+1)} \int_{\mathbb{R}^d} \frac{1}{\mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})} \mu(\mathrm{d}\mathbf{x}) + L^2\varepsilon_\ell^2 \\
&\qquad + \int_{\mathbb{R}^d} T^2(\mathbf{r}_k(\mathbf{x})) \exp\left(-\ell\mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})\right) \mu(\mathrm{d}\mathbf{x}) \\
&\quad \le \frac{8R^2}{(\ell+1)} \int_{\mathbb{R}^d} \frac{1}{\mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})} \mu(\mathrm{d}\mathbf{x}) + L^2\varepsilon_\ell^2 \\
&\qquad + \Bigg(\sup_{\mathbf{x}\in\mathbb{R}^d} T^2(\mathbf{r}_k(\mathbf{x})) \max_{\mathbf{u}\in\mathbb{R}^+} \mathbf{u}e^{-\mathbf{u}} \\
&\qquad\qquad \times \int_{\mathbb{R}^d} \frac{1}{\ell\mu(\bigcap_{m=1}^M \{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X})| \le \varepsilon_\ell\})} \mu(\mathrm{d}\mathbf{x}) \Bigg).
\end{aligned}$$

Introducing a bounded interval I as in the proof of Proposition 4.2.2, we observe that the boundedness of the $\mathbf{r}_k$ yields that

$$\mu\left(\bigcup_{m=1}^M r_{k,m}^{-1}(I^c)\right) = 0,$$

as soon as I is sufficiently large, independently of k. Then, proceeding as in the proof of Proposition 4.2.2, we obtain

$$\begin{aligned}
&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \\
&\quad \le 8R^2 \left\lceil \frac{|I|}{\varepsilon_\ell} \right\rceil^M \frac{1}{\ell+1} + L^2\varepsilon_\ell^2 + R^2 \max_{\mathbf{u}\in\mathbb{R}^+} \mathbf{u}e^{-\mathbf{u}} \left\lceil \frac{|I|}{\varepsilon_\ell} \right\rceil^M \frac{1}{\ell} \\
&\quad \le C_1 \frac{R^2}{\ell\varepsilon_\ell^M} + L^2\varepsilon_\ell^2,
\end{aligned}$$

for some positive constant C_1, independent of k. Hence, for the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$, we obtain

$$\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 \le C\ell^{-\frac{2}{M+2}},$$

for some positive constant C depending on L, R and independent of k, as desired.

Chapitre 5

Estimating the Location and Shape of Hybrid Zones

Abstract. We propose a new model to make use of geo-referenced genetic data for inferring the location and shape of a hybrid zone. The model output includes the posterior distribution of a parameter that quantifies the width of the hybrid zone. The model proposed is implemented in the GUI and command-line versions of the GENELAND program versions ≥ 3.3.0. Information about the program can be found on http://www2.imm.dtu.dk/~gigu/Geneland/

This chapter benefited greatly from review comments of Stuart J. E. Baird, two anonymous reviewers and the Subject Editor. This work was partially supported by a grant of the Danish Centre for Scientific Computing and a grant of Agence Nationale de la Recherche (ANR-09-BLAN-0145-01).

Contents

This chapter has been published in *Molecular Ecology Resources* (Guedj and Guillot, 2011).

5.1 Background

Hybrid zones have been the object of considerable attention as they are seen as *windows on the evolutionary process* (Harrison, 1990) and inference about genetic structure in their neighbourhood can provide valuable insights about the intensity of selection. This is made possible through the existence of explicit models of cline shapes as a function of selection (Haldane, 1948; Bazykin, 1969; Kruuk et al., 1999). To analyse hybrid zones, scientists have relied on a variety of approaches. They can use hybrid zones models that predict patterns of

allele frequencies and fit corresponding parametric curves (ANALYSE program, Barton and Baird, 1998) or nonparametric curves (Macholán et al., 2008). They can also use general purpose computer programs such as STRUCTURE (Pritchard et al., 2000) that seek patterns in ancestries of individuals without reference to any model of hybrid zones. Here we propose a new spatial model that combines features of both approaches: it explicitly accounts for the presence of a cline without making restrictive assumption about the shape of the cline path and it also retains the flexibility of the admixture model of STRUCTURE.

5.2 Model

We assume that individuals in the dataset at hand have alleles with origins in K distinct gene pools characterised by different allele frequencies. We denote by $z = (z_{i\ell})$ the matrix of genotype data where $z_{i\ell}$ denotes the genotype of individual i at locus ℓ and by $f_{k\ell a}$ the frequency of allele a at locus ℓ in the k-th gene pool. We introduce the matrix $q = (q_{ik})$, where q_{ik} refers to individual i's genome proportion originating from cluster k. For diploid individuals and assuming statistical independence of the two alleles harboured on the same locus of homologous chromosomes we have

$$\mathcal{L}(z_{i\ell}|f,q) = \sum_{k=1}^{K} q_{ik} f_{k\ell z_{i\ell 1}} f_{k\ell z_{i\ell 2}} (2 - \delta_{z_{i\ell 1}}^{z_{i\ell 2}}),$$

where δ_a^b is the Kronecker symbol, *i.e.*, $\delta_a^b = 1$ if $a = b$ and 0 otherwise. For haploid data we have

$$\mathcal{L}(z_{i\ell}|f,q) = \sum_{k=1}^{K} q_{ik} f_{k\ell z_{i\ell}}.$$

Further, assuming independence across the different loci, we have

$$\mathcal{L}(z|f,q) = \prod_{i=1}^{n} \prod_{\ell=1}^{L} \mathcal{L}(z_{i\ell}|f,q).$$

This is the classical admixture likelihood assumed in the STRUCTURE program and related works. We assume further that each gene pool (or cluster) occupies a certain fraction of the spatial domain. The spatial domain of each cluster is assumed to display a certain organisation in the sense that the various clusters do not overlap too much in space. This is accounted for by a so-called coloured Poisson-Voronoi tessellation which is the spatial model implemented in the GENELAND program. An example is given on Figure 5.1. The reader unfamiliar with this model is invited to refer to Guillot et al. (2005) and Guillot et al. (2009) for a detailed presentation. See Section 5.A for details about how the novel part of the model connects to earlier versions of the GENELAND program. The model introduced here differs from earlier versions of GENELAND in that it models admixture and from STRUCTURE in that it is spatial. Those two features are accounted for as follows: each vector of admixture proportions $q_{i.} = (q_{ik})_{k=1}^{K}$ is assumed to follow a Dirichlet distribution $\mathcal{D}(\alpha_{i1}, ..., \alpha_{iK})$. We denote by d_{ik} the distance of individual i to cluster k (in particular, $d_{ik} = 0$ if individual i has been sampled in cluster k) and we assume a deterministic relationship

$$\alpha_{ik} = a \exp(-d_{ik}/b). \tag{5.1}$$

By a standard property of the Dirichlet distribution, under equation (5.1) the expected value of q_{ik} is

$$\mathbb{E}[q_{ik}] = \frac{e^{-d_{ik}/b}}{\sum_k e^{-d_{ik}/b}}.$$

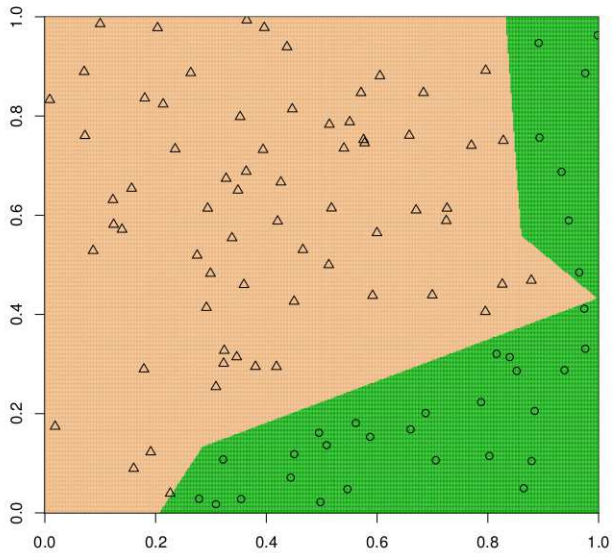


FIGURE 5.1 – Example of $K = 2$ spatial clusters simulated from a coloured Poisson-Voronoi prior model. The points represent putative sampling sites of individuals (shapes represent cluster membership). The realization of the Poisson process governing the tessellation is not shown for clarity.

In the presence of $K = 2$ clusters in contact along a hybrid zone, and if individual i belongs to cluster 1, then by definition $d_{i1} = 0$ and we get

$$\begin{aligned}\mathbb{E}[q_{i1}] &= \frac{e^{-d_{i1}/b}}{e^{-d_{i1}/b} + e^{-d_{i2}/b}} \\ &= \frac{1}{1 + e^{-d_{i2}/b}},\end{aligned}$$

i.e., the well known sigmoid function (or logistic function , cf e.g. Cramer, 2003) familiar to people studying hybrid zones, which is also equivalent to the hyperbolic tangent cline model described by Bazykin (1969):

$$\frac{1}{2}(1 + \tanh(d)) = \frac{1}{1 + e^{-2d/b}}.$$

Under this model, the width of the cline (defined as the inverse of the maximum gradient) is $w = 4b$. The variation of the expected admixture coefficients is illustrated in Figure 5.2.

Parameter a is a-dimensional, it does not affect the expected value of q_{ik} but controls its variance with $V[q_{ik}] \propto 1/a$. Large a values correspond to datasets with individuals displaying pretty similar admixture proportions within clusters. Parameter b is a spatial scale parameter, it has the dimension of a distance and is expressed in the same unit as spatial coordinates. Large b values correspond to situations where admixture coefficients are loosely structured in space. At the limit where $b = +\infty$, the vector $q_{i.}$ follows a flat Dirichlet distribution and the model does not display spatial features at all. Conversely, at the limit value $b = 0$, all individuals display admixture proportions that are 0 or 1 with a spatial pattern mirroring exactly the underlying Poisson-Voronoi tessellation. In all subsequent analyses and in our program, we place a uniform prior on a and b and assume independence of these two parameters. Section 5.A contains details on the inference algorithm.

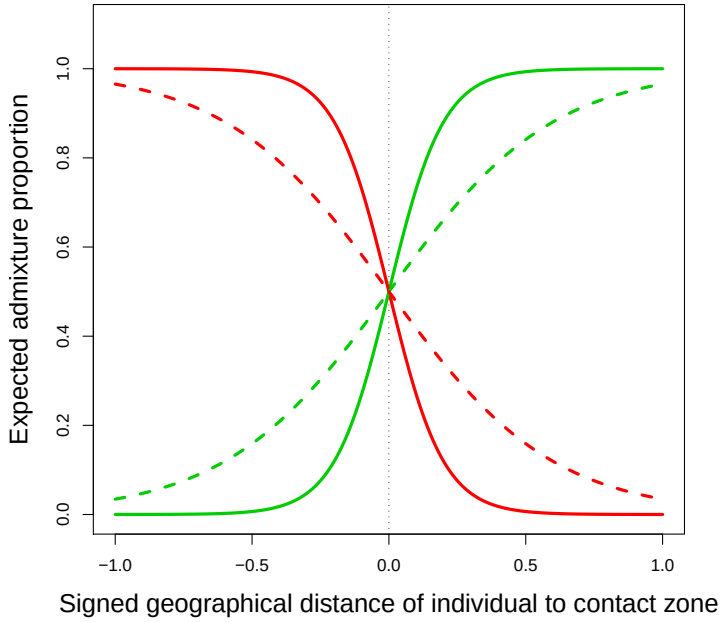


FIGURE 5.2 – Examples of spatial variation of expected admixture proportions in presence of two clusters. Individuals whose proportions are displayed here are assumed to be continuously located along a linear transect crossing perpendicularly a hybrid zone. Decreasing curves: expected admixture proportion q_{i1}. Increasing curves: expected admixture proportion q_{i2}. Continuous lines: $a = 1$, $b = 0.1$, dashed lines: $a = 1$, $b = 0.3$. Note that the curves are exactly sigmoid (logistic) functions.

5.3 Test of the method on simulated data

To test the efficiency of our approach, we carried out MCMC inference on data produced by simulation under our model. We explored various situations in terms of variance ($\propto 1/a$) and spatial scale (b) of admixture proportions but also in terms of number of loci L. In all cases the dataset consists of 200 individuals belonging to $K = 2$ different clusters located on a $[0,1] \times [0,1]$ square. We explored a broad range of pairwise cluster differentiations as measured by F_{ST} (lower quartile $F_{\text{ST}} = 0.003$, upper quartile 0.03). Some graphical examples of inference are presented in Figure 5.3 and Figure 5.4. Our main numerical results are summarized on Table 5.1. It appears that our method is accurate even for moderate to small numbers of loci ($L = 20$ or $L = 10$). We also note that the accuracy decreases when b increases (*i.e.*, in case of loose spatial structure) which is the price to pay for using a spatial model. Another observed loss of accuracy (not shown here) occurs when the spatial scale of the cline is smaller than the resolution of the spatial sampling. In the extreme case when the width of the cline is smaller than the smallest inter-distance between individuals, no reliable inference of b can be made. This means that users must have an idea of the characteristic scale of the cline before sampling.

TABLE 5.1 – Mean square error in the inference of admixture proportions. Data are generated by simulation from our prior-likelihood model. Each number is obtained as an average over ten independent datasets.

	L = 10			L = 20			L = 50		
	a = 1	a = 2	a = 5	a = 1	a = 2	a = 5	a = 1	a = 2	a = 5
b = 0.05	0.028	0.028	0.027	0.019	0.015	0.010	0.005	0.012	0.006
b = 0.1	0.038	0.030	0.029	0.019	0.021	0.015	0.007	0.012	0.015
b = 0.3	0.036	0.036	0.027	0.022	0.035	0.018	0.009	0.010	0.008
b = 0.7	0.046	0.031	0.020	0.022	0.022	0.016	0.010	0.009	0.015

5.4 Discussion

Our model of clinal variation is the same model as in MacCallum et al. (1998), the equivalent options in ANALYSE are for 2D spatial analysis with constant cline width along the course of the zone centre and a sigmoid cline cross section. The difference between this existing method and our global model is that the former is constrained by a very simple model of the path of the zone centre through space, the limitations of which are discussed at length by Bridle et al. (2001). This difference highlights one of the properties of our work: placing an explicit clinal admixture model in the context of the GENELAND Voronoi tessellation approach—which is reminiscent of the approach taken by Macholán et al. (2011)—removes the existing unrealistic restriction for modelling the course of a hybrid zone centre through a 2D field area (although it does not allow for cline width to vary along the course of the hybrid zone).

ANALYSE requires the user to *a priori* reduce multi-allelic loci to two states, corresponding to origin in two source clusters. The frequency of these two states in the source clusters can be co-estimated with cline parameters, however, the reduction to two states very much reduces ANALYSE's applicability to, for example, micro-satellite data, as a posteriori the user cannot for example quantify which allelic states are most associated with each source. In contrast, STRUCTURE co-assigns allelic state to source while estimating their frequencies in clusters, making micro-satellites easy to use, but of course there is no spatial model. In this sense, our work combines aspects of each approach, allowing frequencies of multi-allelic allelic states to be co-estimated with cline parameters in a spatial explicit way.

Our model has the direct advantage over STRUCTURE to explicitly model the presence of a hybrid zone and therefore to allow one to estimate its width and the intensity of selection or the age of contact, at least in the case of sigmoid clines. For a discussion of sigmoid versus stepped clines (see Kruuk et al., 1999). However we note that in contrast with STRUCTURE that explicitly models admixture linkage disequilibrium (Falush et al., 2003), our model assumes independence among loci. Associations (LD) between loci under selection lead to a different class of clinal model—stepped clines—not considered here (see Kruuk et al., 1999).

Durand et al. (2009) proposed an admixture model also based on spatially varying admixture coefficients involving the Dirichlet distribution. It is a general-purpose model that can be justified whenever spatial structure of admixture coefficient is expected. However, it is not specifically tailored for the study of hybrid zones (even though it has been presented in the context). Indeed, their approach does not explicitly model the presence of a contact zone, or to use a mathematical phrasing, their model does not account for the existence of a singularity in space (the contact zone) of genetic variation. What makes the potential use-

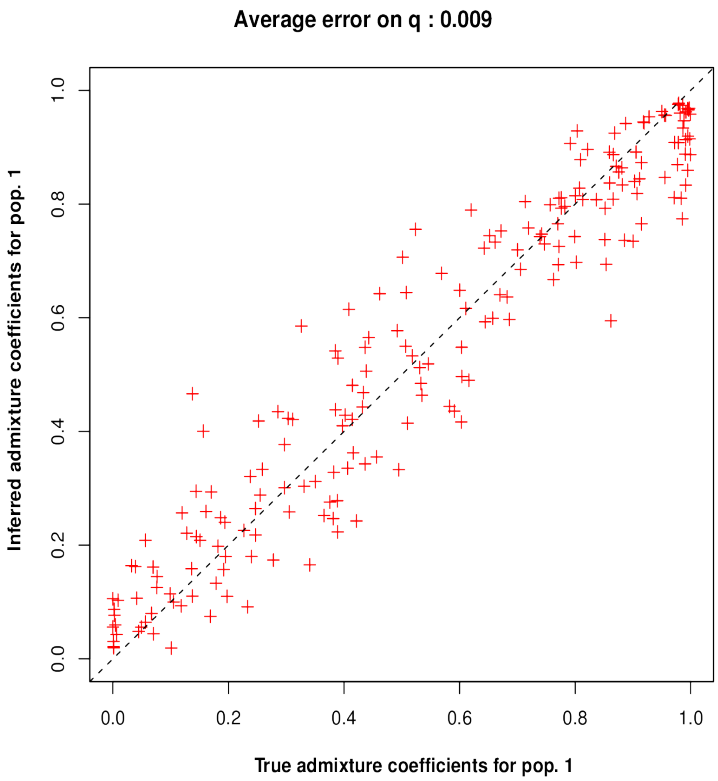


FIGURE 5.3 – Examples of result of inference: estimated versus true admixture proportions. The hyper-parameters of the admixture proportions were $a = 1$, $b = 0.3$.

fulness of their approach for the study of hybrid zones is its extreme flexibility, but it does not offer a straightforward way to estimate the width of the hybrid zone and the intensity of selection.

A second salient difference between our approach and that of Durand et al. (2009) is the inference machinery. We try to rely as much as possible on Bayesian estimators and therefore on MCMC, including for the estimation of the number of clusters (admittedly with a degree of approximation here) while they resort to likelihood or penalised likelihood methods. In this respect, the initial version of the TESS program (Chen et al., 2007) suffered from a number of flaws pointed out by Guillot (2009a,b). Even if the updated model of Durand et al. (2009) is an improvement in many respects over Chen et al. (2007), it still has some limitations. An obvious one is the impossibility to compare the scenario $K = 1$ against $K > 1$ which makes it impossible to test the null hypothesis of absence of structure. A recent study by Safner et al. (2011) suggests also that the new admixture model of Durand et al. (2009) may be less accurate than the old no-admixture model of Chen et al. (2007).

Our method allows evolutionists to make inference about the location and shape of hybrid zones. It should prove useful in particular in the case of secondary contact between weakly differentiated populations. However, as a final note, we stress that the spatial regression of admixture proportions does not capture all the complexity of hybrid zones: their semi-permeable nature, the fine scale discordance of clines, the interplay of various component of reproductive isolation etc... Admixture proportions and cline width are only a rough summary of how genomes intermix in hybrid zones and hybrid zones cannot simply be summarized by logistic variation of admixture proportions. We think the present model will be

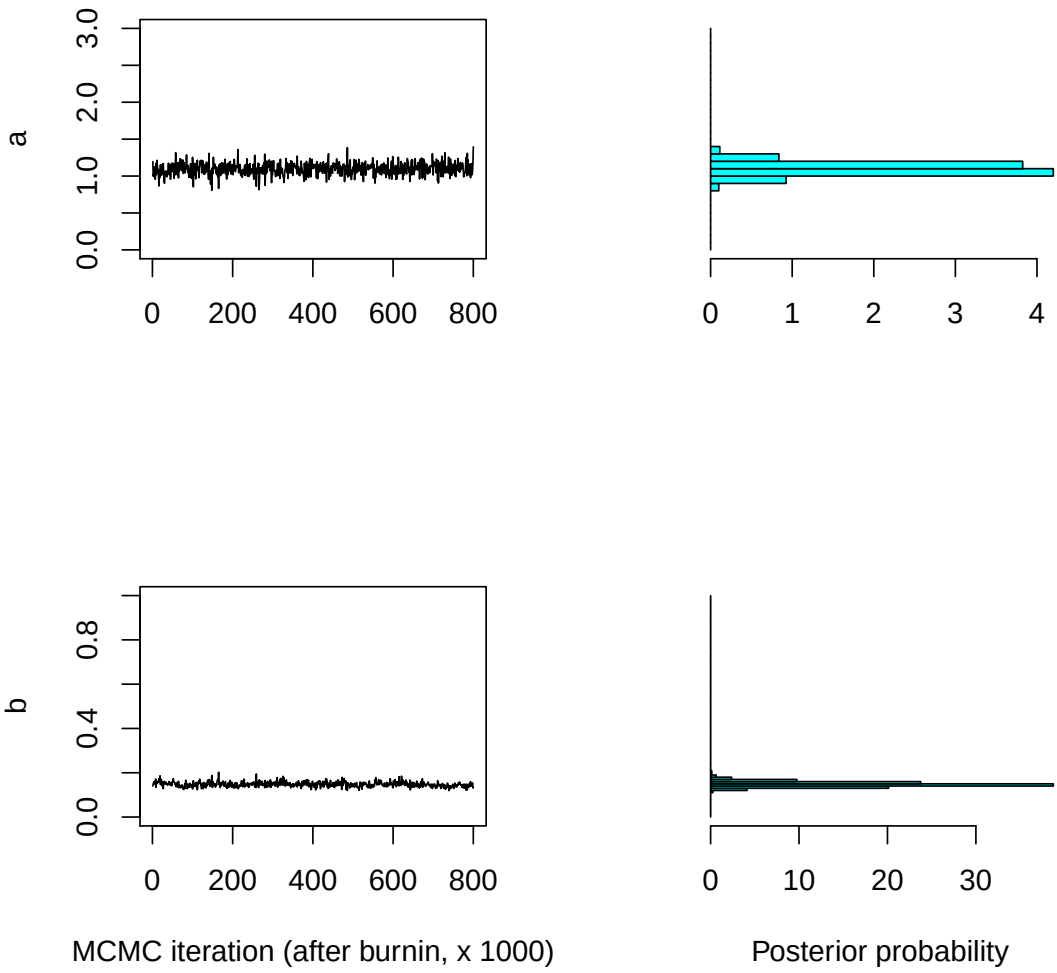


FIGURE 5.4 – Example of result of inference: MCMC trace (left) and posterior distribution (right) of parameters a and b.

of great help as a complementary procedure to estimate the course of hybrid zone centres and selection acting, at least in the case of sigmoid clines. However we also believe that it will not substitute for detailed analyses of cline shapes and departure from Hardy-Weinberg or linkage disequilibria traditionally conducted in hybrid zones.

5.A Supplements

5.A.1 Inference algorithm

The novel part of the model involves three blocks of parameters: the matrix of admixture proportions $q = (q_{ik})$, and the vector of parameters (a, b). An exact Bayesian inference would estimate them by joint MCMC simulation of (q, a, b) together with any other parameters involved in the model (number of gene pools, tessellation parameters and allele frequencies). We believe that the implementation of this strategy would offer a number of

numerical challenges, caused by the joint estimation of q and the number of clusters.

For this reason, we implement an alternative approximate two-stage strategy: first we estimate allele frequencies and cluster locations under the non-admixture model of Geneland. In a second step, we estimate (q,a,b) by MCMC simulation from the distribution of (q,a,b) conditioned by the data and the parameters estimates obtained from the non-admixture Geneland run.

5.A.2 Updates of q

We perform updates of $q_{i.}$ into $q^*_{i.}$ where $q^*_{i.}$ is obtained by perturbating two randomly chosen components, *i.e.*, $q^*_{ik_1} = q_{ik_1} + \delta$ and $q^*_{ik_2} = q_{ik_2} - \delta$. When δ is sampled from a symmetric distribution, the Metropolis-Hastings ratio is

$$R = \frac{\pi(z|q^*,...)}{\pi(z|q,...)} \frac{\pi(q^*|\alpha)}{\pi(q|\alpha)}.$$

The function $\pi(z|q,...)$ refers to the full conditional distribution of the data. The function $\pi(q|\alpha)$ is a product of Dirichlet densities.

5.A.3 Updates of a and b

We perform Metropolis-Hastings updates of a. With a symmetric proposal, the acceptance ratio is

$$R = \frac{\pi(q|\alpha^*)}{\pi(q|\alpha)} \frac{\pi(a^*)}{\pi(a)},$$

where $\alpha^*_{ik} = a^* \exp(-d_{ik}/b)$. We proceed similarly to update b.

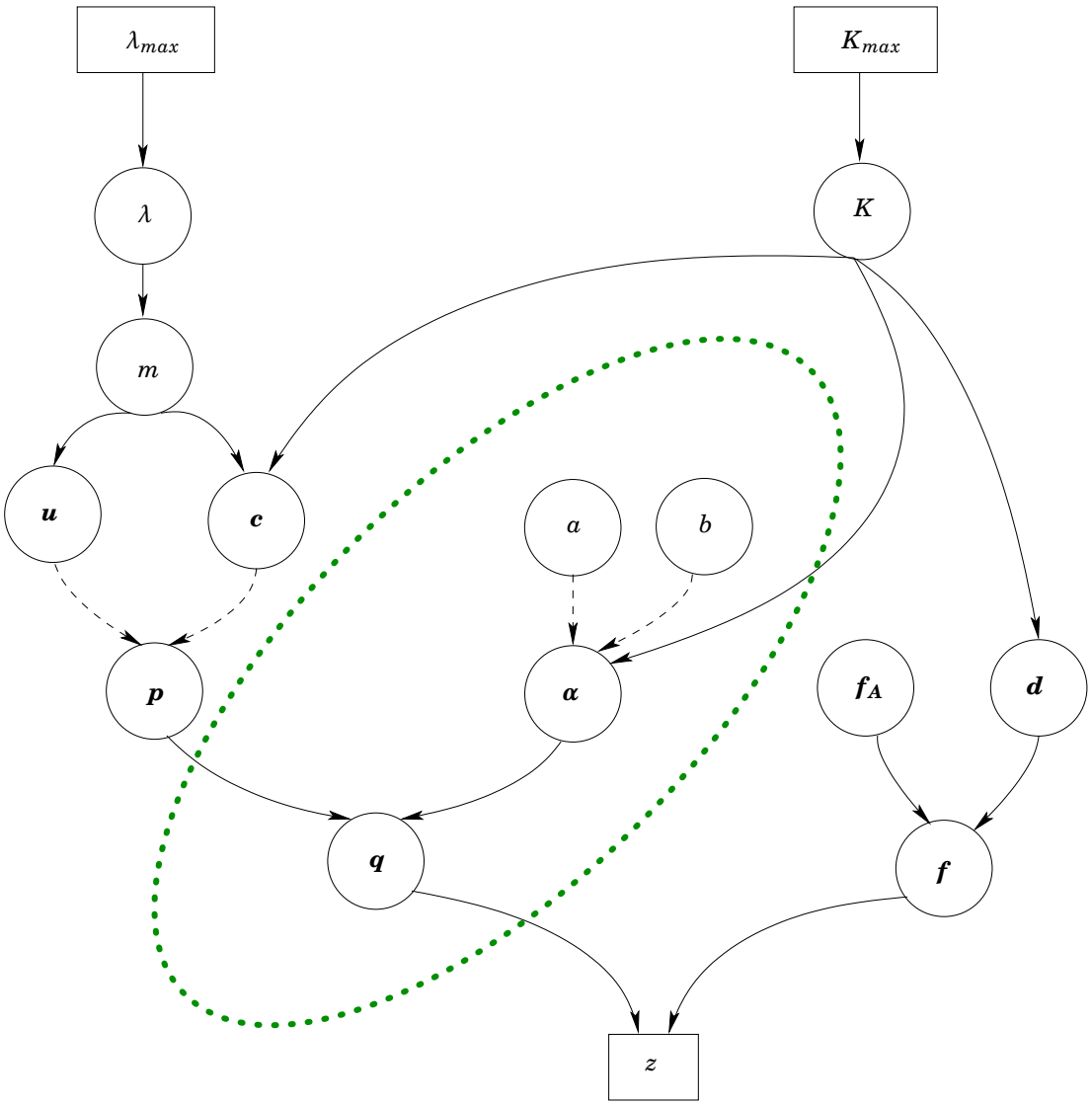


FIGURE 5.5 – Directed acyclic graph of proposed model. Continuous lines represent stochastic dependencies, dashed lines represent deterministic dependencies. Squared boxes enclose data or fixed hyper-parameters, rounded boxes enclose inferred parameters. The thick dotted line enclose the part of the model proposed that is novel. The other parts are borrowed to STRUCTURE or GENELAND.

Annexe A

Lemmes techniques

Une large part des résultats oracles démontrés dans les chapitres 2 et 3 repose sur le Lemme A.1, considérant la transformée de Legendre de la divergence de Kullback-Leibler, dont nous rappelons ici la définition : pour deux mesures μ_1 et μ_2, la divergence de μ_1 par rapport à μ_2 est notée $\mathcal{KL}(\mu_1,\mu_2)$ et définie par :

$$\mathcal{KL}(\mu_1,\mu_2) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)\mathrm{d}\mu_1 & \text{si } \mu_1 \ll \mu_2, \\ \infty & \text{sinon.} \end{cases}$$

Pour tout ensemble mesurable $(A,\mathcal{A})$ muni d'une mesure de probabilité μ, on désignera par $\mathcal{M}^1_{+,\mu}(A,\mathcal{A})$ l'ensemble des mesures de probabilité absolument continues par rapport à μ. Le résultat suivant a été formalisé par Catoni (2004, Equation 5.2.1). Nous en reproduisons la preuve dans un souci de complétude.

Lemme A.1. *Soit $(A,\mathcal{A})$ un ensemble mesurable. Pour toute mesure de probabilité μ sur $(A,\mathcal{A})$ et toute fonction mesurable $h : A \to \mathbb{R}$ telle que $\int(\exp\circ h)\mathrm{d}\mu < \infty$, il vient*

$$\log\int(\exp\circ h)\mathrm{d}\mu = \sup_{m\in\mathcal{M}^1_{+,\mu}(A,\mathcal{A})}\left\{\int h\mathrm{d}m - \mathcal{KL}(m,\mu)\right\},$$

avec la convention $\infty-\infty=-\infty$. De plus, si h est bornée sur le support de μ, le supremum par rapport à m dans le terme de droite est atteint pour la distribution de Gibbs g définie par

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int(\exp\circ h)\mathrm{d}\mu}, \quad a\in A.$$

Preuve. Soit m une mesure de probabilité sur $(A,\mathcal{A})$. En notant que $\mathcal{KL}(\cdot,\cdot)$ ne prend pas de valeurs négatives, il est clair que $m \mapsto -\mathcal{KL}(m,g)$ atteint son supremum en $m=g$ et ce supremum est nul. Par suite,

$$\begin{aligned} -\mathcal{KL}(m,g) &= -\int \log\left(\frac{\mathrm{d}m}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}g}\right)\mathrm{d}m \\ &= -\int \log\left(\frac{\mathrm{d}m}{\mathrm{d}\mu}\right)\mathrm{d}m + \int \log\left(\frac{\mathrm{d}g}{\mathrm{d}\mu}\right)\mathrm{d}m \\ &= -\mathcal{KL}(m,\mu) + \int h\mathrm{d}m - \log\int(\exp\circ h)\mathrm{d}\mu. \end{aligned}$$

En prenant le supremum sur toutes les mesures de probabilité sur $(A,\mathcal{A})$ des deux côtés de l'égalité, on obtient le résultat désiré. □

Les techniques PAC-bayésiennes reposent en grande partie sur l'utilisation d'inégalités de concentration pour dériver des inégalités oracles en probabilité. Dans le Chapitre 2, nous utilisons l'inégalité suivante, issue de Massart (2007, Proposition 2.9) et dont nous reproduisons également la preuve. Pour $x \in \mathbb{R}$, notons $(x)_+ = \max(x, 0)$.

Lemme A.2. *Soit $(T_i)_{i=1}^n$ une suite de variables aléatoires indépendantes à valeurs réelles. Supposons qu'il existe deux constantes positives v et w telles que*

$$\sum_{i=1}^n \mathbb{E}T_i^2 \leq v,$$

et pour tout entier $k \geq 3$,

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \leq \frac{k!}{2} v w^{k-2}.$$

Alors, pour tout $\gamma \in \left(0, \frac{1}{w}\right)$,

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^n (T_i - \mathbb{E}T_i)\right)\right] \leq \exp\left(\frac{v\gamma^2}{2(1-w\gamma)}\right).$$

Preuve. Considérons la fonction $\phi : u \mapsto \exp(u) - u - 1$. De façon triviale, pour tout $u \leq 0$, $\phi(u) \leq \frac{u^2}{2}$. Par suite, pour tout $\gamma > 0$ et tout entier $i = 1, \ldots, n$,

$$\phi(\gamma T_i) \leq \frac{\gamma^2 T_i^2}{2} + \sum_{k=3}^{\infty} \frac{\gamma^k (T_i)_+^k}{k!},$$

et

$$\mathbb{E}\phi(\gamma T_i) \leq \frac{\gamma^2 \mathbb{E}T_i^2}{2} + \sum_{k=3}^{\infty} \frac{\gamma^k \mathbb{E}[(T_i)_+^k]}{k!},$$

d'où

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\phi(\gamma T_i) &\leq \frac{\gamma^2}{2} \sum_{i=1}^n \mathbb{E}T_i^2 + \sum_{k=3}^{\infty} \frac{\gamma^k}{k!} \sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \\ &\leq \frac{v}{2} \sum_{k=2}^{\infty} \gamma^k w^{k-2}.\end{aligned}$$

Ce dernier terme est une série convergente si et seulement si $\gamma \in (0, 1/w)$ (les cas $\gamma = 0$ et $\gamma = 1/w$ impliquent des inégalités triviales), ce qui prouve l'intégrabilité des variables $\exp(\gamma T_i)$ pour les entiers $i = 1, \ldots, n$ dans ce cas. Comme $w\gamma < 1$, il vient

$$\sum_{k=2}^{\infty} \gamma^k w^{k-2} = \frac{\gamma^2}{1 - w\gamma}.$$

En utilisant l'inégalité élémentaire $\log u \leq u - 1$ quand $u > 0$, nous obtenons

$$\begin{aligned}\sum_{k=1}^n \mathbb{E}\phi(\gamma T_i) &= \sum_{k=1}^n \left(\mathbb{E}\exp(\gamma T_i) - \gamma \mathbb{E}T_i - 1\right) \\ &\geq \sum_{k=1}^n \left(\log \mathbb{E}\exp(\gamma T_i) - \gamma \mathbb{E}T_i\right) \\ &= \sum_{k=1}^n \log \mathbb{E}\exp(\gamma(T_i - \mathbb{E}T_i)).\end{aligned}$$

En assemblant ce qui précède, nous avons finalement

$$\sum_{k=1}^{n} \log \mathbb{E} \exp(\gamma(T_i - \mathbb{E}T_i)) \leq \frac{v\gamma^2}{2(1-w\gamma)}.$$

Comme $(T_i)_{1\leq i\leq n}$ est une suite de variables indépendantes, en prenant l'exponentielle de chaque côté de la dernière inégalité, on obtient le résultat. □

Enfin, les résultats du Chapitre 3 nécessitent une version du Lemme A.2 adaptée à la fonction de perte logistique.

Lemme A.3. *Considérons la fonction $\phi : u \mapsto \exp(u) - u - 1$. Soit $(T_i)_{i=1}^n$ une suite de variables aléatoires indépendantes à valeurs réelles. Supposons qu'il existe deux constantes positives v et w telles que*

$$\sum_{i=1}^{n} \mathbb{E}[T_i^2] \leq v,$$

et pour tout entier $k \geq 3$,

$$\sum_{i=1}^{n} \mathbb{E}[(T_i)_+^k] \leq v w^{k-2}.$$

Alors pour tout $\gamma > 0$,

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^{n} (T_i - \mathbb{E}T_i)\right)\right] \leq \exp\left(\frac{v\phi(\gamma w)}{w^2}\right).$$

Preuve. Il est clair que pour tout $u \leq 0$, $\phi(u) \leq \frac{u^2}{2}$. Par suite, pour tout $\gamma > 0$ et tout entier $i = 1, \ldots, n$,

$$\phi(\gamma T_i) \leq \frac{\gamma^2 T_i^2}{2} + \sum_{k=3}^{\infty} \frac{\gamma^k (T_i)_+^k}{k!},$$

ce qui entraîne

$$\sum_{i=1}^{n} \mathbb{E}\phi(\gamma T_i) \leq \frac{\gamma^2}{2} \sum_{i=1}^{n} \mathbb{E}T_i^2 + \sum_{k=3}^{\infty} \frac{\gamma^k}{k!} \sum_{i=1}^{n} \mathbb{E}[(T_i)_+^k] \leq v \sum_{k=2}^{\infty} \frac{\gamma^k w^{k-2}}{k!} = \frac{v}{w^2}\phi(\gamma w).$$

En utilisant l'inégalité élémentaire $\log u \leq u - 1$ quand $u > 0$, nous obtenons

$$\begin{aligned}\sum_{k=1}^{n} \mathbb{E}\phi(\gamma T_i) = \sum_{k=1}^{n} \left[\mathbb{E}\exp(\gamma T_i) - \gamma \mathbb{E}T_i - 1\right] &\geq \sum_{k=1}^{n} \left[\log \mathbb{E}\exp(\gamma T_i) - \gamma \mathbb{E}T_i\right] \\ &= \sum_{k=1}^{n} \log \mathbb{E}\exp\left[\gamma(T_i - \mathbb{E}T_i)\right].\end{aligned}$$

Finalement, en utilisant ce qui précède,

$$\sum_{k=1}^{n} \log \mathbb{E}\exp\left[\gamma(T_i - \mathbb{E}T_i)\right] \leq \frac{v\phi(\gamma w)}{w^2}.$$

Comme $(T_i)_{i=1}^n$ est une suite de variables indépendantes, en prenant l'exponentielle de chaque côté de la dernière inégalité, on obtient le résultat. □

Liste des Figures

Liste des Tableaux

Bibliographie

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19** 716–723. 16

ALQUIER, P. (2006). *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 22, 38, 41, 60

ALQUIER, P. (2008). PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics*, **17** 279–304. 22, 38, 41, 49, 51, 60

ALQUIER, P. and BIAU, G. (2013). Sparse Single-Index Model. *Journal of Machine Learning Research*, **14** 243–280. 22, 24, 39, 44, 49, 60, 64, 72

ALQUIER, P., LI, X. and WINTENBERGER, O. (2012). Prediction of time series by statistical learning: General losses and fast rates. Preprint, URL http://arxiv.org/abs/1211.1847. 22

ALQUIER, P. and LOUNICI, K. (2011). PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics*, **5** 127–145. 20, 22, 38, 39, 60, 63

ALQUIER, P. and WINTENBERGER, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli*, **18** 883–913. 22

AMIT, Y. and GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, **9** 1545–1588. 13

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, **18** 343–373. 23

ARIAS-CASTRO, E. and LOUNICI, K. (2012). Variable selection with exponential weights and ℓ_0-penalization. Preprint, URL http://arxiv.org/abs/1208.2635. 20, 60

ARLOT, S. (2007). *Rééchantillonage et sélection de modèles*. Ph.D. thesis, Université Paris-Sud 11. 16

ARLOT, S. and BARTLETT, P. L. (2011). Margin-adaptive model selection in statistical learning. *Bernoulli*, **17** 687–713. 16

ARLOT, S. and CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4** 40–79. 16

ARLOT, S. and MASSART, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, **10** 245–279. 16

AUDIBERT, J.-Y. (2004a). Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré : Probabilités et Statistiques*, **40** 685–736. 22, 38, 41, 51, 60, 72

AUDIBERT, J.-Y. (2004b). *Théorie statistique de l'apprentissage : une approche PAC-Bayésienne*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 22, 38, 60

AUDIBERT, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, **37** 1591–1646. 38

AUDIBERT, J.-Y. and CATONI, O. (2010). Robust linear regression through PAC-Bayesian truncation. Preprint, URL http://arxiv.org/abs/1010.0072. 38, 60

AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, **39** 2766–2794. 38, 60

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, **35** 608–633. 13, 59

BACH, F. R. (2008a). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*. 17

BACH, F. R. (2008b). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, **9** 1179–1225. 17

BACH, F. R. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, **4** 384–414. 60

BARAUD, Y., GIRAUD, C. and HUET, S. (2013). Estimator selection in the Gaussian setting. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*. 72

BARRON, A. R., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113** 301–413. 16

BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, **101** 138–156. 59

BARTLETT, P. L., MENDELSON, S. and PHILIPS, P. (2004). Local complexities for empirical risk minimization. Tech. rep., UC Berkeley. 12

BARTON, N. H. and BAIRD, S. J. E. (1998). *Analyse*. Version 1.1, URL http://helios.bto.ed.ac.uk/evolgen/Mac/Analyse/. 102

BAZYKIN, A. D. (1969). Hypothetical mechanism of speciation. *Evolution*, **23** 685–687. 101, 103

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, **98** 791–806. 16

BIAU, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, **13** 1063–1095. 13, 77

BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, **101** 2499–2518. 13

BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, **9** 2015–2033. 13

BIAU, G., FISCHER, A., GUEDJ, B. and MALLEY, J. D. (2013). COBRA: A Nonlinear Aggregation Strategy. Preprint, URL http://arxiv.org/abs/1303.2236. 9, 31

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37** 1705–1732. 17, 18, 38, 60

BIRGÉ, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **42** 273–325. 72

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, **3** 203–268. 16

BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, **138** 33–73. 16

BLANCHARD, G., LUGOSI, G. and VAYATIS, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, **4** 861–894. 59

BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*. 59

BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2004). *Advanced lectures on Machine Learning*, chap. Introduction to Statistical Learning Theory. Springer-Verlag, 169–207. 12

BREIMAN, L. (1996). Bagging predictors. *Machine Learning*, **24** 123–140. 13

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45** 5–32. 13

BRIDLE, J. R., BAIRD, S. J. E. and BUTLIN, R. K. (2001). Spatial structure and habitat variation in a grasshopper hybrid zone. *Evolution*, **55** 1832–1843. 105

BUNEA, F. and NOBEL, A. (2008). Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Transactions on Information Theory*, **54** 1725–1735. 19, 72

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of the 19th annual conference on Computational Learning Theory* (G. Lugosi and H. U. Simon, eds.), vol. 35. Springer-Verlag, 379–391. 38, 72

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007a). Aggregation for gaussian regression. *The Annals of Statistics*, **35** 1674–1697. 17, 72

BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007b). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, **35** 169–194. 17, 72, 78

BÜHLMANN, P. and VAN DE GEER, S. A. (2011). *Statistics for High-Dimensional Data*. Springer. 37, 39, 60

CANDÈS, E. J. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, **35** 2313–2351. 17

CARLIN, B. P. and CHIB, S. (1995). Bayesian Model choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B*, **57** 473–484. 27, 39, 61, 63

CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour XXXI – 2001, Springer. 13, 18, 19, 20, 22, 38, 41, 46, 49, 51, 60, 66, 72, 111

CATONI, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, vol. 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics. 22, 38, 41, 51, 60

CAWLEY, G. C. and TALBOT, N. L. C. (2006). Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, **22** 2348–2355. 59

CERVONENKIS, A. J. and VAPNIK, V. N. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Daukladi Akademii Nauk USSR*, **181**. 11

CESA-BIANCHI, N. and LUGOSI, G. (1999). On prediction of individual sequences. *The Annals of Statistics*, **27** 1865–1895. 18

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, Learning and Games*. Cambridge University Press. 18

CHEN, C., DURAND, E., FORBES, F. and FRANÇOIS, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Molecular Ecology Notes*, **7** 747–756. 106

CORTEZ, P., CERDEIRA, A., ALMEIDA, F., MATOS, T. and REIS, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47** 547–553. 82

CRAMER, J. S. (2003). *Logit models from economics and other fields*, chap. The origins and development of the *logit* model. Cambridge University Press. 59, 103

CULE, E. (2012). *ridge: Ridge Regression with automatic selection of the penalty parameter*. R package version 2.1-2, URL `http://CRAN.R-project.org/package=ridge`. 79

DALALYAN, A. and TSYBAKOV, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory, COLT 2007* (N. H. Bshouty and C. Gentile, eds.), vol. 4539 of *Lecture Notes in Computer Science*. Springer, Berlin, 97–111. 19

DALALYAN, A. S. (2012). SOCP based variance free Dantzig selector with application to robust estimation. *Comptes Rendus Mathématiques de l'Académie des Sciences de Paris*, **350** 785–788. 16

DALALYAN, A. S., HEBIRI, M., MEZIANI, K. and SALMON, J. (2013). Learning heteroscedastic models by convex programming under group sparsity. In *Journal of Machine Learning Research - W & CP (ICML 2013)*, vol. 28. 379–387. 16

DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, **40** 2327–2355. 60

DALALYAN, A. S. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, **72** 39–61. 19, 22, 38, 39, 41, 44, 49, 60, 63, 72, 74

DALALYAN, A. S. and TSYBAKOV, A. B. (2012a). Mirror averaging with sparsity priors. *Bernoulli*, **18** 914–944. 19

DALALYAN, A. S. and TSYBAKOV, A. B. (2012b). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, **78** 1423–1443. 19, 23, 38, 39, 41, 44, 46, 49, 60, 64

DE CASTRO, Y. (2011). *Constructions déterministes pour la régression parcimonieuse*. Ph.D. thesis, Université Toulouse III - Paul Sabatier. 21

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A probabilistic theory of pattern recognition*. Springer. 12, 75

DONOHO, D. L. and JOHNSTON, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** 425–455. 14

DONOHO, D. L. and JOHNSTON, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90** 1200–1224. 14

DONOHO, D. L., JOHNSTON, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, **57** 301–369. 14

DURAND, E., JAY, F., GAGGIOTTI, O. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, **26** 1963–1973. 34, 105, 106

EFRON, B., HASTIE, T., JOHNSTON, I. M. and TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics*, **32** 407–499. 17

FALUSH, D., STEPHENS, M. and PRITCHARD, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164** 1567–1587. 105

FREUND, Y. (1990). Boosting a weak learning algorithm by majority. In *Proceedings of the third annual workshop on Computational Learning Theory*. 202–216. 13

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, **28** 337–407. With discussion. 59

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 17

GAÏFFAS, S. and LECUÉ, G. (2011). Hyper-sparse optimal aggregation. *Journal of Machine Learning Research*, **12** 1813–1833. 22

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC. 59

GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, **49** 291–304. 59

GENUER, R. (2010). *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. Ph.D. thesis, Université Paris-Sud 11. 13

GERCHINOVITZ, S. (2011). *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*. Ph.D. thesis, Université Paris-Sud 11. 18

GIRAUD, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli*, **14** 1089–1107. 16, 22

GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statistical Science*, **27** 500–518. 16, 22, 38

GOLUBEV, Y. (2010). On universal oracle inequalities related to high-dimensional linear models. *The Annals of Statistics*, **38** 2751–2780. 22

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82** 711–732. 24, 39

GUEDJ, B. (2013a). *COBRA: COmBined Regression Alternative*. R package version 0.99.4, URL `http://cran.r-project.org/web/packages/COBRA/index.html`. 9, 31, 74, 78

GUEDJ, B. (2013b). *pacbpred: PAC-Bayesian Estimation and Prediction in Sparse Additive Models*. R package version 0.92.2, URL `http://cran.r-project.org/web/packages/pacbpred/index.html`. 9, 27, 44, 61

GUEDJ, B. and ALQUIER, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, **7** 264–291. 9, 37, 60, 61, 63, 72

GUEDJ, B. and GUILLOT, G. (2011). Estimating the location and shape of hybrid zones. *Molecular Ecology Resources*, **11** 1119–1123. First published online : 07/07/2011. 9, 32, 101

GUILLOT, G. (2009a). On the inference of spatial structure from population genetics data. *Bioinformatics*, **25** 1796–1801. 106

GUILLOT, G. (2009b). Response to comment on 'On the inference of spatial structure from population genetics data'. *Bioinformatics*, **25** 1805–1806. 106

GUILLOT, G., ESTOUP, A., MORTIER, F. and COSSON, J.-F. (2005). A spatial statistical model for landscape genetics. *Genetics*, **170** 1261–1280. 33, 102

GUILLOT, G., LEBLOIS, R., COULON, A. and FRANTZ, A. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, **18** 4734–4756. 33, 102

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics, Springer. 72, 75, 92, 93, 96

HALDANE, J. B. S. (1948). The theory of a cline. *Journal of Genetics*, **48** 277–284. 101

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association*, **102** 507–516. 27, 39, 43, 61, 63

HARRISON, R. G. (1990). Hybrid zones: windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, **7** 69–128. 101

HASTIE, T. and EFRON, B. (2012). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.1, URL http://CRAN.R-project.org/package=lars. 79

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized Additive Models. *Statistical Science*, **1** 297–318. 25, 38

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*, vol. 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC. 25, 38

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning – Data mining, Inference, and Prediction*. 2nd ed. Springer. 12, 17, 37

HEBIRI, M. and VAN DE GEER, S. A. (2010). The smooth Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electronic Journal of Statistics*, **5** 1184–1226. 17, 60

HÄRDLE, W. K. (1990). *Applied nonparametric regression*. Cambridge University Press. 25, 38

JUDITSKY, A., NAZIN, A. V., TSYBAKOV, A. B. and VAYATIS, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Problems of Information Transmission*, **41** 368–384. 18, 72

JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric estimation. *The Annals of Statistics*, **28** 681–712. 72

JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. B. (2008). Learning by mirror averaging. *The Annals of Statistics*, **36** 2183–2206. 19

KIVINEN, J. and WARMUTH, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, **132** 1–63. 18

KNAUS, J. (2010). *snowfall: Easier cluster computing (based on snow)*. R package version 1.84, URL http://CRAN.R-project.org/package=snowfall. 78

KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **45** 7–57. 72

KOLTCHINSKII, V. (2010). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15** 799–828. 17

KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. École d'Été de Probabilités de Saint-Flour XXXVIII – 2008, Springer. 17

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics*, **38** 3660–3695. 25, 38, 42

KRUUK, L. E. B., BAIRD, S. J. E. and BARTON, N. H. (1999). A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153** 1959–1971. 101, 105

KWEMOU, M. (2012). Non-asymptotic oracle inequalities for the Lasso and Group Lasso in high dimensional logistic model. Preprint, URL http://arxiv.org/abs/1206.0710. 28, 60

LECUÉ, G. (2007a). *Méthodes d'agrégation : optimalité et méthodes rapides*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 13, 16

LECUÉ, G. (2007b). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, **13** 1000–1022. 13

LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, **52** 3396–3410. 19

LI, S. (2013). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1, URL http://CRAN.R-project.org/package=FNN. 79

LIAW, A. and WIENER, M. (2002). Classification and regression by randomforest. *R News*, **2** 18–22. URL http://CRAN.R-project.org/doc/Rnews/. 79

LITTLESTONE, N. and WARMUTH, M. K. (1994). The weighted majority algorithm. *Information and Computation*, **108** 212–261. 13, 18

LOUNICI, K. (2007). Generalized mirror averaging and D-convex aggregation. *Mathematical Methods of Statistics*, **16** 246–259. 19

LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, **2** 90–102. 17

LOUNICI, K. (2009). *Estimation statistique en grande dimension, parcimonie et inégalités d'oracle*. Ph.D. thesis, Université Denis Diderot - Paris VII. 19

LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, **32** 30–55. 59

MACCALLUM, C. J., NURNBERGER, B., BARTON, N. H. and SZYMURA, J. M. (1998). Habitat preference in the *Bombina* hybrid zone in Croatia. *Evolution*, **52** 227–239. 105

MACHOLÁN, M., BAIRD, S. J. E., DUFKOVÁ, P., MUNCLINGER, P., BÍMOVÁ, B. V. and PIÁLEK, J. (2011). Assessing multilocus introgression patterns: a case study on the mouse x chromosome in Central Europe. *Evolution*, **65** 1428–1446. 105

MACHOLÁN, M., BAIRD, S. J. E., MUNCLINGER, P., DUFKOVÁ, P., BÍMOVÁ, B. V. and PIÁLEK, J. (2008). Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evolutionary Biology*, **8**. 102

MALLEY, J. D., KRUPPA, J., DASGUPTA, A., MALLEY, K. G. and ZIEGLER, A. (2012). Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, **51** 74–81. 79, 82

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics*, **15** 661–675. 16

MARIN, J.-M. and ROBERT, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer. 23, 39

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. École d'Été de Probabilités de Saint-Flour XXXIII – 2003, Springer. 16, 46, 67, 112

MASSART, P. and MEYNET, C. (2011). The lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, **5** 669–687. 17

MCALLESTER, D. A. (1999). Some PAC-Bayesian Theorems. *Machine Learning*, **37** 355–363. 22, 38, 60

MEIER, L., VAN DE GEER, S. A. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, **70** 53–71. 28, 60

MEIER, L., VAN DE GEER, S. A. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, **37** 3779–3821. 25, 38, 46, 79

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*. 17

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, **37** 246–270. 38, 60

MEYN, S. and TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*. 2nd ed. Cambridge University Press. 23, 39

MOJIRSHEIBANI, M. (1999). Combining classifiers via discretization. *Journal of the American Statistical Association*, **94** 600–609. 29, 72, 75

NEMIROVSKI, A. (2000). *Topics in Non-Parametric Statistics*. Springer. 13, 72

NEMIROVSKI, A. and YUDIN, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley. 18

NG, A. Y. and JORDAN, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Proceedings of Neural Information Processing Systems*. 59

PEKALSKA, E. and DUIN, R. P. W. (2005). *The Dissimilarity Representation For Pattern Recognition: Foundations and Applications*, vol. 64 of *Machine Perception and Artificial Intelligence*. World Scientific. 75

PETRALIAS, A. (2010). *Bayesian model determination and nonlinear threshold volatility models*. Ph.D. thesis, Athens University of Economics and Business. 27, 39, 43, 64

PETRALIAS, A. and DELLAPORTAS, P. (2012). An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation*, **0** 1–19. 27, 39, 43, 61, 64

POLLEY, E. C. and VAN DER LAAN, M. J. (2010). Super learner in prediction. Tech. rep., UC Berkeley. 32, 80

POLLEY, E. C. and VAN DER LAAN, M. J. (2012). *SuperLearner: Super Learner Prediction*. R package version 2.0-9, URL http://CRAN.R-project.org/package=SuperLearner. 80

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155** 945–959. 33, 102

R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 44, 74

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, **13** 389–427. 25, 40

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B*, **71** 1009–1030. 25, 38

RIGOLLET, P. (2006). *Inégalités d'oracle, agrégation et adaptation*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 13, 15, 38

RIGOLLET, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, **40** 639–655. 60

RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, **16** 260–280. 13

RIGOLLET, P. and TSYBAKOV, A. B. (2011). Exponential screening and optimal rates of sparse aggregation. *The Annals of Statistics*, **39** 731–771. 16, 24

RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science*, **27** 558–575. 20, 21, 38

RINALDO, A. (2009). Properties and refinements of the fused Lasso. *The Annals of Statistics*, **37** 2922–2952. 17, 60

RIPLEY, B. (2012). *tree: Classification and regression trees*. R package version 1.0-32, URL http://CRAN.R-project.org/package=tree. 79

ROBERT, C. P. (2007). *The Bayesian Choice*. Springer-Verlag. 21

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer. 23

ROGERS, L. and WILLIAMS, D. (1987). *Diffusions, Markov processes, and martingales*, vol. 2. John Wiley & Sons Inc., New York. 23

SAFNER, T., MILLER, M. P., MCRAE, B., FORTIN, M. J. and MANEL, S. (2011). Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *International Journal of Molecular Sciences*, **12** 865–889. 106

SALMON, J. (2010). *Agrégation d'estimateurs et méthodes à patchs pour le débruitage d'images numériques*. Ph.D. thesis, Université Denis Diderot - Paris VII. 23

SALMON, J. and LE PENNEC, E. (2009a). An aggregator point of view on NL-Means. In *SPIE*, vol. 7446. 74461E. 23

SALMON, J. and LE PENNEC, E. (2009b). NL-Means and aggregation procedures. In *ICIP*. 2977–2980. 23

SCHAPIRE, R. E. (1990). The strength of weak learnability. *Machine Learning*, **5** 197–227. 13

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** 461–464. 16

SEEGER, M. (2002). PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, **3** 233–269. 22

SEEGER, M. (2003). *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. Ph.D. thesis, University of Edinburgh. 22

SELDIN, Y., LAVIOLETTE, F., CESA-BIANCHI, N., SHAWE-TAYLOR, J. and AUER, P. (2012). Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, **58** 7086–7093. 22

SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (1997). A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*. ACM, 2–9. 22, 38, 60

STOLTZ, G. (2005). *Incomplete information and internal regret in prediction of individual sequences*. Ph.D. thesis, Université Paris-Sud 11. 18

STOLTZ, G. (2010). Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société française de Statistique*, **151** 66–106. 18

STONE, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13** 689–705. 25, 38

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, **99** 879–898. 16

SUZUKI, T. (2012). PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model. In *Proceedings of the 25th annual conference on Computational Learning Theory*. 22, 39, 60

SUZUKI, T. and SUGIYAMA, M. (2013). Fast learning rates of multiple kernel learning: trade-off between sparsity and smoothness. *The Annals of Statistics*, **41** 1381–1405. 25, 38, 42

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267–288. 17, 37, 60

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society, Series B*, **67** 91–108. 17, 60

TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Computational Learning Theory and Kernel Machines* (B. Schölkopf and M. K. Warmuth, eds.). Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg, Springer, Heidelberg, 303–313. 15, 72

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, **32** 135–166. 59

TSYBAKOV, A. B. (2008). Agrégation d'estimateurs et optimisation stochastique. *Journal de la Société française de Statistique*, **149** 3–26. 13

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Statistics, Springer. 42

TSYBAKOV, A. B. (2013). Aggregation and high-dimensional statistics. Lecture notes for the course given at the École d'été de Probabilités in Saint-Flour, URL http://www.crest.fr/ckfinder/userfiles/files/Pageperso/ATsybakov/Lecture_notes_SFlour.pdf. 16, 20

VALIANT, L. G. (1984). A theory of the learnable. *Communications of the ACM*, **27** 1134–1142. 21

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, **36** 614–645. 17, 28, 38, 60, 72, 77, 78

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, **3** 1360–1392. 18, 60

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**. 32, 74, 79, 80

VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience. 12, 16

VAPNIK, V. N. (2000). *The nature of Statistical Learning Theory*. Springer. 12

VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, **6** 38–90. 22

VOVK, V. (1990). Aggregating strategies. In *Proceedings of the third annual workshop on Computational Learning Theory*. 372–383. 13, 18

WEGKAMP, M. H. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, **31** 252–273. 72

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25** 714–721. 59

YANG, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, **74** 135–161. 18, 72

YANG, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, **96** 574–588. 72

YANG, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, **15** 783–809. 13

YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10** 25–47. 72

YANG, Y. and BARRON, A. R. (1999). Information theoretic determination of minimax rates of convergence. *The Annals of Statistics*, **27** 1564–1599. 15

YEH, I.-C. (1998). Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, **28** 1797–1808. 82

YEH, I.-C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, **29** 474–480. 82

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68** 49–67. 17, 60

ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, **32** 56–134. 59

ZHOU, X., LIU, K.-Y. and WONG, S. T. (2004). Cancer classification and prediction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics*, **37** 249 – 259. 59

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67** 301–320. 17

Benjamin GUEDJ

AGRÉGATION D'ESTIMATEURS ET DE CLASSIFICATEURS : THÉORIE ET MÉTHODES

Ce manuscrit de thèse est consacré à l'étude des propriétés théoriques et méthodologiques de différentes procédures d'agrégation d'estimateurs. Un premier ensemble de résultats vise à étendre la théorie PAC-bayésienne au contexte de la grande dimension, dans les modèles de régression additive et logistique. Nous prouvons dans ce contexte l'optimalité, au sens minimax et à un terme logarithmique près, de nos estimateurs. La mise en œuvre pratique de cette stratégie, par des techniques MCMC, est étayée par des simulations numériques. Dans un second temps, nous introduisons une stratégie originale d'agrégation non linéaire d'estimateurs de la fonction de régression. Les qualités théoriques et pratiques de cette approche — dénommée COBRA — sont étudiées, et illustrées sur données simulées et réelles. Enfin, nous présentons une modélisation bayésienne — et l'implémentation MCMC correspondante — d'un problème de génétique des populations. Les différentes approches développées dans ce document sont toutes librement téléchargeables depuis le site de l'auteur.

Mots-clés : Agrégation, régression, classification, inégalités oracles, théorie PAC-bayésienne, COBRA, MCMC, parcimonie.

AGGREGATION OF ESTIMATORS AND CLASSIFIERS: THEORY AND METHODS

This thesis is devoted to the study of both theoretical and practical properties of various aggregation techniques. We first extend the PAC-Bayesian theory to the high dimensional paradigm in the additive and logistic regression settings. We prove that our estimators are nearly minimax optimal, and we provide an MCMC implementation, backed up by numerical simulations. Next, we introduce an original nonlinear aggregation strategy. Its theoretical merits are presented, and we benchmark the method—called COBRA—on a lengthy series of numerical experiments. Finally, a Bayesian approach to model admixture in population genetics is presented, along with its MCMC implementation. All approaches introduced in this thesis are freely available on the author's website.

Keywords: Aggregation, regression, classification, oracle inequalities, PAC-Bayesian theory, COBRA, MCMC, sparsity.