Estimation adaptative par sélection de partitions en rectangles dyadiques

Nathalie Akakpo

N° d'ordre : 9674

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES DE
L'UNIVERSITÉ PARIS-SUD XI

Spécialité : Mathématiques

par

Nathalie AKAKPO

Estimation adaptative par sélection de partitions en rectangles dyadiques

Soutenue le Lundi 7 décembre 2009 devant la Commission d'examen :

Mme	Fabienne COMTE	
Mme	Cécile DUROT	(Directrice de thèse)
Mme	Béatrice LAURENT-BONNEAU	(Présidente du jury)
M.	Oleg LEPSKI	
M.	Pascal MASSART	

Rapporteurs :

Mme	Fabienne COMTE
M.	Enno MAMMEN

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

Remerciements

Mes premiers remerciements s'adressent à Cécile Durot, tout d'abord pour avoir bien voulu se lancer il y a un peu plus de trois ans dans l'aventure de la direction de thèse. Ce travail n'aurait pu ainsi aboutir sans sa patience, sa sérénité, son exigence de rigueur et de clarté. Pour m'avoir convertie aux statistiques, je remercie – et félicite – non seulement Cécile, mais aussi Pascal Massart, pour son cours de M2 enthousiasmant sur la sélection de modèles.

Je suis très reconnaissante envers mes rapporteurs Fabienne Comte et Enno Mammen d'avoir consacré de leur temps à la lecture de ce manuscrit. Je remercie également Fabienne Comte, Béatrice Laurent et Oleg Lepski d'avoir accompli ce petit périple jusqu'à Orsay pour faire partie de mon jury.

Cette dernière année de thèse a été éclairée par la collaboration avec Claire Lacour, que je remercie aussi pour ses conseils de jeune ex-doctorante. Merci également à Anne-Sophie Tocquet pour son aide dans la réalisation de la monstrueuse figure 5.6 du Chapitre II et à Vincent Rivoirard pour sa relecture à la fois rapide et active de mon introduction.

De ces dernières années à Orsay, je garderai un heureux souvenir grâce aux doctorants traversant cette période de dur labeur avec humour et humilité, avec une pensée en particulier pour mes compagnons de M2, de bureau, de CESFO ou de pause-thé : Merlin, Mahendra, Pierre, Nicolas, Wilson, Dominique, Benoît, Robin, Pierre, Camille, Sébastien, Cathy et Jean-Patrick, pour ne citer qu'eux.

A mes proches, enfin, pour leur inébranlable confiance en moi et leur soutien inconditionnel.

Table des matières

Chapitre I

Introduction

Dans cette thèse, nous nous intéressons à divers problèmes d'estimation fonctionnelle par sélection de modèles construits sur des partitions en intervalles ou rectangles dyadiques. La procédure statistique que nous étudions s'inscrit plus généralement parmi les procédures non-paramétriques possédant des propriétés d'adaptation spatiale au sens minimax, notions que nous rappelons en début d'introduction. Puis nous exposons le principe de sélection de modèle sous-jacent à notre procédure et rappelons les principales collections de modèles utilisées jusqu'ici. Nous décrivons alors les collections de modèles sur lesquelles sont basés les travaux des chapitres suivants, en indiquant les quelques résultats déjà établis à leur sujet. Nous poursuivons par un état de l'art des diverses procédures spatialement adaptatives. Enfin, nous présentons notre contribution, au regard des différentes procédures existantes.

I.1 Cadre général et exemples de référence

Dans cette introduction, nous nous placerons dans le cadre général suivant. Etant donné un entier $n \geq 1$ fixé, on observe n variables aléatoires $Y_1, \ldots, Y_n$ définies sur un même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, à valeurs dans un borélien $\mathcal{Y}$ de $\mathbb{R}^d$, où $d \in \mathbb{N}^\star$. La loi de probabilité du vecteur $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ est inconnue, mais appartient à une famille de lois $\{P_s\}_{s \in \mathcal{S}}$ qu'on appelle modèle. On suppose de plus que pour $(s,t) \in \mathcal{S}^2$, $P_s = P_t$ si et seulement si $s = t$. Sous cette hypothèse d'identifiabilité, il existe un unique élément s dans $\mathcal{S}$ tel que P_s soit la loi de $\boldsymbol{Y}$, et on utilisera indifféremment le terme de modèle pour désigner $\{P_s\}_{s \in \mathcal{S}}$ ou $\mathcal{S}$. Par ailleurs, le modèle est supposé non-paramétrique, au sens où $\mathcal{S}$ n'est pas une partie d'un sous-espace vectoriel de dimension finie et indépendante de n. Notre objectif est alors d'estimer l'élément s de $\mathcal{S}$ tel que P_s soit la loi de $\boldsymbol{Y}$, c'est-à-dire de construire à partir de l'observation de $\boldsymbol{Y}$ une « bonne approximation » de s. Plus précisément, il s'agit de déterminer $\hat{s}$, fonction mesurable de $\boldsymbol{Y}$ définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et à valeurs dans $\mathcal{S}$. Une telle fonction est appelée estimateur de s. Nous considèrerons en fait des estimateurs de s à valeurs dans un sous-ensemble M de $\mathcal{S}$, sans pour autant supposer a priori que s appartienne à M : ce sous-ensemble est un modèle approché pour s, que nous appellerons encore, pour simplifier, modèle, et qui lui aussi pourra être non-paramétrique.

Décrivons deux problèmes fréquemment étudiés en estimation non-paramétrique, et permettant de ce fait de comparer les performances de différentes procédures statistiques. Dans chacun de ces deux problèmes, il s'agit d'estimer une fonction s à valeurs réelles définie sur $[0,1]$. Etant donnée une partition m de $[0,1]$ en un nombre fini d'intervalles, on peut considérer comme modèle approché pour s l'espace S_m des fonctions à valeurs réelles, définies sur $[0,1]$ et constantes sur chaque intervalle de m, l'estimation de s se ramenant alors à celle d'un nombre

fini de paramètres. Bien que S_m soit de dimension finie, ce modèle peut être non-paramétrique dans la mesure où m peut dépendre de n, notamment via le nombre ou la longueur des intervalles qui la composent. Un estimateur usuel à valeurs dans le modèle approché S_m, auquel nous ferons régulièrement référence dans la suite, est l'histogramme construit sur la partition m. Nous en rappelons la définition pour chacun des problèmes suivants.

Problème 1 : Estimation de densité.
Soit $\mathcal{S}$ l'ensemble des densités de probabilité par rapport à la mesure de Lebesgue μ sur $[0,1]$. On observe $Y_1, \ldots, Y_n$ variables aléatoires indépendantes et de même loi, admettant une densité $s \in \mathcal{S}$ inconnue, que l'on souhaite estimer. L'histogramme construit sur la partition m est défini par

$$\hat{s}_m = \sum_{I \in m} \left(\frac{1}{n\mu(I)} \sum_{i=1}^{n} \mathbb{1}_I(Y_i) \right) \mathbb{1}_I,$$

et à valeurs dans le modèle approché $\mathcal{S} \cap S_m$.

L'histogramme $\hat{s}_m$ est en fait une version empirique de la projection orthogonale de s sur S_m pour la norme $\mathbb{L}_2$.

Problème 2 : Estimation de la fonction de régression.
Soient $\mathcal{S}$ l'ensemble des fonctions définies sur $[0,1]$ à valeurs réelles et $(x_1, \ldots, x_n)$ un vecteur déterministe de $[0,1]^n$ donné. On observe un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de $\mathbb{R}^n$ à coordonnées indépendantes tel que, pour une certaine fonction $s \in \mathcal{S}$,

$$Y_i = s(x_i) + \sigma\varepsilon_i, i = 1, \ldots, n, \tag{I.1.1}$$

avec $\sigma \geq 0$ supposé connu et $(\varepsilon_i)_{1 \leq i \leq n}$ variables aléatoires non observables, supposées indépendantes et de loi normale centrée réduite. Il s'agit alors d'estimer la fonction de régression s. L'analogue dans ce cadre de l'histogramme est le régressogramme construit sur la partition m défini par

$$\hat{s}_m = \sum_{I \in m} \left(\frac{1}{\sum_{i=1}^{n} \mathbb{1}_I(x_i)} \sum_{i=1}^{n} Y_i \mathbb{1}_I(x_i) \right) \mathbb{1}_I.$$

Afin de pouvoir juger de la qualité d'un estimateur de s, il est d'usage de se donner une fonction de perte ℓ, c'est-à-dire une fonction définie sur $\mathcal{S} \times \mathcal{S}$ à valeurs dans $\mathbb{R}_+$. Généralement, $\mathcal{S}$ est muni d'une semi-distance d et on considère une fonction de perte de la forme d^p, où p est un entier naturel non nul. Lorsque $\boldsymbol{Y}$ suit la loi P_s, on note $\mathbb{P}_s$ la loi de probabilité sur $(\Omega, \mathcal{A})$ telle que, pour tout borélien B de $\mathcal{Y}$,

$$P_s(B) = \mathbb{P}_s\big(\{\omega \in \Omega \text{ t.q. } \boldsymbol{Y}(\omega) \in B\}\big).$$

On utilise alors comme critère de qualité d'un estimateur $\hat{s}$ de s son risque pour la fonction de perte ℓ, défini comme

$$s \in \mathcal{S} \mapsto \mathbb{E}_s\left[\ell(s, \hat{s})\right],$$

où $\mathbb{E}_s$ désigne l'espérance sous la loi $\mathbb{P}_s$. Par commodité, sous l'hypothèse que $\mathcal{S} \subset \mathbb{L}_2([0,1])$, on considère souvent le risque quadratique intégré

$$s \in \mathcal{S} \mapsto \mathbb{E}_s\left[\|s - \hat{s}\|^2\right],$$

où $\|.\|$ désigne la norme usuelle sur $\mathbb{L}_2([0,1])$. De manière générale, pour $q \geq 1$, on définit le risque $\mathbb{L}_q$ intégré par

$$s \in \mathcal{S} \mapsto \mathbb{E}_s\left[\|s - \hat{s}\|_q^q\right],$$

où $\|.\|_q$ désigne la norme usuelle sur $\mathbb{L}_q([0,1])$.

Dans l'ensemble de cette thèse, la lettre C désigne un réel positif non nul, dont la valeur peut changer d'une ligne à l'autre. La notation $C(\theta)$ indique que ce réel dépend éventuellement d'un paramètre θ.

I.2 Adaptation au sens minimax, adaptation spatiale

Cette partie est consacrée au point de vue minimax en estimation non-paramétrique. Nous évoquons en particulier quelques points fondamentaux tels que la décomposition biais-variance du risque d'un estimateur basé sur un modèle, les limites de l'estimation basée sur un seul modèle, la notion d'oracle, et soulignons les liens étroits entre estimation minimax et théorie de l'approximation.

I.2.1 Estimation basée sur un modèle et compromis biais-variance

Plaçons-nous tout d'abord dans le cadre d'estimation de densité introduit au paragraphe I.1, en supposant de plus les éléments de $\mathcal{S}$ de carré intégrable. Fixons une partition m de $[0,1]$, notons D_m la dimension de l'espace vectoriel S_m défini au paragraphe précédent (qui n'est autre que le nombre d'intervalles de m) et s_m la projection orthogonale de s sur S_m. D'après le théorème de Pythagore, le carré de la distance entre s et l'histogramme $\hat{s}_m$ est la somme d'une erreur déterministe et d'une erreur stochastique :

$$\|s - \hat{s}_m\|^2 = \|s - s_m\|^2 + \|\hat{s}_m - s_m\|^2.$$

Etant donnée l'expression de $\hat{s}_m$ et puisque les $(Y_i)_{1\leq i\leq n}$ sont indépendantes, on en déduit la décomposition du risque

$$\begin{aligned}\mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] &= \|s-s_m\|^2 + \frac{1}{n}\sum_{I\in m}\frac{\mathrm{Var}_s(\mathbb{1}_I(Y_1))}{\mu(I)} \\ &= \|s-s_m\|^2 + \frac{1}{n}\sum_{I\in m}\frac{\int_I s\left(1-\int_I s\right)}{\mu(I)}.\end{aligned}$$

Le premier terme, appelé terme de biais, correspond à une erreur d'approximation par le modèle S_m, et le second, appelé terme de variance, à une erreur d'estimation au sein du modèle S_m. Sous certaines hypothèses sur s, le terme de variance est exactement de l'ordre de D_m/n. Plus précisément, on peut montrer que

$$\|s-s_m\|^2 + \left(\inf_{[0,1]} s\right)\frac{D_m-1}{n} \leq \mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] \leq \|s-s_m\|^2 + \mathcal{N}(s)\frac{D_m-1}{n}, \qquad \text{(I.2.2)}$$

où $\mathcal{N}(s) = 1$ si la partition est régulière (par un argument de convexité) et $\mathcal{N}(s) = \|s\|_\infty$ sinon. L'encadrement ci-dessus donne lieu à plusieurs commentaires. Afin que le terme de biais soit faible, il est tentant de considérer S_m de grande dimension, d'où un terme de variance élevé. En revanche, si S_m est de petite dimension, c'est le terme de biais qui est susceptible d'être grand. Aussi, choisir un modèle S_m pour lequel le risque est faible nécessite de réaliser un bon compromis entre ces deux erreurs, qui, typiquement, varient en sens contraire lorsque la dimension de S_m croît. Soulignons que la nature du modèle S_m a également son importance pour contrôler le terme de biais. En effet, à dimension fixée, un modèle associé à une partition irrégulière, c'est-à-dire en intervalles de longueurs différentes, peut présenter une meilleure qualité d'approximation pour s qu'un modèle associé à une partition régulière. Par ailleurs,

même si s appartient à un modèle $S_{m'}$, auquel cas le risque $\mathbb{E}_s\left[\|s-\hat{s}_{m'}\|^2\right]$ se réduit à l'erreur d'estimation de l'ordre de $D_{m'}/n$, il est parfois préférable de considérer un modèle approché S_m ne contenant pas s. En effet, quitte à introduire une erreur d'approximation, on peut espérer gagner en terme de risque en considérant un modèle S_m pour lequel l'erreur d'estimation est significativement plus faible que dans le vrai modèle $S_{m'}$.

L'exemple précédent est en fait tout à fait représentatif d'une situation courante. Etant donné un modèle M inclus dans $\mathcal{S}$, un estimateur $\hat{s}_M$ à valeurs dans M, et une fonction de perte ℓ, il est fréquent d'obtenir, sinon une décomposition exacte du risque en une erreur d'approximation et une erreur d'estimation proportionnelle à la dimension de M, du moins un encadrement du type

$$C_1\left(\inf_{t\in M}\ell(s,t)+\frac{\dim(M)}{n}\right)\leq\mathbb{E}_s\left[\ell(s,\hat{s}_M)\right]\leq C_2\left(\inf_{t\in M}\ell(s,t)+\frac{\dim(M)}{n}\right). \tag{I.2.3}$$

Différentes notions de dimension sont envisageables selon la nature de M, qui, toutes, permettent de mesurer la difficulté à estimer au sein de ce modèle. Des majorations du risque comme dans l'inégalité de droite sont démontrées par exemple pour les estimateurs linéaires par ondelettes [HKPT98], les estimateurs par minimum de contraste [BM93; BM98a], les T-estimateurs introduits par Birgé [Bir06a]. De même que (I.2.2), les bornes de risque (I.2.3) conduisent à chercher un modèle M réalisant un bon compromis entre la fidélité au vrai paramètre s et la difficulté à estimer au sein de ce modèle.

I.2.2 Estimation minimax

Pour juger de la performance d'un estimateur $\tilde{s}$ de s, nous adopterons le point de vue minimax, qui consiste à procéder de la manière suivante. On choisit tout d'abord un sous-ensemble $\mathcal{F}$ de $\mathcal{S}$. Typiquement, si $\mathcal{S}$ est un espace fonctionnel, $\mathcal{F}$ est un sous-ensemble de fonctions de $\mathcal{S}$ présentant la même régularité. On définit alors le risque maximal de $\tilde{s}$ pour s appartenant à $\mathcal{F}$, *i.e.* $\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\tilde{s})\right]$, que l'on compare au risque minimax sur $\mathcal{F}$

$$\inf_{\hat{s}}\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\hat{s})\right]$$

où l'infimum est pris sur l'ensemble des estimateurs $\hat{s}$ de s. Comme cet infimum n'est pas nécessairement atteint et qu'on ne dipose généralement que de bornes pour le risque minimax, nous nous fixerons pour objectif de trouver un estimateur $\tilde{s}$ *approximativement* minimax sur $\mathcal{F}$, c'est-à-dire tel que

$$\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\tilde{s})\right]\leq C(\mathcal{F})\inf_{\hat{s}}\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\hat{s})\right],$$

où $C(\mathcal{F})$ est un réel positif qui peut dépendre de $\mathcal{F}$ mais pas de n. Nous rencontrerons également des estimateurs minimax sur $\mathcal{F}$ à un facteur logarithmique près, c'est-à-dire tels que

$$\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\tilde{s})\right]\leq C(\mathcal{F})\ln^{\delta}(n)\inf_{\hat{s}}\sup_{s\in\mathcal{F}}\mathbb{E}_s\left[\ell(s,\hat{s})\right],$$

pour un certain $\delta>0$.

Pourvu que la régularité des éléments de $\mathcal{F}$ soit suffisante, la construction d'un estimateur $\tilde{s}$ approximativement minimax sur $\mathcal{F}$ repose essentiellement sur le choix d'un modèle linéaire adapté à $\mathcal{F}$. Donnons-en un exemple en revenant au problème d'estimation de densité. Fixons $0<\alpha\leq 1, p\geq 2, R\geq 0$ et $\rho>0$, et notons $\mathcal{L}(\alpha,p,R,\rho)$ le sous-ensemble de $\mathcal{S}$ composé des densités t telles que $t\geq\rho$,

$$t\in\mathbb{L}_p([0,1])\text{ et, pour tout }0<h<1,\left(\int_0^{1-h}|t(x+h)-t(x)|^p\mu(\mathrm{d}x)\right)^{1/p}\leq Rh^{\alpha}. \tag{I.2.4}$$

Ainsi, les éléments de $\mathcal{L}(\alpha, p, R, \rho)$ présentent une régularité lipschitzienne, d'ordre α, mesurée dans la norme $\mathbb{L}_p$. Pour $D \in \mathbb{N}^\star$, notons m_D la partition régulière de $[0,1]$ en D intervalles, c'est-à-dire la partition de $[0,1]$ en D intervalles de même longueur. D'après [DeV98] (inégalité (3.12)), on dispose sur $\mathcal{L}(\alpha, p, R, \rho)$ d'une majoration uniforme du terme de biais :

$$\sup_{s \in \mathcal{L}(\alpha,p,R,\rho)} \|s - s_{m_D}\|^2 \leq CR^2 D^{-2\alpha}. \tag{I.2.5}$$

Il découle alors de la majoration donnée en (I.2.2) que

$$\sup_{s \in \mathcal{L}(\alpha,p,R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}_m\|^2\right] \leq C\left(R^2 D^{-2\alpha} + \frac{D}{n}\right).$$

Pour réaliser approximativement le meilleur compromis entre le terme de biais $R^2 D^{-2\alpha}$ et le terme de variance, il suffit de choisir D le plus grand possible tel que $D/n \leq R^2 D^{-2\alpha}$. Si $nR^2 \geq 1$, on peut définir $D_\diamondsuit$ plus grand entier non nul inférieur ou égal à $(nR^2)^{1/(1+2\alpha)}$ et choisir la partition régulière $m_\diamondsuit$ en $D_\diamondsuit$ intervalles. On obtient alors

$$\sup_{s \in \mathcal{L}(\alpha,p,R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}_{m_\diamondsuit}\|^2\right] \leq C(\alpha)(Rn^{-\alpha})^{2/(1+2\alpha)}.$$

Or on dispose de la minoration suivante du risque minimax sur $\mathcal{L}(\alpha, p, R, \rho)$, déduite par exemple de [Mas07] (Proposition 7.16),

$$\inf_{\hat{s}} \sup_{s \in \mathcal{L}(\alpha,p,R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}\|^2\right] \geq C(\alpha, p, \rho)(Rn^{-\alpha})^{2/(1+2\alpha)}, \tag{I.2.6}$$

pourvu que $n^{-1/2} \leq R \leq n^\alpha$. Aussi, pour $n^{-1/2} \leq R \leq n^\alpha$, l'histogramme $\hat{s}_{m_\diamondsuit}$ est un exemple d'estimateur approximativement minimax sur $\mathcal{L}(\alpha, p, R, \rho)$.

I.2.3 Adaptation au sens minimax et inégalité d'oracle

Un inconvénient de l'estimateur $\hat{s}_{m_\diamondsuit}$ précédemment défini est que le choix de $m_\diamondsuit$ nécessite la connaissance d'un paramètre (α, R) tel que pour un certain $p \geq 2$ et un certain $\rho > 0$, $s \in \mathcal{L}(\alpha, p, R, \rho)$. Il s'agit là d'une condition assez restrictive et peu réaliste. De plus, quand bien même un tel paramètre serait connu, il est possible que s appartienne également à l'ensemble $\mathcal{L}(\alpha', p', R, \rho)$ avec $\alpha' > \alpha$, $p' \geq 2$. Autrement dit, s peut présenter une régularité plus grande, éventuellement mesurée dans une norme $\mathbb{L}_{p'}$ plus faible que la norme $\mathbb{L}_p$ au sens où $p' \leq p$ (*cf.* paragraphe I.3 ci-dessous). Si l'histogramme construit sur la partition $m_\diamondsuit(\alpha', R)$ atteint bien approximativement le risque minimax sur $\mathcal{L}(\alpha', p', R, \rho)$, de l'ordre de $(Rn^{-\alpha'})^{2/(1+2\alpha')}$, la partition $m_\diamondsuit(\alpha, R)$ n'est quant à elle plus adaptée d'un point de vue minimax. En effet, pourvu que $R \geq 2^{(\alpha+1/2)} n^{-1/2}$, il résulte de la minoration donnée en (I.2.2) que

$$\sup_{s \in \mathcal{L}(\alpha',p',R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}_{m_\diamondsuit(\alpha,R)}\|^2\right] \geq \rho \frac{D_\diamondsuit - 1}{n} \geq C(\alpha, \rho)(Rn^{-\alpha})^{2/(1+2\alpha)},$$

de sorte que $\sup_{s \in \mathcal{L}(\alpha',p',R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}_{m_\diamondsuit(\alpha,R)}\|^2\right] / \inf_{\hat{s}} \sup_{s \in \mathcal{L}(\alpha',p',R,\rho)} \mathbb{E}_s \left[\|s - \hat{s}\|^2\right]$ est minoré par une fonction non bornée de n. Il serait donc souhaitable de construire un estimateur de s qui soit approximativement minimax sur chacun des ensembles $\mathcal{L}(\alpha, p, R, \rho)$ pour un large choix de valeurs de (α, p, R, ρ). Aussi, nous nous intéresserons essentiellement dans la suite à des estimateurs dits adaptatifs au sens minimax, c'est-à-dire simultanément approximativement minimax sur chacun des éléments d'une famille $\{\mathcal{F}_\theta, \theta \in \Theta\}$ de sous-ensembles de $\mathcal{S}$ choisie a priori. De tels estimateurs ont l'avantage d'être presque aussi performants que si tous les paramètres $\theta \in \Theta$ tels que $s \in \mathcal{F}_\theta$ étaient connus. De manière générale, la construction d'un

estimateur adaptatif au sens minimax repose sur une procédure de sélection parmi une famille d'estimateurs approximativement minimax sur différentes classes de fonctions. Etant donnée une telle famille d'estimateurs $\{\hat{s}_m\}_{m\in\mathcal{M}}$, il suffit de construire un estimateur $\tilde{s}$ vérifiant une inégalité de la forme

$$\mathbb{E}_s\left[\ell(s,\tilde{s})\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\ell(s,\hat{s}_m)\right]. \tag{I.2.7}$$

Ce type d'inégalité est appelé inégalité d'oracle, pour reprendre la terminologie introduite par Donoho et Johnstone [DJ94b]. Pourvu qu'on dispose de majorations du type (I.2.3) pour chaque estimateur de la collection, une telle inégalité suffit à assurer que $\tilde{s}$ réalise approximativement le meilleur compromis biais-variance parmi les estimateurs de la collection, d'où ses propriétés d'adaptation au sens minimax.

I.2.4 Adaptation spatiale et non-linéarité

Le lecteur attentif aura noté que, dans l'exemple des paragraphes I.2.2 et I.2.3, nous nous sommes limités à $p \geq 2$. Supposons maintenant $1 \leq p < 2$, et $\alpha > 1/p - 1/2$ de telle sorte que l'ensemble $L(\alpha,p,R)$ des fonctions vérifiant la condition (I.2.4) est un sous-ensemble de $\mathbb{L}_2([0,1])$. Pour $s \in \mathbb{L}_2([0,1])$ et M sous-espace vectoriel de $\mathbb{L}_2([0,1])$, notons s_M la projection orthogonale de s sur M. D'après [LGM96] (Chapitre 14, Théorème 1.1), pour tout $D \in \mathbb{N}^\star$ et tout sous-espace vectoriel M de $\mathbb{L}_2([0,1])$ de dimension D,

$$\sup_{s\in L(\alpha,p,R)} \|s - s_M\|^2 \geq C(\alpha,p) R^2 D^{-2(\alpha+1/2-1/p)}, \tag{I.2.8}$$

de sorte que

$$\sup_{s\in L(\alpha,p,R)} \|s - s_M\|^2 + \frac{D}{n} \geq C(\alpha,p)\left(Rn^{-(\alpha+1/2-1/p)}\right)^{1/(\alpha+1-1/p)}, \tag{I.2.9}$$

où la seconde inégalité est obtenue par minimisation sur D. Compte tenu de la minoration (I.2.2), on ne peut guère espérer construire un seul histogramme, ni même un estimateur basé sur un seul modèle linéaire de dimension finie, dont le risque quadratique intégré atteigne approximativement sur $\mathcal{L}(\alpha,p,R,\rho)$ la vitesse $(Rn^{-\alpha})^{1/(1+2\alpha)}$. Comme suggéré par la minoration (I.2.8) ci-dessus, cela tient essentiellement aux limites de l'approximation par un modèle linéaire. La fonction de perte considérée étant la perte $\mathbb{L}_2$, on dit des fonctions dont la régularité est mesurée dans une norme $\mathbb{L}_p$ avec $p < 2$ qu'elles présentent une régularité non homogène. Cette définition vaut également pour un risque mesuré dans une norme $\mathbb{L}_q$ et une régularité mesurée dans une norme $\mathbb{L}_p$ avec $p < q$. Nous donnerons dans le paragraphe I.3 des exemples illustrant la pertinence de ce terme. Dans divers cadres statistiques, des résultats établissent rigoureusement sur de telles classes de fonctions la sous-optimalité des estimateurs linéaires, dont font partie les estimateurs basés sur un modèle linéaire usuels tels que l'histogramme (voir par exemple [DJKP96] en densité ou [DJ98] en régression pour une borne inférieure similaire à (I.2.9)). Nous dirons d'un estimateur qu'il s'adapte spatialement s'il s'adapte au sens minimax sur une famille de sous-ensembles de $\mathcal{S}$ contenant notamment des fonctions de régularité non homogène. De manière générale, la construction d'un estimateur spatialement adaptatif repose essentiellement sur deux ingrédients : une inégalité d'oracle telle que (I.2.7) et un peu de non-linéarité. Par non-linéarité, nous entendons typiquement la possibilité de choisir, à dimension fixée, entre plusieurs modèles linéaires de même dimension. Nous verrons par exemple au paragraphe I.5 qu'à dimension D fixée, on sait construire une famille finie $\mathcal{M}_D$ de partitions de $[0,1]$ en D intervalles, éventuellement irrégulières, possédant les qualités d'approximation adéquates. En effet, chaque famille $\mathcal{M}_D$ permet de retrouver une majoration uniforme du biais sur la classe $\mathcal{L}(\alpha,p,R,\rho)$, de la forme

$$\sup_{s\in\mathcal{L}(\alpha,p,R,\rho)} \inf_{m\in\mathcal{M}_D} \|s - s_m\|^2 \leq C(\alpha,p) R^2 D^{-2\alpha}, \tag{I.2.10}$$

pour $p < 2$ et $\alpha > 1/p - 1/2$ comme pour $p \geq 2$ et $\alpha > 0$. Contrairement à (I.2.5), qui repose uniquement sur le modèle linéaire S_{m_D}, cette majoration fait intervenir le modèle *non-linéaire* $\cup_{m\in\mathcal{M}_D} S_m$. Puis en choisissant $D_\diamondsuit(\alpha, R)$ comme au paragraphe I.2.2, on en déduit que

$$\sup_{s\in\mathcal{L}(\alpha,p,R,\rho)} \inf_{m\in\mathcal{M}_{D_\diamondsuit}} \mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] \leq C(\alpha,p)(Rn^{-\alpha})^{2/(1+2\alpha)}$$

pourvu que $R \geq n^{-1/2}$. Par conséquent, si l'on sait construire un estimateur $\tilde{s}$ vérifiant l'inégalité d'oracle

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right]$$

où $\mathcal{M} = \cup_{D\in\mathbb{N}^\star}\mathcal{M}_D$, alors $\tilde{s}$ vérifie également, pour tout (α, p) tel que $p < 2$ et $\alpha > 1/p - 1/2$ ou $p \geq 2$ et $\alpha > 0$, et tout $n^{-1/2} \leq R \leq n^\alpha$,

$$\begin{aligned}\sup_{s\in\mathcal{L}(\alpha,p,R,\rho)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] &\leq C \sup_{s\in\mathcal{L}(\alpha,p,R,\rho)} \inf_{m\in\mathcal{M}_{D_\diamondsuit(\alpha,R)}} \mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] \\ &\leq C(\alpha,p)(Rn^{-\alpha})^{2/(1+2\alpha)}.\end{aligned}$$

La minoration (I.2.6) étant toujours valable pour les valeurs de α, p, R considérées ici, cet estimateur $\tilde{s}$ s'adapte donc spatialement.

I.3 Quelques classes de régularité usuelles

Rappelons tout d'abord la définition des espaces de Besov et des fonctions à α-variations bornées. Soient $\alpha > 0$, $0 < p, q \leq \infty$ et $r = \lfloor\alpha\rfloor + 1$, où $\lfloor\alpha\rfloor$ est le plus petit entier inférieur ou égal à α. Pour $t \in \mathbb{L}_p([0,1])$, on définit les différences d'ordre r

$$\Delta_h^r(t,x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} t(x+kh), \text{ pour } h \geq 0 \text{ et } 0 \leq x \leq 1-rh$$

et le module de continuité d'ordre r

$$\omega_r(t,y)_p = \begin{cases} \sup_{0<h\leq y} \left(\int_0^{1-rh} |\Delta_h^r(t,x)|^p \,\mu(\mathrm{d}x)\right)^{1/p} & \text{si } 0 < p < \infty \\ \sup_{0<h\leq y} \sup_{0\leq x\leq 1-rh} |\Delta_h^r(t,x)| & \text{si } p = \infty \end{cases}$$

en supposant de plus t continue sur $[0,1]$ lorsque $p = \infty$. L'espace de Besov $B_q^\alpha(\mathbb{L}_p([0,1]))$ est l'ensemble des fonctions $t \in \mathbb{L}_p([0,1])$ telles que

$$|t|_{B_q^\alpha(\mathbb{L}_p([0,1]))} = \begin{cases} \left(\int_0^\infty (y^{-\alpha}\omega_r(t,y)_p)^q \,\frac{\mu(\mathrm{d}y)}{y}\right)^{1/q} & \text{si } 0 < q < \infty \\ \sup_{y>0} y^{-\alpha}\omega_r(t,y)_p \text{ si } q = \infty \end{cases}$$

est finie, et l'on pose

$$\|t\|_{B_q^\alpha(\mathbb{L}_p([0,1]))} = \|t\|_p + |t|_{B_q^\alpha(\mathbb{L}_p([0,1]))}.$$

On notera que, pour $0 < \alpha < 1$, $B_\infty^\alpha(\mathbb{L}_p([0,1]))$ n'est autre que l'ensemble des fonctions vérifiant la condition de Lipschitz (I.2.4). Les propriétés de ces espaces qui nous seront utiles sont regroupées dans la proposition ci-dessous, et démontrées par exemple dans [DeV98; Tri83].

Proposition 1 *Soient $\alpha > 0$, $0 < p, q \leq \infty$ et $r = \lfloor\alpha\rfloor + 1$.*

i) Si $1 \leq p, q \leq \infty$, $\|.\|_{B_q^\alpha(\mathbb{L}_p([0,1]))}$ est une norme, et une quasi-norme sinon, l'inégalité triangulaire n'étant plus vérifiée qu'à constante près.

ii) *En remplaçant* ω_r *par* ω_k *avec* $k > r$, *on obtient une (quasi-)norme équivalente sur* $B_q^\alpha(\mathbb{L}_p([0,1]))$.

iii) *Si* $\alpha_1 < \alpha_2$, *alors quels que soient* $0 < q_1, q_2 \leq \infty$,

$$\|t\|_{B_{q_1}^{\alpha_1}(\mathbb{L}_p([0,1]))} \leq C(\alpha_1, \alpha_2, q_1, q_2)\|t\|_{B_{q_2}^{\alpha_2}(\mathbb{L}_p([0,1]))}.$$

iv) *À* α *et* p *fixés, et pour tout* $q > 0$,

$$|t|_{B_\infty^\alpha(\mathbb{L}_p([0,1]))} \leq C(\alpha, p, q)|t|_{B_q^\alpha(\mathbb{L}_p([0,1]))}.$$

v) *Si* $\alpha > \max\{1/p - 1/2, 0\}$, *alors* $B_q^\alpha(\mathbb{L}_p([0,1]))$ *est inclus dans* $\mathbb{L}_2([0,1])$ *et*

$$\|t\| \leq C(\alpha, p, q)\|t\|_{B_q^\alpha(\mathbb{L}_p([0,1]))}.$$

vi) *Si* $\alpha > 1/p$, *les fonctions de* $B_q^\alpha(\mathbb{L}_p([0,1]))$ *sont continues, et*

$$\|t\|_\infty \leq C(\alpha, p, q)\|t\|_{B_q^\alpha(\mathbb{L}_p([0,1]))}.$$

Les fonctions de $B_q^\alpha(\mathbb{L}_p([0,1]))$ présentent une régularité d'ordre α, mesurée dans la norme $\mathbb{L}_p$. Le paramètre q n'est qu'un paramètre secondaire, d'après les points *iii)* et *iv)* ci-dessus, et la propriété *iv)* justifie l'intérêt porté à l'espace $B_q^\alpha(\mathbb{L}_p([0,1]))$ pour $q = \infty$. Il existe essentiellement deux manières de généraliser la définition des espaces de Besov à des fonctions à valeurs réelles définies sur $[0,1]^d$, $d \geq 2$, selon que l'on autorise la régularité de la fonction à changer selon la direction (espace de Besov anisotrope) ou non (espace de Besov isotrope). Ainsi, en adoptant par exemple la définition de [Tri06], un espace de Besov anisotrope est caractérisé par la donnée d'un d-uplet de réels strictement positifs $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, α_i indiquant le degré de régularité dans la i^{e} direction, d'un paramètre p indiquant la norme $\mathbb{L}_p$ dans laquelle la régularité est mesurée, et d'un paramètre secondaire q. Pour $0 < \alpha \leq 1$, l'espace $BV(\alpha)$ des fonctions à α-variations bornées est l'ensemble des fonctions $t : [0,1] \to \mathbb{R}$ telles que

$$V_\alpha(t) = \sup_{i \geq 1} \sup_{0 \leq x_0 < \ldots < x_i \leq 1} \left(\sum_{j=1}^{i} |t(x_j) - t(x_{j-1})|^{1/\alpha} \right)^\alpha$$

est finie. Cette échelle de régularité est liée à l'échelle des espaces de Besov par la propriété suivante (*cf.* [Pee76], Théorème 7).

Proposition 2 *Pour tout* $0 < \alpha \leq 1$,

$$C_1(\alpha)|.|_{B_\infty^\alpha(\mathbb{L}_{1/\alpha}([0,1]))} \leq V_\alpha(.) \leq C_2(\alpha)|.|_{B_1^\alpha(\mathbb{L}_{1/\alpha}([0,1]))}.$$

L'échelle des espaces de Besov et des fonctions à α-variations bornées contient la plupart des échelles de régularité « classiques ». Ainsi, pour $\alpha \in \mathbb{N}^\star$, $B_p^\alpha(\mathbb{L}_p([0,1]))$ contient l'espace de Sobolev $W^\alpha(\mathbb{L}_p([0,1]))$ des fonctions admettant α dérivées dans $\mathbb{L}_p$. L'espace des fonctions α-hölderiennes n'est autre que $B_\infty^\alpha(\mathbb{L}_\infty([0,1]))$. Par ailleurs, pour tout $0 < \alpha \leq 1$, toute fonction constante par morceaux et toute fonction α-höldérienne appartiennent à $BV(\alpha)$.

Supposons maintenant que l'on mesure la qualité d'approximation d'une fonction via la norme $\mathbb{L}_2$. Parmi les espaces de fonctions précédemment cités, les espaces de Hölder décrivent une régularité homogène, de même que les espaces de de Besov $B_q^\alpha(\mathbb{L}_p([0,1]))$ avec $p \geq 2$ ou $BV(\alpha)$ avec $\alpha \leq 1/2$ (compte-tenu de la Proposition 2). En revanche, les espaces $B_q^\alpha(\mathbb{L}_p([0,1]))$ avec $p < 2$ ou $BV(\alpha)$ avec $1/2 < \alpha \leq 1$ décrivent une régularité non-homogène. Donnons deux exemples illustrant la pertinence de cette dernière expression. Une fonction bornée sur $[0,1]$,

présentant un nombre fini de discontinuités et höldérienne d'ordre $\sigma > 0$ entre ces discontinuités, appartient à tous les espaces $B_q^\alpha(\mathbb{L}_p([0,1]))$, pourvu que $\alpha < \min\{\sigma, 1/p\}$ et $1 \leq p, q \leq \infty$ (*cf.* [Ren99] Lemme 2.2 ou [Mal98] Proposition 9.4). En particulier, l'indice de régularité α peut prendre des valeurs d'autant plus grandes que le paramètre p est petit. Par ailleurs, en pratique, l'espace $BV(1)$ des fonctions à variations bornées, ou son analogue en dimension 2, est couramment utilisé en théorie du signal et de l'image.

I.4 Sélection de modèle

Présentons maintenant le principe de sélection de modèle introduit par Birgé et Massart [BM97], qui sera utilisé pour construire les estimateurs étudiés dans cette thèse. Nous abordons également dans cette partie la question du choix de la collection de modèles et décrivons les collections de modèles usuelles.

I.4.1 Principe et objectif

La procédure de sélection de modèle de [BM97] peut être décrite de la manière suivante. On se donne une famille finie de modèles approchés $\{S_m\}_{m\in\mathcal{M}}$ inclus dans $\mathcal{S}$, où $\mathcal{M}$ dépend éventuellement de n, et l'on choisit un contraste γ, c'est-à-dire une fonction mesurable de $\boldsymbol{Y}$ telle que $t \mapsto \mathbb{E}_s[\gamma(t)]$ admette un minimum sur $\mathcal{S}$ en s. La perte considérée est la fonction $\ell : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}^+$ telle que $\ell(s,t) = \mathbb{E}_s\left[\gamma(t) - \gamma(s)\right]$. Puis on définit sur chaque modèle un estimateur $\hat{s}_m$ obtenu par minimisation du contraste γ sur S_m. L'idéal serait alors de choisir, parmi la famille $\{S_m, m \in \mathcal{M}\}$, le modèle $S_{m_{or}}$ pour lequel le risque de l'estimateur associé est minimal, c'est-à-dire tel que

$$\mathbb{E}_s\left[\ell(s, \hat{s}_{m_{or}})\right] = \min_{m\in\mathcal{M}} \mathbb{E}_s\left[\ell(s, \hat{s}_m)\right].$$

Ce modèle idéal $S_{m_{or}}$, malheureusement impossible à déterminer puisqu'il dépend du paramètre s inconnu, sera baptisé oracle. L'idée consiste alors à choisir un modèle en se basant sur les données. Pour cela, on se donne une fonction pen : $\mathcal{M} \to \mathbb{R}^+$, appelée pénalité, et on considère la procédure de sélection aléatoire, dépendant de $\boldsymbol{Y}$,

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{\gamma(\hat{s}_m) + \mathrm{pen}(m)\right\}.$$

Puis on définit

$$\tilde{s} = \hat{s}_{\hat{m}}, \tag{I.4.11}$$

appelé estimateur pénalisé, qui n'est plus un estimateur de la collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$ puisque le modèle $\hat{m}$ sélectionné peut changer selon les données. L'objectif est de déterminer une pénalité telle que le risque de $\tilde{s}$ soit proche du risque de l'oracle, c'est-à-dire vérifie l'inégalité d'oracle

$$\mathbb{E}_s\left[\ell(s, \tilde{s})\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\ell(s, \hat{s}_m)\right],$$

qualifiée de non asymptotique puisque n est fixé et quelconque. Soulignons qu'une telle inégalité assure *sans aucune hypothèse de régularité* sur la fonction s que l'estimateur pénalisé $\tilde{s}$ est presque aussi bon que le meilleur estimateur de s parmi la collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$. Par ailleurs, comme expliqué au paragraphe I.2, une inégalité d'oracle est un outil fort utile pour démontrer que, sous certaines hypothèses de régularité et pourvu que la collection de modèles soit bien choisie, l'estimateur $\tilde{s}$ s'adapte également au sens minimax.

En vue de prouver une inégalité d'oracle, on démontre tout d'abord un théorème de sélection de modèles, en proposant une forme de pénalité permettant de réaliser approximativement le

meilleur compromis biais-variance parmi les estimateurs de la collection. Typiquement, une pénalité convenable est telle que pen(m) se comporte comme une erreur d'estimation au sein du modèle S_m, c'est-à-dire croît avec la dimension du modèle. Ce genre de théorème repose essentiellement sur des inégalités de concentration pour le supremum d'un processus empirique inspirées de l'inégalité de Talagrand [Tal96]. Donnons un exemple de théorème de sélection de modèles dans le cadre de régression décrit au paragraphe I.1. Soit $\mathcal{S}$ l'ensemble des fonctions à valeurs réelles définies sur $[0,1]$, muni de la semi-norme $\|.\|_n$ définie par

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(x_i).$$

On considère le contraste

$$\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(x_i))^2 \tag{I.4.12}$$

associé à la fontion de perte

$$\ell(s,t) = \|s-t\|_n^2.$$

Etant donnée une famille finie $\{S_m, m \in \mathcal{M}\}$ de sous-espaces vectoriels de $\mathcal{S}$, on définit, pour tout $m \in \mathcal{M}$, $\hat{s}_m = \operatorname{argmin}_{t\in S_m}\gamma(t)$ où γ est donné par (I.4.12), puis $\tilde{s}$ comme en (I.4.11). Le théorème suivant est une version extrêmement simplifiée du Théorème 2 de [BM01], démontré dans un cadre gaussien beaucoup plus général.

Théorème 1 *On considère le cadre de régression défini au paragraphe I.1. Soit $\{S_m, m \in \mathcal{M}\}$ une famille finie de sous-espaces vectoriels de $\mathcal{S}$ de dimension finie. Soit $\{L_m\}_{m\in\mathcal{M}}$ une famille de réels positifs tels que*

$$\Sigma := \sum_{m\in\mathcal{M}} \exp(-D_m L_m) \leq 1, \tag{I.4.13}$$

où $D_m = \dim(S_m)$. Si la pénalité est de la forme

$$\operatorname{pen}(m) = \sigma^2(k_1 + k_2 L_m)\frac{D_m}{n},$$

où k_1, k_2 sont des réels positifs suffisamment grands, alors l'estimateur pénalisé $\tilde{s}$ vérifie

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C(k_1,k_2)\left(\min_{m\in\mathcal{M}}\left(d_n^2(s,S_m) + \sigma^2(1+L_m)\frac{D_m}{n}\right) + \frac{1}{n}\right),$$

où $d_n(s,S_m) = \inf_{t\in S_m}\|s-t\|_n$.

Par ailleurs, sur chaque modèle, le risque quadratique associé à la perte $\|.\|_n$ admet la décomposition biais-variance

$$\mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2\right] = d_n^2(s,S_m) + \sigma^2\frac{D_m}{n}.$$

Sous les hypothèses du Théorème 1, on obtient donc l'inégalité

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_1(k_1,k_2)\left(1+\max_{m\in\mathcal{M}} L_m\right)\min_{m\in\mathcal{M}}\mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2\right] + \frac{C_2(k_1,k_2)}{n}. \tag{I.4.14}$$

Il s'agit bien là d'une inégalité d'oracle, à un reste près $C_2(k_1,k_2)/n$ qui devient négligeable dès que n est suffisamment grand, et à un facteur près : $1+\max_{m\in\mathcal{M}} L_m$.

Il existe d'autres procédures de sélection de modèle, dont l'objectif est toujours d'établir une inégalité de type oracle, mais avec une règle de sélection différente. Birgé [Bir06a] propose par exemple de sélectionner le meilleur modèle par une procédure de tests multiples entre les estimateurs de la collection (voir aussi [Bir07; BB09]).

I.4.2 Choix de la famille de modèles

Le choix d'une famille de modèles doit notamment prendre en compte les éléments suivants. Reprenons l'exemple de la régression ci-dessus. Afin d'interpréter la contrainte (I.4.13), décomposons la famille de modèles en sous-familles de modèles de même dimension

$$\mathcal{M}_D = \{m \in \mathcal{M} \text{ t.q. } D_m = D\}, \text{ pour } D \in \mathbb{N}^\star.$$

Choisissons des poids $\{L_m\}_{m\in\mathcal{M}}$ qui ne dépendent du modèle que via sa dimension et notons, pour tout $D \in \mathbb{N}^\star$ et $m \in \mathcal{M}_D, L_m = L(D)$. Nous pouvons alors réécrire Σ comme la somme, finie par hypothèse sur $\mathcal{M}$,

$$\Sigma = \sum_{D\in\mathbb{N}^\star} \exp\Big(- D\Big(L(D) - D^{-1}\ln_+(|\mathcal{M}_D|)\Big)\Big),$$

où $\ln_+(x) = \ln(x)$ pour $x \geq 1$ et $\ln_+(0) = 0$. Pour que la condition (I.4.13) soit réalisée, il suffit que, pour tout $D \in \mathbb{N}^\star$,

$$L(D) \geq D^{-1}\ln_+(|\mathcal{M}_D|) + \ln 2.$$

Aussi, pour obtenir une inégalité d'oracle, il suffit que le nombre de modèles par dimension soit sous-exponentiel, autrement dit qu'il existe une constante absolue $\kappa > 0$ telle que, pour tout $D \in \mathbb{N}^\star$,

$$|\mathcal{M}_D| \leq \kappa^D. \tag{I.4.15}$$

En effet, un choix de poids convenable est alors

$$L(D) = \ln(2\kappa), \text{ pour tout } D \in \mathbb{N}^\star.$$

De manière générale, la quantité $\sup_{D\in\mathbb{N}^\star} D^{-1}\ln_+(|\mathcal{M}_D|)$ peut être considérée comme un indice de complexité de la collection de modèles, pour reprendre un terme déjà employé dans [BBM99] ou [Bir06b] par exemple. Plus cet indice est élevé, et plus la collection est complexe. Nous dirons d'une collection de modèles qu'elle présente une complexité sous-exponentielle si la condition (I.4.15) est vérifiée. Concernant maintenant la qualité d'estimation de $\tilde{s}$ sur un sous-ensemble $\mathcal{F}$ donné de $\mathcal{S}$, on déduit de la majoration (I.4.14) que

$$\sup_{s\in\mathcal{F}} \mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_1(k_1,k_2) \inf_{D\in\mathbb{N}^\star}\left(\sup_{s\in\mathcal{F}}\inf_{m\in\mathcal{M}_D} d_n^2(s,S_m) + \sigma^2(1+L(D))\frac{D}{n}\right) + \frac{C_2(k_1,k_2)}{n}.$$

Cette borne supérieure nous invite donc à considérer une famille de modèles à la fois suffisamment riche, c'est-à-dire contenant suffisamment de représentants de chaque dimension pour bien contrôler, à D fixé, les erreurs d'approximation, mais pas trop complexe, afin de ne pas trop influer sur la taille des poids. C'est le rôle de la théorie de l'approximation de proposer des familles de modèles réalisant un bon compromis entre complexité et qualité d'approximation. Nous reviendrons sur ces questions aux paragraphes I.4.3 et I.5. Enfin, soulignons que la complexité algorithmique liée au calcul des estimateurs de la collection impose également en pratique des limites sur le nombre de modèles.

I.4.3 Collections de modèles usuelles

Décrivons les collections de modèles usuelles permettant d'obtenir des résultats d'adaptation en estimation fonctionnelle. Elles sont liées à deux types d'approximation : approximation par des polynômes par morceaux ou approximation de certains termes de la décomposition dans une base orthonormée de $\mathbb{L}_2([0,1])$. Toutes ces collections sont des familles finies $\{S_m\}_{m\in\mathcal{M}}$ de sous-espaces vectoriels de $\mathbb{L}_2([0,1])$ de dimension finie, comportant un modèle de dimension au

plus n contenant tous les autres modèles de la famille. On désigne par D_m la dimension de S_m et, comme dans le paragraphe précédent, par $\mathcal{M}_D$ la sous-famille des modèles de dimension D.

Collections de modèles réguliers. Ces collections vérifient au moins

- une hypothèse sur le nombre de modèles de même dimension, satisfaite typiquement lorsque celui-ci est au plus polynomial, c'est-à-dire lorsqu'il existe une constante absolue κ telle que
$$|\mathcal{M}_D| \leq D^\kappa;$$
- une hypothèse liant les normes $\mathbb{L}_2$ et $\mathbb{L}_\infty$ sur ces modèles, qui assure l'existence d'une constante absolue $\Phi > 0$ telle que pour tout $m \in \mathcal{M}$ et tout $t \in S_m$,
$$\|t\|_\infty \leq \Phi\sqrt{D_m}\|t\|.$$

Les modèles de polynômes par morceaux construits sur les partitions régulières de $[0,1]$, notamment, sont des modèles réguliers. La complexité de ces collections est *a fortiori* sous-exponentielle. Cependant, tous les exemples connus ne possèdent de bonnes qualités d'approximation que pour des fonctions de régularité homogène. Par ailleurs, comme il n'y a en général qu'un seul modèle par dimension, déterminer tous les estimateurs de la collection puis l'estimateur pénalisé ne requiert qu'une complexité algorithmique linéaire en n. Nous renvoyons par exemple à Baraud [Bar00], Castellan [Cas00; Cas03], [BM97; Mas07], ainsi qu'à la bibliographie de Fabienne Comte pour des résultats de sélection parmi des modèles réguliers généraux dans différents cadres statistiques.

Collection de polynômes par morceaux construits sur une grille dyadique. On se fixe deux entiers naturels r et $J_\star$, avec $2^{J_\star} \leq n$. On définit $\mathcal{M}$ comme l'ensemble des partitions construites sur la partition régulière de $[0,1]$ en $2^{J_\star}$ intervalles, c'est-à-dire l'ensemble des partitions en intervalles de la forme $[k2^{-J_\star}, l2^{-J_\star}]$, où k, l sont des entiers, $0 \leq k < l \leq 2^{J_\star}$. Chaque modèle S_m est alors décrit comme l'espace des fonctions polynomiales de degré au plus r sur chaque intervalle de la partition m. Construire une partition de m en D intervalles équivaut à choisir $D-1$ points sur la grille $\{k2^{-J_\star}; k = 1, \ldots, 2^{J_\star} - 1\}$, de sorte que

$$|\mathcal{M}_D| = \binom{2^{J_\star} - 1}{D-1} \leq \left(\frac{e2^{J_\star}}{D}\right)^D \leq (en)^D,$$

où la majoration résulte par exemple de [Mas07] (Proposition 2.5). La complexité de cette collection n'est plus sous-exponentielle au sens utilisé dans le paragraphe précédent puisque la constante κ est ici remplacée par en, ce qui incite par exemple à choisir des poids constants égaux $\ln(2en)$. On n'obtient donc avec cette collection que des inégalités de type oracle à un facteur $\ln(n)$ près. On pourra pour cela consulter par exemple [BM97; BBM99; Cas00; Cas03; RB03; Sau02]. Cependant, à D fixé, les modèles de dimension D possèdent de bonnes qualités d'approximation relativement à certaines fonctions de régularité éventuellement non homogène, telles que les fonctions à α-variations bornées (*cf.* [BBM99], Corollaire 1). Cette collection permet donc d'obtenir un estimateur spatialement adaptatif, souvent à un facteur $\ln(n)$ près. Un estimateur pénalisé basé sur cette collection et un contraste de type moindres carrés a été implémenté par E. Lebarbier [Leb02] dans le cadre de régression décrit au paragraphe I.1. Il nécessite $\mathcal{O}(n^3)$ calculs par un algorithme de programmation dynamique et pour une pénalité bien choisie.

Collection de polynômes par morceaux construits sur des partitions obtenues par l'algorithme CART. La procédure Classification And Regression Trees (CART) de Breiman *et al.* [BFOS84] peut être présentée comme une procédure de sélection de modèle en adoptant le point de vue de Gey et Lebarbier [GL08] ou Lebarbier et Nédélec [LN07]. Nous dirons dans la suite d'un arbre binaire qu'il est complet si tous ses noeuds internes ont deux enfants, et parfait si de plus

toutes ses feuilles sont à la même distance de la racine. Plaçons-nous par exemple dans le cadre de régression du Problème 2, avec $0 = x_1 < \ldots < x_n = 1$. La première étape de CART consiste à construire un arbre binaire parfait de racine $\{x_1, \ldots, x_n\}$, dont chaque noeud est ensuite obtenu par partitionnement récursif, en minimisant un critère local basé sur les données. On considère alors la collection de toutes les partitions de $\{x_1, \ldots, x_n\}$ correspondant aux feuilles d'un quelconque sous-arbre binaire complet obtenu par élagage de cet arbre binaire parfait. La deuxième étape de CART correspond en fait à la sélection d'une meilleure partition parmi cette collection en utilisant un critère pénalisé. Comme le rappellent par exemple [GL08], le nombre d'arbres binaires complets à D feuilles, $D \in \mathbb{N}^\star$, ainsi construits n'est autre que le nombre de Catalan

$$\frac{1}{D}\binom{2(D-1)}{D-1} \leq 4^D.$$

Aussi, la complexité de cette collection de partitions est sous-exponentielle. Par ailleurs, la complexité algorithmique de la procédure de sélection d'une meilleure partition peut, se réduit à $\mathcal{O}(n\ln(n))$ calculs dans le meilleur des cas(*cf.* [GL08]). Cependant, le fait que cette collection de partitions dépende, par construction, des données, rend son étude difficile d'un point de vue théorique. En partageant l'échantillon, de manière à utiliser une partie des données pour l'étape de construction, et l'autre pour l'étape de sélection, il est possible d'obtenir des inégalités de type oracle, mais conditionnellement au premier échantillon. Nous renvoyons à [GN05; Sau02] en régression et à [LN07] en estimation de loi discrète pour ce type de résultats. Cependant, les qualités d'approximation de cette collection de partitions aléatoire demeurent inconnues, d'où l'absence de résultat d'adaptation.

Collection exhaustive des espaces engendrés par les sous-ensembles d'un système orthonormé. Etant donnée une famille orthonormale $\{\phi_\lambda\}_{\lambda\in\Lambda}$ de $\mathbb{L}_2([0,1])$, avec $|\Lambda|$ au plus d'ordre n, on considère la famille $\mathcal{M}$ de tous les sous-ensembles de Λ. Chaque modèle S_m est alors défini comme l'espace vectoriel engendré par les $\{\phi_\lambda\}_{\lambda\in m}$. On vérifie aisément que cette collection a la même complexité que la collection de polynômes par morceaux construits sur une grille dyadique. Les qualités d'approximation et donc d'adaptation dépendent du choix de la base. Nous renvoyons par exemple à [BM97; BM01; Mas07] pour des résultats dans le modèle gaussien ou le modèle de densité.

Collection de modèles d'ondelettes inspirée de l'algorithme de compression de Birgé et Massart [BM00]. Nous emploierons par la suite l'expression « stratégie Birgé-Massart » pour faire référence au choix de cette collection. Soit $\{\phi_\lambda\}_{\lambda\in\Lambda}$ une base d'ondelettes orthonormale de $\mathbb{L}_2([0,1])$. Dans cette base, il est d'usage d'écrire la décomposition de s par blocs

$$s = \sum_{j\geq -1}\sum_{\lambda\in\Lambda(j)} \beta_\lambda\phi_\lambda,$$

où $\Lambda = \cup_{j\geq -1}\Lambda(j)$, avec $\Lambda(-1)$ ensemble fini et, pour $j \geq 0$, $|\Lambda(j)|$ de l'ordre de 2^j. Pour $J \geq -1$, les $J+1$ premiers blocs

$$\sum_{j=-1}^{J-1}\sum_{\lambda\in\Lambda(j)} \beta_\lambda\phi_\lambda$$

donnent une approximation de s à la résolution 2^{-J}. Pourvu que J soit assez grand, une fonction s globalement régulière avec quelques singularités isolées peut être fidèlement reconstruite en rajoutant à cette approximation grossière certains « détails » perçus grâce aux résolutions plus fines, c'est-à-dire quelques termes $\beta_\lambda\phi_\lambda$ correspondant aux niveaux de résolution $\Lambda(j), j \geq J$. Notons de nouveau $J_\star$ un entier naturel fixé *a priori* tel que $2^{J_\star} \leq n$. La collection de modèles inspirée de [BM00] exploite les propriétés de l'analyse multirésolution en considérant la famille

$\mathcal{M}$ des parties de Λ de la forme

$$m = \left[\bigcup_{j=-1}^{J-1} \Lambda(j) \right] \cup \left[\bigcup_{k=0}^{J_\star - J - 1} \Lambda'(J+k) \right]$$

où $\Lambda'(J+k)$ est un sous-ensemble quelconque de $\Lambda(J+k)$ de cardinal $\lfloor 2^J/(k+1)^3 \rfloor$, l'entier J étant autorisé à varier entre 0 et $J_\star - 1$. Chaque S_m est alors défini comme l'espace vectoriel engendré par les $\{\phi_\lambda\}_{\lambda \in m}$. Non seulement cette collection ne compte qu'un nombre sous-exponentiel de modèles de même dimension (voir par exemple [Mas07], paragraphe 4.3.5), mais en plus elle dispose des qualités d'approximations adéquates en norme $\mathbb{L}_q$ pour des fonctions présentant une régularité de type Besov, homogène ou non, comme démontré par Birgé et Massart [BM00]. Par ailleurs, l'implémentation de l'estimateur pénalisé basé sur cette collection et un contraste de type moindres carrés pénalisé ne nécessite que $\mathcal{O}(n \ln n)$ opérations (voir par exemple [DLT09]). Des inégalités de type oracle ainsi que des résultats d'adaptation ont été obtenues grâce à ce type de collection dans différents cadres statistiques : densité [BM97], régression à pas aléatoire avec erreurs sous-gaussiennes [Bar02; BCV01], intensité d'un processus de Poisson [RB03], modèle gaussien général [Mas07], loi discrète [DLT09].

I.5 Collections de modèles basés sur des partitions en intervalles, cubes, rectangles dyadiques

Présentons maintenant les collections de modèles auxquelles nous nous intéressons dans cette thèse. Comme dans le paragraphe précédent, nous en donnons les principales caractéristiques, fournissant ainsi les premiers éléments de comparaison avec les autres collections de modèles. Par ailleurs, nous décrivons les résultats déjà existants à leur sujet en estimation fonctionnelle, pour la plupart très récents, puisqu'obtenus parallèlement à la préparation de cette thèse. Nous exposerons dans la partie I.7 notre contribution.

I.5.1 Description des collections

Soit $J_\star$ un entier fixé a priori, dont la valeur dépend du cadre statistique. Pour l'estimation de fonctions à valeurs réelles définies sur $[0,1]$, nous utiliserons des modèles de fonctions constantes, voire polynomiales, par morceaux sur les partitions de $[0,1]$ en intervalles dyadiques de longueur au moins $2^{-J_\star}$. Ce sont toutes les partitions en intervalles de la forme

$$I_{(j,0)} = [0, 2^{-j}]$$

où $j \in \{0, \dots, J_\star\}$, ou

$$I_{(j,k)} =]k2^{-j}, (k+1)2^{-j}],$$

où $j \in \{1, \dots, J_\star\}$, $k \in \{0, \dots, 2^j - 1\}$, j pouvant varier d'un intervalle à un autre au sein de la même partition. L'ensemble de ces partitions peut également être décrit à l'aide de l'arbre binaire $\mathcal{A}$ de racine $(0,0)$ tel que

- pour tout $j \in \{1, \dots, J_\star\}$, les noeuds du niveau j sont indexés par les éléments de $\Lambda(j) = \{(j,k); k = 0, \dots, 2^j - 1\}$;
- pour tout $j \in \{1, \dots, J_\star\}$ et tout $k \in \{0, \dots, 2^j - 1\}$, les branches gauche et droite issues du noeud (j,k) conduisent respectivement aux noeuds $(j+1, 2k)$ et $(j+1, 2k+1)$, qu'on appellera les enfants du noeud (j,k).

L'ensemble des noeuds de $\mathcal{A}$ est $\Lambda = \cup_{j=0}^{J_\star} \Lambda(j)$, où $\Lambda(0) = \{(0,0)\}$. Pour $j \in \{0, \dots, J_\star\}$ et $k \in \{0, \dots, 2^j - 1\}$, l'intervalle dyadique $I_{(j,k)}$ s'identifie alors au noeud (j,k) de $\mathcal{A}$, et ses

enfants sont les intervalles dyadiques associés aux enfants de (j,k). Toute partition de $[0,1]$ en intervalles dyadiques de longueur $\geq 2^{-J_\star}$ correspond aux feuilles d'un arbre binaire complet obtenu par élagation de l'arbre $\mathcal{A}$, la correspondance étant même bijective. Précisons que par « complet » nous faisons référence à un arbre binaire dont tous les noeuds, sauf les feuilles, ont deux enfants. La figure I.5.1 représente l'arbre binaire $\mathcal{A}$ de tous les intervalles dyadiques de longueur au moins $2^{-J_\star}$ pour $J_\star = 3$. Les arbres des figures I.5.2 et I.5.3 fournissent des exemples de partitions de $[0,1]$ en intervalles dyadiques de longueur au moins $2^{-J_\star}$. Cette collection de partitions est donc plus riche qu'une collection de partitions régulières, sans pour autant contenir toutes les partitions construites sur la grille dyadique $\{k2^{-J_\star}; k = 0, \ldots, 2^{J_\star}\}$. Elle possède la même structure en arbre qu'une collection de partitions produite par l'algorithme CART, d'où une complexité sous-exponentielle. Cependant, elle ne dépend pas des données et sa structure, assez proche de celle de la collection de modèles associée à la stratégie Birgé-Massart, laisse présager de qualités d'approximation similaires.

Pour l'estimation de fonctions à valeurs réelles définies sur $[0,1]^d$, avec $d \geq 2$, il existe essentiellement deux extensions possibles de la définition d'un intervalle dyadique. Ou bien l'on considère les cubes dyadiques, c'est-à-dire les produits de d intervalles dyadiques de mêmes longueurs, ou bien l'on considère les rectangles dyadiques, c'est-à-dire les produits de d intervalles dyadiques, de longueurs éventuellement différentes. Les partitions en cubes dyadiques de côté au moins $2^{-J_\star}$ sont en correspondance bijective avec les sous-arbres 2^d-aires complets de l'arbre 2^d-aire contenant tous les cubes de côté au moins $2^{-J_\star}$, un arbre 2^d-aire complet étant un arbre dont tous les noeuds, sauf les feuilles, ont 2^d enfants. Par ailleurs, toute partition en rectangles dyadiques s'obtient à partir du cube unité $[0,1]^d$ par partitionnement récursif, en choisissant à chaque étape une direction de coupure et en remplaçant l'intervalle dyadique correspondant par ses deux enfants. Une partition en D rectangles dyadiques, $D \in \mathbb{N}^\star$, est donc caractérisée par la donnée d'une structure d'arbre binaire et d'une suite ordonnée de $D-1$ entiers compris entre 1 et d correspondant à des directions de coupure. Dans les Figures I.5.4 à I.5.7, la direction 1 correspond à la verticale et la suite d'entiers se lit sur les feuilles de l'arbre de gauche à droite et de haut en bas. Ces figures représentent deux partitions en rectangles dyadiques différentes associées à la même structure d'arbre.

I.5.2 Résultats existants

Même s'ils ne relèvent pas de la sélection de modèle, les travaux d'Engel sont la source d'une première approche pour ce type de résultat concernant les collections de partitions en intervalles, cubes ou rectangles dyadiques. Ils mettent en évidence et exploitent le lien entre histogrammes dyadiques et analyse multirésolution, via l'estimation par projection sur la base de Haar. En estimation de densité par exemple, Engel [Eng97] donne une condition nécessaire et suffisante pour qu'un estimateur par projection sur un espace vectoriel engendré par un ensemble d'ondelettes de Haar soit un histogramme : ce sous-ensemble doit correspondre à un sous-arbre binaire, pas nécessairement complet, de l'arbre $\mathcal{A}$ décrit dans le paragraphe précédent. Il définit un « histogramme multirésolution » possédant cette double nature, qui par conséquent est également un histogramme dyadique. Dans le cadre de la régression gaussienne, Engel [Eng94] obtient l'un des premiers résultats sur la vitesse de convergence des régressogrammes dyadiques, pour le risque quadratique intégré et sur des classes de fonctions lipschitziennes, cette vitesse correspondant bien à la vitesse minimax. Dans les deux cas, il propose un algorithme, justifié heuristiquement, pour sélectionner un estimateur adéquat à partir des données.

Les premiers résultats d'adaptation spatiale en relation avec la collection des partitions en rectangles dyadiques sont démontrés par Donoho [Don97]. Ce dernier s'inspire essentiellement de l'argument de Engel et de travaux auxquels il a contribué sur la sélection d'une meilleure base

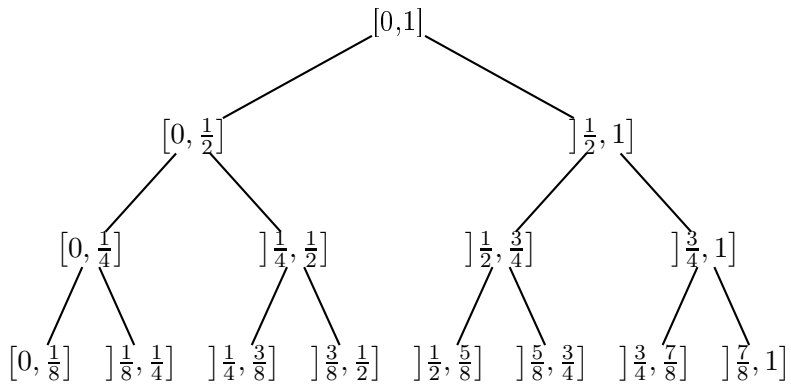

FIG. I.5.1 – Arbre binaire $\mathcal{A}$ des intervalles dyadiques de $[0, 1]$ de longueur au moins $2^{-J_\star}$, pour $J_\star = 3$.

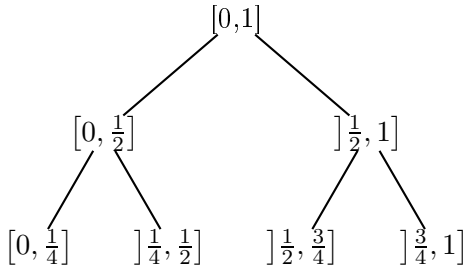

FIG. I.5.2 – Exemple de partition régulière de $[0, 1]$ en intervalles dyadiques.

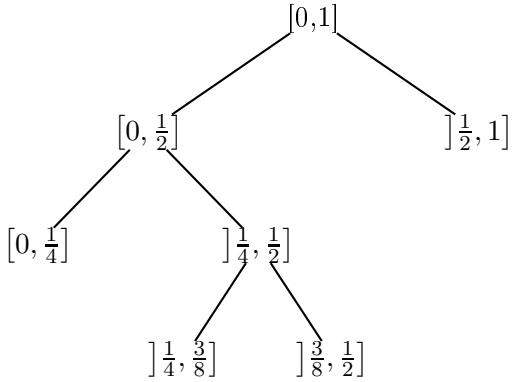

FIG. I.5.3 – Exemple de partition irrégulière de $[0, 1]$ en intervalles dyadiques.

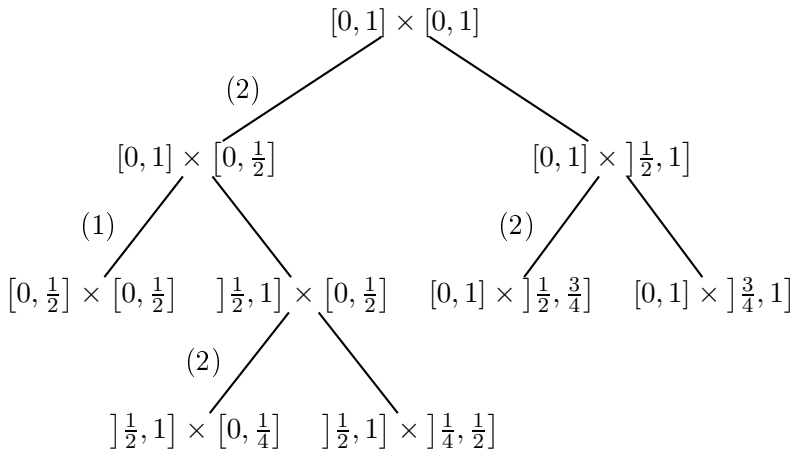

FIG. I.5.4 – Arbre dyadique dont les feuilles représentent une partition de $[0,1]^2$ en rectangles dyadiques.

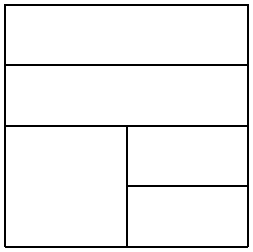

FIG. I.5.5 – Partition de $[0,1]^2$ en rectangles dyadiques correspondant aux feuilles de l'arbre I.5.4 et à la suite $(2,1,2,2)$.

orthogonale [Don93; Don96; DJ94a], aussi bien en estimation qu'en théorie de l'approximation. Donoho [Don97] se place dans un cadre de régression gaussienne, semblable au Problème 2 décrit dans le paragraphe I.1, si ce n'est que la fonction à estimer est définie sur le carré unité et échantillonnée uniformément en $x_{i_1,i_2} = (i_1/n, i_2/n)$ pour $0 \leq i_1, i_2 \leq n-1$. Il baptise « CART dyadique » une procédure de sélection parmi les modèles de fonctions constantes par morceaux sur une partition de $[0,1]^2$ en rectangles dyadiques, basée sur un critère de type moindres carrés pénalisé. La pénalité associée à chaque partition correspond au nombre de rectangles de la partition à un facteur près, proportionnel à $\sigma^2 \ln(n)/n$. Par ailleurs, il introduit une famille de fonctions de Haar anisotropes et la collection des modèles engendrés par les sous-familles de fonctions de Haar vérifiant certaines contraintes. Ces contraintes, qualifiées d'héréditaires, sont une extension directe au cadre multi-dimensionnel de celles décrites par Engel [Eng97]. L'astuce utilisée par Donoho tient dans l'équivalence entre la procédure CART dyadique et une certaine procédure de sélection parmi cette seconde collection, procédure dénommée « Best Orthogonal Basis with hereditary constraints ». En effet, alors que la définition même de la procédure CART dyadique la rend implémentable en temps linéaire, ses propriétés théoriques — inégalité d'oracle et adaptation sur des espaces de Besov anisotropes de régularité non homogène à un facteur $\ln(n)$ près — sont déduites de celles de la seconde procédure.

Si leur approche est inspirée de celles de Donoho, puisque basée sur des fonctions de type Haar, Kolaczyk et Nowak [KN04] s'intéressent à un sujet différent. Ils proposent en effet un cadre unifié pour l'étude de divers problèmes de régression par maximum de vraisemblance pénalisé. Ils considèrent une collection de modèles de fonctions constantes par morceaux basées sur des « partitions dyadiques récursives » qui ne sont autres que des partitions en intervalles dyadiques. Pour l'estimation de la moyenne d'un signal gaussien, poissonien ou multinomial,

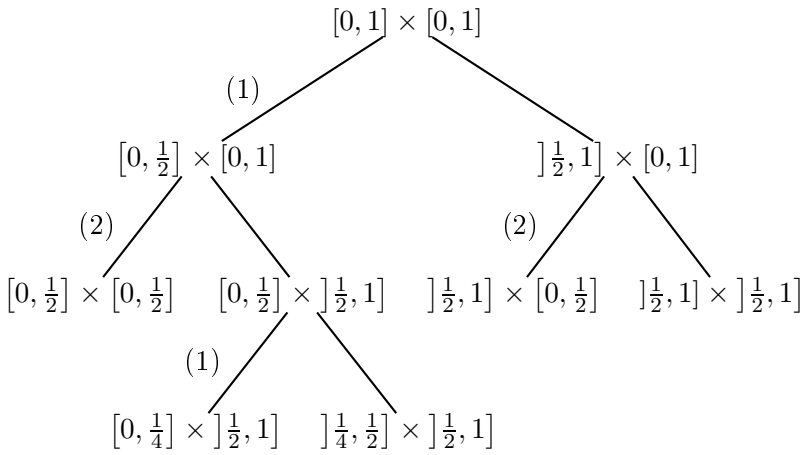

FIG. I.5.6 – Arbre dyadique dont les feuilles représentent une partition de $[0,1]^2$ en rectangles dyadiques.

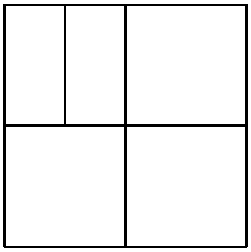

FIG. I.5.7 – Partition de $[0,1]^2$ en rectangles dyadiques correspondant aux feuilles de l'arbre I.5.6 et à la suite $(1,2,2,1)$.

ils obtiennent des résultats d'adaptation au sens minimax pour le risque quadratique associé à la perte de Hellinger ou le risque quadratique intégré. Ces résultats, toujours à un facteur logarithmique près, concernent des fonctions à variations bornées ou présentant une régularité de type Besov, éventuellement non homogène.

Tout récemment, Klemelä [Kle09] emprunte lui-aussi la même démarche que Donoho pour l'estimation d'une densité multi-dimensionnelle. Il considère un critère de types moindres carrés pénalisé pour sélectionner un meilleur histogramme construit sur une partition en rectangles dyadiques. Pour le risque quadratique intégré, il obtient des résultats d'adaptation au sens minimax, encore à un facteur logarithmique près, pour des classes de fonctions anisotropes similaires à celles considérées par [Don97], la régularité pouvant cependant être mesurée dans des normes $\mathbb{L}_p$ qui diffèrent selon la direction.

Les travaux de Willett et Nowak [WN07] font partie des premiers à se démarquer de l'approche de Donoho. Ils portent sur l'estimation de densité et l'estimation de l'intensité d'un processus de Poisson, dans les cas uni- et multi-dimensionnels, par maximum de vraisemblance pénalisé. Ils reposent uniquement sur la collection des partitions en intervalles ou cubes dyadiques, aussi il n'y est pas question d'anisotropie. L'originalité de leur procédure tient au fait qu'elle permet de choisir non seulement la meilleure partition en cubes dyadiques à partir des données, mais aussi un meilleur polynôme sur chaque cube dont le degré maximal dépend lui aussi des données et non d'une limite fixée *a priori* par l'utilisateur. En dimension 1 par exemple, leur estimateur reste implémentable, et nécessite $\mathcal{O}(n\ln(n))$ calculs. D'un point de vue théorique, ils étudient le risque quadratique associé à la distance de Hellinger en adoptant clairement un point de vue asymptotique. Dans le cas unidimensionnel, leur estimateur n'atteint la vitesse minimax qu'à un facteur logarithmique près, sur des classes de Besov dont la

régularité peut être arbitrairement grande pourvu que le nombre de données n soit suffisamment grand. Ces classes de Besov contiennent certes des fonctions de régularité non homogène, mais nécessairement continues. Dans le cadre multidimensionnel, [WN07] proposent un résultat d'adaptation sur des classes de Hölder, toujours à un facteur logarithmique près, mais pas de résultat d'adaptation spatiale.

Blanchard *et al.* [BSRM07] démontrent des inégalités de type oracle, éventuellement à un facteur logarithmique près, pour des histogrammes construits sur des partitions en rectangles dyadiques dans les cadres suivants : en classification, pour l'estimation de la loi de probabilité conditionnellement à la classe par moindres carrés ou maximum de vraisemblance pénalisé, et l'estimation de densité par maximum de vraisemblance pénalisé. Leurs résultats découlent d'un théorème de sélection de modèles général pour fonction de perte bornée dû à Blanchard *et al.* [BBM08]. Ils obtiennent certes un résultat d'adaptation à l'anisotropie, mais uniquement pour des fonctions höldériennes, n'exploitant donc pas les capacités d'adaptation spatiale de la collection. En revanche, ils proposent une amélioration de l'algorithme présenté par [Don97; Kle09] dans un cadre multivarié. Leur étude par simulations montre notamment que leur procédure se compare favorablement à l'algorithme CART en classification.

Enfin, Birgé [Bir06b] propose d'effectuer de la sélection d'histogrammes construits sur des partitions en intervalles dyadiques en utilisant sa procédure de sélection basée sur des tests multiples. Seront ainsi traitées l'estimation de densité [Bir06b], de l'intensité d'un processus de Poisson [Bir07] et de l'intensité d'une mesure aléatoire par Baraud et Birgé [BB09]. Dans chacun de ces cas, la performance de l'estimateur est mesurée via le risque quadratique associé à une distance de type Hellinger. C'est à Birgé [Bir06b] que l'on doit le rapprochement entre la sélection de modèle et le résultat d'approximation de DeVore et Yu [DY90] pour les espaces de Besov, étendu dans [Bir07] aux fonctions à α-variations bornées. L'utilisation de cette famille de partitions est en fait beaucoup plus ancienne en théorie de l'approximation, et remonte au moins à Birman et Solomjak [BS67]. En effet, une manière naturelle de construire une approximation, disons constante par morceaux, d'une fonction à valeurs réelles définie sur $[0,1]$ consiste à partir de la partition triviale $\{[0,1]\}$, puis à raffiner successivement la partition en remplaçant un intervalle bien choisi de la partition par ses deux enfants. Si les procédures définies dans [Bir06b; Bir07; BB09] possèdent des propriétés théoriques intéressantes (inégalité de type oracle et adaptation spatiale, sans facteur logarithmique), en pratique, leur implémentation semble à ce jour difficilement réalisable en raison d'une complexité algorithmique quadratique en la taille de la collection de partitions (*cf.* [BB09]). Or, celle-ci est de l'ordre de 4^n lorsqu'on se restreint à des intervalles de longueur au moins $1/n$.

I.6 Autres procédures spatialement adaptatives au sens minimax

Parmi les procédures spatialement adaptatives au sens minimax autres que la sélection de modèle décrite aux paragraphes I.4 et I.5, nous distinguerons les procédures par seuillage dans une base d'ondelettes, les procédures avec paramètre de lissage variable inspirées de la méthode dite de Lepski, et les procédures par régularisation. Nous nous attarderons un peu plus sur les premières, qui sont devenues des procédures de référence, tant du point de vue théorique que pratique, et font l'objet d'une abondante littérature.

Seuillage dans une base d'ondelettes. Ecrivons le développement de $s \in \mathbb{L}_2([0,1])$ dans une

base d'ondelettes caractérisée par une ondelette-père ϕ et une ondelette-mère ψ sous la forme

$$s = \sum_{k\in\Lambda(-1)} \alpha_k\phi_k + \sum_{j\in\mathbb{N}}\sum_{k\in\Lambda(j)} \beta_{j,k}\psi_k,$$

où $\Lambda(-1)$ est un ensemble fini et, pour $j \geq 0$, $|\Lambda(j)|$ est de l'ordre de 2^j. Les procédures de seuillage coefficient par coefficient conduisent à des estimateurs de s de la forme

$$\tilde{s} = \sum_{k\in\Lambda(-1)} \hat{\alpha}_k\phi_k + \sum_{j=0}^{J_\star}\sum_{k\in\Lambda(j)} \eta(\hat{\beta}_{j,k},\lambda_j)\psi_k, \tag{I.6.16}$$

où les $\hat{\alpha}_k$ et $\hat{\beta}_{j,k}$ sont les coefficients empiriques, auxquels on applique éventuellement une fonction de seuillage η. La fonction η a pour effet de remplacer par 0 les coefficients empiriques du niveau de résolution j inférieurs en valeur absolue à un certain seuil λ_j et agit comme une application contractante sur les autres coefficients. L'indice $J_\star$ correspondant à la résolution la plus fine est choisi *a priori* et joue le même rôle que celui défini au paragraphe I.4.3. Dans le cadre de régression décrit dans le Problème 2, avec des x_i déterministes uniformément répartis sur $[0,1]$, Donoho et Johnstone [DJ94b] (voir aussi [DJKP95]) proposent l'estimateur *VisuShrink*, basé sur la fonction de seuillage doux et un seuil universel, ne dépendant ni des données ni du niveau de résolution, choisi de manière à vérifier une inégalité d'oracle. Cet estimateur, qui ne nécessite que $\mathcal{O}(n)$ calculs, possède, pour le risque quadratique intégré, des propriétés d'adaptation spatiale sur des boules de Besov de régularité α, à un facteur $(\ln(n))^{2\alpha/(1+2\alpha)}$ près. Dans ce même cadre, l'estimateur *NeighCoeff* de Cai et Silverman [CS01], qui diffère légèrement par le type de seuillage et le choix du seuil, possède des propriétés similaires. Des extensions de *VisuShrink* à des fonctions anisotropes sont proposées par Neumann [Neu00] et Neumann et Von Sachs [NvS97], dans le modèle de bruit blanc gaussien et pour l'estimation du spectre évolutif, toujours avec des propriétés d'adaptation spatiale à un facteur logarithmique près. En estimation de densité, Donoho *et al.* [DJKP96] proposent un estimateur par seuillage fort, avec un seuil dépendant du niveau de résolution mais pas des données. Pour le risque $\mathbb{L}_q$ intégré, $q \geq 1$, cet estimateur s'adapte à un facteur $(\ln(n))^{q\alpha/(1+2\alpha)}$ sur des boules de Besov de régularité α, pourvu que les paramètres de l'estimateur soient bien choisis en fonction de q et du rayon de la boule. Il existe des extensions de cet estimateur à des données censurées (*cf.* Li [Li07]), ainsi qu'à des données dépendantes, proposées par Clémençon [Clé00a; Clé00b] (densité stationnaire et densité de transition d'une chaîne de Markov) et Gannaz et Wintenberger [GW09] (densité pour données faiblement dépendantes). Enfin, dans le cadre de régression évoqué en début de paragraphe, deux procédures par seuillage s'adaptent spatialement sans facteur logarithmique : l'estimateur *SureShrink* de Donoho et Johnstone [DJ95] et celui défini par Juditsky [Jud97]. Dans les deux cas, le choix du seuil dépend à la fois du niveau de résolution et des données, mais l'estimateur reste implémentable avec une complexité algorithmique en $\mathcal{O}(n\ln(n))$. Alors que la définition de *SureShrink* est basée sur l'estimation sans biais du risque quadratique de Stein, l'estimateur de Juditsky repose sur la méthode de Lepski [Lep91] et ses performances sont mesurées via le risque $\mathbb{L}_q$ intégré, $q \geq 1$.

Une alternative au seuillage coefficient par coefficient est le seuillage par blocs : dans l'expression (I.6.16), au sein de chaque niveau de résolution $\Lambda(j)$, les coefficients empiriques sont d'abord regroupés par blocs, et c'est à ces derniers qu'on applique la fonction de seuillage. De tels estimateurs sont initialement étudiés par Hall *et al.* en densité [HKP98] et régression [HKP99], pour le risque quadratique intégré, avec un seuil constant et des blocs de longueur $\mathcal{O}(\ln^2(n))$. Ils s'adaptent spatialement sur des classes de fonctions définies par superposition d'une fonction présentant des irrégularités d'un certain type à une fonction de régularité homogène. Cai [Cai02] en régression et Chicken et Cai [CC05] en densité montrent que des estimateurs similaires, avec des blocs de longueur $\mathcal{O}(\ln(n))$, s'adaptent aussi localement.

L'estimateur *BlockJS* de Cai [Cai99], autre variante de [HKP99] avec des blocs de longueur $\mathcal{O}(\ln(n))$, et l'estimateur *NeighBlock* de [CS01] se calculent en temps linéaire et possèdent des propriétés d'adaptation spatiale sur les traditionnelles boules de Besov. Cependant, pour le risque quadratique intégré et sur des boules de régularité α non homogène, mesurée dans la norme $\mathbb{L}_p$, l'adaptation n'est obtenue qu'à un facteur $(\ln(n))^{2(1/p-1/2)/(1+2\alpha)}$ près, qui se détériore donc lorsque p décroît. Mentionnons également l'approche unifiée de Chesneau [Che08] – et ses conséquences [Che09; Che07]– qui étudie une variante de *BlockJS* pour le risque $\mathbb{L}_q$ intégré. Néanmoins, il apparaît toujours un facteur logarithmique dans le cas où la régularité n'est pas homogène. Nous terminerons la revue des méthodes d'ondelettes par la toute récente procédure *SureBlock* de Cai et Zhou [CZ09] en régression, qui, comme son nom l'indique, est une version « seuillage par blocs » de la procédure *SureShrink* [DJ95]. La taille des blocs ainsi que les seuils dépendent du niveau de résolution et sont choisis à partir des données de manière à vérifier une inégalité d'oracle. La complexité algorithmique n'est pas mentionnée mais cet estimateur est capable de s'adapter spatialement et sans facteur logarithmique.

Méthode de Lepski. Lepski *et al.* [LMS97], dans le modèle de bruit blanc gaussien, et Goldenshluger et Nemirovksi [GN97], dans le modèle de régression gaussien, proposent des estimateurs avec paramètre de lissage variable inspirés de la méthode de Lepski [Lep91]. Etant donnée une famille d'estimateurs à noyaux, le paramètre de lissage est choisi en fonction des données de manière à assurer des propriétés d'adaptation locale. Ces estimateurs s'adaptent égalemant spatialement, sur des boules de Sobolev [GN97] ou de Besov [LMS97], à un facteur logarithmique près. Une extension de [LMS97] dans un cadre multidimensionnel est étudiée par Kerkyacharian *et al.* [KLP01], qui obtiennent des résultats d'adaptation spatiale à facteur logarithmique près sur des boules de Besov anisotropes.

Régularisation. Etant donné un entier $k \in \mathbb{N}^\star$ choisi *a priori*, Mammen et Van de Geer([MvdG97]) proposent, dans le cadre de régression du Problème 2, d'estimer s en minimisant le critère de type moindres carrés pénalisé

$$\sum_{i=1}^{n}(Y_i - t(x_i))^2 + \lambda V_1(t^{(k-1)})$$

sur une certaine classe de fonctions. Ici, λ désigne une constante, V_1 la semi-norme dans l'espace des fonctions à variations bornées, et $t^{(k-1)}$ la dérivée d'ordre $k-1$ de t. Ils obtiennent ainsi un estimateur polynomial par morceaux sur une partition choisie en fonction des données. Cet estimateur atteint notamment la vitesse minimax à constante près pour les fonctions dont la dérivée d'ordre $k-1$ est à variations bornées. La complexité algorithmique est précisée dans le cas $k=1$, de l'ordre de $n\ln(n)$ pour une valeur donnée de λ et de l'ordre de n^2 pour toutes les valeurs de λ.

I.7 Présentation des résultats de la thèse

Dans cette thèse, nous nous proposons d'estimer diverses fonctions par sélection de polynômes par morceaux construits sur des partitions en intervalles, cubes ou rectangles dyadiques, à l'aide d'un critère de type moindres carrés pénalisé, les deux premiers chapitres étant consacrés à la sélection d'histogrammes. Les cadres statistiques que nous étudions sont les suivants. Dans le Chapitre II, nous nous intéressons à l'estimation d'une loi de probabilité discrète. De plus, l'estimateur que nous construisons est utilisé au cours de l'étape préliminaire d'une procédure de détection de ruptures. Dans les deux chapitres suivants, nous sommes confrontés à

des paramètres de nuisance. En effet, le Chapitre III présente une approche unifiée pour l'estimation fonctionnelle basée sur des données éventuellement censurées. Le Chapitre IV porte sur l'estimation de densité conditionnelle pour des données dépendantes, la structure de dépendance étant en partie inconnue. Ces différents cadres statistiques ont été peu étudiés jusqu'ici par des méthodes adaptatives, voire par des méthodes non-paramétriques, comme indiqué dans la bibliographie propre à chaque chapitre. Cependant, certains problèmes de référence, tels que l'estimation de la fonction de régression et l'estimation de densité univariée ou multivariée, correspondent à des cas particuliers des cadres considérés dans les Chapitres III et IV, ce qui autorise la comparaison avec les diverses procédures statistiques existantes. Nous parvenons dans tous les cas à établir des inégalités de type oracle ainsi que des propriétés d'adaptation spatiale, pour une pénalité simplement linéaire en la dimension des modèles. Ces résultats sont obtenus *sans* le facteur logarithmique qui apparaît dans la grande majorité des travaux évoqués dans cette introduction, dont ceux portant déjà sur cette collection de partitions. Pour cela, nous utilisons de manière essentielle la structure en arbre de la collection de partitions et les propriétés combinatoires qui en découlent, des arguments de concentration propres aux modèles non réguliers, ainsi que le résultat d'approximation de DeVore et Yu [DY90]. Cependant, chaque nouveau cadre nécessite de démontrer de nouveaux résultats d'approximation, qui ont leur propre intérêt. Par ailleurs, nous décrivons un algorithme de calcul exact de notre estimateur, que nous formulons comme un simple algorithme de plus court chemin, et dont la complexité est seulement linéaire en la taille de l'échantillon.

Présentons plus précisément les différents problèmes que nous étudions ainsi que les principaux résultats établis.

Chapitre II. Les travaux de ce chapitre font l'objet d'une publication dans la revue *ESAIM : Probability and Statistics*, intitulée *Estimating a discrete distribution via histogram selection* [Aka09].

À partir de l'observation de n variables aléatoires indépendantes $Y_1, \dots, Y_n$ à valeurs dans l'ensemble $\{1, \dots, r\}$, où r est un entier naturel connu, $r \geq 2$, il s'agit d'estimer la loi de probabilité s du vecteur $(Y_1, \dots, Y_n)$. Cette loi s est vue comme une fonction à valeurs dans $\mathbb{R}^r$ dont la l^{e} coordonnée, $l = 1, \dots, r$, est la fonction

$$i \in \{1, \dots, n\} \mapsto \mathbb{P}(Y_i = l).$$

Nous considérons ici la collection $\mathcal{M}$ des partitions de $\{1, \dots, n\}$ en intervalles dyadiques, en supposant que n est une puissance de 2, chaque modèle S_m, $m \in \mathcal{M}$, étant constitué des fonctions définies sur $\{1, \dots, n\}$ à valeurs dans $\mathbb{R}^r$ constantes par morceaux sur la partition m. On obtient une collection d'estimateurs $\{\hat{s}_m\}_{m \in \mathcal{M}}$ par minimisation sur chaque modèle d'un contraste de type moindres carrés γ adéquat, comme expliqué dans le paragraphe I.4. Notre étude porte alors sur l'estimateur pénalisé $\tilde{s} = \hat{s}_{\hat{m}}$ où $\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}} \{\gamma(\hat{s}_m) + \operatorname{pen}(m)\}$, pour une pénalité de la forme

$$\operatorname{pen}(m) = c D_m$$

où c désigne une constante et D_m le nombre d'intervalles de la partition m. Nous justifions d'un point de vue théorique cette forme de pénalité en prouvant que $\tilde{s}$ vérifie une inégalité de type oracle et atteint effectivement la vitesse d'estimation minimax sur des classes de fonctions présentant une régularité de type Besov ou à α-variations bornées. Pour ce faire, nous démontrons notamment un résultat d'approximation similaire à celui évoqué en (I.2.10), l'extension de [DY90] à notre cadre n'étant pas immédiate puisque les fonctions que nous étudions sont à valeurs dans $\mathbb{R}^r$, avec $r \geq 2$. Pour le calcul de l'estimateur $\tilde{s}$, il n'est pas nécessaire de déterminer tous les estimateurs de la collection, dont le nombre est exponentiel en n. En effet, nous expliquons comment tirer profit de l'additivité du contraste et de la forme de la pénalité pour

obtenir comme nouvelle caractérisation de la partition à sélectionner

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \sum_{I \in m} (\mathcal{L}(I) + c)$$

où les quantités $\mathcal{L}(I)$ sont exprimées explicitement en fonction des données pour chaque intervalle dyadique I. Cette caractérisation permet d'utiliser un algorithme de plus court chemin pour déterminer $\hat{m}$, dont la complexité en $\mathcal{O}(n)$ résulte essentiellement du nombre d'intervalles dyadiques – également en $\mathcal{O}(n)$ – servant à construire toutes les partitions de la collection. Grâce à une étude de simulations, nous proposons une manière de choisir, en pratique, une constante de pénalité c pertinente. De plus, pour ce choix de constante, notre estimateur se compare assez favorablement à celui basé sur la stratégie Birgé-Massart étudié dans [DLT09].

Dans un deuxième temps, nous proposons une procédure hybride pour la détection de ruptures dans la loi s. Une difficulté de ce problème réside dans le fait que nous ne disposons *a priori* d'aucune information sur le nombre ou la position de ces ruptures. L'estimateur $\tilde{s}$ seul n'est pas tout à fait adapté à ce problème étant donnée la contrainte sur la structure des partitions. Aussi, nous définissons une procédure hybride comme suit. On détermine tout d'abord la partition $\hat{m}$ associée à $\tilde{s}$. Puis on considère la collection (aléatoire) des partitions construites sur $\hat{m}$, et l'on procède à la sélection de la meilleure partition parmi cette collection à l'aide d'un critère de type moindres carrés pénalisé. On en déduit également un nouvel estimateur $\tilde{s}_{hyb}$ de s. Une qualité majeure de cette procédure est son temps de calcul extrêmement réduit. Par ailleurs, nous montrons que cet estimateur hybride conserve des propriétés d'adaptation similaires à celles de l'estimateur initial. Nous implémentons cette procédure aussi bien sur des données simulées que sur des données réelles de grande dimension. En effet, nous proposons une application à la détection de ruptures dans une vraie séquence d'ADN.

Chapitre III. Ce chapitre présente une approche unifiée pour l'estimation fonctionnelle basée sur des données éventuellement censurées, qui interviennent fréquemment en analyse des données de survie ou en fiabilité par exemple. Nous nous plaçons dans le cadre général suivant. Notre but est d'estimer une fonction s à valeurs réelles sur la base d'observations indépendantes $Z_1 = (X_1, Y_1)$, ..., $Z_n = (X_n, Y_n)$, $n \geq 3$, où $X_i \in [0,1]$ peut être considéré comme un instant d'observation et Y_i est à valeurs dans un espace $\mathcal{Y}$ donné. On désigne par $\mathbb{P}_s$ la mesure de probabilité sous-jacente et par $\mathbb{E}_s$ l'espérance associée, et on note $\mathbb{L}_2([0,1])$ l'espace des fonctions de carré intégrable par rapport à une mesure de probabilité μ, muni du produit scalaire et de la norme usuels notés $\langle \,.\,, \,.\, \rangle$ et $\| \,.\, \|$. Nous considérons ici dans un premier temps une collection finie quelconque $\mathcal{M}^\star$ de partitions de $[0,1]$ et définissons chaque modèle S_m comme l'ensemble des fonctions constantes par morceaux sur la partition m. Nous supposons que s appartient à un sous-espace donné $\mathcal{S}$ de $\mathbb{L}_2([0,1])$ et que, pour tout $t \in \mathcal{S}$,

$$\langle t, s \rangle = \mathbb{E}_s \left[\frac{1}{n} \sum_{i=1}^{n} w(Z_i) t(X_i) \right]$$

où $w : [0,1] \times \mathcal{Y} \to \mathbb{R}$ est une fonction mesurable donnée qui dépend éventuellement de paramètres inconnus (par exemple, d'un paramètre de nuisance dû à la censure). Alors, s minimise sur $\mathcal{S}$

$$t \mapsto \mathbb{E}_s \left[\|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} w(Z_i) t(X_i) \right] = \|s - t\|^2 - \|s\|^2,$$

donc nous considérons des estimateurs définis par minimisation du critère

$$\gamma(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{w}(Z_i) t(X_i)$$

sur les modèles S_m, $m \in \mathcal{M}^\star$, pour un estimateur donné $\hat{w}$ de w. L'estimateur $\tilde{s}$ est alors défini comme en I.4, avec une fonction de pénalité qu'il reste à choisir. Nous illustrons la pertinence de ce cadre général en détaillant les exemples suivants : estimation de la fonction de régression, estimation de densité avec données censurées ou non, estimation du taux de survie avec données censurées. Nous formulons de manière concise des hypothèses générales permettant de démontrer que l'estimateur $\tilde{s}$ vérifie une inégalité de type oracle, éventuellement à un facteur près dépendant de n, pour une pénalité adéquate. Ces hypothèses portent sur la structure de la collection de partitions, l'existence de moments exponentiels de certaines variables aléatoires et la qualité de $\hat{w}$ en tant qu'estimateur de w.

Hypothèse (P) *Il existe une partition $m^\star$ sur laquelle sont construites toutes les partitions de $\mathcal{M}^\star$ et telle que $\inf_{I\in m^\star} n\mu(I) \geq 1$. De plus, pour tout $I \in m^\star$, $\mathbb{1}_I \in \mathcal{S}$.*

Hypothèse (M) *Il existe une constante positive a et une constante strictement positive v telles que, pour tout $I \in m^\star$ et tout $\lambda \in]-1/a, 1/a[$,*

$$\ln \mathbb{E}_s\left[\exp\left(\lambda \sum_{i=1}^{n}\Big(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w(Z_i)\mathbb{1}_I(X_i)\big]\Big)\right)\right] \leq \frac{nv\mu(I)\lambda^2}{2(1-a|\lambda|)}.$$

Hypothèse (W) *Il existe un réel positif C_Δ, qui dépend éventuellement de w et s, tel que pour tout $I \in m^\star$,*

$$\mathbb{E}_s\left[\left(\frac{1}{n\mu(I)}\sum_{i=1}^{n}(\hat{w}-w)(Z_i)\mathbb{1}_I(X_i)\right)^2\right] \leq \frac{C_\Delta}{n}.$$

Dans le cas particulier où $\mathcal{M}^\star$ est la collection des partitions en intervalles dyadiques de longueur suffisamment grande, une pénalité adéquate peut s'écrire sous la forme

$$\mathrm{pen}(m) = c\frac{\widehat{M}D_m}{n},$$

où c est une constante, $\widehat{M}$ est un estimateur d'un terme de variance et D_m, le nombre d'intervalles de m. Nous explicitons dans de nombreux cas $\widehat{M}$ en fonction des données. Nous montrons alors des résultats d'adaptation spatiale sur des classes de fonctions à α-variations bornées ou des classes de fonctions présentant une régularité de type Besov plus larges que celles habituellement considérées. En effet, la limite inférieure sur la régularité α est usuellement en $1/p$ pour les fonctions présentant une régularité non-homogène mesurée dans une norme $\mathbb{L}_p$, et nous obtenons la condition moins restrictive

$$\alpha > \frac{1}{2}\left(\frac{1}{p}-\frac{1}{2}\right)\left(1+\sqrt{\frac{2+3p}{2-p}}\right)$$

pourvu que les fonctions d'une même classe soient uniformément bornées. D'autre part, contrairement à [DY90; Bir06a; Bir07; BB09], nous devons prendre en compte la contrainte sur la taille minimale des intervalles pour établir les qualités d'approximation de la collection de modèles et la vitesse d'estimation qui en résulte.

Chapitre IV. Nous proposons dans ce chapitre une étude simultanée de l'estimation de la densité et de la densité conditionnelle, lorsque les données sont dépendantes. Aussi, la fonction s à estimer est définie dans ce chapitre sur $[0,1]^d$, où $d \geq 2$. Etant donné un entier naturel r, nous

considérons tout d'abord une collection $\mathcal{M}^\star$ de partitions de $[0,1]^d$ assez générale, et définissons chaque modèle S_m, $m \in \mathcal{M}^\star$, comme l'ensemble des fonctions polynomiales par morceaux sur m de degré au plus r en chaque coordonnée. Nous proposons différentes hypothèses de dépendance, basées sur les coefficients de mélange ou une condition d'indépendance conditionnelle, sous lesquelles les estimateurs sur un modèle ou l'estimateur pénalisé se comportent presque aussi bien que dans le cas où les données sont dépendantes. Puis, nous nous intéressons plus particulièrement aux collections de partitions en cubes dyadiques et en rectangles dyadiques. Pour chacune de ces collections, nous proposons une pénalité de la forme

$$\mathrm{pen}(m) = c\vartheta \ln^\delta(n) \frac{\mathcal{N}(s,f) D_m}{n}.$$

Ici, c désigne une constante positive et D_m, la dimension du modèle S_m. Le réel positif ϑ ainsi que δ, qui vaut soit 0, soit 1, changent selon l'hypothèse de dépendance. La quantité $\mathcal{N}(s,f)$ dépend de s et de la densité f des variables de conditionnement, mais peut en pratique être remplacée par un estimateur adéquat. Chacune de ces collections permet d'obtenir des résultats d'adaptation spatiale, sur des boules de Besov isotropes pour la première, et plus généralement sur des boules de Besov anisotropes pour la seconde. Ce dernier résultat repose notamment sur une extension de [DY90] à des fonctions présentant une régularité anisotrope. Bien que [Don97; Kle09] aient étudié l'adaptation sur des espaces de ce genre, rappelons que leur ligne de preuve, basée sur des ondelettes de Haar, est différente de celle que nous proposons et se limite à des fonctions dont les degrés de régularité dans chacune des directions ne peuvent être supérieurs à 1.

En étoffant ainsi la liste des cadres statistiques étudiés, nous contribuons à montrer l'universalité de la procédure de sélection de partitions en rectangles dyadiques. Tout en étant implémentable avec une complexité on ne peut plus raisonnable, cette procédure possède des propriétés d'adaptation qui s'étendent jusqu'aux fonctions de régularité à la fois non homogène et anisotrope. A notre connaissance, il n'existe pas d'autre estimateur possédant ces qualités. De plus, cette procédure semble pertinente en pratique. Ces différents résultats la rendent donc tout à fait compétitive par rapport aux méthodes déjà existantes, dont les méthodes d'ondelettes. D'autre part, les travaux présentés ici ouvrent des perspectives de recherche décrites en fin de thèse.

Chapter II

Estimating a discrete distribution via dyadic histogram selection

This chapter is a slightly modified version of the article *Estimating a discrete distribution via histogram selection* by the author, to appear in *ESAIM: Probability and Statistics.*

Abstract

In this chapter, our aim is to estimate the joint distribution of a finite sequence of independent categorical variables. We consider the collection of partitions into dyadic intervals and the associated histograms, and we select from the data the best histogram by minimizing a penalized least-squares criterion. The choice of the collection of partitions is inspired from approximation results due to DeVore and Yu. Our estimator satisfies a nonasymptotic oracle-type inequality and adaptivity properties in the minimax sense. Moreover, its computational complexity is only linear in the length of the sequence. We also use that estimator during the preliminary stage of a hybrid procedure for detecting multiple change-points in the joint distribution of the sequence. That second procedure still satisfies adaptivity properties and can be implemented efficiently. We provide a simulation study and apply the hybrid procedure to the segmentation of a DNA sequence.

II.1 Introduction

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables taking values in the finite set $\{1, \ldots, r\}$, where r is an integer and $r \geq 2$. Let s be the joint distribution of $(Y_1, Y_2, \ldots, Y_n)$, that we consider as the $\mathbb{R}^r$-valued function defined on $\{1, \ldots, n\}$ with l-th coordinate function

$$i \in \{1, \ldots, n\} \mapsto \mathbb{P}(Y_i = l),$$

for $l = 1, \ldots, r$. The aim of this chapter is to study a nonparametric estimator of the distribution s. References treating about this problem are so scarce that we can only cite three of them. Aerts and Veraverbeke [AV95] propose a kernel estimator, whose convergence rate is given under a Lipschitz regularity condition. More recently, Lebarbier and Nédélec [LN07] and then Durot, Lebarbier and Tocquet [DLT09] have studied procedures based on the model selection principle introduced by Barron, Birgé and Massart [BBM99]. Thus, all their results are nonasymptotic. In both cases, a family of linear spaces of real-valued functions defined on $\{1, \ldots, n\}$ is given, and the procedures allow to select from the data, by minimizing some penalized criterion, one space among that family in which all the coordinate functions of s are estimated. The choice of the penalty is supported by an oracle-type inequality. Lebarbier *et al.* [LN07] consider two different penalized criteria, one based on least-squares, the other one on maximum-likelihood, and spaces of piecewise constant functions. Durot *et al.* [DLT09] consider only a penalized least-squares criterion, but provide an oracle-type inequality that is valid for almost all finite families of linear spaces. Moreover, they are particularly interested in three families of spaces. The so-called exhaustive indicator strategy corresponds with the family made up of all spaces of functions piecewise constant on some partition of $\{1, \ldots, n\}$, a family already encountered in [LN07]; the exhaustive Haar and non-exhaustive Haar (or *neH*) strategies are based on families made up of spaces generated by some Haar wavelets. In these three cases, the resulting estimator is proved to have adaptivity properties. Due to the richness of the underlying families of spaces, both exhaustive strategies yield estimators that only satisfy an oracle-type inequality up to a $\ln(n)$ factor, but the non-exhaustive one does not have the same drawback. Besides, implementing the first strategy requires $\mathcal{O}(n^3)$ computations, against only $\mathcal{O}(n \ln(n))$ for the other two.

In this chapter, we study the penalized least-squares estimator defined as in [DLT09] but based on a fourth family of linear spaces: in our case, each space is composed of functions piecewise constant on a partition of $\{1, \ldots, n\}$ into *dyadic* intervals. Thus, we will refer to our estimator as the d-estimator. The collection of linear spaces we consider has been chosen for its potential qualities of approximation, as suggested by approximation results for real-valued functions due to DeVore and Yu [DY90] and DeVore (*cf.* [Bir07]). Adapting the proofs to our framework, we prove that our collection of spaces has indeed good approximation qualities with respect to $\mathbb{R}^r$-valued functions defined on $\{1, \ldots, n\}$ that either belong to Besov bodies – some discrete analogues of balls in a Besov space – or have bounded variation. On the other hand, the number of spaces per dimension is low enough to yield an oracle-type inequality with no extra logarithmic factor. The conjunction of both properties of our collection allows to prove adaptivity results in the minimax sense. From a theoretical point of view, the d-estimator thus satisfies properties similar to those of the *neH*-estimator, and is also proved to be adaptive for functions with bounded variation. Moreover, the d-estimator can be implemented with only $\mathcal{O}(n)$ computations. Notice that a similar collection of linear spaces has lately been used by Birgé [Bir06a; Bir07] and Baraud and Birgé [BB09] for estimation by model selection in various statistical frameworks.

As an application of our estimation procedure, we address the problem of multiple change-point detection in the distribution s. Our aim is then to estimate s by a function that is piecewise constant on some partition of $\{1, \ldots, n\}$ with a number of intervals much smaller

than n. That issue has attracted much attention due to its application to the segmentation of DNA sequences into regions of homogeneous composition (*cf.* the review [BM98b] by Braun and Müller). Owing to the length of sequences such as DNA ones, a special attention must be paid to the computational complexity of the statistical procedures. Braun, Braun and Müller [BBM00] prove consistency results for the estimation of the change-points and the number of change-points when using a penalized quasi-deviance criterion, but their estimator suffers from a heavy computational complexity, of order $\mathcal{O}(n^3)$. The two-stage procedure proposed by Gey and Lebarbier [GL08] in a Gaussian regression framework can be adapted to the framework considered here (*cf.* [Leb02], Chapter 7). The preliminary stage uses CART algorithm to select a partition. In order to reduce the size of the partition, the second stage consists in selecting a partition among the rougher partitions built on the previous one, by minimizing a penalized least-squares criterion. In the best case, the number of computations falls down to only $\mathcal{O}(n\ln(n))$ for the first stage of the procedure. Last, a few linear time procedures exist, such as the one proposed by Fu and Curnow [FC90] (*cf.* [Csű04] for the implementation) and the one studied by Szpankowski, Szpankowski and Ren [SRS05]. We propose in this chapter a hybrid procedure similar to that of [GL08], where the first stage consists this time in selecting a partition into dyadic intervals. In practice, our hybrid procedure can be implemented quite efficiently. Moreover, unlike the CART-based hybrid estimator, our hybrid estimator is proved to enjoy some adaptivity properties, which are similar to those of the d-estimator, up to a multiplicative constant. Notice that, contrary to [BBM00], our aim is not to detect all the change-points, but only the most relevant ones.

The chapter is organized as follows. In Section II.2, we describe the statistical framework and introduce notation used throughout the chapter. The next section is devoted to the theoretical study of the d-estimator . Then, we present the subsequent hybrid procedure. The performance of these procedures are illustrated in Section II.5 through a simulation study. In particular, we discuss there the practical choice of the penalties constants. Besides, we compare the d-estimator with the neH-estimator introduced in [DLT09], and apply the hybrid procedure to a DNA sequence. The chapter ends with the proof of the approximation result needed to derive one of the adaptivity properties.

II.2 Framework and notation

II.2.1 Framework

We observe n independent random variables $Y_1, \dots, Y_n$ defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and with values in $\{1, \dots, r\}$, where r is an integer and $r \geq 2$. We assume that n is a power of 2, $n \geq 2$, and write $n = 2^N$. The distribution of the n-uple $(Y_1, \dots, Y_n)$ is represented by the $r \times n$ matrix s whose i-th column is

$$s_i = \big(\mathbb{P}(Y_i = 1) \dots \mathbb{P}(Y_i = r)\big)^T, \text{ for } i = 1, \dots, n.$$

Observing $(Y_1, \dots, Y_n)$ is equivalent to observing the random $r \times n$ matrix X whose i-th column is

$$X_i = \big(\mathbb{1}_{Y_i=1} \dots \mathbb{1}_{Y_i=r}\big)^T, \text{ for } i = 1, \dots, n.$$

It should be noticed that the distribution s to estimate is in fact the expectation of X.

II.2.2 Notation

All along the chapter, we identify real-valued functions defined on $\{1, \dots, n\}$ with $\mathbb{R}^n$-vectors, so that $u = (u_1 \dots u_n) \in \mathbb{R}^n$ represents the function $u : i \in \{1, \dots, n\} \mapsto u_i$. In particular,

for any subset I of $\{1,\dots,n\}$, we call indicator function of I, and denote by $\mathbb{1}_I$, the $\mathbb{R}^n$-vector whose i-th coordinate is equal to 1 if $i \in I$, and null otherwise. In the same way, we identify $\mathbb{R}^r$-valued functions defined on $\{1,\dots,n\}$ with elements of $\mathscr{M}(r,n)$, the set of real matrices with r rows and n columns. Given an element $t \in \mathscr{M}(r,n)$, we denote by $t^{(l)}$ its l-th row and by t_i its i-th column. Thus $t \in \mathscr{M}(r,n)$ represents the function, also denoted by t, defined on $\{1,\dots,n\}$, whose value in i is the $\mathbb{R}^r$-vector t_i, while $t^{(1)},\dots,t^{(r)}$ are the coordinate functions of t.

The space $\mathscr{M}(r,n)$ is endowed with the inner product defined by

$$\langle t,u\rangle = \sum_{i=1}^{n}\sum_{l=1}^{r} t_i^{(l)} u_i^{(l)}.$$

That product is linked with the standard inner products on $\mathbb{R}^r$ and $\mathbb{R}^n$, denoted respectively by $\langle .,.\rangle_r$ and $\langle .,.\rangle_n$, by the relations

$$\langle t,u\rangle = \sum_{i=1}^{n}\langle t_i,u_i\rangle_r = \sum_{l=1}^{r}\langle t^{(l)},u^{(l)}\rangle_n.$$

The norms induced by these products on $\mathscr{M}(r,n)$, $\mathbb{R}^r$ and $\mathbb{R}^n$ are respectively denoted by $\|.\|$, $\|.\|_r$ and $\|.\|_n$. Another norm on $\mathscr{M}(r,n)$ appearing in this chapter is

$$\|t\|_\infty := \max\left\{|t_i^{(l)}|; 1\le i\le n, 1\le l\le r\right\}.$$

Let us now define some subsets of $\mathscr{M}(r,n)$ of special interest. The set composed of the $r\times n$ matrices whose columns are probability distributions on $\{1,\dots,r\}$ is denoted by $\mathscr{P}$. Given a linear subspace S of $\mathbb{R}^n$, the notation $\mathbb{R}^r \otimes S$ stands for the linear subspace of $\mathscr{M}(r,n)$ composed of the matrices whose rows all belong to S.

When the distribution of $(Y_1,\dots,Y_n)$ is given by s, we denote respectively by $\mathbb{P}_s$ and $\mathbb{E}_s$ the underlying probability distribution on $(\Omega^{\otimes n},\mathcal{A}^{\otimes n})$ and the associated expectation.

Last, in the many inequalities we shall encounter, the letters $C, C_1, c_1, \dots$ stand for positive constants. Sometimes, their dependence on one or several parameters will be indicated. For instance, the notation $C(\alpha,p)$ means that C only depends on α and p. The only constant whose value is allowed to change from one line to another is denoted by C, with no index.

II.3 The d-estimator

We study in this section the d-estimator of the distribution s, thus called because it takes values in the set of piecewise constant functions on some partition of $\{1,\dots,n\}$ into *dyadic* intervals. We begin with the definition of the estimator, explain the underlying model selection principle and justify the form of the involved penalty thanks to [DLT09]. Then, we present the main result of this chapter, about the adaptivity of the d-estimator. They greatly rely on an approximation result that will be proved later in the article. Last, we describe the algorithm used to implement that procedure and give its computational complexity.

II.3.1 Definition of the d-estimator

A partition of $\{1,\dots,n\}$ into dyadic intervals is a partition of $\{1,\dots,n\}$ into sets of the form $\left\{kn2^{-j}+1,\dots,(k+1)n2^{-j}\right\}$, where $j\in\{0,\dots,N\}$ is allowed to change from one interval

of the partition to another, and $k \in \{0, \dots, 2^j - 1\}$. We denote by $\mathcal{M}$ the family of all such partitions of $\{1, \dots, n\}$. We consider the collection of linear spaces of the form $\mathbb{R}^r \otimes S_m$, where $m \in \mathcal{M}$ and S_m is the linear subspace of $\mathbb{R}^n$ generated by the indicator functions $\{\mathbb{1}_I, I \in m\}$. In the sequel, the term 'model' refers indifferently to such a subspace of $\mathscr{M}(r,n)$ or to the associated partition in $\mathcal{M}$. For all $m \in \mathcal{M}$, the least-squares estimator of s in $\mathbb{R}^r \otimes S_m$ is defined by

$$\hat{s}_m = \operatorname*{argmin}_{t \in \mathbb{R}^r \otimes S_m} \|X - t\|^2.$$

Over each interval $I \in m$, $\hat{s}_m$ is constant and equal to the mean of the $\mathbb{R}^r$-vectors $(X_i)_{i \in I}$.

Ideally, we would like to choose a model among the collection $\mathcal{M}$ such that the risk of the associated estimator is minimal. However, determining such a model requires the knowledge of s. Therefore the challenge is to define a procedure $\hat{m}$, based solely on the data, that selects a model for which the risk of $\hat{s}_{\hat{m}}$ almost reaches the minimal one. In other words, the estimator $\hat{s}_{\hat{m}}$ should satisfy a so-called oracle inequality

$$\mathbb{E}_s\big[\|s - \hat{s}_{\hat{m}}\|^2\big] \leq C \inf_{m \in \mathcal{M}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big].$$

Besides, as often, the risk of each estimator $\hat{s}_m$ breaks down into an approximation error and an estimation error roughly proportional to the dimension of the model. Indeed, for all $m \in \mathcal{M}$, the estimator $\hat{s}_m$ satisfies

$$\|s - s_m\|^2 + \big(1 - \|s\|_\infty\big) D_m \leq \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big] \leq \|s - s_m\|^2 + \Big(1 - \frac{1}{r}\Big) D_m, \tag{II.3.1}$$

where s_m is the orthogonal projection of s on $\mathbb{R}^r \otimes S_m$ and D_m is the dimension of S_m (*cf.* [DLT09], proof of Corollary 1). Reaching the minimal risk among the estimators of the collection thus amounts to realizing the best trade-off between the approximation error and the dimension of the model, which vary in opposite ways. Therefore, we consider the procedure

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\|X - \hat{s}_m\|^2 + \mathrm{pen}(m)\big\},$$

where $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ is called penalty function. The *d*-estimator $\tilde{s}$ of s is then defined as

$$\tilde{s} = \hat{s}_{\hat{m}}.$$

Our choice of penalty, that relies on results proved in [DLT09], is justified by an oracle inequality, up to a quantity that depends on $\|s\|_\infty$ (*cf.* Inequality (II.3.4) below).

Proposition 3 *Let* $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ *be a penalty of the form*

$$\mathrm{pen}(m) = c_0 D_m, \tag{II.3.2}$$

where, for $m \in \mathcal{M}$, D_m *is the dimension of* S_m. *If* c_0 *is positive and large enough, then*

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0) \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}. \tag{II.3.3}$$

Moreover, if $\|s\|_\infty < 1$, *then*

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0)(1 - \|s\|_\infty)^{-1} \inf_{m \in \mathcal{M}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big]. \tag{II.3.4}$$

Proof. For all $1 \leq D \leq n$, we introduce the subcollection of models of dimension D:

$$\mathcal{M}_D = \{m \in \mathcal{M} \text{ s.t. } D_m = D\}.$$

In order to evaluate the cardinal of $\mathcal{M}_D$, let us describe $\mathcal{M}$ in a more constructive way. Let $\mathcal{T}$ be the complete binary tree with root $(0,0)$ such that:

- for all $j \in \{1, \dots, N\}$, the nodes at level j are indexed by the elements of the set $\Lambda(j) = \{(j,k), k = 0, \dots, 2^j - 1\}$;
- for all $j \in \{0, \dots, N-1\}$ and all $k \in \{0, \dots, 2^j - 1\}$, the left branch that stems from node (j,k) leads to node $(j+1, 2k)$, and the right one, to node $(j+1, 2k+1)$.

The node set of $\mathcal{T}$ is $\mathcal{N} = \cup_{j=0}^{N} \Lambda(j)$, where $\Lambda(0) = \{(0,0)\}$. The dyadic intervals of $\{1, \dots, n\}$ are the sets

$$I_{(j,k)} = \left\{k2^{N-j} + 1, \dots, (k+1)2^{N-j}\right\}$$

indexed by the elements of $\mathcal{N}$. Hence we deduce a one-to-one correspondence between the partitions of $\{1, \dots, n\}$ that belong to $\mathcal{M}$ and the subsets of $\mathcal{N}$ composed of the leaves of any complete binary tree resulting from an elagation of $\mathcal{T}$. The cardinal of $\mathcal{M}_D$ is thus equal to the number of complete binary trees with D leaves resulting from an elagation of $\mathcal{T}$. So it is given by the Catalan number $D^{-1}\binom{2(D-1)}{D-1}$ and upper-bounded by 4^D. Therefore, choosing for instance $L = \ln(8)$, we get

$$\sum_{m \in \mathcal{M}} \exp(-LD_m) = \sum_{D=1}^{n} |\mathcal{M}_D| \exp(-LD) \leq 1.$$

Inequality (II.3.3) thus follows from Theorem 1 in [DLT09]. Inequality (II.3.4) results from the upper-bound (II.3.3) and the lower-bound given in (II.3.1). ■

From now on, we will always assume that the d-estimator derives from a penalty of the form $\text{pen}(m) = c_0 D_m$, where the constant c_0 is positive and large enough to yield an oracle-type inequality. Choosing in practice an adequate value of c_0 is an issue that will be treated in Section II.5. By way of comparison, let us mention that the *neH*-procedure studied in [DLT09] satisfies the same kind of oracle-type inequality (*cf.* [DLT09], Proposition 3). But the similar procedure based on the exhaustive collection of partitions of $\{1, \dots, n\}$ only satisfies an oracle-type inequality such as (II.3.4) within a $\ln(n)$ factor, owing to the greater number of models per dimension (*cf.* [DLT09], Proposition 1).

Last, notice that $\tilde{s}$ does not necessarily belong to $\mathscr{P}$. Nevertheless, since the vector $(1 \dots 1)$ belongs to any S_m, for $m \in \mathcal{M}$, the elements in a same row of $\tilde{s}$ sum up to 1. In order to get an estimator of s with values in $\mathscr{P}$, we may consider the orthogonal projection of $\tilde{s}$ on the closed convex $\mathscr{P}$, whose risk is even smaller than that of $\tilde{s}$.

II.3.2 Adaptivity of the d-estimator

Though the oracle-type inequality (II.3.4) ensures that, under a minor constraint on s, the estimator $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$, it does not allow to compare $\tilde{s}$ with other estimators of s. Therefore, we now pursue the study of $\tilde{s}$ adopting a minimax point of view. We consider a large family of subsets of $\mathscr{P}$, to be defined in the next paragraph. Let us denote by $\mathcal{S}$ some subset in that family. Our aim is to compare the maximal risk of $\tilde{s}$ when s belongs to $\mathcal{S}$ to the minimax risk over $\mathcal{S}$. We may rewrite the upper-bound (II.3.3) for the risk of $\tilde{s}$ as

$$\mathbb{E}_s\left[\|s - \tilde{s}\|^2\right] \leq C(c_0) \inf_{1 \leq D \leq n} \left\{ \inf_{m \in \mathcal{M}_D} \|s - s_m\|^2 + D \right\}, \tag{II.3.5}$$

where we recall that $\mathcal{M}_D = \{m \in \mathcal{M} \text{ s.t. } D_m = D\}$ and s_m is the orthogonal projection of s on $\mathbb{R}^r \otimes S_m$. Thus, the approximation qualities of our family of models with respect to each subset $\mathcal{S}$ remain to be evaluated. More precisely, for each subset $\mathcal{S}$, and each dimension D, we shall provide upper-bounds for the approximation error $\inf_{m \in \mathcal{M}_D} \|s - s_m\|^2$ when $s \in \mathcal{S}$.

On the one hand, we consider subsets of $\mathscr{P}$ introduced in [DLT09], whose definition is inspired from the characterization in terms of wavelet coefficients of balls in Besov spaces. In order to define them, we equip $\mathbb{R}^n$ with an orthonormal wavelet basis, the Haar basis.

Definition 1 *Let $\varphi : \mathbb{R} \to \{-1, 1\}$ be the function with support $(0, 1]$ that takes value 1 on $(0, 1/2]$ and -1 on $(1/2, 1]$. Let $\Lambda = \cup_{j=-1}^{N-1} \Lambda(j)$, where $\Lambda(-1) = \{(-1, 0)\}$ and*

$$\Lambda(j) = \{(j, k), k = 0, \dots, 2^j - 1\}, \text{ for } j = 0, \dots, N-1.$$

If $\lambda = (-1, 0)$, ϕ_λ is the $\mathbb{R}^n$-vector whose coordinates are all equal to $1/\sqrt{n}$.
If $\lambda = (j, k)$, where $j \neq -1$ and $k \in \Lambda(j)$, ϕ_λ is the $\mathbb{R}^n$-vector whose i-th coordinate is

$$\phi_{\lambda\, i} = \frac{2^{j/2}}{\sqrt{n}} \varphi\left(2^j \frac{i}{n} - k\right), \text{ for } i = 1, \dots, n.$$

The functions $\{\phi_\lambda\}_{\lambda \in \Lambda}$ are called the Haar functions. They form an orthonormal basis of $\mathbb{R}^n$ called the Haar basis.

Any element $t \in \mathscr{M}(r, n)$ can be decomposed into

$$t = \sum_{j=-1}^{N-1} \sum_{\lambda \in \Lambda(j)} \beta_\lambda \phi_\lambda$$

where, for all $\lambda \in \Lambda$, β_λ is the column-vector in $\mathbb{R}^r$ whose l-th coefficient is $\beta_\lambda^{(l)} = \langle t^{(l)}, \phi_\lambda \rangle_n$, for $l = 1, \dots, r$. So, we improperly refer to the β_λ's as the wavelet coefficients of t. Besov bodies are then defined as follows.

Definition 2 *Let $\alpha > 0$, $p > 0$ and $R \geq 0$. The set composed of all the elements $t \in \mathscr{M}(r, n)$ such that*

$$\frac{1}{\sqrt{n}} \left(\sum_{j=0}^{N-1} 2^{jp(\alpha + 1/2 - 1/p)} \sum_{\lambda \in \Lambda(j)} \|\beta_\lambda\|_r^p \right)^{1/p} \leq R,$$

where, for $l = 1, \dots, r$, $\beta_\lambda^{(l)} = \langle t^{(l)}, \phi_\lambda \rangle_n$, is denoted by $\mathscr{B}(\alpha, p, R)$ and called a Besov body. The set of all the elements of $\mathscr{P}$ that belong to $\mathscr{B}(\alpha, p, R)$ is denoted by $\mathscr{BP}(\alpha, p, R)$.

We also consider subsets of $\mathscr{P}$ whose definition is inspired from functions of bounded α-variation.

Definition 3 *Let $\alpha > 0$ and $R \geq 0$. For $t \in \mathscr{M}(r, n)$, let*

$$V_\alpha(t) = \sup_{1 \leq i \leq n-1} \sup_{\substack{x_0 < \dots < x_i \\ s.t.\ 1 \leq x_0 < x_i \leq n}} \left\{ \sum_{j=1}^{i} \|t_{x_j} - t_{x_{j-1}}\|_r^{1/\alpha} \right\}^\alpha.$$

The set composed of all the elements $t \in \mathscr{M}(r, n)$ such that $V_\alpha(t) \leq R$ is denoted by $\mathscr{V}(\alpha, R)$. The set of all the elements of $\mathscr{P}$ that belong to $\mathscr{V}(\alpha, R)$ is denoted by $\mathscr{VP}(\alpha, R)$.

Notice that, for all $t \in \mathscr{M}(r, n)$, when $\alpha \geq 1$,

$$V_\alpha(t) = \left\{ \sum_{i=2}^{n} \|t_i - t_{i-1}\|_r^{1/\alpha} \right\}^\alpha$$

so that $V_\alpha(t)$ may be interpreted as the $\ell_{1/\alpha}$-norm of the 'jumps' of t.

For a wide range of values of the parameters (α, p, R) or (α, R), we are able to bound the approximation errors appearing in (II.3.5) uniformly over $\mathscr{BP}(\alpha, p, R)$ and $\mathscr{VP}(\alpha, R)$.

Theorem 1 *Let $p \in (0,2]$, $\alpha > 1/p - 1/2$ and $R \geq 0$. For all $s \in \mathscr{BP}(\alpha,p,R)$ and all $D \in \{1,\dots,n\}$, there exists a partition $m \in \mathcal{M}$ such that $D_m = D$ and*

$$\|s - s_m\|^2 \leq C(\alpha,p) n R^2 D^{-2\alpha}.$$

That result will be proved in Section II.6.

Theorem 2 *Let $\alpha > 0$ and $R \geq 0$. Let $k_1(\alpha) = \big(1 - 2^{-(1+2\alpha)/(2\alpha)}\big)/\big(1 - 2^{-1/(2\alpha)}\big)$. For all $s \in \mathscr{VP}(\alpha,R)$ and all $0 \leq j \leq N$, there exists a partition $m \in \mathcal{M}$ such that $1 \leq D_m \leq k_1(\alpha)2^j$ and*

$$\|s - s_m\|^2 \leq C(\alpha) n R^2 2^{-2\alpha j}.$$

Proof. The proof of Proposition 3 in [Bir07] can be readily adapted to our framework, whatever $\alpha > 0$. In the proof of that proposition, the assumption $\alpha \in (0,1]$ is only used to bound $k_1(\alpha)$ and $C(\alpha)$. ■

Let us now come back to our initial problem, that is comparing the performance of $\tilde{s}$ with that of any other estimator of s. For $\alpha > 0$, $p > 0$ and $R \geq 0$, the minimax risk over $\mathscr{BP}(\alpha,p,R)$ is given by

$$\mathcal{R}_{\mathscr{B}}(\alpha,p,R) = \inf_{\hat{s}} \sup_{s \in \mathscr{BP}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \hat{s}\|^2\big]$$

where the infimum is taken over all the estimators $\hat{s}$ of s. We denote by $\mathcal{R}_{\mathscr{V}}(\alpha,R)$ the minimax risk over $\mathscr{VP}(\alpha,R)$. Thanks to the above approximation results, we obtain, as stated below, that, for a whole range of values of (α,p,R) or (α,R), the estimator $\tilde{s}$ reaches the minimax risk over $\mathscr{BP}(\alpha,p,R)$ and $\mathscr{VP}(\alpha,R)$ within a multiplicative constant. Therefore, $\tilde{s}$ is adaptive in the minimax sense not only over the same range of Besov bodies as the *neH*-estimator (*cf.* [DLT09], Corollary 4) but also on a wide range of sets in the scale $\{\mathscr{VP}(\alpha,R)\}_{\alpha>0,R\geq 0}$.

Theorem 3 *For all $p \in (0,2]$ and $\alpha > 1/p - 1/2$, if $n^{-1/2} \leq R < n^{\alpha}$, then*

$$\sup_{s \in \mathscr{BP}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha,p)\mathcal{R}_{\mathscr{B}}(\alpha,p,R).$$

For all $\alpha > 0$, there exists a real $k_2(\alpha) \in (0,1)$ such that, if $R \geq n^{-1/2}$ and $R \leq k_2(\alpha)n^{\alpha}$, then

$$\sup_{s \in \mathscr{VP}(\alpha,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha)\mathcal{R}_{\mathscr{V}}(\alpha,R).$$

Proof. Let us fix $p \in (0,2], \alpha > 1/p - 1/2$ and $n^{-1/2} \leq R < n^{\alpha}$. Combining Inequality (II.3.5) and Theorem 1 leads to

$$\sup_{s \in \mathscr{BP}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha,p) \inf_{1 \leq D \leq n} \big\{nR^2D^{-2\alpha} + D\big\}. \tag{II.3.6}$$

In order to realize approximately the best trade-off between the terms $nR^2D^{-2\alpha}$ and D, which vary in opposite ways when D increases, we choose D as large as possible under the constraint $D \leq nR^2D^{-2\alpha}$. Let us denote by $D^{\star}$ the largest integer D such that $D \leq (nR^2)^{1/(1+2\alpha)}$. One can easily check that, given the hypotheses linking n and R, $D^{\star}$ does belong to $\{1,\dots,n\}$. Since $2D^{\star} > (nR^2)^{1/(1+2\alpha)}$, we deduce from Inequality (II.3.6) the upper-bound

$$\sup_{s \in \mathscr{BP}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha,p)(nR^2)^{1/(2\alpha+1)}.$$

The matching lower-bound for the minimax risk over $\mathscr{BP}(\alpha,p,R)$ is proved in [DLT09] (Theorem 3).

Table II.3.1: Algorithm for computing $\tilde{s}$

***Step 1**: Initialization*
Set $d(1) = 0$ and $p(1) = +\infty$.
For $i = 2, \dots, n+1$,
  set $d(i) = +\infty$ and $p(i) = +\infty$.

***Step 2**: Determining the lengths of the shortest paths with origin 1*
For $i = 1, \dots, n$,
  for $j \in \Gamma_i$,
    if $d(j) > d(i) + \mathcal{L}(i,j)$,
      then do $d(j) \leftarrow d(i) + \mathcal{L}(i,j)$ and $p(j) \leftarrow i$.

***Step 3**: Determining a shortest path P from 1 to $n+1$*
Set $pred = p(n+1)$ and $P = (n+1)$.
While $pred \neq +\infty$,
  replace P with the concatenation of $pred$ followed by P,
  do $pred \leftarrow p(pred)$.

***Step 4**: Computing the d-estimator*
Set $\tilde{D} = length(P) - 1$.
For $k = 1, \dots, \tilde{D}$,
  for $i = P(k), \dots, P(k+1) - 1$,
    set $\tilde{s}_i = \bar{X}(P(k) : P(k+1))$.

In the same way, for $\alpha > 0$ and $n^{-1/2} \leq R \leq n^{\alpha}$, we get

$$\sup_{s \in \mathscr{VP}(\alpha,R)} \mathbb{E}_s\left[\|s - \tilde{s}\|^2\right] \leq C(c_0, \alpha)(nR^2)^{1/(2\alpha+1)}.$$

The definition of $k_2(\alpha)$ and the matching lower-bound for the minimax risk over $\mathscr{VP}(\alpha, R)$, for $n^{-1/2} \leq R \leq k_2(\alpha) n^{\alpha}$, are given in Proposition 8. ■

II.3.3 Computing the d-estimator

Since the penalty only depends on the dimension of the models, we denote by $\mathrm{pen}(D)$ the penalty assigned to all models in $\mathcal{M}_D$, for $1 \leq D \leq n$. A way to compute $\tilde{s}$ could rely on the equality

$$\min_{m \in \mathcal{M}} \left\{\|X - \hat{s}_m\|^2 + \mathrm{pen}(m)\right\} = \min_{1 \leq D \leq n} \left\{ \min_{m \in \mathcal{M}_D} \|X - \hat{s}_m\|^2 + \mathrm{pen}(D) \right\}.$$

We should thus compute the best estimator for each dimension $D \in \{1, \dots, n\}$, and choose one among them by taking into account the penalty term, as in [Leb02] (Chapter 3) or [BBM00]. But, even with Bellman's algorithm, that requires polynomial time. Here, we shall see that we can avoid such a computationaly intensive way by taking advantage of the form of the penalty.

Let us express more explicitly the criterion to minimize. The dyadic intervals of a given partition $m \in \mathcal{M}$ are denoted by $\{i_k, \dots, i_{k+1} - 1\}$, $k = 1, \dots, D_m$, with $1 = i_1 < i_2 \dots < i_{D_m+1} = n+1$. For all $1 \leq k \leq D_m$, any column of $\hat{s}_m$ whose index belongs to $\{i_k, \dots, i_{k+1} - 1\}$ is equal to the mean $\bar{X}(i_k : i_{k+1})$ of the columns of X whose indices belong to the interval

$\{i_k, \ldots, i_{k+1} - 1\}$. Owing to the form of the penalty, and to the additivity of the least-squares criterion, the whole criterion to minimize breaks down into a sum:

$$\|X - \hat{s}_m\|^2 + \mathrm{pen}(m) = \sum_{k=1}^{D_m} \mathcal{L}(i_k, i_{k+1}), \tag{II.3.7}$$

where, for all $1 \leq k \leq D_m$,

$$\mathcal{L}(i_k, i_{k+1}) = c_0 + \sum_{i=i_k}^{i_{k+1}-1} \|X_i - \bar{X}(i_k : i_{k+1})\|_r^2.$$

By comparison with the method suggested in the previous paragraph, we are left with only one minimization problem, with no dimension constraint, instead of n. We now turn to graph theory where our minimization problem finds a natural interpretation. We consider the weighted directed graph G having $\{1, \ldots, n+1\}$ as vertex set and whose edges are the pairs (i, j) such that $\{i, \ldots, j-1\}$ is a dyadic interval of $\{1, \ldots, n\}$ assigned with the weight $\mathcal{L}(i, j)$. We say that a vertex j is a successor to a vertex i if (i, j) is an edge of the graph G and we associate to each vertex i its successor list Γ_i. For all $1 \leq D \leq n$, a $D+1$-uple $(i_1, i_2, \ldots, i_{D+1})$ of vertices of G such that $i_1 = 1$, $i_{D+1} = n+1$ and each vertex is a successor to the previous one, will be called a path leading from 1 to $n+1$ in D steps. The length of such a path is defined as $\sum_{k=1}^{D} \mathcal{L}(i_k, i_{k+1})$. Determining $\hat{m}$ thus amounts to finding a shortest path leading from 1 to $n+1$ in the graph G. That problem can be solved by using a simple shortest-path algorithm dedicated to acyclic directed graphs, presented for instance in [CLRS01] (Section 24.2). For the sake of completeness, we also describe it in Table II.3.1. We have to underline that there are only $2n-1$ dyadic intervals of $\{1, \ldots, n\}$. Therefore, the graph G, with $n+1$ vertices and $2n-1$ edges, can be represented by only $\mathcal{O}(n)$ data: the weights $\mathcal{L}(i, j)$, for $1 \leq i \leq n$ and $j \in \Gamma_i$, and the successor lists Γ_i, for $1 \leq i \leq n$. In the key step of the algorithm, *i.e.* step 2, each edge is considered only once. When the time comes to consider the edges with origin i, the variables $d(i)$ and $p(i)$ respectively contain the length of a shortest path from 1 to i and a predecessor of i in such a path. Just before the edge (i, j), where $j \in \Gamma_i$, be processed, the variables $d(j)$ and $p(j)$ contain respectively the length of a shortest path leading from 1 to j and a predecessor of j in such a path, based solely on the edges that have already been encountered. Then dealing with the edge (i, j) consists in testing whether the length of the path leading from 1 to j can be shortened by going via i and updating, if necessary, $d(j)$ and $p(j)$. What clearly appears from the above description of the algorithm is that its complexity is only *linear* in the length n of the sequence.

II.4 Hybrid procedure

We shall now apply the previous procedure to the detection of multiple change-points in the distribution s. Let us give a first glimpse of what can be expected from the d-estimator for that problem. In Figure II.4, we plot the first coordinate function of a distribution $s_a \in \mathscr{M}(2, 1024)$ that is piecewise constant over a partition with only 3 segments together with the first coordinate of a realization of $\widetilde{s_a}$. The value of c_0 has been chosen so as to minimize the distance between s_a and its estimator. If both change-points in s_a are indeed detected, the selected partition, due to its special nature, also points at irrelevant ones. In order to get rid of them, we propose a two-stage procedure, that we name hybrid procedure. After describing it, we provide an adaptivity result for that procedure and end this section with computational issues.

In the sequel, we suppose that $n \geq 4$. We shall work with the set $\mathscr{M}(r, n/2)$ of $r \times (n/2)$ real matrices and introduce other notation. For all $t \in \mathscr{M}(r, n)$, we denote by $t^\bullet$ (resp. t°)

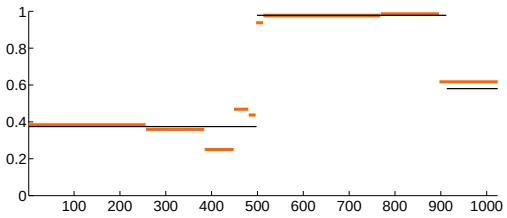

Figure II.4.1: First coordinate functions of the distribution s_a (thin black line) and of its d-estimator $\widetilde{s_a}$ (thick yellow line).

the element of $\mathscr{M}(r, n/2)$ composed of the columns of t whose indices are even (resp. odd). We equip $\mathscr{M}(r, n/2)$ with the norm analogous to the norm $\|.\|$ on $\mathscr{M}(r, n)$. For the sake of simplicity, we also denote by $\|.\|$ that norm on $\mathscr{M}(r, n/2)$. For a partition m of $\{1, \dots, n/2\}$, we denote by S'_m the linear subspace of $\mathbb{R}^{n/2}$ generated by the indicator functions of the intervals $I \in m$ and by D'_m its dimension. We are now able to describe the hybrid procedure. First, the previous procedure based on $X^\bullet$ provide us with a random partition of $\{1, \dots, n/2\}$ into dyadic intervals denoted by $\hat{m}^\bullet$. Then, we consider the random collection $\widehat{\mathcal{M}}^\bullet$ of all the partitions of $\{1, \dots, n/2\}$ that are built on $\hat{m}^\bullet$. For each partition m of $\{1, \dots, n/2\}$, we define the least-squares estimator of s° in $\mathbb{R}^r \otimes S'_m$ by

$$\widehat{s^\circ}_m = \operatorname*{argmin}_{t \in \mathbb{R}^r \otimes S'_m} \|X^\circ - t\|^2.$$

Then we select

$$\hat{m}^\circ = \operatorname*{argmin}_{m \in \widehat{\mathcal{M}}^\bullet} \left\{ \|X^\circ - \widehat{s^\circ}_m\|^2 + \widehat{\mathrm{pen}}^\circ(m) \right\},$$

where the penalty $\widehat{\mathrm{pen}}^\circ$ will be chosen in the next paragraph. That partition provide us with the estimated change-points in the distribution s. As a matter of fact, we define the hybrid estimator $\tilde{s}_{hyb}$ of s as the random $r \times n$ matrix whose submatrices composed respectively of columns with even indices and of columns with odd indices are both equal to $\widehat{s^\circ}_{\hat{m}^\circ}$. The application of this procedure to s_a is illustrated by Figure II.5.4. Notice that other ways of splitting the sample could be considered. This one has been chosen for ease of notation.

We obtain the following upper-bound for the risk of $\tilde{s}_{hyb}$.

Theorem 4 *Let $\widehat{D}$ be the cardinal of $\hat{m}^\bullet$ and $\widehat{\mathrm{pen}}^\circ : \widehat{\mathcal{M}}^\bullet \to \mathbb{R}^+$ be a penalty of the form*

$$\widehat{\mathrm{pen}}^\circ(m) = \left(c_1 + c_2 \ln\left(\widehat{D}/D'_m\right) \right) D'_m, \tag{II.4.8}$$

where c_1 and c_2 are positive. If c_0, c_1 and c_2 are large enough, then

$$\mathbb{E}_s\left[\|s - \tilde{s}_{hyb}\|^2\right] \le C(c_0, c_1, c_2) \left[\inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\} + \|s^\circ - s^\bullet\|^2 \right].$$

Thus, if s also satisfies $\|s^\circ - s^\bullet\|^2 \le \lambda \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\}$, then

$$\mathbb{E}_s\left[\|s - \tilde{s}_{hyb}\|^2\right] \le C(c_0, c_1, c_2, \lambda) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\}. \tag{II.4.9}$$

Inequality (II.4.9) must be compared with Inequality (II.3.3). In particular, provided s° and $s^\bullet$ are close enough, the adaptivity properties of the hybrid estimator are similar to those of the d-estimator. The constant $C(c_0, c_1, c_2, \lambda)$ in (II.4.9) is expected to be larger than the constant $C(c_0)$ in (II.3.3), but we will see in Section II.5.3 that, in practice, provided the penalty constants are well chosen, the risk of $\tilde{s}_{hyb}$ is not so far from that of $\tilde{s}$.

Proof. For all $1 \leq D \leq \widehat{D}$, the number $\widehat{N}_D$ of partitions in $\widehat{\mathcal{M}}^\bullet$ with D pieces satisfies

$$\widehat{N}_D = \binom{\widehat{D}-1}{D-1} \leq \left(\frac{e\widehat{D}}{D}\right)^D.$$

The above inequality results from a property of binomial coefficients that may be found in [Mas07] (Proposition 2.5) for instance. So the weights defined by

$$\widehat{L}(D) = \ln(2e) + \ln(\widehat{D}/D), \text{ for } 1 \leq D \leq \widehat{D},$$

are such that

$$\sum_{D=1}^{\widehat{D}} \widehat{N}_D \exp(-D\widehat{L}(D)) \leq 1.$$

Moreover, given $X^\bullet$, the penalty $\widehat{\mathrm{pen}}^\circ$ given by (II.4.8) fulfills the hypotheses of Theorem 1 in [DLT09] provided c_1 and c_2 are large enough. With a slight abuse of notation, for any partition m of $\{1,\dots,n/2\}$, we still denote by t_m the orthogonal projection of an element $t \in \mathscr{M}(r, n/2)$ on $\mathbb{R}^r \otimes S'_m$. Working conditionally to $X^\bullet$, the collection $\widehat{\mathcal{M}}^\bullet$ is deterministic, so we deduce from Theorem 1 of [DLT09] applied to the estimator $\widehat{s}^\circ_{\hat{m}^\circ}$ of s° that

$$\mathbb{E}_{s^\circ}\big[\|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2 | X^\bullet\big] \leq C(c_1,c_2)\big[\|s^\circ - s^\circ_{\hat{m}^\bullet}\|^2 + \widehat{\mathrm{pen}}^\circ(\hat{m}^\bullet)\big]. \tag{II.4.10}$$

We recall that the d-estimator of $s^\bullet$ is $\widetilde{s}^\bullet = \widehat{s}^\bullet_{\hat{m}^\bullet}$. So, thanks to the triangle inequality, and since an orthogonal projection is a shrinking map, we get

$$\|s^\circ - s^\circ_{\hat{m}^\bullet}\|^2 \leq C\big(\|s^\circ - s^\bullet\|^2 + \|s^\bullet - \widetilde{s}^\bullet\|^2\big).$$

By definition, $\widehat{D} = D'_{\hat{m}^\bullet}$, so

$$\widehat{\mathrm{pen}}^\circ(\hat{m}^\bullet) = c_1\widehat{D}.$$

Taking into account the last two inequalities and integrating with respect to $X^\bullet$ leads from (II.4.10) to

$$\mathbb{E}_s\big[\|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2\big] \leq C(c_1,c_2)\Big[\|s^\circ - s^\bullet\|^2 + \mathbb{E}_{s^\bullet}\big[\|s^\bullet - \widetilde{s}^\bullet\|^2\big] + \mathbb{E}_{s^\bullet}(\widehat{D})\Big].$$

Besides, it follows from the definition of $\tilde{s}_{hyb}$ that

$$\|s - \tilde{s}_{hyb}\|^2 = \|s^\bullet - \widehat{s}^\circ_{\hat{m}^\circ}\|^2 + \|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2.$$

Applying the triangle inequality, we then get

$$\|s - \tilde{s}_{hyb}\|^2 \leq C\big(\|s^\bullet - s^\circ\|^2 + \|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2\big).$$

Consequently,

$$\mathbb{E}_s\big[\|s - \tilde{s}_{hyb}\|^2\big] \leq C(c_1,c_2)\Big[\|s^\circ - s^\bullet\|^2 + \mathbb{E}_{s^\bullet}\big[\|s^\bullet - \widetilde{s}^\bullet\|^2\big] + \mathbb{E}_{s^\bullet}(\widehat{D})\Big]. \tag{II.4.11}$$

Let us denote by $\mathcal{M}'$ the set of all partitions of $\{1,\dots,n/2\}$ into dyadic intervals. For the risk of $\widetilde{s}^\bullet$, Inequality (II.3.3) provides

$$\mathbb{E}_{s^\bullet}\big[\|s^\bullet - \widetilde{s}^\bullet\|^2\big] \leq C(c_0) \inf_{m\in\mathcal{M}'} \big\{\|s^\bullet - s^\bullet_m\|^2 + D'_m\big\}. \tag{II.4.12}$$

In order to bound the term $\mathbb{E}_{s^\bullet}(\widehat{D})$, we need to go back to the proof of Theorem 1 in [DLT09] (Section 8.1). As already seen during the proof of Proposition 3, we can choose a positive constant L such that $\sum_{m\in\mathcal{M}'} \exp(-LD'_m) \leq 1$. Let us fix a partition $m \in \mathcal{M}'$ and $\xi > 0$.

Using the same notation as in [DLT09], we deduce from the proof of Theorem 1 in [DLT09] that there exists an event $\Omega_\xi(m)$ such that $\mathbb{P}_{s^\bullet}(\Omega_\xi(m)) \geq 1 - \exp(-\xi)$ and on which

$$c_0 \widehat{D} \leq C_1 \|s^\bullet - s_m^\bullet\|^2 + C_2(c_0) D'_m + C_3 \widehat{D} + C_4 \xi.$$

Therefore, if $c_0 > C_3$, then

$$\widehat{D} \leq C(c_0)\big(\|s^\bullet - s_m^\bullet\|^2 + D'_m + \xi\big).$$

Integrating this inequality and taking the infimum over $m \in \mathcal{M}'$ then yields

$$\mathbb{E}_{s^\bullet}(\widehat{D}) \leq C(c_0) \inf_{m \in \mathcal{M}'} \big\{\|s^\bullet - s_m^\bullet\|^2 + D'_m\big\}. \tag{II.4.13}$$

Moreover, one can check that

$$\inf_{m \in \mathcal{M}'} \big\{\|s^\bullet - s_m^\bullet\|^2 + D'_m\big\} \leq \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}. \tag{II.4.14}$$

Combining Inequalities (II.4.11) to (II.4.14), we finally get

$$\mathbb{E}_s\big[\|s - \tilde{s}_{hyb}\|^2\big] \leq C(c_0, c_1, c_2)\Big[\|s^\circ - s^\bullet\|^2 + \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}\Big].$$

■

Regarding the computation of $\tilde{s}_{hyb}$, we know from Section II.3.3 that determining $\widetilde{s^\bullet}$ only requires $\mathcal{O}(n)$ computations. On the other hand, since $\widehat{\text{pen}}^\circ$ is not linear in the dimension of the models, $\hat{m}^\circ$ has to be determined following the method suggested at the beginning of Section II.3.3 and using Bellman's algorithm. Thus, the second stage requires $\mathcal{O}(\hat{D}^3)$ computations. However, if s belongs to $\mathscr{BP}(\alpha, p, R)$ or $\mathscr{VP}(\alpha, p, R)$, it follows from Inequalities (II.4.13) and (II.4.14) and the proof of Theorem 3 that the expectation of $\hat{D}$ is of order $n^{1/(1+2\alpha)}$. In such a case, the second stage of the hybrid procedure is thus expected to require much less than $\mathcal{O}(n^3)$ computations.

II.5 Simulation study

In the previous sections, our main concern has been to propose a form of penalty yielding, in theory, a performant estimator. In this section, we study some practical choice of the penalty for each procedure. Besides, we compare the d-estimator with the neH-estimator proposed in [DLT09] on several simulated examples. We also compare on a DNA sequence our hybrid procedure with that based on CART (*cf.* [Leb02]) and that based on the neH-procedure (*cf.* [DLT09], Section 7).

II.5.1 Choosing the penalty constant for the d-estimator

We have examined some examples for $r = 2$ and $r = 4$, with different values of $n = 2^N$. For $r = 2$, the distribution s is entirely determined by its first coordinate function, that is the only one to be plotted (*cf.* Figure II.5.2). For $r = 4$, examples s_d to s_f are plotted in Figure II.5.3 (left column).

As already said in Section II.3.1, the d-estimator has been designed for satisfying an oracle inequality, what it almost does according to Proposition 3. Therefore, the risk of the oracle, *i.e.* $\inf_{m \in \mathcal{M}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big]$, serves as a benchmark in order to judge of the quality of $\tilde{s}$, and also of the quality of a method for choosing a penalty constant. The different quantities introduced in

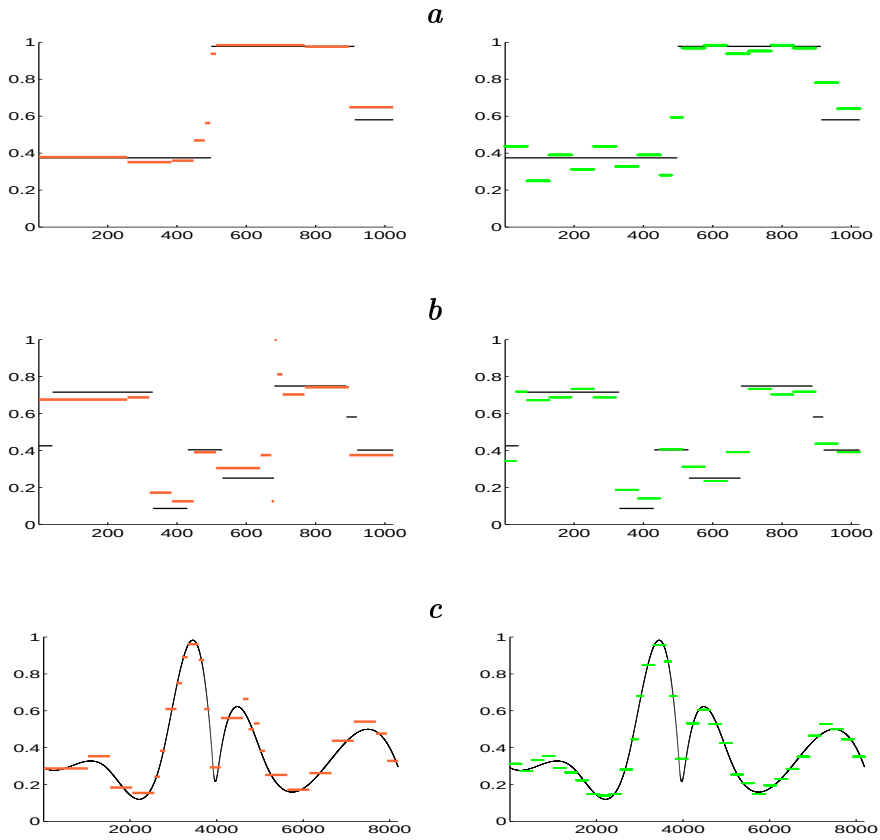

Figure II.5.2: Left column: first coordinate functions of s (thin black line) and $\tilde{s}$ (thick orange line) for $s \in \{s_a, s_b, s_c\}$; Right column: first coordinate functions of s (thin black line) and $\tilde{s}_{neH}$ (thick green line) for $s \in \{s_a, s_b, s_c\}$.

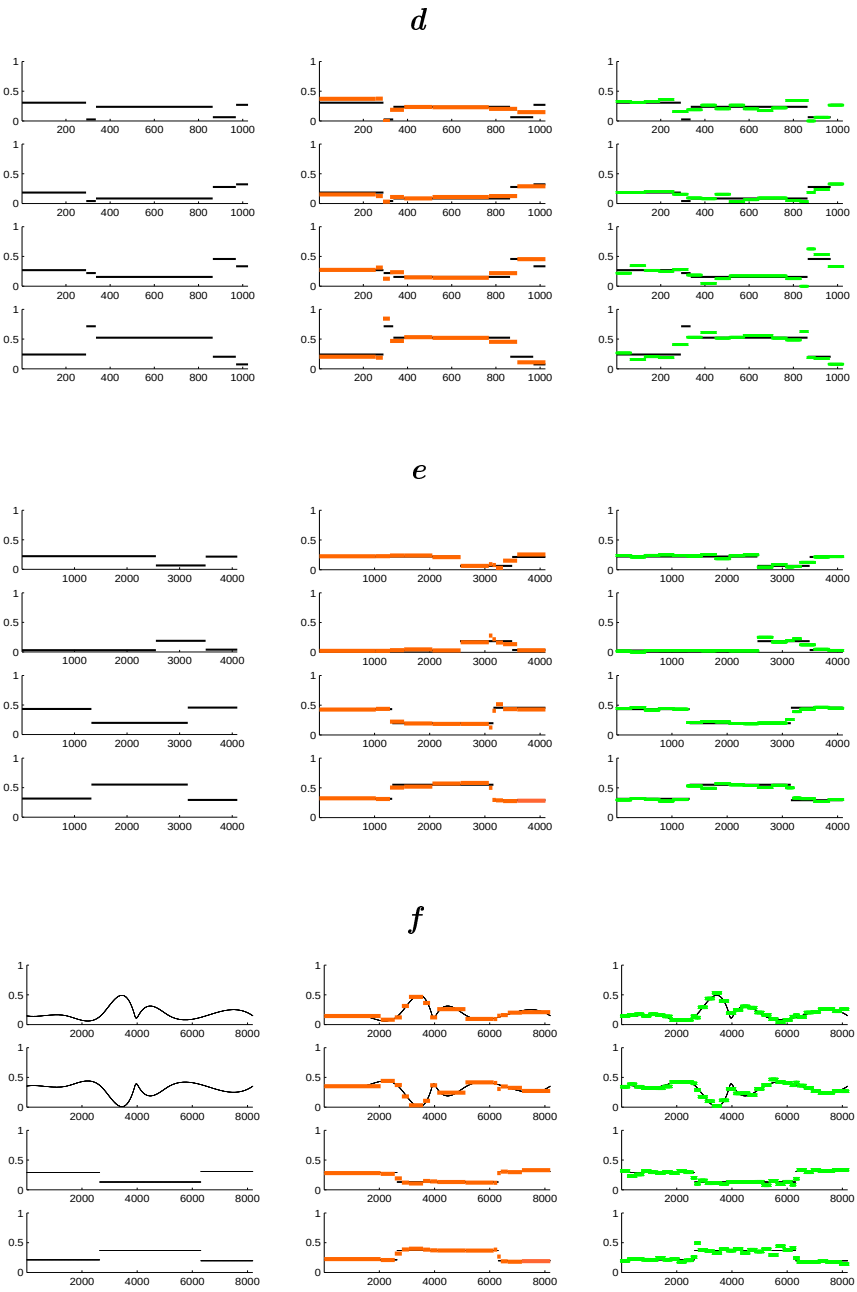

Figure II.5.3: For $s \in \{s_d, s_e, s_f\}$. Left column: Four coordinate functions of s; Center column: four coordinate functions of s (thin black line) and $\tilde{s}$ (thick orange line); Right column: four coordinate functions of s (thin black line) and $\tilde{s}_{neH}$ (thick green line).

Table II.5.2: Performance of the d-estimator for different choices of the penalty constant.

s	r	N	$c^\star$	$Q^\star$	$\bar{c}_j$	σ_j	Q_j
s_a	2	10	1.7	2.4	1.9	0.2	2.7
s_b	2	10	1.7	1.9	2.0	0.2	2.1
s_c	2	13	2.2	1.7	2.0	0.1	1.8
s_d	4	10	2.1	1.4	2.4	0.2	1.4
s_e	4	12	2.5	1.3	2.3	0.1	1.3
s_f	4	13	2.7	1.3	2.5	0.1	1.3

the sequel have been estimated over 500 simulations. Denoting by $\tilde{s}(c)$ the d-estimator when c_0 takes the value c, we have first estimated

$$c^\star(s) := \operatorname*{argmin}_c \mathbb{E}_s\big[\|s - \tilde{s}(c)\|^2\big],$$

where, in practice, we have varied c from 0 to 4, by step 0.1, and from 4 to 6 by step 0.5. We plot in Table II.5.2 an estimation of $c^\star$ and the ratio $Q^\star$ between an estimation of $\mathbb{E}_s\big[\|s - \tilde{s}(c^\star)\|^2\big]$ and the estimated risk of the oracle. In view of the results obtained here, it seems difficult to propose a value of c_0 that would be convenient for any s. Therefore, as in [DLT09], Section 8, we have tried a data-driven method, inspired from results proved by Birgé and Massart in a Gaussian framework (*cf.* [BM07]). Given a simulation of $(Y_1, \ldots, Y_n)$, the procedure we have followed can be decomposed into three steps:

- determine the dimension $\hat{D}(c)$ of the selected partition for each value c of the penalty constant c_0, where c increases from 0, by step 0.1, until $\hat{D}(c) = 1$;
- compute the difference between the dimensions of the selected partitions for two consecutive values of c_0 and retain the value $\hat{c}$ corresponding to the biggest jump in dimension under the constraint $\hat{D}(\hat{c}) \leq D_{max}$, where D_{max} is a prescribed maximal dimension;
- set $\hat{c}_j = 2\hat{c}$ and compute the d-estimator with $\mathrm{pen}(D) = \hat{c}_j D$.

Here we have taken $D_{max} = 60$ when $N = 10$, $D_{max} = 200$ when $N = 12$ and $D_{max} = 300$ when $N = 13$. We give in Table II.5.2 the ratio Q_j between the estimated risk of $\tilde{s}$ for that procedure and the estimated risk of the oracle. We also give estimations of the mean value and standard-error of $\hat{c}_j$, denoted respectively by $\bar{c}_j$ and σ_j. One realization of each d-estimator computed with that method is plotted in Figures II.5.2 (left column) and II.5.3 (center column).

Let us analyze the results of the simulations. The data-driven method really seems to adapt to the unknown distribution s: in terms of risk, it is almost as good as if we knew the constant that minimizes the risk of $\tilde{s}$. Let us now compare the different values of $Q^\star$ (or Q_j). As foreseen by the oracle-type inequality (II.3.4), the ratio between the risk of the d-estimator and that of the oracle depends on s. In particular, the ratios $Q^\star$ or Q_j reach their highest value for s_a. It should be noted that the first coordinate function of this example takes values very close to 1 on a large segment (*cf.* Figure II.5.2), a critical case according to the oracle-type inequality. However, for all examples studied here, the values of those ratios remain quite low, inferior or close to 2, except for s_a.

II.5.2 Comparing the d-estimator with the neH-estimator

For examples s_a to s_f, we have realized 500 simulations of the d-estimator and the neH-estimator, using a data-driven penalty (*cf.* [DLT09], Section 8, and the previous paragraph).

We provide in Table II.5.3 the estimated risks of each procedure, denoted by risk_d and risk_{neH}. Thanks to MATLAB 'tic' and 'toc' functions, we have measured the computational time of those 500 simulations for each estimator, denoted by time_d and time_{neH}. The ratio of those computational times is given in Table II.5.3. The neH-estimators of examples s_a to s_f are plotted in Figures II.5.2 and II.5.3 (right columns).

Table II.5.3: Comparison between the d-estimator and the neH-estimator.

s	r	N	risk_d	risk_{neH}	$\text{risk}_d/\text{risk}_{neH}$	$\text{time}_d/\text{time}_{neH}$
s_a	2	10	11.5	16.4	0.7	8.6
s_b	2	10	22.1	26.5	0.8	1.1
s_c	2	13	40.2	36.2	1.1	0.5
s_d	4	10	14.0	19.4	0.7	1.2
s_e	4	12	18.3	23.8	0.8	0.6
s_f	4	13	33.0	37.7	0.9	0.3

Those results confirm that both procedures have about the same quality of estimation, with a slight advantage though for the d-procedure for almost all the examples. As to their computational time, let us recall that the neH-procedure requires $\mathcal{O}(n\ln(n))$ computations, against only $\mathcal{O}(n)$ for the d-procedure. That difference clearly appears through our simulations. The d-estimator seems faster to compute if n is large enough, and else requires roughly the same computational time as the neH-estimator. The only exception here occurs with s_a, but 500 simulations of $\tilde{s}_a$ can be computed within a few minutes only.

II.5.3 Choosing the penalty for the hybrid procedure

For the first stage of the hybrid procedure, the d-estimator has been computed using the data-driven penalty. For the second stage, the practical choice of an adequate penalty is more delicate, since the theoretical penalty depends in this case on two constants and on the dimension $\widehat{D}$ of the partition selected during the first stage. Since those two constants seem difficult to determine, we have assigned to all partitions of $\{1,\dots,n/2\}$ into D intervals the penalty

$$\widehat{\text{pen}}^{\circ}(D) = \hat{\beta}D.$$

The value of $\hat{\beta}$ is determined once again according to the same process as $\hat{c}_j$ (*cf.* Section II.5.1), varying the value of the constant by step 1, and taking $D_{max} = \hat{D}$. Since that penalty is a linear funtion of D, the second stage of the hybrid procedure can be implemented in that case with the same algorithm as the d-procedure(*cf.* Section II.3.3). As the graph associated with all the partitions built on a partition with $\hat{D}$ intervals has $\mathcal{O}(\hat{D}^2)$ vertices, the second stage thus requires $\mathcal{O}(\hat{D}^2)$ computations, instead of $\mathcal{O}(\hat{D}^3)$ if we had used a penalty with two constants.

We have tested the hybrid procedure on examples s_a, s_b, s_d and s_e. The hybrid estimators of these examples are plotted in Figures II.5.4 and II.5.5. In order to draw a comparison between the hybrid procedure and the d-procedure, we give in Table II.5.4 the following information for distributions s_a, s_b, s_d and s_e, still computed over 500 simulations. We first recall the dimension D of the partition on which s is built. Then we indicate the estimated mean of the dimensions $\hat{D}_d$ and $\hat{D}_{hyb}$ of the partitions selected respectively by the d-procedure and the hybrid procedure with data-driven penalties, and give between parentheses their estimated standard errors. We also give the ratio $Q_{hyb:d}$ between the estimated risk of the hybrid estimator and the estimated risk of the d-estimator. The estimated means of $\hat{D}_{hyb}$ and $\hat{D}_d$ indicate that the dimension of the partition selected by the hybrid procedure is much closer to the true one. Moreover,

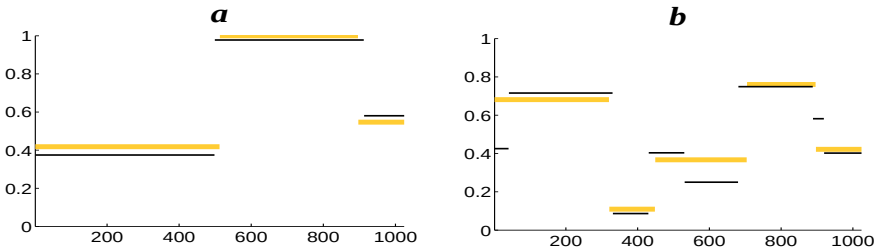

Figure II.5.4: First coordinate functions of s (thin black line) and $\tilde{s}_{hyb}$ (thick yellow line) for $s \in \{s_a, s_b\}$.

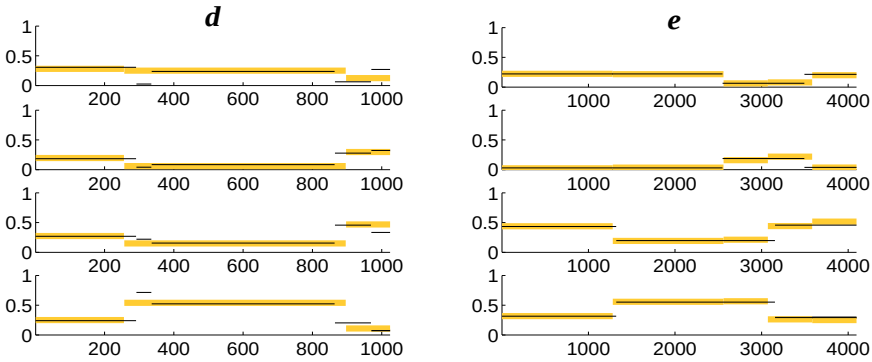

Figure II.5.5: Coordinate functions of s (thin black line) and $\tilde{s}_{hyb}$ (thick yellow line) for $s \in \{s_d, s_e\}$.

Figures II.5.4 and II.5.5 show that the most significant change-points are still detected and quite close to the true ones, and that irrelevant change-points are much fewer with the hybrid procedure. The only price to pay is an increase in risk, but only by a factor of the order of 2.

Table II.5.4: Comparison between the hybrid estimator and the d-estimator.

s	D	$\hat{D}_d$	$\hat{D}_{hyb}$	$Q_{hyb:d}$
s_a	3	9.5 (3.3)	2.9 (0.4)	1.6
s_b	8	16.1 (3.1)	4.5 (1.0)	1.6
s_d	5	8.9 (2.0)	3.0 (0.8)	1.7
s_e	5	12.3 (2.1)	4.9 (1.1)	1.7

II.5.4 Application to the segmentation of a DNA sequence

A DNA sequence of length n can be considered as a realization of a n-uple $(Y_1, \ldots, Y_n)$ of independent categorical variables with values in $\{1, \ldots, 4\}$, when coding the set of bases {A,C,G,T} by $\{1, \ldots, 4\}$ for instance. We have tested our hybrid procedure on a DNA sequence taken from the *Bacillus subtilis* genome. The whole genome of that bacterium (available on the NCBI website, under accession number NC_000964 in the Genome database) is composed of two complementary strands, each counting approximately 4 millions of bases. We have applied our procedure on the DNA sequence composed of the first $2^{21} = 2\ 097\ 152$ bases of the strand usually referred to as the (+) strand. For the sake of readability, we only represent on Fig-

Table II.5.5: Estimated proportions of bases A, C, G, T (in percentage) in the 18 first segments obtained by applying our hybrid procedure to *B. subtilis* (+) strand.

Segment	1	2	3	4	5	6	7	8	9	10
Begin	1	9730	14850	16386	22530	26626	29698	30210	35330	35842
A	32	26	30	31	27	33	30	26	37	33
C	22	31	18	25	22	24	22	30	19	23
G	26	21	33	26	31	24	33	21	29	26
T	20	23	18	18	20	18	15	23	15	18

Segment	11	12	13	14	15	16	17	18
Begin	49154	90114	101378	102402	114690	158722	159746	160770
A	31	26	37	31	31	24	26	26
C	24	30	20	25	23	20	21	30
G	26	22	27	26	27	40	34	21
T	19	22	16	18	19	17	19	22

ure II.5.6, realized with MuGeN software [HNB03], the genes corresponding to the first 178 000 base pairs (bp) of *B. subtilis* genome. We shall mainly distinguish between two kinds of genes: those coding for proteins and those coding for structural RNA. The first ones are represented by cyan or magenta arrows, depending on their orientation, an unfilled arrow indicating that the protein function is still unknown. The other ones are represented by red arrows if they code for ribosomal RNA (rRNA), dark blue arrows if they code for transfer RNA (tRNA), and by an empty box if they code for a small cytoplasmic RNA (scRNA). The rest of the sequence corresponds to intergenic regions, that do not contain any gene.

Let us first analyze our results for the subsequence represented on Figure II.5.6. Our hybrid procedure delineates 19 segments : on Figure II.5.6, the 18 corresponding change-points are represented by the highest vertical bars, and numbered from 2 to 19 so that the number i indicates the beginning of the i-th segment. The estimated proportions of bases A,C,G,T in each segment are given in Table II.5.5. Segments 2, 8, 12 and 18 clearly correspond with the 4 regions of the sequence composed at the same time of genes coding for rRNA and of genes coding for tRNA. Table II.5.5 shows that these segments have almost the same composition, that differs from the composition of any other segment. Segments 3, 5, 16 and 17 correspond with 4 regions mainly composed of protein coding genes oriented in the negative sense. We detect all such regions except for the smallest one of about 300 bases (near 45 000 bp). All the other segments are mainly composed of protein coding genes oriented in the positive sense. In particular, segment 15 includes all the genes known to code for ribosomal proteins. Let us also underline that segments 9 and 13 have similar compositions and are both situated just after one of the 4 segments coding for rRNA and tRNA. But, as the function of the protein coded by gene csfB in segment 9 is unknown, we do not know whether such similarities are related to a biological feature.

Let us now compare our results to the aforementionned procedures. The hybrid procedures based on CART and on the *neH*-procedures have been tested on the subsequences composed respectively of the first 200 000 bases of *B. subtilis* (+) strand in [Leb02] (Section 7.2.3) and of the first 2^{21} bases of that same strand in [DLT09] (Section 7.2). On Figure II.5.6, the resulting change-points are represented by the smallest vertical bars, numbered from 2 to 10, for the former procedure, and by the medium height bars, numbered from 2 to 17, for the latter procedure. As [Leb02] and [DLT09], we detect all the regions composed of genes coding for rRNA and tRNA. We recover the same changes of orientation as [DLT09], except for the

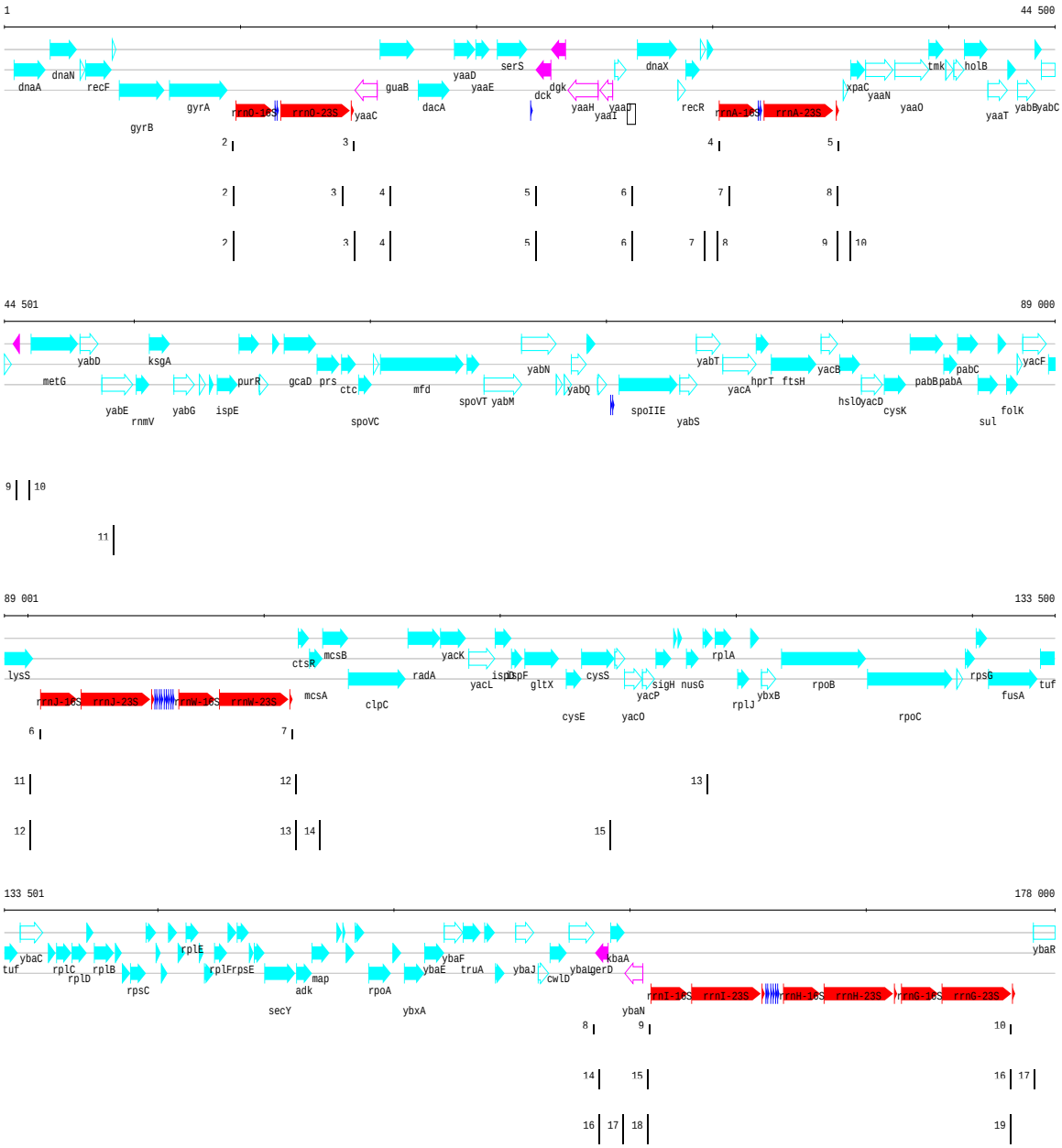

Figure II.5.6: Annotation of the first 178 000 base pairs (bp) of *B. subtilis* genome, and change-points detected by the hybrid procedures based on CART (small bars), on the *neH*-procedure (medium bars) and on the dyadic procedure (tall bars). Protein coding genes are represented by cyan or magenta arrows, depending on their orientation, unfilled when the protein function is unknown. Red and dark blue arrows represent genes coding respectively for rRNA and tRNA. The empty box stands for a gene coding for scRNA.

shortest region, and also detect another change (near 160 000 bp). The 15th segment obtained with our dyadic based hybrid procedure can be compared with the 13th segment obtained by [DLT09], that contains all genes known to code for ribosomal proteins except for the one following gene sigH. Consequently, as [DLT09], we slightly improve on the results obtained by [Leb02]. Besides, unlike [Leb02] or [DLT09], we detect two segments that might be relevant to the biologist. Moreover, our method is expected to be the fastest since its first stage has the lowest computational complexity.

Let us end with a comparison of our results with those obtained in [NBM+02] by using hidden Markov chain models on the whole (+) strand of *B. subtilis* genome. At the level of gene detection, our procedure, that relies on the assumption that the bases are independent, cannot rival with that used in [NBM+02]. But we can compare the biological features of the groups of genes that our procedure highlights with those associated with the hidden states of the most complex model fitted by [NBM+02] (see their Figure 3). Our method does not seem to detect neither intergenic regions and protein coding genes having a similar composition (called atypical genes in [NBM+02]), nor genes coding for hydrophobic proteins. But we detect the other four features, since we delineate large groups of genes coding for structural RNA, groups of protein coding genes with negative orientation, groups of protein coding genes with positive orientation, and among them the group of genes coding for ribosomal proteins, described in [NBM+02] as the main region composed of highly expressed genes. Notice also that the distinction between those four features is not made by any of the less complex models tested by [NBM+02].

II.6 Proof of the approximation result over Besov bodies

This section is devoted to the proof of Theorem 1, that extends the approximation result of DeVore and Yu [DY90] (Section 3) to the approximation of $\mathbb{R}^r$-valued functions defined on $\{1,\dots,n\}$ by piecewise constant functions. For $r=1$, that extension simply results from [DY90] (Corollary 3.2) and Proposition 7 (*cf.* Section II.6.2), which is not the case anymore for $r \geq 2$. We first describe the approximation algorithm adapted from [DY90]. Then, we give the main lines of the proof and also demonstrate the key result, which is a direct consequence of the approximation algorithm. The proofs of more technical points are postponed to the next subsections.

II.6.1 Approximation algorithm

Let us fix $p \in (0,2]$, $\alpha > 1/p - 1/2$, $R > 0$ and $D \in \{1,\dots,n\}$. In order to prove Theorem 1, we look for an upper bound for

$$\inf_{m \in \mathcal{M}_D} \|t - t_m\|^2$$

uniformly over $t \in \mathscr{B}(\alpha,p,R)$. Let I be a dyadic interval of $\{1,\dots,n\}$. The restriction of the norm $\|.\|$ to I is denoted by $\|.\|_I$. Let U be the linear subspace of $\mathbb{R}^n$ generated by the vector $(1\dots1)$, we denote by $\mathcal{E}_2(t,I)$ the error in approximating t on I by an element of $\mathbb{R}^r \otimes U$, *i.e.*

$$\mathcal{E}_2(t,I) = \inf_{c \in \mathbb{R}^r \otimes U} \|t - c\|_I.$$

Besides, both intervals obtained by dividing I into two intervals of same length are called the children of I. The algorithm proposed by DeVore and Yu [DY90] (Section 2) proceeds as follows. We fix a threshold $\epsilon > 0$. At the beginning, the set $\mathcal{I}^1(t,\epsilon)$ contains $I_{(0,0)} = \{1,\dots,n\}$. If $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, then the algorithm stops. Else, $I_{(0,0)}$ is replaced in the partition $\mathcal{I}^1(t,\epsilon)$ with his children, hence a new partition $\mathcal{I}^2(t,\epsilon)$ of $\{1,\dots,n\}$. In the same way, the k-th step starts

with a partition $\mathcal{I}^k(t,\epsilon)$ of $\{1,\dots,n\}$ into k dyadic intervals. If $\sup_{I\in\mathcal{I}^k(t,\epsilon)}\mathcal{E}_2(t,I)\leq\epsilon$, then the algorithm stops, else an interval I such that $\mathcal{E}_2(t,I)>\epsilon$ is chosen in $\mathcal{I}^k(t,\epsilon)$ and replaced with his children, hence a new partition $\mathcal{I}^{k+1}(t,\epsilon)$ of $\{1,\dots,n\}$ into $k+1$ dyadic intervals. The algorithm finally stops, giving a partition $\mathcal{I}(t,\epsilon)$. Denoting by $S(t,\epsilon)$ the linear space composed of the functions that are piecewise constant on $\mathcal{I}(t,\epsilon)$, the approximation $A(t,\epsilon)$ of t associated with this partition is defined as the orthogonal projection of t on $\mathbb{R}^r\otimes S(t,\epsilon)$. So, the approximation error of t by $A(t,\epsilon)$ satisfies

$$\|t-A(t,\epsilon)\|^2=\sum_{I\in\mathcal{I}(t,\epsilon)}\big(\mathcal{E}_2(t,I)\big)^2\leq|\mathcal{I}(t,\epsilon)|\epsilon^2.$$

For any $\epsilon>0$ such that the algorithm stops at the latest at step D, the approximation of t that we get belongs to the collection $\{\mathbb{R}^r\otimes S_m\}_{m\in\mathcal{M}_D}$. Therefore

$$\inf_{m\in\mathcal{M}_D}\|t-t_m\|^2\leq|\mathcal{I}(t,\epsilon)|\epsilon^2.$$

Let us denote by $\mathcal{E}_D(t)$ the infimum of $|\mathcal{I}(t,\epsilon)|\epsilon^2$ taken over all $\epsilon>0$ satisfying $|\mathcal{I}(t,\epsilon)|\leq D$. This is in fact the quantity that we shall bound, as indicated in Theorem 5 below.

Theorem 5 *Let $p\in(0,2]$, $\alpha>1/p-1/2$ and $R>0$. For all $D\in\{1,\dots,n\}$ and $t\in\mathscr{B}(\alpha,p,R)$,*

$$\mathcal{E}_D(t)\leq C(\alpha,p)nR^2D^{-2\alpha}.$$

We then get Theorem 1 as a straightforward consequence of Theorem 5.

II.6.2 Proof of Theorem 5: the main lines

We shall prove Theorem 5 by following the path of DeVore and Yu in [DY90] (Section 3). Here are the notions and notation that we will need along the proof. Let $p>0$, $\alpha>0$ and $t\in\mathscr{M}(r,n)$. For every subset I of $\{1,\dots,n\}$, let

$$\mathcal{E}_p(t,I)=\inf_{v\in\mathbb{R}^r}\left(\sum_{k\in I}\|t_k-v\|_r^p\right)^{1/p}.$$

We define the vector $t^{\sharp,\alpha,p}$ in $\mathbb{R}^n$ whose coordinates are

$$t_i^{\sharp,\alpha,p}=\sup_{I\ni i}|I|^{-(\alpha+1/p)}\mathcal{E}_p(t,I),\ \text{for } i=1,\dots,n,$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ that contain i. We denote by $\|.\|_{\ell_p}$ the (quasi-)norm defined on $\mathbb{R}^n$ by

$$\|u\|_{\ell_p}=\left(\sum_{i=1}^n|u_i|^p\right)^{1/p}$$

(that is a norm only for $p\geq1$) and by $\|.\|_{\ell_p,I}$ its restriction to a subset I of $\{1,\dots,n\}$. We define on $\mathbb{R}^n$ the discrete Hardy-Littlewood maximal function M_p by

$$\big(M_p(u)\big)_i=\sup_{I\ni i}|I|^{-1/p}\|u\|_{\ell_p,I},\ \text{for } i=1,\dots,n,$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ containing i. Last, we recall that every vector $u\in\mathbb{R}^n$ is identified with the function $u:i\in\{1,\dots,n\}\mapsto u_i$, hence the meaning of notation such as $u\leq v$ or u^q, for $u\in\mathbb{R}^n$, $v\in\mathbb{R}^n$ and $q>0$.

The beginning of the proof directly results from the way the algorithm works out. A dimension D being fixed, choosing $\epsilon > 0$ as small as possible such that the algorithm generates a partition with at most D intervals leads to a first comparison between the quantities $\mathcal{E}_D(t)$ and $D^{-2\alpha}$, without making use of any particular hypothesis on t.

Proposition 4 *Let $\alpha > 0$ and $p(\alpha) = (\alpha + 1/2)^{-1}$. For all $1 \leq D \in \leq n$ and $t \in \mathscr{M}(r, n)$,*

$$\mathcal{E}_D(t) \leq C(\alpha) \|t^{\sharp,\alpha,2}\|^2_{\ell_{p(\alpha)}} D^{-2\alpha}.$$

Proof. If $t^{\sharp,\alpha,2} = 0$, then, whatever $\epsilon > 0$, $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, so $\mathcal{E}_D(t) = 0$, which completes the proof in that case. Let us now assume that $t^{\sharp,\alpha,2}$ is non-null, and let $\epsilon > 0$. If $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, then $|\mathcal{I}(t, \epsilon)| = 1$. Else, let I be a dyadic interval that belongs to $\mathcal{I}(t, \epsilon)$, then I is a child of a dyadic interval $\tilde{I}$ such that

$$\epsilon < \mathcal{E}_2\big(t, \tilde{I}\big).$$

Using the definition of $t^{\sharp,\alpha,2}$, we get, for all $i \in \tilde{I}$,

$$\mathcal{E}_2\big(t, \tilde{I}\big) \leq \big|\tilde{I}\big|^{\alpha+1/2} t_i^{\sharp,\alpha,2}.$$

Since $I \subset \tilde{I}$, $|\tilde{I}| = 2|I|$ and $p(\alpha) = (\alpha + 1/2)^{-1}$, the last two inequalities lead, for all $i \in I$, to

$$\epsilon < 2^{1/p(\alpha)} |I|^{1/p(\alpha)} t_i^{\sharp,\alpha,2},$$

hence

$$\epsilon^{p(\alpha)} < 2 \sum_{i \in I} \big(t_i^{\sharp,\alpha,2}\big)^{p(\alpha)}.$$

Then we deduce by summing over all the intervals I in the partition $\mathcal{I}(t, \epsilon)$ that

$$|\mathcal{I}(t, \epsilon)| \leq 2 \|t^{\sharp,\alpha,2}\|^{p(\alpha)}_{\ell_{p(\alpha)}} \epsilon^{-p(\alpha)}.$$

Whether $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$ or not, by choosing $\epsilon = 2^{1/p(\alpha)} \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}} D^{-1/p(\alpha)}$, we get a partition $\mathcal{I}(t, \epsilon)$ that contains at most D elements and satisfies

$$|\mathcal{I}(t, \epsilon)| \epsilon^2 \leq D^{1-2/p(\alpha)} 2^{2/p(\alpha)} \|t^{\sharp,\alpha,2}\|^2_{\ell_{p(\alpha)}}.$$

As $p(\alpha) = (\alpha + 1/2)^{-1}$, we conclude that

$$|\mathcal{I}(t, \epsilon)| \epsilon^2 \leq 4^{\alpha+1/2} \|t^{\sharp,\alpha,2}\|^2_{\ell_{p(\alpha)}} D^{-2\alpha}.$$

■

The proof of Theorem 5 now relies upon three inequalities. The first one allows to draw a comparison between $\mathcal{E}_D(t)$ and $D^{-2\alpha}$ via a term that does not depend on $t^{\sharp,\alpha,2}$ anymore but on $t^{\sharp,\alpha,p(\alpha)}$. It is the discrete analogue of a particular case of Theorem 4.3. of [DS84].

Proposition 5 *Let $\alpha > 0$ and $p(\alpha) = (\alpha + 1/2)^{-1}$. For all $t \in \mathscr{M}(r, n)$,*

$$t^{\sharp,\alpha,2} \leq C(\alpha) M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big).$$

For $\alpha > 0$, $p(\alpha) = (\alpha + 1/2)^{-1}$ and $D \in \{1, \ldots, n\}$, Propositions 4 and 5 immediately lead to

$$\mathcal{E}_D(t) \leq C(\alpha) \big\| M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big) \big\|^2_{\ell_{p(\alpha)}} D^{-2\alpha}.$$

Let us now fix $p \in (0,2]$. By Jensen's inequality, we have

$$\left\|M_{p(\alpha)}\left(t^{\sharp,\alpha,p(\alpha)}\right)\right\|_{\ell_{p(\alpha)}} \leq n^{1/p(\alpha)-1/p}\left\|M_{p(\alpha)}\left(t^{\sharp,\alpha,p(\alpha)}\right)\right\|_{\ell_p}$$

and

$$t^{\sharp,\alpha,p(\alpha)} \leq t^{\sharp,\alpha,p},$$

hence

$$\mathcal{E}_D(t) \leq C(\alpha) n^{2(\alpha+1/2-1/p)}\left\|M_{p(\alpha)}(t^{\sharp,\alpha,p})\right\|_{\ell_p}^2 D^{-2\alpha}.$$

The following maximal inequality (Inequality (II.6.15) below) ensures a control of u over its maximal functions. It is in fact the discrete version of the Hardy-Littlewood maximal inequality, that may be found in [BS88] (Theorem 3.10, p.125).

Proposition 6 *Let $q > 1$. For all $u \in \mathbb{R}^n$,*

$$\|M_1(u)\|_{\ell_q} \leq C(q)\|u\|_{\ell_q}.$$

Since the maximal function M_q, $q > 0$, is related to M_1 by the property

$$M_q(u) = \left(M_1(u^q)\right)^{1/q}, \text{for all } u \in \mathbb{R}^n,$$

Proposition 6 yields, for all $r > q > 0$ and $u \in \mathbb{R}^n$,

$$\|M_q(u)\|_{\ell_r} \leq C(r,q)\|u\|_{\ell_r}. \tag{II.6.15}$$

Thus, when applied with $u = t^{\sharp,\alpha,p}$, $r = p$ and $q = p(\alpha)$, this inequality leads to

$$\mathcal{E}_D(t) \leq C(\alpha,p) n^{2(\alpha+1/2-1/p)}\|t^{\sharp,\alpha,p}\|_{\ell_p}^2 D^{-2\alpha}.$$

Last, Proposition 7 below provides the adequate control of the ℓ_p-(quasi-)norm of $t^{\sharp,\alpha,p}$ by the size of the wavelet coefficients of t and allows to complete immediately the proof of Theorem 5.

Proposition 7 *Let $p \in (0,2]$ and $\alpha > 1/p - 1/2$. For all $t \in \mathscr{M}(r,n)$,*

$$\|t^{\sharp,\alpha,p}\|_{\ell_p} \leq C(\alpha,p) n^{-(\alpha+1/2-1/p)}\left(\sum_{j=0}^{N-1} 2^{jp(\alpha+1/2-1/p)} \sum_{\lambda\in\Lambda(j)} \|\beta_\lambda\|_r^p\right)^{1/p},$$

where, for all $\lambda \in \Lambda$, β_λ stands for the column-vector of $\mathbb{R}^r$ with l-th line $\beta_\lambda^{(l)} = \langle t^{(l)}, \phi_\lambda\rangle_n$, for $l = 1,\ldots, r$.

II.6.3 Proof of Proposition 5

The proof of Proposition 5 mostly relies on a lemma that we demonstrate after introducing some notation. Let I be a dyadic interval of $\{1,\ldots,n\}$, $t \in \mathscr{M}(r,n)$, and $p > 0$. By a compactness argument, there exists at least one vector in $\mathbb{R}^r$, denoted by $v_p(t,I)$, satisfying

$$\mathcal{E}_p(t,I) = \left(\sum_{k\in I} \|t_k - v_p(t,I)\|_r^p\right)^{1/p}.$$

We define the vectors $u_p(t,I)$ and $t^{\sharp,\alpha,p,I}$ in $\mathbb{R}^n$ whose coordinates are null outside of I and given otherwise respectively by

$$\left(u_p(t,I)\right)_i = \|t_i - v_p(t,I)\|_r, \text{ for } i \in I,$$

and

$$t_i^{\sharp,\alpha,p,I} = \sup_{I \supset J \ni i} |J|^{-(\alpha+1/p)} \mathcal{E}_p(t,J), \text{ for } i \in I,$$

where the supremum is taken over all dyadic intervals J of $\{1,\dots,n\}$ that are contained in I and contain i. Last, for $u \in \mathbb{R}^n$, we denote by $u^\star$ its decreasing rearrangement, *i.e.* the $\mathbb{R}^n$-vector satisfying

$$u_1^\star \geq u_2^\star \geq \ldots \geq u_n^\star \text{ and } \{u_i^\star; 1 \leq i \leq n\} = \{|u_i|; 1 \leq i \leq n\}.$$

Lemma 1 *Let $\alpha > 0$, $p > 0$ and $t \in \mathscr{M}(r,n)$. Let I be a dyadic interval of $\{1,\dots,n\}$ containing at least two elements. For all $j \in \{1,\dots,|I|/2\}$,*

$$\big(u_p(t,I)\big)_j^\star \leq C(\alpha,p)\left(\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)_k^\star + j^\alpha\big(t^{\sharp,\alpha,p,I}\big)_j^\star\right).$$

Proof. We fix $j \in \{1,\dots,|I|/2\}$. Let E be the set composed of all the indices i in $\{1,\dots,n\}$ satisfying $(t^{\sharp,\alpha,p,I})_i > (t^{\sharp,\alpha,p,I})_j^\star$. As $|E| \leq j-1$, we only have to prove that

$$\big(u_p(t,I)\big)_i \leq C(\alpha,p)\left(\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)_k^\star + j^\alpha\big(t^{\sharp,\alpha,p,I}\big)_j^\star\right) \tag{II.6.16}$$

for all the indices $i \in \{1,\dots,n\}$, except maybe for those belonging to E. Consider $i \in \{1,\dots,n\}$ such that $i \notin E$. If $i \notin I$, then $\big(u_p(t,I)\big)_i = 0$, so Inequality (II.6.16) is trivial. Suppose now that $i \in I$ and $i \notin E$, and let $\{I_l\}_{1 \leq l \leq m}$ be the sequence of dyadic intervals defined by

$$I_1 = I,\ I_{l+1} \text{ is the child of } I_l \text{ containing } i, \text{ and } I_m = \{i\},$$

where $m \geq 2$ because $|I| \geq 2$. Notice that, for all $l \in \{0,\dots,m-1\}$, $|I_{l+1}| = 2^{-l}|I|$. Let q be the strictly positive integer such that

$$2^{-(q+1)}|I| < j \leq 2^{-q}|I|.$$

That definition implies, in particular, that $2^{-q}|I| \geq 1$,hence $q < m$. From the triangle inequality,

$$\big(u_p(t,I)\big)_i \leq \sum_{l=2}^{q} \|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r + \sum_{l=q+1}^{m} \|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r, \tag{II.6.17}$$

with the convention that the first sum in Inequality (II.6.17) is null for $q = 1$. Let us fix $l \in \{2,\dots,m\}$ and determine an upper-bound for the term $\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r$. We recall that $I_l \subset I_{l-1}$ and $|I_{l-1}| = 2|I_l|$. Besides, for all $p > 0$, the (quasi-)norm $\|.\|_{\ell_p}$ satisfies a triangle inequality within a multiplicative constant $C(p)$, where we can take $C(p) = 1$ for $p \geq 1$, and $C(p) = 2^{1/p}$ for $0 < p < 1$. Therefore, we get

$$\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \leq C(p)|I_l|^{-1/p}\Big(\mathcal{E}_p(t,I_{l-1}) + \mathcal{E}_p(t,I_l)\Big),$$

which leads to

$$\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \leq C(\alpha,p)|I_l|^\alpha \min_{k \in I_l} t_k^{\sharp,\alpha,p,I}. \tag{II.6.18}$$

Let us bound the first sum appearing in (II.6.17). For all $l \in \{2,\dots,m\}$, we have

$$\min_{k \in I_l} t_k^{\sharp,\alpha,p,I} \leq \big(t^{\sharp,\alpha,p,I}\big)_{|I_l|}^\star = \min_{1 \leq k \leq |I_l|} \big(t^{\sharp,\alpha,p,I}\big)_k^\star,$$

and, as $|I_{l+1}| = |I_l|/2$,

$$|I_l|^\alpha = C(\alpha)\int_{|I_{l+1}|}^{|I_l|} x^{\alpha-1}\,\mathrm{d}x \le C(\alpha)\sum_{k=|I_{l+1}|}^{|I_l|} k^{\alpha-1}.$$

Consequently, when $q \ge 2$, Inequality (II.6.18) yields

$$\begin{aligned}\sum_{l=2}^{q}\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r &\le C(\alpha,p)\sum_{l=2}^{q}\sum_{k=|I_{l+1}|}^{|I_l|} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)^\star_k\\ &\le C(\alpha,p)\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)^\star_k.\end{aligned}$$

Regarding the second sum appearing in (II.6.17), we now use Inequality (II.6.18) combined with the following remarks. For all l such that $q+1 \le l \le m$, we have $\min_{k\in I_l} t_k^{\sharp,\alpha,p,I} \le t_i^{\sharp,\alpha,p,I}$, since I_l contains i, and we recall that $|I_l| = 2^{-(l-1)}|I|$. Therefore,

$$\sum_{l=q+1}^{m}\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \le C(\alpha,p)|I|^\alpha\big(t^{\sharp,\alpha,p,I}\big)_i\sum_{l=q+1}^{m} 2^{-(l-1)\alpha}.$$

Furthermore, remember that $2^{-(q+1)}|I| < j$ and $i \notin E$, so we finally obtain

$$\sum_{l=q+1}^{m}\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \le C(\alpha,p)j^\alpha\big(t^{\sharp,\alpha,p,I}\big)^\star_j.$$

We have thus proved inequality (II.6.16) and Lemma 1. ■

Let us now prove Proposition 5. Let $\alpha > 0$, $p(\alpha) = (\alpha+1/2)^{-1}$, $t \in \mathscr{M}(r,n)$ and $i \in \{1,\dots,n\}$. From the definition of $\mathcal{E}_2(t,I)$ for a subset I of $\{1,\dots,n\}$, and due to the fact that $\mathcal{E}_2(t,\{i\}) = 0$, we have

$$t_i^{\sharp,\alpha,2} \le \sup_{I\ni i}|I|^{-1/p(\alpha)}\|u_{p(\alpha)}(t,I)\|_{\ell_2},$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ that contain i, except for $\{i\}$. We fix such an interval I. The sequence $\big\{\big(u_{p(\alpha)}(t,I)\big)^\star_j\big\}_{1\le j\le n}$ decreases and is null for $j \ge |I|+1$, hence

$$\big\|u_{p(\alpha)}(t,I)\big\|_{\ell_2}^2 \le 2\sum_{j=1}^{|I|/2}\Big(\big(u_{p(\alpha)}(t,I)\big)^\star_j\Big)^2.$$

For $0 < p,q < +\infty$, we denote by $\|.\|_{\ell_{p,q}}$ the Lorentz (quasi-)norm defined on $\mathbb{R}^n$ by

$$\|u\|_{\ell_{p,q}} = \left(\sum_{i=1}^{n} i^{-1}(i^{1/p}u_i^\star)^q\right)^{1/q}.$$

For all subset I of $\{1,\dots,n\}$, we denote by $\|.\|_{\ell_{p,q},I}$ the restriction of $\|.\|_{\ell_{p,q}}$ to I. In particular, notice that, for all $u \in \mathbb{R}^n$ and $0 < p,q < +\infty$,

$$\|u\|_{\ell_{p,p}} = \|u\|_{\ell_p} \text{ and } \|u^\star\|_{\ell_{p,q}} = \|u\|_{\ell_{p,q}}.$$

From Lemma 1 and the definition of $p(\alpha)$, we get

$$\big\|u_{p(\alpha)}(t,I)\big\|_{\ell_2}^2 \le C(\alpha)\left(\sum_{j=1}^{|I|/2} j^{-1}\left(j^{1/2}\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p(\alpha),I}\big)^\star_k\right)^2 + \big\|\big(t^{\sharp,\alpha,p(\alpha),I}\big)^\star\big\|_{\ell_{p(\alpha),2}}^2\right).$$

Using a discrete version of Hardy's inequality (*cf.* Proposition 11 in the Appendix) and noticing that $t^{\sharp,\alpha,p(\alpha),I} \leq t^{\sharp,\alpha,p(\alpha)}$, we are led to

$$\left\|u_{p(\alpha)}(t,I)\right\|_{\ell_2} \leq C(\alpha)\|t^{\sharp,\alpha,p(\alpha)}\|_{\ell_{p(\alpha),2},I}.$$

Last, since $p(\alpha) < 2$, we conclude thanks to classical inequalities between Lorentz (quasi-)norms (*cf.* Proposition 10 in the Appendix) that

$$t_i^{\sharp,\alpha,2} \leq C(\alpha) \sup_{I \ni i} |I|^{-1/p(\alpha)} \|t^{\sharp,\alpha,p(\alpha)}\|_{\ell_{p(\alpha)},I}$$

where the supremum is taken over all the dyadic intervals I of $\{1,\ldots,n\}$ that contain i.

II.6.4 Proof of Proposition 6

Let $q > 1$ and $u \in \mathbb{R}^n$. As $M_1(u) = M_1(|u|)$, we can suppose that u has positive or null coordinates. Let us first demonstrate that, for all $i \in \{1,\ldots,n\}$,

$$(M_1(u))_i^\star \leq C\left(i^{-1}\sum_{k=1}^{i} u_k^\star\right). \tag{II.6.19}$$

If $i = 1$, then this inequality easily follows from the definitions of $(M_1(u))_1^\star$ and $u_1^\star$. Let us now fix $i \in \{2,\ldots,n\}$. We can write u as $u = v + w$, where v and w are the $\mathbb{R}^n$-vectors whose respective coordinates are

$$v_k = \max\{u_k - u_i^\star, 0\} \text{ and } w_k = \min\{u_k, u_i^\star\}, \text{ for } k = 1,\ldots,n.$$

From the triangle inequality, we deduce that $M_1(u) \leq M_1(v) + M_1(w)$. Properties of discrete decreasing rearrangements recalled in Proposition 9 then lead to

$$(M_1(u))_i^\star \leq (M_1(v))_{\lceil i/2 \rceil}^\star + (M_1(w))_{\lfloor i/2 \rfloor}^\star.$$

Moreover,

$$(M_1(w))_{\lfloor i/2 \rfloor}^\star \leq \|M_1(w)\|_{\ell_\infty} \leq \|w\|_{\ell_\infty},$$

and, from Proposition 9,

$$(M_1(v))_{\lceil i/2 \rceil}^\star \leq 2i^{-1}\|v\|_{\ell^1}.$$

Consequently,

$$(M_1(u))_i^\star \leq C\big(i^{-1}\|v\|_{\ell^1} + \|w\|_{\ell^\infty}\big). \tag{II.6.20}$$

Let I be the set of all the indices $l \in \{1,\ldots,n\}$ such that $u_l > u_i^\star$. From the definitions of v and w, we get

$$\|v\|_{\ell^1} + i\|w\|_{\ell^\infty} \leq \sum_{k=1}^{|I|} u_k^\star + (i - |I|)u_i^\star = \sum_{k=1}^{i} u_k^\star,$$

which, given Inequality (II.6.20), completes the proof of (II.6.19). We now have

$$\|(M_1(u))^\star\|_{\ell_q}^q \leq C(q)\sum_{i=1}^{n}\left(i^{-1}\sum_{k=1}^{i} u_k^\star\right)^q. \tag{II.6.21}$$

Let us denote by q' the conjugate exponent of q and introduce, for all $k \in \{1,\ldots,n\}$, $u_k^\star = k^{-1/qq'}k^{1/qq'}u_k^\star$. We deduce from Hölder's inequality that

$$\sum_{i=1}^{n}\left(i^{-1}\sum_{k=1}^{i} u_k^\star\right)^q \leq \sum_{i=1}^{n}\left(q' i^{-1/q}\right)^{q/q'}\left(i^{-1}\sum_{k=1}^{i} k^{1/q'}(u_k^\star)^q\right).$$

Interchanging the order of the summations, we obtain

$$\sum_{i=1}^{n}\left(i^{-1}\sum_{k=1}^{i}u_k^\star\right)^q \le C(q)\sum_{k=1}^{n}(u_k^\star)^q.$$

Consequently,

$$\|(M_1(u))^\star\|_{\ell_q} \le C(q)\|u^\star\|_{\ell_q},$$

hence Proposition 6.

II.6.5 Proof of Proposition 7

Let $p\in(0,2]$, $\alpha>1/p-1/2$ and $t\in\mathscr{M}(r,n)$. For all $i\in\{1,\dots,n\}$ and all $0\le J\le N$, we denote by $I(J,i)$ the only dyadic interval of length $n2^{-J}$ that is contained in $\{1,\dots,n\}$ and contains i. From the definition of $t^{\sharp,\alpha,p}$, we deduce

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \le \sum_{J=0}^{N-1}(n^{-1}2^J)^{\alpha p+1}\sum_{i=1}^{n}\Big(\mathcal{E}_p\big(t,I(J,i)\big)\Big)^p. \tag{II.6.22}$$

Let us first suppose that $0<p\le 1$. From the definition of $\mathcal{E}_p(t,I(J,i))$, we have

$$\Big(\mathcal{E}_p\big(t,I(J,i)\big)\Big)^p \le \sum_{k\in I(J,i)}\|t_k-t_i\|_r^p.$$

For all $-1\le j\le N-1$, the functions $\{\phi_\lambda\}_{\lambda\in\Lambda(j)}$ are constant over any dyadic interval of length $n2^{-(j+1)}$. Therefore, if k belongs to $I(J,i)$, then

$$t_k-t_i=\sum_{j=J}^{N-1}\sum_{\lambda\in\Lambda(j)}\beta_\lambda(\phi_{\lambda\,k}-\phi_{\lambda\,i}).$$

As $0<p\le 1$, we deduce from the classical inequality between ℓ_p-quasi-norm and ℓ_1-norm

$$\sum_{i=1}^{n}\Big(\mathcal{E}_p\big(t,I(J,i)\big)\Big)^p \le 2n^{2-p/2}2^{-J}\sum_{j=J}^{N-1}2^{jp(1/2-1/p)}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p.$$

Interchanging the order of the summations, we get

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \le C(\alpha,p)n^{1-p(\alpha+1/2)}\sum_{j=0}^{N-1}2^{jp(\alpha+1/2-1/p)}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p.$$

Let us now consider the case $1<p\le 2$. We fix $0\le J\le N-1$ and define

$$T(J)=\sum_{j=J}^{N-1}\sum_{\lambda\in\Lambda(j)}\beta_\lambda\phi_\lambda.$$

As $t-T(J)$ is constant over any dyadic interval of length $n2^{-J}$,

$$\mathcal{E}_p\big(t,I(J,i)\big)=\mathcal{E}_p\big(T(J),I(J,i)\big).$$

This equality and the definition of $\mathcal{E}_p\big(T(J), I(J,i)\big)$ lead to

$$\sum_{i=1}^{n}\Big(\mathcal{E}_p\big(t, I(J,i)\big)\Big)^p \le \sum_{i=1}^{n}\sum_{k\in I(J,i)}\big\|\big(T(J)\big)_k\big\|_r^p$$
$$\le n2^{-J}\sum_{k=1}^{n}\Big(\sum_{j=J}^{N-1}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r|\phi_{\lambda\,k}|\Big)^p.$$

From (II.6.22) and this last inequality, we get

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \le n^{-\alpha p}\sum_{k=1}^{n}\sum_{J=0}^{N-1}\Big(2^{J\alpha}\sum_{j=J}^{N-1}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r|\phi_{\lambda\,k}|\Big)^p.$$

Then, using one of Hardy's inequalities (*cf.* Proposition 10) and remembering that, for all $j\in\{-1,\dots,N-1\}$, the functions $\{\phi_\lambda\}_{\lambda\in\Lambda(j)}$ have disjoint supports, we conclude that

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \le C(\alpha,p)n^{-\alpha p}\sum_{j=0}^{N-1}2^{j\alpha p}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p\sum_{k=1}^{n}|\phi_{\lambda\,k}|^p,$$

hence Proposition 7.

II.7 Lower bound for the minimax risk over $\mathscr{V}\mathscr{P}(\alpha,R)$

Proposition 8 *For all $\alpha>0$, there exists a real $k_2(\alpha)\in(0,1)$ such that, if $R\ge n^{-1/2}$ and $R\le k_2(\alpha)n^\alpha$, then*

$$\mathcal{R}_{\mathscr{V}}(\alpha,R)\ge C(\alpha)(nR^2)^{1/(2\alpha+1)}.$$

Proof. Let $R\ge 0$ and $0\le J\le N$. For $\theta\in\{0,1\}^{2^J}$, we use the notation $\theta=(\theta_0,\dots,\theta_{2^J-1})$. The Hamming distance between two elements θ and θ' in $\{0,1\}^{2^J}$ is $\delta(\theta,\theta')=\sum_{k=0}^{2^J-1}|\theta_k-\theta'_k|$. The Kullback-Leibler divergence is denoted by K. We shall construct a family $\{s_\theta\}_{\theta\in\Theta}$ of elements of $\mathscr{V}\mathscr{P}(\alpha,R)$ indexed by a properly chosen subset Θ of $\{0,1\}^{2^J}$. According to Birgé's version of Fano's lemma (*cf.* [Mas07], Corollary 2.19), if $\max_{\theta,\theta'\in\Theta}K(s_\theta,s_{\theta'})\le\kappa\ln(|\Theta|)$, where κ is a universal constant that belongs to $(0,1)$, then

$$\mathcal{R}_{\mathscr{V}}(\alpha,R)\ge\frac{1-\kappa}{4}\eta^2$$

where $\eta^2=\min_{\theta,\theta'\in\Theta,\theta\ne\theta'}\|s_\theta-s_{\theta'}\|^2$. Let us fix $\theta\in\{0,1\}^{2^J}$. We define $s_\theta\in\mathscr{M}(r,n)$ by

$$\begin{cases} s_\theta^{(1)}=1/2+ag_\theta\\ s_\theta^{(2)}=1/2-ag_\theta\\ s_\theta^{(l)}=0 \text{ for } 3\le l\le r, \text{ if } r\ge 3\end{cases}$$

where $a=\sqrt{2}R2^{-\alpha J}/4$ and $g_\theta=\sum_{k=0}^{2^J-1}(2\theta_k-1)\mathbb{1}_{I_{(J,k)}}$. For all $1\le i<j\le n$,

$$\|s_{\theta\,j}-s_{\theta\,i}\|_r=a\sqrt{2}|g_\theta(j)-g_\theta(i)|.$$

Since g_θ is constant on any dyadic interval of length $n/2^J$ and takes values in $\{-1,1\}$,

$$V_\alpha^{1/\alpha}(s_\theta)\le 2^J(2\sqrt{2}a)^{1/\alpha}\le R^{1/\alpha},$$

so $s_\theta \in \mathscr{V}(\alpha, R)$. Besides, as $\|g_\theta\|_\infty \leq 1$, if we assume that $R2^{-\alpha J} \leq \sqrt{2}$, then $s_\theta \in \mathscr{P}$. Let us rather assume that $\sqrt{2}R2^{-\alpha J} \leq 1$. We can then apply Lemma 4 of [DLT09] and get, for all $\theta, \theta' \in \{0,1\}^{2^J}$,

$$K(s_\theta, s_{\theta'}) \leq 4\|s_\theta - s_{\theta'}\|^2.$$

Now, according to Varshamov-Gilbert's lemma (*cf.* [Mas07], Lemma 4.7, for instance), we can choose $\Theta \subset \{0,1\}^{2^J}$ such that, for any two distinct elements $\theta, \theta' \in \Theta$,

$$\delta(\theta, \theta') > 2^J/4 \tag{II.7.23}$$

and $\ln(|\Theta|) > 2^J/8$. For all $\theta, \theta' \in \{0,1\}^{2^J}$, $\|s_\theta - s_{\theta'}\|^2 = 8a^2 n 2^{-J}\delta(\theta, \theta')$, so, for all $\theta, \theta' \in \Theta$,

$$K(s_\theta, s_{\theta'}) \leq 2^8 a^2 n 2^{-J} \ln(|\Theta|)$$

and

$$\eta^2 \geq 2na^2.$$

Let us now set $k_2(\alpha) = \min\left(\sqrt{\kappa/2^5}, (2^5/\kappa)^\alpha 2^{-(\alpha+1/2)}\right)$ and assume that $n^{-1/2} \leq R \leq k_2(\alpha)n^\alpha$. We can then define J as the smallest integer in $\{0, \ldots, N\}$ such that $2^8 a^2 n 2^{-J} \leq \kappa$, *i.e.*

$$J = \min\left\{0 \leq j \leq N \text{ s.t. } (\kappa^{-1}2^5 nR^2)^{1/(2\alpha+1)} \leq 2^j\right\}.$$

Such an integer J does satisfy $\sqrt{2}R2^{-\alpha J} \leq 1$ and leads to

$$\mathcal{R}_{\mathscr{V}}(\alpha, R) \geq C(\kappa, \alpha)(nR^2)^{1/(2\alpha+1)}.$$

■

Appendix

We state here, for vectors in $\mathbb{R}^n$, a few inequalities that are similar to classical inequalities for functions of a continuous parameter. The proofs of the latter, which may be found in [BS88], for instance, are easy to transpose to the finite-dimensional case.

Proposition 9 (Some properties of decreasing rearrangements) *Let u and v be two vectors in $\mathbb{R}^n$. For all $\lambda \geq 0$, let $I_u(\lambda)$ be the set of the indices k in $\{1, \ldots, n\}$ such that $|u_k| \geq \lambda$.*

i) For all $i \in \{1, \ldots, n\}$, $u_i^\star = \sup\{\lambda \geq 0 \; s.t. \; |I_u(\lambda)| \geq i\}$.

ii) If, for all $i \in \{1, \ldots, n\}$, $u_i \leq v_i$, then, for all $i \in \{1, \ldots, n\}$, $u_i^\star \leq v_i^\star$.

iii) For all $i, j \in \{1, \ldots, n\}$ such that $1 \leq i + j \leq n$, $(u + v)_{i+j}^\star \leq u_i^\star + v_j^\star$.

iv) For all $i \in \{1, \ldots, n\}$, $(M_1(u))_i^\star \leq i^{-1} \|u\|_{\ell_1}$.

Proof. See, for instance, [BS88], Proposition 1.7. and Theorem 3.3. ■

Proposition 10 (Inequalities between Lorentz (quasi-)norms) *Let p, q and q' be positive reals and let u be a vector in $\mathbb{R}^n$.*

i) If $p \leq q$, then $\|u\|_{\ell_{p,\infty}} \leq C(p, q) \|u\|_{\ell_{p,q}}$.

ii) If $q' \leq q$, then $\|u\|_{\ell_{p,q}} \leq C(p, q, q') \|u\|_{\ell_{p,q'}}$.

Proof. See, for instance, [BS88], Proposition 4.2. ■

Proposition 11 (Hardy's inequalities) *Let $q > 1$ and let ψ be a vector in $\mathbb{R}^n$ whose coordinates are non-negative.*

i) For all $\lambda < 1$,

$$\sum_{i=1}^{n} i^{-1} \left(i^{1-\lambda} \sum_{k=i}^{n} k^{-1} \psi_k \right)^q \leq C(\lambda, q) \sum_{i=1}^{n} i^{-1} (i^{1-\lambda} \psi_i)^q.$$

ii) For all $\alpha > 0$,

$$\sum_{i=1}^{n} \left(2^{i\alpha} \sum_{k=i}^{n} \psi_k \right)^q \leq C(\alpha, q) \sum_{i=1}^{n} (2^{i\alpha} \psi_i)^q.$$

Proof. See, for instance, [BS88], Lemma 3.9. ■

Chapter III

Histogram selection based on possibly censored data

That chapter presents a work in collaboration with Cécile Durot.

Abstract

We propose in this chapter a unified approach to functional estimation problems based on possibly censored data. The general framework that we define allows for instance to handle density and hazard rate estimation based on randomly right-censored data, or regression. Given a collection of histograms, our estimation procedure consists in selecting the best histogram among that collection from the data, by minimizing a penalized least-squares type criterion. For a general collection of histograms, we obtain nonasymptotic oracle-type inequalities. Then, we consider the collection of histograms built on partitions into dyadic intervals, a choice inspired from an approximation result due to DeVore and Yu. In that case, our estimator is also adaptive in the minimax sense over a wide range of smoothness classes, that contain functions of inhomogeneous smoothness. Besides, its computational complexity is only linear in the size of the sample.

III.1 Introduction

Reliability and survival analysis provide many functional estimation problems based on lifetimes that can only be partly observed. For instance, in a clinical study devoted to the survival times after an operation, some patients may live beyond the end of the study or die from a cause unrelated to this operation. Hence, survival times may only be observed up to a so-called censoring event, and are then considered as randomly right-censored data. However, assuming they are identically distributed, one may still want to estimate their common density, or other function of interest such as the hazard rate. The aim of this chapter is to provide a nonparametric estimator of a function based for instance on such censored data, that also adapts to the unknown and possibly inhomogeneous smoothness of the function.

Let us mention nonparametric methods already studied in related frameworks. We shall focus on adaptive methods, in the sense that they do not require prior knowledge about the smoothness of the function to estimate. We refer the reader to [BC05] for a bibliography over nonadaptive estimation with censored data. For estimating the density of i.i.d. survival times under random right-censorship, Li [Li08] proposes an estimator based on wavelet block thresholding, that improves on his wavelet hard thresholding estimator [Li07]. He provides the estimation rates for the $\mathbb{L}_2$-risk over a range of Besov classes composed of continuous functions, but that may be of inhomogeneous smoothness, otherwise said whose smoothness is measured in a $\mathbb{L}_p$-norm with $p < 2$. He also considers Besov-type classes that may contain discontinuous functions. Nevertheless, the threshold constant depends on the unknown survival and censoring distributions. Estimators based on model selection via penalization have also been proposed. Given a criterion and a collection of sets usually called models, such procedures consist in defining the collection of estimators minimizing that criterion over each model and then in choosing the best model by minimizing a penalized criterion. That is how Döhler and Rüschendorf [DR02] estimate the log-hazard function based on censored data in presence of covariates. They use a likelihood-based criterion and a penalty inspired from Vapnik's structural risk minimization approach. With models generated by splines, they obtain the best estimation rate within a logarithmic factor over Hölder smoothness classes, for a risk measured with a distance equivalent to the $\mathbb{L}_2$-distance. They also consider a more general type of models. Contrary to the previous references, the next ones adopt a nonasymptotic point of view. Using the model selection principle developed by Birgé and Massart since [BM97], Reynaud-Bouret [RB06] estimates the intensity of a counting process in the Aalen multiplicative intensity model, that includes hazard rate with censored data. In her work, the selection is performed via a least-squares type criterion based on a random norm and the risk is measured in a weighted $\mathbb{L}_2$-norm. That estimator reaches the minimax estimation rate over Hölderian smoothness classes when the models are spaces of piecewise constant functions on regular partitions. For hazard rate estimation with censored data, Brunel and Comte, inspired from [BM97], propose two estimators based on penalized least-squares type criteria, linked either with the Kaplan-Meier estimator of the survival function (*cf.* [BC08]) or with the Nelson-Aalen estimator of the cumulative hazard function (*cf.* [BC05]). Their results can be applied to various collections of models, provided these models are nested linear spaces, and they obtain the expected estimation rate for the $\mathbb{L}_2$-norm over Besov classes of homogeneous smoothness when the models are nested linear spaces generated by trigonometric or piecewise polynomials, splines or wavelets. They also provide an oracle-type inequality up to a logarithmic factor, without estimation rate, when considering a collection of models made up of piecewise polynomial functions over irregular partitions. Last, a new model selection procedure has recently been proposed by Baraud and Birgé [BB09] for estimating the intensity of a random measure, examples of which are hazard rate estimation with censored data, estimation of the mean of a random vector or density estimation. Given a collection of partitions, they define histogram-type estimators and select the best partition by performing tests between all pairs of histograms. For a risk

measured via a random distance akin to the Hellinger distance and an adequate collection of partitions, they obtain adaptivity results over Besov smoothness classes and for functions with bounded variations. Nevertheless, their procedure seems difficult to implement in practice since it requires a computational time which is quadratic in the size of the collection of partitions.

In this chapter, we propose a universal procedure for functional estimation based on possibly censored data, that enjoys better theoretical performance and lower computational complexity than the aforementioned estimators. We consider a model selection procedure based on a penalized least-squares type criterion generalizing the one used in [BC05], for collections of models made up of piecewise constant functions on some given partitions. We define a general framework and provide several examples illustrating its relevance. Density and hazard rate estimation with randomly right-censored data, for instance, are both embedded in such a framework. As we do not require the data to be identically distributed, we can also deal with regression. Besides, we believe that such a framework also allows to take into account many other estimation problems with partly observed data, such as those studied in [BC05; BC06; BCG09; BC09] for instance. The properties of our estimator are studied for an adequate penalty that does not require any prior knowledge about the function to estimate. Our estimator is proved to be almost as good, for the $\mathbb{L}_2$-risk, as the best histogram in the collection, up to a multiplicative factor that may depend on n. The assumptions required for such a result to hold are expressed in a clear and concise way, thanks to a recent concentration inequality due to Adamczak [Ada08]. Then, we present a special collection of partitions, where each partition is only composed of *dyadic* intervals, hence the name *dyadic histograms* given to the estimators minimizing our least-squares type criterion over each model. Among the aforementioned works, that collection of models has only been used in [BB09], and cannot be handled by [BC08; BC05]. The subsequent penalized estimator is proved to be almost as good as the best dyadic histogram, up to a multiplicative factor that does not depend on n anymore. At the same time, that collection of models has good approximation qualities with respect to functions that belong to some Besov spaces or have bounded variations, as shown in DeVore and Yu [DY90] and Birgé [Bir07]. Thus our estimator reaches in that case a rate which is expected to be the minimax one, up to some constant, over those classes. It should be noticed that the range of smoothness classes we consider is wider than those considered in the aforementioned references and that the estimation rate we obtain over Besov classes of inhomogeneous smoothness is better than in [Li08]. Last, in practice, the procedure based on dyadic histograms can be implemented with a computational complexity only linear in the sample size.

The chapter is organized as follows. We present in Section III.2 a general framework that allows to recover as special cases some classical estimation problems, such as regression and density estimation with uncensored data, and density and hazard rate estimation with randomly right-censored data. We also describe there the histogram selection procedure based on a penalized least-squares type criterion. In Section III.3, we state a histogram selection theorem general enough to be applied in the framework described in Section III.2 and with almost any collection of histograms. Section III.4 is devoted to the study of the penalized estimator based on the collection of dyadic histograms. Most proofs are postponed to Section III.5.

III.2 Estimation procedure

In this section, we present our general statistical framework and the histogram-type estimators. We then describe the histogram selection procedure based on a penalized least-squares criterion. We end with some classical estimation problems covered by the general framework, with censored or uncensored data.

III.2.1 General framework and notation

Our aim is to estimate a real valued function $s \in \mathbb{L}_2([0,1])$ on the basis of independent observations $Z_1 = (X_1, Y_1)$, ..., $Z_n = (X_n, Y_n)$, where $n \geq 3$ and for all $i = 1, \ldots, n$, $X_i \in [0,1]$ can be viewed as an observation time and Y_i takes values in a given space $\mathcal{Y}$. We denote by $(\Omega, \mathcal{A})$ the measurable space on which $Z_1, \ldots, Z_n$ are defined. As is customary, $\mathbb{P}_s$ stands for the underlying probability measure on $(\Omega, \mathcal{A})$ and $\mathbb{E}_s$ for the corresponding expectation. Moreover, we consider a probability measure μ on $[0,1]$ and we denote by $\mathbb{L}_2([0,1])$ the set of all measurable functions $t : [0,1] \to \mathbb{R}$ such that

$$\int_{[0,1]} t^2 \mathrm{d}\mu < \infty.$$

The scalar product and norm on $\mathbb{L}_2([0,1])$ are respectively denoted by $\langle \, . \, , \, . \, \rangle$ and $\| \, . \, \|$, while for a real-valued function f defined and bounded on $[0,1]$, we use the notation

$$\iota(f) = \operatorname*{essinf}_{x \in [0,1]} |f(x)| \quad \text{and} \quad \|f\|_\infty = \operatorname*{essup}_{x \in [0,1]} |f(x)|.$$

We assume that s belongs to a given subspace $\mathcal{S}$ of $\mathbb{L}_2([0,1])$ and that

$$\langle t, s \rangle = \mathbb{E}_s \left[\frac{1}{n} \sum_{i=1}^{n} w(Z_i) t(X_i) \right] \quad \text{for all } t \in \mathcal{S}, \tag{III.2.1}$$

where $w : [0,1] \times \mathcal{Y} \to \mathbb{R}$ is a measurable function that possibly depends on unknown parameters. Then, s minimizes

$$\|t\|^2 - \mathbb{E}_s \left[\frac{2}{n} \sum_{i=1}^{n} w(Z_i) t(X_i) \right] = \|s - t\|^2 - \|s\|^2$$

over $t \in \mathcal{S}$, so we consider estimators defined as the minimizer of the criterion

$$\gamma(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{w}(Z_i) t(X_i) \tag{III.2.2}$$

over a given space, for a given estimator $\hat{w}$ of w. If w is entirely known, we plainly set $\hat{w} = w$. We choose the space over which we minimize γ in such a way that the resulting estimator admits a histogram-like expression. More precisely, we consider a partition m of $[0,1]$ into D_m intervals, we denote by S_m the set of all real-valued functions that are piecewise constant on m and we set

$$\hat{s}_m = \operatorname*{argmin}_{t \in S_m} \gamma(t).$$

To see that $\hat{s}_m$ admits a histogram-like expression, let us denote by $\mathbb{1}_I$ the characteristic function of an interval I, and define

$$\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}(Z_i) \mathbb{1}_{[0,x]}(X_i), \ x \in [0,1].$$

Then, $\hat{P} : [0,1] \to \mathbb{R}$ is a cadlag step process that can jump only at points X_i, $i = 1, \ldots, n$. In view of (III.2.1), it can be considered as an estimator of

$$P(x) = \int_{[0,x]} s \mathrm{d}\mu, \ x \in [0,1], \tag{III.2.3}$$

and one has

$$\hat{s}_m = \sum_{I \in m} \left(\frac{1}{n\mu(I)} \sum_{i=1}^{n} \hat{w}(Z_i) \mathbb{1}_I(X_i) \right) \mathbb{1}_I = \sum_{I \in m} \left(\frac{1}{\mu(I)} \int_I \mathrm{d}\hat{P}(x) \right) \mathbb{1}_I.$$

Here, the Stieltjes integral $\int_I \mathrm{d}\hat{P}(x)$ merely denotes the increment of $\hat{P}$ over I. Thus, in the sequel, we call $\hat{s}_m$ a histogram estimator.

Now that we have defined the histogram estimator $\hat{s}_m$ based on a partition m, we would like to select an adequate partition m. For this task, we consider a family $\mathcal{M}^\star$ of partitions of $[0,1]$ into intervals and we select from the data the partition

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}^\star} \big\{ \gamma(\hat{s}_m) + \mathrm{pen}(m) \big\},$$

where $\mathrm{pen} : \mathcal{M}^\star \to \mathbb{R}^+$ is a so-called penalty function. We are concerned with the subsequent penalized estimator:

$$\tilde{s} = \hat{s}_{\hat{m}}.$$

Our aim is to propose, under fairly general assumptions on $\mathcal{M}^\star$, a penalty function such that the penalized estimator has good performances by comparison with the best histogram among the family $\{\hat{s}_m\}_{m \in \mathcal{M}^\star}$. We will propose in Section III.4 a collection $\mathcal{M}^\star$ such that $\tilde{s}$ is easy to implement and also has good performances with respect to any other estimator of s.

III.2.2 Examples

Let us describe some classical estimation problems that fall within the general framework described in Section III.2.1. In each problem considered below, μ is the Lebesgue measure on $[0,1]$, and we denote the length $\mu(I)$ of an interval $I \subset [0,1]$ by $|I|$. We have at hand a cadlag step estimator $\hat{P}$ of the function P defined by (III.2.3). This estimator can jump only at the n distinct points $X_1, \ldots, X_n$, so we define $\hat{w}(Z_i)/n$ as the height of the jump of $\hat{P}$ at point X_i, and we check that Condition (III.2.1) holds for a given w. Moreover, denoting by s_m the orthogonal projection of s on S_m, we show that there are numbers c and C such that

$$c\frac{D_m}{n} \leq \mathbb{E}_s\big[\|s_m - \hat{s}_m\|^2\big] \leq C\frac{D_m}{n}. \tag{III.2.4}$$

From Pythagoras' equality, this proves

$$\|s_m - s\|^2 + c\frac{D_m}{n} \leq \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big] \leq \|s_m - s\|^2 + C\frac{D_m}{n}.$$

Thus, in each problem under consideration, provided c can be chosen positive, the risk of the histogram $\hat{s}_m$ can decomposed into an approximation error and an estimation error of order D_m/n.

Regression. We observe a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ whose coordinates are independent and we aim at estimating the mean $r = \big(r_1, \ldots, r_n\big)$ of $\boldsymbol{Y}$. This amounts to estimate the function $s : [0,1] \to \mathbb{R}$ that is constant on $((i-1)/n, i/n]$ with $s(i/n) = r_i$ for every $i = 1, \ldots, n$ and $s(0) = s(1/n)$. With this notation one has

$$Y_i = s(i/n) + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

where the ε_i's are independent and centered variables. A cadlag step estimator of P is given by

$$\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbb{1}_{i/n \leq x},$$

so we set $Z_i = (i/n, Y_i)$ and $\hat{w}(x,y) = y$ for all $(x,y) \in [0,1] \times \mathbb{R}$. Moreover, we define $\mathcal{S}$ as the set of real-valued functions defined on $[0,1]$ that are constant on $[0,1/n]$ and on every $((i-1)/n, i/n]$, $i = 2, \ldots, n$, and we assume $S_m \subset \mathcal{S}$. Then, Condition (III.2.1) is satisfied with $w = \hat{w}$ and $\hat{s}_m$ is the usual least-squares estimator of s based on m:

$$\hat{s}_m = \operatorname*{argmin}_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - t(i/n))^2 \right\}.$$

Otherwise said, $\hat{s}_m$ is the regressogram based on m. We get

$$\mathbb{E}_s\big[\|s_m - \hat{s}_m\|^2\big] = \sum_{I \in m} \frac{1}{n^2 |I|} \sum_{i=1}^{n} \mathrm{Var}_s(\varepsilon_i) \mathbb{1}_I(i/n) .$$

Thus, Inequalities (III.2.4) hold provided $\max_i \mathrm{Var}_s(\varepsilon_i) \le C$ and $\min_i \mathrm{Var}_s(\varepsilon_i) \ge c$.

Density estimation with uncensored data. The observed random variables $X_1, \ldots, X_n$ are assumed to be independent and identically distributed, with values in $[0,1]$, and to admit a density $s \in \mathbb{L}_2([0,1])$ with respect to the Lebesgue measure. Considering the empirical distribution function $\hat{P}$, we set $\mathcal{S} = \mathbb{L}_2([0,1])$, $Z_i = (X_i, 1)$ and $\hat{w}(x,1) = 1$ for all $x \in [0,1]$ (note that the choice $Y_i = 1$ is arbitrary). Then, Condition (III.2.1) is plainly satisfied with $w = \hat{w}$. The criterion satisfying Condition (III.2.2) is the well-known contrast

$$\gamma(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t(X_i)$$

and, for every partition m, $\hat{s}_m$ is the usual histogram based on m. We get

$$\mathbb{E}_s\big[\|s_m - \hat{s}_m\|^2\big] = \sum_{I \in m} \frac{1}{n|I|} \mathrm{Var}_s(\mathbb{1}_I(X_i)) = \sum_{I \in m} \frac{1}{n|I|} \langle s, \mathbb{1}_I \rangle \langle s, \mathbb{1}_{[0,1] \setminus I} \rangle.$$

Thus, Inequalities (III.2.4) hold provided $\|s\|_\infty \le C$ and $c \le \iota(s)(1 - D_m^{-1})$.

Density estimation with censored data. Let $T_1, \ldots, T_n$ be independent random variables having the same density f with respect to the Lebesgue measure on $\mathbb{R}^+$. Let $C_1, \ldots, C_n$ be i.i.d. random variables in $\mathbb{R}^+$ independent of the T_i's. The T_i's represent for instance survival times and the C_i's censoring times. The observed random variables are

$$X_i = \min(T_i, C_i) \text{ and } Y_i = \mathbb{1}_{T_i \le C_i}, \text{ for } i = 1, \ldots, n. \tag{III.2.5}$$

Basing ourselves on these censored observations, we aim to estimate the restriction s of f to $[0,1]$, assuming s belongs to $\mathcal{S} = \mathbb{L}_2([0,1])$. Without loss of generality, possibly replacing C_i by $C_i \wedge 1$, we assume $C_i \in [0,1]$. Our choice of $\hat{w}$ relies on the Kaplan-Meier estimator $\hat{P}$ of the distribution function P. Let us denote by $Z_{(1)}, \ldots, Z_{(n)}$ the rearrangement of $Z_1, \ldots, Z_n$ such that $X_{(1)}, \ldots, X_{(n)}$ is the increasing rearrangement of $X_1, \ldots, X_n$. The Kaplan-Meier estimator of P is given by

$$\hat{P}(x) = 1 - \prod_{k | X_{(k)} \le x} \Big(1 - \frac{1}{n-k+1}\Big)^{Y_{(k)}}$$

when $x \ge X_{(1)}$ and equal to 0 otherwise. The Kaplan-Meier estimator $\tilde{G}$ of the distribution function G of the C_i's is defined in the same way, just replacing Y_i with $1 - Y_i$ for all i. In fact, $\hat{P}$ is a cadlag step estimator, with jump

$$\frac{Y_k}{n\big(1 - \tilde{G}^-(X_k)\big)}$$

in X_k, for $k = 1, \ldots, n$, where $\tilde{G}^-$ is the left-continuous version of $\tilde{G}$. Consequently, we assume $G^-(1) < 1$ where G^- denotes the left-continuous version of G and we set

$$w(x,y) = \frac{y}{1-G^-(x)} \text{ and } \hat{w}(x,y) = \frac{y}{1-\tilde{G}^-(x)},$$

for all $(x,y) \in [0,1] \times \{0,1\}$. Then, Condition (III.2.1) is fulfilled and one has:

Proposition 12 *Assume that* $\inf_{I \in m} |I| \geq 1/n$. *If* $G^-(1) < 1$, s *is bounded and* s *is bounded away from zero, then Inequalities (III.2.4) hold provided*

$$C \geq \frac{2\|s\|_\infty}{1-G^-(1)} + \frac{C(s,G)}{D_m},$$

and

$$c \leq \iota(s)\frac{D_m - 1}{2D_m} - \frac{C(s,G)}{D_m},$$

where $C(s,G)$ *only depends on* $\|s\|$, $\|s\|_\infty$, $\mathbb{P}_s(T_1 \leq 1)$ *and* $G^-(1)$.

Hazard rate estimation with censored data. Here again, the observed random variables are given by (III.2.5) where the T_i's are i.i.d. with density function f on $\mathbb{R}^+$ and the C_i's belong to $[0,1]$, are i.i.d. and independent of the T_i's. We denote by F and G respectively the distribution functions of T_i and C_i. We aim to estimate the restriction s to $[0,1]$ of the hazard rate

$$\lambda = \frac{f}{1-F},$$

assuming s belongs to $\mathcal{S} = \mathbb{L}_2([0,1])$ and $F(1) < 1$. Our choice of $\hat{w}$ relies on the Nelson-Aalen estimator $\hat{P}$ of the cumulative hazard rate P. In fact, $\hat{P}$ is a cadlag step estimator, with jump

$$\frac{Y_k}{n\big(1-\hat{H}^-(X_k)\big)}$$

in X_k, for $k = 1, \ldots, n$, where $\hat{H}^-$ is the left-continuous version of the empirical distribution function of the X_i's, that is, for all $x \in \mathbb{R}$,

$$\hat{H}^-(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{[0,x)}(X_j).$$

Consequently, we assume that the left-continuous version H^- of the distribution function H of X_i satisfies $H^-(1) < 1$ and we set

$$w(x,y) = \frac{y}{1-H^-(x)} \text{ and } \hat{w}(x,y) = \frac{y}{1-\hat{H}^-(x)},$$

for all $(x,y) \in [0,1] \times \{0,1\}$. Since

$$1 - H = (1-F)(1-G),$$

Condition (III.2.1) is then fulfilled and the criterion γ in (III.2.2) is similar to the one considered in [BC05]. Moreover, one has:

Proposition 13 *If* s *is bounded and bounded away from zero, then Inequalities (III.2.4) hold provided*

$$C \geq \frac{6\|s\|_\infty}{1-H^-(1)} + 2\|s\|^2\frac{n(H^-(1))^n}{D_m}$$

and

$$c \leq \frac{\iota(s)}{2}\big(1-(H^-(1))^n\big) - \|s\|^2\frac{n\,(H^-(1))^n}{D_m}.$$

III.3 A general histogram selection theorem

In this section, we are concerned with the choice of the penalty for a general collection of partitions $\mathcal{M}^\star$. We would like the resulting penalized estimator to be almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m\in\mathcal{M}^\star}$, and more precisely to satisfy the nonasymptotic oracle inequality

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \le C \inf_{m\in\mathcal{M}^\star} \mathbb{E}_s\big[\|s-\hat{s}_m\|^2\big] \tag{III.3.6}$$

for some positive constant C. In the histogram selection theorem stated hereafter (Theorem 6), we shall in fact provide a form of penalty yielding an inequality close to

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \le C \inf_{m\in\mathcal{M}^\star} \left\{\|s-s_m\|^2 + \frac{D_m}{n}\right\}. \tag{III.3.7}$$

However, as recalled in Section III.2.2, the risk of each histogram $\hat{s}_m$ can usually be decomposed into the approximation error $\|s-s_m\|^2$ and an estimation error roughly proportional to D_m/n. Therefore, an inequality like (III.3.7) is expected to lead to an inequality like (III.3.6). We state our assumptions and results in Section III.3.1, and we show in Section III.3.2 that those assumptions are fulfilled in all the examples introduced in Section III.2.2.

III.3.1 The oracle-type inequality

In order to prove our histogram selection theorem, we require only three assumptions: one about the structure of the collection of partitions, one about the existence of exponential moments of some random variables, and the last one about the quality of $\hat{w}$ as an estimator of w.

Assumption (P) *There exists a partition $m^\star$ such that any partition $m\in\mathcal{M}^\star$ is built on $m^\star$ and $\ell_\star := \inf_{I\in m^\star} n\mu(I) \ge 1$. Moreover, for all $I\in m^\star$, $\mathbb{1}_I \in \mathcal{S}$.*

Assumption (M) *There exist a nonnegative constant a and a positive constant v such that, for all $I\in m^\star$ and all $\lambda\in(-1/a,1/a)$,*

$$\ln \mathbb{E}_s\left[\exp\left(\lambda\sum_{i=1}^n \Big(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w(Z_i)\mathbb{1}_I(X_i)\big]\Big)\right)\right] \le \frac{nv\mu(I)\lambda^2}{2(1-a|\lambda|)}.$$

Assumption (W) *There exists a nonnegative real C_Δ, that may depend on w and s, such that for all $I\in m^\star$,*

$$\mathbb{E}_s\left[\left(\frac{1}{n\mu(I)}\sum_{i=1}^n (\hat{w}-w)(Z_i)\mathbb{1}_I(X_i)\right)^2\right] \le \frac{C_\Delta}{n}.$$

Remarks: *Assumption* **(P)** *ensures that Equality (III.2.1) holds for every $t\in S_m$, whatever $m\in\mathcal{M}^\star$. Assumption* **(M)** *has been chosen with a view to applying the results of this section to the collection of dyadic partitions described in Section III.4. But [Bar00] shows for instance in a regression framework that, for a collection of regular partitions, that assumption can be weakened.*

The properties collected in Proposition 14 below are fundamental in proving the histogram selection theorem to follow. Hereafter, we denote by M a positive number such that for all partitions m built on $m^\star$ and all $(a_I)_{I\in m} \in [0,1]^{D_m}$ such that $\sum_{I\in m} a_I^2 = 1$,

$$M \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_s\left(\sum_{I\in m} \frac{a_I}{\sqrt{\mu(I)}} w(Z_i)\mathbb{1}_I(X_i)\right). \tag{III.3.8}$$

It should be mentioned that in every example introduced in Section III.2.2, we are able to choose M independent of n. If the X_i's are possibly random, then the upper bounds we obtain below also depend on a real number b that may depend on n. It should be noticed that b can also be chosen independent of n in those examples (see Section III.3.2).

Proposition 14 *Assumptions* **(P)** *and* **(M)** *are supposed to be fulfilled. Let*

$$b_0 = \inf\Big\{\lambda > 0 \text{ s.t. } \mathbb{E}_s\Big[\exp\Big(\max_{1\leq i\leq n} |w(Z_i)|/\lambda\Big)\Big] \leq 2\Big\},$$

and, for all partitions m of $[0,1]$, let

$$\check{s}_m = \sum_{I\in m}\left(\frac{1}{n\mu(I)}\sum_{i=1}^{n} w(Z_i)\mathbb{1}_I(X_i)\right)\mathbb{1}_I.$$

Then, b_0 is finite and there exist nonnegative absolute constants k_1, k_2, k_3, k_4, k_5, a nonnegative real k_6 and some set $\Omega_\star \in \mathcal{A}$ such that, for all partitions m built on $m^\star$, and all $x > 0$,

$$\mathbb{P}_s\Big(n\|s_m - \check{s}_m\|^2\mathbb{1}_{\Omega_\star} \geq M\Big(k_1 D_m + k_2 x + 2k_3\sqrt{k_1 k_2 D_m x}\Big)\Big) \leq k_4\exp(-x) \tag{III.3.9}$$

and

$$\mathbb{P}_s(\Omega_\star^c) \leq \frac{k_5}{\ell_\star} n\exp\left(-k_6\ell_\star\right). \tag{III.3.10}$$

For instance, the values $k_1 = k_2 = k_4 = 4$, $k_3 = 1$, $k_5 = 2$ and

$$k_6 = C\frac{M^2}{b(bv + aM)},$$

where $b \geq b_0$ and C is a small enough absolute constant, suit. If the X_i's are non-random, then the values $k_1 = k_4 = 1$, $k_2 = 4$, $k_3 = 2\sqrt{2}$, $k_5 = 2$ and $k_6 = v/(2a^2)$ suit, and one can set $k_5 = 0$ if, moreover, $a = 0$.

Let us now state the histogram selection theorem. The penalty we suggest involves an estimator $\widehat{M}$ of M which, obviously, can be taken equal to M if M is entirely known. In that case, one can take $A = \Omega$, $C_A = 0$ and $k_7 = k_8 = 1$ in Theorem 6.

Theorem 6 *Assumptions* **(P)**, **(M)** *and* **(W)** *are supposed to be fulfilled and the notation are those of Proposition 14. We also assume that there exist an estimator $\widehat{M}$ of M, a set $A \in \mathcal{A}$, a positive constant k_7 and a nonnegative real C_A that may depend on w and s such that*

$$M \leq k_7\widehat{M} \text{ on } A \quad \text{and} \quad \mathbb{P}_s\big(A^c\big) \leq \frac{C_A}{n^2}. \tag{III.3.11}$$

Let $\{L_m\}_{m\in\mathcal{M}^\star}$ be a family of nonnegative reals, that may depend on n, such that

$$\Sigma := \sum_{m\in\mathcal{M}^\star} \exp(-L_m D_m) \leq 1. \tag{III.3.12}$$

i) If pen *satisfies, for all* $m \in \mathcal{M}^\star$,

$$\mathrm{pen}(m) \geq K(k_7, L_m)\frac{\widehat{M}D_m}{n},$$

where

$$K(k_7, L_m) = 6k_7(1+k_3)\big(k_1 + k_2 L_m\big) + 2k_3\sqrt{k_1 k_2 L_m},$$

then

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \leq \inf_{m\in\mathcal{M}^\star}\Big\{C_1\|s-s_m\|^2 + C_2\mathbb{E}_s\big[\mathrm{pen}(m)\mathbb{1}_A\big]\Big\} + \frac{C_3 M + C_4}{n},$$

where C_1, C_2, C_3 *are positive constants and*

$$C_4 = \Big(C(a,v) + \|s\|^2\Big)\sqrt{\frac{k_5}{\ell_\star}n^3\exp(-k_6\ell_\star) + C_A} + 12C_\Delta.$$

Here, $C(a,v)$ *only depends on* a *and* v*, and increases with* a *and* v*.*

ii) Let us also assume that, for some positive constant k_8*,*

$$\widehat{M} \leq k_8 M \text{ on } A. \tag{III.3.13}$$

If pen *satisfies, for all* $m \in \mathcal{M}^\star$,

$$\mathrm{pen}(m) = \Big(c_0 + c_1\sqrt{L_m} + c_2 L_m\Big)\frac{\widehat{M}D_m}{n}, \tag{III.3.14}$$

where c_0*,* c_1*,* c_2 *are nonnegative and large enough constants, then*

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \leq C_5\Big(1 + \max_{m\in\mathcal{M}^\star} L_m\Big)\inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2 + \frac{MD_m}{n}\right\} + \frac{C_4}{n}, \tag{III.3.15}$$

where C_5 *is a positive real that only depends on* k_8*,* c_0*,* c_1*,* c_2*.*

Inequality (III.3.15) shows that, provided the penalty is well-chosen, and M and C_4 are bounded independently of n, $\tilde{s}$ satisfies an oracle-type inequality akin to (III.3.7), up to a multiplicative factor that may depend on n through the weights $\{L_m\}_{m\in\mathcal{M}^\star}$, and may also depend on w and s.

III.3.2 Examples (continued)

Let us now come back to the examples presented in Section III.2.2. Hereafter, we assume that there exists a partition $m^\star$ such that any partition $m \in \mathcal{M}^\star$ is built on $m^\star$ and we consider a penalty of the form (III.3.14), where c_0, c_1, c_2 are nonnegative constants, the weights L_m's satisfy (III.3.12), and $\widehat{M}$ remains to be chosen. In all the examples under consideration but regression, we will use the following proposition in order to check that (III.3.15) holds.

Proposition 15 *Assume that*

- $Z_1, \ldots, Z_n$ *are independent and identically distributed;*
- w *is nonnegative, bounded,* $C_w := \inf_{z\in[0,1]\times\mathcal{Y}}\{w(z) \text{ s.t. } w(z)\neq 0\} > 0$*, and* $\hat{w}$ *vanishes whenever* w *does;*
- s *is lower bounded by a positive real and upper bounded;*

- *Assumption* **(P)** *is satisfied and, for all* $I \in m^\star$, $n|I| \geq \ln^2(n)$;
- *the estimator* $\hat{w}$ *is such that*

$$\mathbb{E}_s^{1/2}[\Delta^4] \leq \frac{C'_\Delta}{n} \quad \textit{where} \quad \Delta = \max_{1\leq i\leq n} \frac{|\hat{w}(Z_i) - w(Z_i)|}{w(Z_i)}, \tag{III.3.16}$$

with the convention $0/0 = 0$.

Setting

$$M = \max_{I\in m^\star} \frac{1}{|I|}\mathbb{E}_s\big[w^2(Z_i)\mathbb{1}_I(X_i)\big] \quad \textit{and} \quad \widehat{M} = \max_{I\in m^\star} \frac{1}{n|I|}\sum_{i=1}^n \hat{w}^2(Z_i)\mathbb{1}_I(X_i), \tag{III.3.17}$$

and choosing a penalty of the form (III.3.14) *with large enough nonnegative constants* c_0, c_1, c_2, *we get (III.3.15) where* C_5 *only depends on* c_0, c_1, c_2 *and* C_4 *only depends on* C_w, $\|w\|_\infty$, $\iota(s)$, $\|s\|_\infty$, C'_Δ, *is a decreasing function of* C_w *and* $\iota(s)$ *and an increasing function of* $\|w\|_\infty$, $\|s\|_\infty$ *and* C'_Δ. *Moreover,* $M \leq \|w\|_\infty\|s\|_\infty$.

In the sequel, we denote by $C(\theta)$ some positive real whose value may change from one line to another and that only depends on the parameter θ.

Regression. In the regression setting, we provide an oracle-type inequality under a condition on the exponential moments of the errors that includes subgaussian errors (case $a = 0$ in Condition (III.3.18) below). More precisely, we assume that there exist a positive real v and a nonnegative real a such that, for all $1 \leq i \leq n$ and all $\lambda \in (-1/a, 1/a)$,

$$\ln \mathbb{E}_s\big[\exp(\lambda\varepsilon_i)\big] \leq \frac{v\lambda^2}{2(1 - a|\lambda|)}. \tag{III.3.18}$$

Moreover, we assume that $\inf_{I\in m^\star} n|I| \geq 1$ when $a = 0$ and $\inf_{I\in m^\star} n|I| \geq \ln^2(n)$ otherwise, and that $\mathbb{1}_I \in \mathcal{S}$ for all $I \in m^\star$. Thus, Assumptions **(P)** and **(W)** are satisfied with $C_\Delta = 0$. The Lebesgue measure $|I|$ of any interval $I \in m^\star$ coincides here with $\sum_{i=1}^n \mathbb{1}_I(i/n)/n$, so Assumption **(M)** follows from Condition (III.3.18) and the independence of $\varepsilon_1, \ldots, \varepsilon_n$.

Since the X_i's are non-random and (III.3.18) holds, we have for all partitions m built on $m^\star$ and all $I \in m$,

$$\mathrm{Var}_s\left(\sum_{i=1}^n w(Z_i)\mathbb{1}_I(X_i)\right) \leq n\sigma^2|I|,$$

where

$$\sigma^2 := \max_{1\leq i\leq n} \mathrm{Var}_s(\varepsilon_i) \leq v.$$

Hence, Condition (III.3.8) is fulfilled with $M = \sigma^2$, and, if σ^2 is known, then we can set $\widehat{M} = \sigma^2$ and choose pen of the form

$$\mathrm{pen}(m) = \Big(c_0 + c_1\sqrt{L_m} + c_2 L_m\Big)\frac{\sigma^2 D_m}{n}.$$

Consequently, Theorem 6 with the values of k_5, k_6 provided in Proposition 14 and large enough nonnegative constants c_0, c_1, c_2 yields

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0, c_1, c_2)\Big(1 + \max_{m\in\mathcal{M}^\star} L_m\Big)\inf_{m\in\mathcal{M}^\star}\left\{\|s - s_m\|^2 + \frac{\sigma^2 D_m}{n}\right\} + \frac{C'}{n}\Big(1 + \|s\|^2\Big), \tag{III.3.19}$$

where $C' = 0$ if $a = 0$ and C' only depends on v and a otherwise. In particular, the lower bound for the risk of each histogram given in Section III.2.2 yields

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C''\bigg(1 + \max_{m \in \mathcal{M}^\star} L_m\bigg) \inf_{m \in \mathcal{M}^\star} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big].$$

When $a = 0$, C'' only depends on c_0, c_1, c_2, $\max_i \mathrm{Var}_s(\varepsilon_i)$ and $\min_i \mathrm{Var}_s(\varepsilon_i)$. When a is non null, C'' also depends on a, v and $\|s\|$. It should be noticed that we thus improve on the result proved by Sauvé [Sau09]. As a matter of fact, when the $\varepsilon_i' s$ are not subgaussian, the penalty suggested by Theorem 1 in [Sau09] to obtain an inequality similar to (III.3.19) also depends on $\|s\|_\infty$.

Density estimation with uncensored data. In this setting, $w = \hat{w} = 1$ and the assumptions of Proposition 15 are fulfilled with $C_w = 1$ and $C'_\Delta = 0$, provided $\iota(s) > 0$, $\|s\|_\infty < \infty$ and $n|I| \geq \ln^2(n)$ for all $I \in m^\star$. Defining M and $\widehat{M}$ as in (III.3.17), *i.e.*

$$M = \max_{I \in m^\star} \frac{\langle s, \mathbb{1}_I \rangle}{|I|} \quad \text{and} \quad \widehat{M} = \max_{I \in m^\star} \frac{1}{n|I|} \sum_{i=1}^n \mathbb{1}_I(X_i),$$

we get with a penalty of the form (III.3.14) with large enough nonnegative constants c_0, c_1, c_2

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0, c_1, c_2)\bigg(1 + \max_{m \in \mathcal{M}^\star} L_m\bigg) \inf_{m \in \mathcal{M}^\star} \bigg\{\|s - s_m\|^2 + \frac{\|s\|_\infty D_m}{n}\bigg\} + \frac{C(s)}{n},$$

where $C(s)$ only depends on $\iota(s)$ and $\|s\|_\infty$, is a decreasing function of $\iota(s)$ and an increasing function of $\|s\|_\infty$. In particular, the lower bound for the risk of each histogram given in Section III.2.2 yields

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0, c_1, c_2, \iota(s), \|s\|_\infty)\bigg(1 + \max_{m \in \mathcal{M}^\star} L_m\bigg) \inf_{m \in \mathcal{M}^\star \setminus \{m_0\}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big],$$

where m_0 is the trivial partition of $[0, 1]$.

Density estimation with censored data. Assume that $G^-(1) < 1$, $\int_0^1 s < 1$, s is bounded and bounded away from zero, and $\inf_{I \in m^\star} n|I| \geq \ln^2(n)$. Here, w satisfies

$$C_w = 1 \quad \text{and} \quad \|w\|_\infty = \frac{1}{1 - G^-(1)},$$

and, according to the proof of Proposition 12, Assumption (III.3.16) on Δ is satisfied with

$$C'_\Delta = \frac{C}{(1 - H^-(1))^6},$$

where C is a positive constant and $1 - H^-(1) = (1 - \int_0^1 s)(1 - G^-(1))$. Let us define M and $\widehat{M}$ as in (III.3.17), *i.e.*

$$M = \max_{I \in m^\star} \frac{1}{|I|} \Big\langle \frac{s}{1 - G^-}, \mathbb{1}_I \Big\rangle \quad \text{and} \quad \widehat{M} = \max_{I \in m^\star} \frac{1}{n|I|} \sum_{i=1}^n \frac{Y_i}{(1 - \tilde{G}^-(X_i))^2} \mathbb{1}_I(X_i),$$

and still choose pen of the form (III.3.14) with large enough nonnegative constants c_0, c_1, c_2. We then deduce from Proposition 15 that

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0, c_1, c_2)\bigg(1 + \max_{m \in \mathcal{M}^\star} L_m\bigg) \inf_{m \in \mathcal{M}^\star} \bigg\{\|s - s_m\|^2 + \frac{\|s\|_\infty D_m}{(1 - G^-(1))n}\bigg\} + \frac{C(s, G)}{n},$$

where $C(s,G)$ only depends on $\int_0^1 s$, $\iota(s)$, $\|s\|_\infty$ and $G^-(1)$, $C(s,G)$ is an increasing function of $\int_0^1 s$, $\|s\|_\infty$, $G^-(1)$, and a decreasing function of $\iota(s)$.

Hazard rate estimation with censored data. Assume that $G^-(1) < 1$, $F(1) < 1$, s is bounded and bounded away from zero, and $\inf_{I\in m^\star} n|I| \geq \ln^2(n)$. Here, w satisfies

$$C_w = 1 \quad \text{and} \quad \|w\|_\infty = \frac{1}{1-H^-(1)}.$$

Using Dvoretzky-Kiefer-Wolfovitz inequality [DKW56; Mas90], we obtain as in the proof of Proposition 12 that Assumption (III.3.16) on Δ is satisfied with

$$C'_\Delta = \frac{C}{(1-H^-(1))^6},$$

where C is a positive constant. In that case, we can also write M and $\widehat{M}$ defined in (III.3.17) as

$$M = \max_{I\in m^\star} \frac{1}{|I|}\left\langle \frac{s}{1-H^-}, \mathbb{1}_I \right\rangle \quad \text{and} \quad \widehat{M} = \max_{I\in m^\star} \frac{1}{n|I|} \sum_{i=1}^n \frac{Y_i}{(1-\hat{H}^-(X_i))^2} \mathbb{1}_I(X_i).$$

Choosing anew pen of the form (III.3.14) with large enough nonnegative constants c_0, c_1, c_2, we deduce from Proposition 15 that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(c_0,c_1,c_2)\left(1+\max_{m\in\mathcal{M}^\star} L_m\right) \inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2 + \frac{\|s\|_\infty D_m}{(1-H^-(1))n}\right\} + \frac{C(s,F,G)}{n}.$$

where $C(s,F,G)$ only depends on $\iota(s), \|s\|_\infty, F(1), G^-(1)$, is an increasing function of $\|s\|_\infty$, $F(1)$, $G^-(1)$, and a decreasing function of $\iota(s)$. Besides, given the lower bound for the risk of each histogram given in Section III.2.2, we immediately obtain, when n is large enough,

$$\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(c_0,c_1,c_2,s,F,G)\left(1+\max_{m\in\mathcal{M}^\star} L_m\right) \inf_{m\in\mathcal{M}^\star} \mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right].$$

III.4 Dyadic histogram selection

From now on, we turn our attention to the collection of partitions of $[0,1]$ into dyadic intervals. In that case, the penalized estimator admits a simple description that allows to implement it with a low computational complexity. Besides, we are able not only to provide an oracle-type inequality but also to evaluate the estimation rate of the estimator over various classes of smoothness. We illustrate the qualities of our estimator on the examples introduced in Section III.2.2.

III.4.1 Presentation

In the rest of the chapter, μ is the Lebesgue measure on $[0,1]$ and, as in Section III.2.2, the length $\mu(I)$ of a subinterval I of $[0,1]$ is denoted by $|I|$. We consider the collection $\mathcal{M}^\star$ made up of all the partitions of $[0,1]$ into *dyadic* intervals whose measure is at least $2^{-J_\star}$, where $J_\star$ is an integer such that $2^{J_\star}$ is at most of order n. Thus, each partition in $\mathcal{M}^\star$ only contains intervals of the form $\left[0, 2^{-j}\right]$ or $\left(k2^{-j}, (k+1)2^{-j}\right]$, where $j \in \{0,\dots,J_\star\}$ may change from one interval to another, and $k \in \{1,\dots,2^j-1\}$. Besides, each partition in $\mathcal{M}^\star$ is built on the regular partition of $[0,1]$ into $2^{J_\star}$ intervals, that we will denote by $m^\star$. We call *dyadic* histogram any histogram $\hat{s}_m$ built on some partition $m \in \mathcal{M}^\star$.

Let K be some nonnegative constant. When choosing the penalty defined on $\mathcal{M}^\star$ by

$$\mathrm{pen}(m) = K\frac{\widehat{M}D_m}{n},$$

whose form we will justify in the next section, the penalized estimator can be determined as follows. First, straightforward computations lead to

$$\gamma(\hat{s}_m) = -\|\hat{s}_m\|^2, \text{ for all } m \in \mathcal{M}^\star.$$

Therefore, the partition to select is

$$\begin{aligned}\hat{m} &= \operatorname*{argmin}_{m\in\mathcal{M}^\star} \{\gamma(\hat{s}_m) + \mathrm{pen}(m)\} \\ &= \operatorname*{argmin}_{m\in\mathcal{M}^\star} \sum_{I\in m} \mathcal{L}(I),\end{aligned}$$

where, for all $m \in \mathcal{M}^\star$ and all $I \in m$,

$$\mathcal{L}(I) = -\frac{1}{n|I|}\left(\sum_{i=1}^{n} \hat{w}(Z_i)\mathbb{1}_I(X_i)\right)^2 + K\widehat{M}.$$

That allows to turn the determination of $\hat{m}$ into a shortest-path problem, that can be solved with the algorithm described for instance in [CLRS01], Section 24.2., or in [Aka09], Section 3.3. Since all the partitions in $\mathcal{M}^\star$ are built from $2^{J_\star} - 1$ intervals of $[0, 1]$, that algorithm only requires here $\mathcal{O}(n)$ computations. Of course, an adequate value of the penalty constant K remains to be chosen. One way consists in using the heuristic method explained in [BM07] (Section 4), which has already been successfully implemented in various frameworks.

III.4.2 Performance

We adopt in both theorems below the notation introduced in Section III.3, and in particular in Proposition 14 and Theorem 6.

Theorem 7 *Assumptions* **(P)**, **(M)** *and* **(W)** *are supposed to be fulfilled. Let* $\widehat{M}$ *be an estimator of* M *satisfying Conditions* (III.3.11) *and* (III.3.13). *If the penalty satisfies, for all* $m \in \mathcal{M}^\star$,

$$\mathrm{pen}(m) = K\frac{\widehat{M}D_m}{n},$$

where K *is a nonnegative and large enough constant, then*

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \le C_6 \inf_{m\in\mathcal{M}^\star}\left\{\|s - s_m\|^2 + \frac{MD_m}{n}\right\} + \frac{C_4}{n},$$

where C_4 *is defined in Theorem 6 and* C_6 *is a positive constant that only depends on* k_8 *and* K.

Proof. For all $D \in \{1, \dots, 2^{J_\star}\}$, let $\mathcal{M}^\star_D = \{m \in \mathcal{M}^\star \text{ s.t. } D_m = D\}$ and let $L_m = \ln 8$ for all $m \in \mathcal{M}^\star$. We have

$$\Sigma = \sum_{D=1}^{2^{J_\star}} \exp\left(-D\left(\ln 8 - \frac{\ln(|\mathcal{M}^\star_D|)}{D}\right)\right).$$

In order to evaluate the cardinal of $\mathcal{M}^\star_D$, let us describe the collection $\mathcal{M}^\star$ with the help of a complete binary tree. Let $\mathcal{T}$ be the binary tree with root $(0, 0)$ such that

- for all $j \in \{1, \dots, J_\star\}$, the nodes at level j are indexed by the elements of the set $\Lambda(j) = \{(j,k), k = 0, \dots, 2^j - 1\}$;
- for all $j \in \{1, \dots, J_\star\}$ and all $k \in \{0, \dots, 2^j - 1\}$, the left branch that stems from node (j,k) leads to node $(j+1, 2k)$, and the right one, to node $(j+1, 2k+1)$.

The node set of $\mathcal{T}$ is $\Lambda_{J_\star} = \cup_{j=0}^{J_\star} \Lambda(j)$, where $\Lambda(0) = \{(0,0)\}$. The dyadic intervals of $[0,1]$ whose measure is at least $2^{-J_\star}$ are the sets

$$I_{(j,0)} = \left[0, 2^{-j}\right] \text{ and } I_{(j,k)} = \left(k2^{-j}, (k+1)2^{-j}\right] \text{ for } k \geq 1 \tag{III.4.20}$$

indexed by the elements of $\Lambda_{J_\star}$. The leaves of $\mathcal{T}$ correspond with the intervals in $m^\star$. More generally, there is a one-to-one correspondence between the partitions in $\mathcal{M}^\star$ and the subsets of $\Lambda_{J_\star}$ composed of the leaves of any complete binary tree resulting from an elagation of $\mathcal{T}$. So the cardinal of $\mathcal{M}_D^\star$ is equal to the number of complete binary trees with D leaves resulting from an elagation of $\mathcal{T}$. It is given by the Catalan number $D^{-1}\binom{2(D-1)}{D-1}$, and thus upper-bounded by 4^D, so $\Sigma \leq 1$. Theorem 7 then follows from Theorem 6. ■

Let us assume that M and C_4 can be bounded independently of n, which is indeed the case in all the examples introduced in Section III.2.2, as proved in Section III.3.2. Then, we obtain here an oracle inequality akin to (III.3.7), up to a multiplicative factor that may depend on s and several nuisance parameters, but does *not* depend on n.

We can also evaluate here the approximation qualities of $\bigcup_{m \in \mathcal{M}_D^\star} S_m$ with respect to functions with bounded variations or with Besov-type smoothness, and then deduce from Theorem 7 the estimation rate of the penalized estimator for such functions. Let us define those smoothness classes. For $\alpha > 0$, $p \geq 1$ and $s : [0,1] \to \mathbb{R}$, let $V_\alpha(s)$ and $|s|_{B^\alpha_{p,\infty}}$ be the α-variations and the Besov semi-norm of s defined by

$$V_\alpha(s) = \sup_{i \geq 1} \sup_{0 \leq x_0 < \dots < x_i \leq 1} \left(\sum_{j=1}^{i} |s(x_j) - s(x_{j-1})|^{1/\alpha} \right)^\alpha$$

and

$$|s|_{B^\alpha_{p,\infty}} = \sup_{x>0} x^{-\alpha} \sup_{0 \leq h \leq x} \left[\int_0^{1-h} |s(y+h) - s(y)|^p \, \mathrm{d}y \right]^{1/p}.$$

For $0 < \alpha \leq 1$, and positive R, we consider the sets

$$\mathcal{V}(\alpha, R) = \{s : [0,1] \to \mathbb{R} \text{ s.t. } V_\alpha(s) \leq R\}$$

and for $p \geq 1$, $\max\{1/p - 1/2, 0\} < \alpha < 1$ and positive R,

$$\mathcal{B}(\alpha, p, R) = \{s : [0,1] \to \mathbb{R} \text{ s.t. } |s|_{B^\alpha_{p,\infty}} \leq R\}.$$

Let us also introduce some parameter sets of interest for studying the estimation rates. For $\delta \in \{0,1\}$, let

$$\Theta\mathcal{V}_1(\delta) = \{(\alpha, R); 0 < \alpha \leq 1/2, n^{-1/2} \leq R \leq (\ln(n))^{-\delta(1+2\alpha)} n^\alpha\},$$

$$\Theta\mathcal{V}_2(\delta) = \{(\alpha, R); 1/2 < \alpha \leq 1, n^{-1/2} \leq R \leq (\ln(n))^{-\delta(1+1/(2\alpha))} n^{1/(4\alpha)}\},$$

$$\Theta\mathcal{B}_1(\delta) = \{(\alpha, p, R); p \geq 2, 0 < \alpha < 1, n^{-1/2} \leq R \leq (\ln(n))^{-\delta(1+2\alpha)} n^\alpha\},$$

$$\Theta\mathcal{B}_2(\delta) = \{(\alpha, p, R); 1 \leq p < 2, \alpha_p \leq \alpha < 1, n^{-1/2} \leq R \leq (\ln(n))^{-\delta q(\alpha,p)} n^{(q(\alpha,p)-1)/2}\},$$

$$\Theta\mathcal{B}_3(\delta) = \{(\alpha, p, R); 1 \leq p < 2, 1/p - 1/2 < \alpha < \alpha_p, (\ln(n))^{-\delta q(\alpha,p)} n^{(q(\alpha,p)-1)/2} \leq R \leq (\ln(n))^{-\delta(1+2\alpha)} n^\alpha\},$$

where, for $1 \leq p < 2$ and $1/p - 1/2 < \alpha < 1$,

$$q(\alpha, p) = \left(\alpha + \frac{1}{2} - \frac{1}{p}\right)\frac{1+2\alpha}{\alpha} \text{ and } \alpha_p = \frac{1}{2}\left(\frac{1}{p} - \frac{1}{2}\right)\left(1 + \sqrt{\frac{2+3p}{2-p}}\right).$$

Let us underline that, for $1 \leq p < 2$ and $1/p - 1/2 < \alpha < 1$, we have $\alpha_p \in (1/p - 1/2, 1/p)$, $q(\alpha, p) \in (0, 1 + 2\alpha)$, and that $q(\alpha, p) > 1$ if and only if $\alpha > \alpha_p$. Notice that $\mathcal{V}(\alpha, R)$ and $\mathcal{B}(\alpha, p, R)$ with $\alpha \leq 1/p$ contain discontinuous functions. Besides, $\mathcal{V}(\alpha, R)$ for $(\alpha, R) \in \Theta\mathcal{V}_2(\delta)$, and $\mathcal{B}(\alpha, p, R)$ with $(\alpha, p, R) \in \Theta\mathcal{B}_2(\delta) \cup \Theta\mathcal{B}_3(\delta)$ are composed of functions whose smoothness is measured in a $\mathbb{L}_p$-norm, with $p < 2$. It is known that such functions cannot be well approximated in the $\mathbb{L}_2$-norm when considering the collection of nested models described in [BC08]. In particular, we could not reach the adequate rate over such classes by considering only regular partitions. We provide in Theorem 8 below the estimation rate over the previous classes.

Theorem 8 *Assume that M and C_4 are bounded by positive reals that do not depend on n. Let $\delta = 0$ when $k_5 = 0$ and $\delta = 1$ otherwise, and choose $J_\star$ such that*

$$cn \ln^{-2\delta}(n) \leq 2^{J_\star} \leq Cn \ln^{-2\delta}(n)$$

for some absolute positive constants c and C. Under the assumptions of Theorem 7,

- *if $s \in \mathcal{V}(\alpha, R)$ with $(\alpha, R) \in \Theta\mathcal{V}_1(\delta) \cup \Theta\mathcal{V}_2(\delta)$, then*

$$\mathbb{E}_s\left[\|s - \tilde{s}\|^2\right] \leq C_7 C(\alpha)(Rn^{-\alpha})^{2/(1+2\alpha)};$$

- *if $s \in \mathcal{B}(\alpha, p, R)$ with $(\alpha, p, R) \in \Theta\mathcal{B}_1(\delta) \cup \Theta\mathcal{B}_2(\delta)$, then*

$$\mathbb{E}_s\left[\|s - \tilde{s}\|^2\right] \leq C_7 C(\alpha, p)(Rn^{-\alpha})^{2/(1+2\alpha)};$$

- *if $s \in \mathcal{B}(\alpha, p, R)$ with $(\alpha, p, R) \in \Theta\mathcal{B}_3(\delta)$, then*

$$\mathbb{E}_s\left[\|s - \tilde{s}\|^2\right] \leq C_7 C(\alpha, p) R^2 (\ln^{2\delta}(n) n^{-1})^{2(\alpha+1/2-1/p)},$$

where $C_7 = C_6(1 + M + C_4)$ and $C(\alpha)$ (resp. $C(\alpha, p)$) is a positive real that only depends on α (resp. α and p).

Let us underline that, on classes akin to $\mathcal{B}(\alpha, p, R)$ with $1 \leq p < 2$ and $\alpha > 1/p$, [Li08] only reaches the estimation rate $n^{-2\alpha/(1+2\alpha)}$ within a logarithmic factor. Besides, when considering the collection of *all* the partitions built on $m^\star$, one only obtains the aforementionned rates within a logarithmic factor (*cf.* [BC05], Section 4.4). In the case of Besov balls with $1 \leq p < 2$, we observe a change in the rate of approximation that results in a change in the rate of estimation, already reported by Birgé [Bir04] in a Gaussian regression framework.

III.4.3 Examples (end)

Let us end with the examples presented in Section III.2.2. In each case, we work under the same assumptions as in Section III.3.2, with the same values for M and $\widehat{M}$. We just replace all the weights L_m, $m \in \mathcal{M}^\star$, with the same constant, obtaining thus a penalty of the form $\text{pen}(m) = K\widehat{M}D_m/n$. Since we have already proved that M and C_4 are bounded independently of n, there is indeed not much left to be done, except for choosing the number $2^{J_\star}$ of intervals in the regular dyadic partition $m^\star$. Then, the reader will easily write the resulting oracle-type inequalities and upper-bounds for the estimation rate of $\tilde{s}$.

Regression. We assume that $n = 2^N$ where $N \geq 4$. We choose $J_\star$ such that $2^{J_\star} = n2^{-N_0}$ where $N_0 = \max\{k \in \mathbb{N}\ s.t.\ 2^k \leq N^2\}$. Then, $S_{m^\star} \subset \mathcal{S}$ and $\inf_{I \in m^\star} n|I| \geq 2^{N_0}$, so that Assumption **(P)** is satisfied. Besides, $2^{J_\star}$ is of order $n/\ln^2(n)$ so that Theorem 8 applies with $\delta = 1$ and C_7 that only depends on $K, a, v, \sigma^2, \|s\|$ and increases with $\|s\|$. If we know that $a = 0$ (case when the errors are subgaussian), then we can choose $2^{J_\star} = n$ and Theorem 8 applies with $\delta = 0$ and C_7 that only depends on K and σ^2.

Density and hazard rate estimation with censored or uncensored data. In the other three examples, the following proposition, which is a straightforward consequence of Proposition 15, ensures that Theorem 8 applies.

Proposition 16 *Let $J_\star = \max\{j \in \mathbb{N}\ s.t.\ 2^j \leq n/\ln^2(n)\}$. Under the assumptions of Proposition 15, the resulting penalized estimator $\tilde{s}$ satisfies Theorem 8 with $\delta = 1$ and C_7 that only depends on $K, C_w, \|w\|_\infty, \iota(s), \|s\|_\infty, C'_\Delta$, is a decreasing function of C_w and $\iota(s)$ and an increasing function of $\|w\|_\infty, \|s\|_\infty, C'_\Delta$.*

III.5 Proofs

All along the section, C denotes an absolute constant and for a given parameter θ, $C(\theta)$ denotes a positive real number that only depends on θ. The values of C and $C(\theta)$ are allowed to change from one line to another.

III.5.1 A useful lemma

Lemma 2 *Let $X_1, \ldots, X_n$ be random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $[0,1]$, $\hat{h} : [0,1] \to [0,1]$ a cadlag process on Ω, and $h : [0,1] \to [0,1]$ a cadlag non-decreasing function. Denote by $\hat{h}^-$ and h^- respectively the left-continuous versions of $\hat{h}$ and h. Assume $1 - \hat{h}^-(X_i) \geq 1/n$ for all $i = 1, \ldots, n$, $h^-(1) < 1$, and there are positive numbers K and κ such that*

$$\mathbb{P}\left(\|\hat{h} - h\|_\infty > x\right) \leq K \exp(-\kappa n x^2) \tag{III.5.21}$$

for all $x \geq 0$. If Δ is a random variable such that

$$0 \leq \Delta \leq \|\hat{h} - h\|_\infty \max_{1 \leq i \leq n} \frac{1}{1 - \hat{h}^-(X_i)}, \tag{III.5.22}$$

then, for all $q \geq 1$, there exists $C(q)$ that only depends on q such that

$$\mathbb{E}\left[\Delta^{2q}\right] \leq C(q) \frac{K}{n^q} \left(\frac{1}{\kappa^{3q}(1 - h^-(1))^{6q}} + \frac{1}{\kappa^q (1 - h^-(1))^{2q}} \right).$$

Proof. Fix $q \geq 1$. We proceed in the same way as Brunel and Comte in [BC05] (proof of Proposition 4.1) and introduce the subset of Ω

$$\Omega_h = \left\{2(\hat{h}^- - h^-) < 1 - h^-(1)\right\}.$$

On Ω_h, $2(1 - \hat{h}^-) \geq 1 - h^-(1)$, so

$$\Delta \mathbb{1}_{\Omega_h} \leq \frac{2}{1 - h^-(1)} \|\hat{h} - h\|_\infty.$$

From (III.5.21), there exists $C(q)$ such that

$$\mathbb{E}\big[\|\hat{h}-h\|_\infty^{2q}\big] = \int_0^\infty 2qx^{2q-1}\mathbb{P}(\|\hat{h}-h\|_\infty > x)\,\mathrm{d}x \le C(q)\frac{K}{(\kappa n)^q},$$

hence, there exists $C(q)$ such that

$$\mathbb{E}\big[\Delta^{2q}\mathbb{1}_{\Omega_h}\big] \le C(q)\frac{K}{(\kappa n)^q(1-h^-(1))^{2q}}.$$

In order to bound $\mathbb{E}\big[\Delta^{2q}\mathbb{1}_{\Omega_h^c}\big]$, we use that $1-\hat{h}^-(X_i) \ge 1/n$ for all i, $\|\hat{h}-h\|_\infty \le 2$ and $\Omega_h^c \subset \big\{2\|\hat{h}-h\|_\infty \ge 1-h^-(1)\big\}$. We then deduce from (III.5.21) and (III.5.22) that

$$\mathbb{E}\big[\Delta^{2q}\mathbb{1}_{\Omega_h^c}\big] \le K(2n)^{2q}\exp\big(-\kappa n(1-h^-(1))^2/4\big).$$

For all positive u, $\exp(-u) < (3q/u)^{3q}$, so there exists $C(q)$ such that

$$\mathbb{E}\big[\Delta^{2q}\mathbb{1}_{\Omega_h^c}\big] \le C(q)\frac{K}{\kappa^{3q}n^q(1-h^-(1))^{6q}},$$

and the result follows. ■

III.5.2 Proof of Proposition 12

In the density estimation problem with censored data,

$$\begin{aligned}\mathbb{E}_s\big[\|s_m-\check{s}_m\|^2\big] &= \sum_{I\in m}\frac{1}{n|I|}\Big(\langle s(1-G^-)^{-1},\mathbb{1}_I\rangle - \langle s,\mathbb{1}_I\rangle^2\Big)\\ &= \sum_{I\in m}\frac{1}{n|I|}\Big(\langle s,\mathbb{1}_I\rangle(1-\langle s,\mathbb{1}_I\rangle) + \langle sG^-(1-G^-)^{-1},\mathbb{1}_I\rangle\Big)\end{aligned}$$

hence

$$\iota(s)\frac{D_m-1}{n} \le \mathbb{E}_s\big[\|s_m-\check{s}_m\|^2\big] \le \frac{\|s\|_\infty}{1-G^-(1)}\frac{D_m}{n}. \tag{III.5.23}$$

Let Δ be defined by

$$\Delta = \max_{1\le i\le n,\ s.t. T_i\le C_i}\left|\frac{\hat{w}(Z_i)-w(Z_i)}{w(Z_i)}\right|.$$

Since $w \ge 0$, for every $x \in I$,

$$(\hat{s}_m(x)-\check{s}_m(x))^2 \le \Delta^2\,(\check{s}_m(x))^2 \le 2\Delta^2\left[(s_m(x)-\check{s}_m(x))^2 + s_m^2(x)\right].$$

It thus follows from Cauchy-Schwarz inequality that

$$\mathbb{E}_s\big[\|\hat{s}_m-\check{s}_m\|^2\big] \le 2\Big(\|s\|^2 + \mathbb{E}_s^{1/2}\big[\|s_m-\check{s}_m\|^4\big]\Big)\mathbb{E}_s^{1/2}\big[\Delta^4\big].$$

But Δ satisfies $\mathbb{P}_s$-almost surely (III.5.22) with $h=G$ anf $\hat{h}=\tilde{G}$. Moreover, according to the Dworetzky-Kiefer-Wolfowitz type inequality proved by [BLM99], for all nonnegative x,

$$\mathbb{P}\big(\|(\tilde{G}-G)(1-F)\|_\infty > x\big) \le 2.5\exp(-2nx^2+\kappa_0\sqrt{n}x),$$

where κ_0 is an absolute constant and F is the distribution function of T_i. Therefore, there exists an absolute constant C such that, for all $x \ge 0$,

$$\mathbb{P}\big(\|(\tilde{G}-G)(1-F)\|_\infty > x\big) \le C\exp(-nx^2),$$

whence (III.5.21) holds for all $x \geq 0$ with K an absolute constant and $\kappa = (1 - F(1))^2$. From Lemma 2, one gets

$$\mathbb{E}\big[\Delta^{2q}\big] \leq \frac{C(q)}{n^q (1 - H^-(1))^{6q}}$$

for all $q \geq 1$, where $1 - H^-(1) = (1 - F(1))(1 - G^-(1))$. By Jensen's inequality,

$$\mathbb{E}_s\big[\|s_m - \check{s}_m\|^4\big] \leq \int_0^1 \mathbb{E}_s |s_m(x) - \check{s}_m(x)|^4 \mathrm{d}x.$$

The random variables $\{w(Z_i)\mathbb{1}_I(X_i)\}_{1 \leq i \leq n}$ are i.i.d. so we derive from (III.2.1) and the definition of s_m and $\check{s}_m$ that

$$\mathbb{E}_s\big[\|s_m - \check{s}_m\|^4\big] \leq \sum_{I \in m} \frac{1}{n^4 |I|^3} \Big(3n(n-1)\mathrm{Var}_s^2(w(Z_i)\mathbb{1}_I(X_i)) + n\mathbb{E}_s \left(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s(w(Z_i)\mathbb{1}_I(X_i))\right)^4 \Big).$$

But s is bounded and $|I| \geq 1/n$ for all $I \in m$, so

$$\mathbb{E}_s^{1/2}\big[\|s_m - \check{s}_m\|^4\big] \leq \frac{\|s\|_\infty}{(1 - G^-(1))^2} \sqrt{\frac{6D_m}{n}}. \tag{III.5.24}$$

Since $D_m \leq n$, we get

$$\mathbb{E}_s\big[\|\hat{s}_m - \check{s}_m\|^2\big] \leq \frac{C(s, G)}{n}$$

where $C(s, G)$ only depends on $\|s\|$, $\|s\|_\infty$, $G^-(1)$ and $F(1)$. But for all real numbers a and b, one has $(a+b)^2 \leq 2a^2 + 2b^2$ and $(a+b)^2 \geq a^2/2 - b^2$, so Proposition 12 then follows from Inequalities (III.5.23) and (III.5.25).

III.5.3 Proof of Proposition 13

For all $x \in [0, 1]$, let

$$s^\bullet(x) = s(x)\mathbb{1}_{\{X_{(n)} \geq x\}} \quad \text{and} \quad P^\bullet(x) = \int_0^x s^\bullet(u)\,\mathrm{d}u.$$

Let $s_m^\bullet$ be the orthogonal projection of $s^\bullet$ onto S_m. From Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\mathbb{E}_s\big[\|s_m^\bullet - s_m\|^2\big] &= \sum_{I \in m} \frac{1}{|I|} \mathbb{E}_s \left[\Big(\int_I s(u)\mathbb{1}_{\{X_{(n)} < u\}}\,\mathrm{d}u \Big)^2 \right] \\
&\leq \sum_{I \in m} \Big(\int_I s^2(u)\,\mathrm{d}u \Big) \mathbb{E}_s \left[\frac{1}{|I|} \int_I \mathbb{1}_{\{X_{(n)} < 1\}}\,\mathrm{d}u \right] \\
&\leq \|s\|^2 \left(H^-(1)\right)^n. \qquad \text{(III.5.25)}
\end{aligned}$$

From Theorem 2 p. 312 in [SW86], for instance, $\hat{P} - P^\bullet$ is a square integrable and centered martingale, with predictable variation process

$$\langle \hat{P} - P^\bullet \rangle(x) = \int_0^x \frac{\mathbb{1}_{\{X_{(n)} \geq y\}}}{n\big(1 - \hat{H}^-(y)\big)} s(y)\,\mathrm{d}y, \text{ for } x \in [0, 1].$$

Therefore, for every fixed $I \in m$ and every $x \in I$,

$$\begin{aligned}
\mathbb{E}_s\left[\left(\hat{s}_m(x) - s_m^\bullet(x)\right)^2 \right] &= \frac{1}{n|I|^2} \int_I \mathbb{E}_s \left[\frac{\mathbb{1}_{\{X_{(n)} \geq y\}}}{1 - \hat{H}^-(y)} \right] s(y)\,\mathrm{d}y \qquad \text{(III.5.26)} \\
&\leq \frac{1}{n|I|^2} \mathbb{E}_s \left[\sup_{[0, X_{(n)}]} \frac{1 - H}{1 - \hat{H}^-} \right] \int_I \frac{s(y)}{1 - H(y)}\,\mathrm{d}y.
\end{aligned}$$

By integrating the inequality

$$\mathbb{P}_s\left(\sup_{[0,X_{(n)}]}\frac{1-H}{1-\hat{H}^-}\geq x\right)\leq ex\exp(-x),\ \text{for } x\geq 1,$$

that may be found in [SW86] (Inequality (d), p. 318), we get

$$\mathbb{E}_s\left[\sup_{[0,X_{(n)}]}\frac{1-H}{1-\hat{H}^-}\right]\leq 3. \tag{III.5.27}$$

On the other hand, for all $y\in[0,1]$,

$$\mathbb{E}_s\left[\frac{\mathbb{1}_{\{X_{(n)}\geq y\}}}{1-\hat{H}^-(y)}\right]\geq \mathbb{P}_s\big(X_{(n)}\geq y\big)\geq 1-(H^-(1))^n. \tag{III.5.28}$$

We then deduce from (III.5.26) to (III.5.28) that

$$\iota(s)\big(1-(H^-(1))^n\big)\frac{D_m}{n}\leq \mathbb{E}_s\left[\|\hat{s}_m-s_m^\bullet\|^2\right]\leq 3\frac{\|s\|_\infty}{1-H^-(1)}\frac{D_m}{n}. \tag{III.5.29}$$

But for all real numbers a and b, one has $(a+b)^2\leq 2a^2+2b^2$ and $(a+b)^2\geq a^2/2-b^2$. The result thus follows from (III.5.25) and (III.5.29).

III.5.4 Proof of Proposition 14

The proof of Proposition 14 relies on Theorem 9 below. This theorem extends a recent concentration inequality due to Adamczak [Ada08], that itself extends Talagrand's concentration inequality (Theorem 1.4. in [Tal96]) to families of functions which are not uniformly bounded. In the sequel, the norm $\|.\|_{\psi_1}$ is defined for every real-valued random variable U by

$$\|U\|_{\psi_1}=\inf\Big\{\lambda>0\text{ s.t. }\mathbb{E}\big[\exp(|U|/\lambda)\big]\leq 2\Big\}.$$

Theorem 9 *Let $U_1,\ldots,U_n$ be independent random variables taking values in a measurable space $(\mathcal{U},\mathcal{B})$. Let $\mathcal{F}$ be a countable family of real-valued and measurable functions defined on $\mathcal{U}$ such that, for all $1\leq i\leq n$, $\|\sup_{f\in\mathcal{F}}|f(U_i)|\|_{\psi_1}$ is finite. Let*

$$Z=\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\Big(f(U_i)-\mathbb{E}[f(U_i)]\Big)\right|,$$

$$\sigma^2=\sup_{f\in\mathcal{F}}\sum_{i=1}^n\mathrm{Var}\big(f(U_i)\big)\quad\text{and}\quad B=\Big\|\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}|f(U_i)|\Big\|_{\psi_1}.$$

Then B is finite and there exists a constant κ such that, for all $0<x\leq(2\sigma/\kappa B)^2$,

$$P\Big(Z\geq 2\mathbb{E}[Z]+2\sigma\sqrt{x}\Big)\leq 4\exp(-x),$$

and, for all $x>(2\sigma/\kappa B)^2$,

$$P\Big(Z\geq 2\mathbb{E}[Z]+\kappa Bx\Big)\leq 4\exp(-x).$$

The proof of Theorem 9 is not given here since it follows almost the same lines as that of Theorem 4 in [Ada08].

Let us prove that b_0 is finite. For all $1 \leq i \leq n$, and all positive λ, a convexity argument leads to

$$\begin{aligned}\exp\big(\lambda|w(Z_i)|\big) &= \prod_{I\in m^\star} \exp\big(\lambda|w(Z_i)|\mathbb{1}_I(X_i)\big) \quad \mathbb{P}_s - a.s.\\ &\leq \sum_{I\in m^\star} \exp\big(\lambda D_{m^\star}|w(Z_i)|\mathbb{1}_I(X_i)\big)/D_{m^\star}.\end{aligned}$$

Besides, it follows from Assumption **(M)** that there exists $\lambda > 0$ such that for all $1 \leq i \leq n$ and all $I \in m^\star$, $\mathbb{E}_s\big[\exp\big(\lambda D_{m^\star}|w(Z_i)|\mathbb{1}_I(X_i)\big)\big]$ is finite. Lemma 2.2.2. in [vdVW96] then allows to conclude that b_0 is finite since

$$b_0 \leq C\ln(n)\max_{1\leq i\leq n}\big\|w(Z_i)\big\|_{\psi_1}$$

for some absolute constant C.

Let us fix some partition m built on $m^\star$ and note that $s_m = \mathbb{E}_s(\check{s}_m)$. Let $\mathscr{S}_m$ be some countable and dense subset of $\big\{(a_I) \in \mathbb{R}^{D_m} \text{ s.t. } \sum_{I\in m} a_I^2 = 1\big\}$ and, for all $a \in \mathscr{S}_m$, let f_a be defined on $[0,1]\times\mathcal{Y}$ by

$$f_a(x,y) = \sum_{I\in m} \frac{a_I}{n\sqrt{\mu(I)}} w(x,y)\mathbb{1}_I(x).$$

From Cauchy-Schwarz inequality and its equality case,

$$\begin{aligned}\|\check{s}_m - s_m\| &= \sqrt{\sum_{I\in m}\mu(I)(\check{s}_m - s_m)^2\mathbb{1}_I}\\ &= \sup_{a\in\mathscr{S}_m}\left|\sum_{I\in m} a_I\sqrt{\mu(I)}(\check{s}_m - s_m)\mathbb{1}_I\right|\\ &= \sup_{a\in\mathscr{S}_m}\left|\sum_{i=1}^n\Big(f_a(Z_i) - \mathbb{E}_s\big[f_a(Z_i)\big]\Big)\right|,\end{aligned}$$

where the supremum is reached by $\check{a} = (\check{a}_I)_{I\in m}$ such that, for all $I \in m$,

$$\check{a}_I = \sqrt{\mu(I)}(\check{s}_m - s_m)\mathbb{1}_I/\|\check{s}_m - s_m\|.$$

Yet, we shall not apply Theorem 9 to $\|\check{s}_m - s_m\|$ but to some truncated version of that variable: as Castellan [Cas00] (Proposition 4.3), our aim is to exhibit the subgaussian behaviour of $\|\check{s}_m - s_m\|$ on some subset of Ω. Let

$$\Omega_\star = \Big\{\|\check{s}_{m^\star} - s_{m^\star}\|_\infty \leq \varepsilon\Big\},$$

where $\varepsilon = 4M/\kappa b$ for some $b \geq b_0$. Let z be a positive real, to be chosen later, and $\mathscr{A}_m$ be a subset of $\big\{a \in \mathscr{S}_m \text{ s.t. } \max_{I\in m}|a_I|/\sqrt{\mu(I)} \leq \varepsilon/z\big\}$. We consider

$$W(m) = \sup_{a\in\mathscr{A}_m}\left|\sum_{i=1}^n\Big(f_a(Z_i) - \mathbb{E}_s\big[f_a(Z_i)\big]\Big)\right|.$$

Since $W(m) \leq \|\check{s}_m - s_m\|$, we obtain by concavity of the square-root function that

$$\mathbb{E}_s[W(m)] \leq \sqrt{\mathbb{E}_s[\|\check{s}_m - s_m\|^2]}.$$

It follows from the definition of M and the independence of $Z_1, \dots, Z_n$ that

$$\max_{I\in m} \mathbb{E}_s\big[n\|(\check{s}_m - s_m)\mathbb{1}_I\|^2\big] = \max_{I\in m} \frac{1}{n\mu(I)} \sum_{i=1}^{n} \mathrm{Var}_s\big[w(Z_i)\mathbb{1}_I(X_i)\big] \le M,$$

hence

$$\mathbb{E}_s[W(m)] \le \sqrt{\frac{MD_m}{n}}.$$

Besides,

$$\sigma_m^2 := \sup_{a\in\mathscr{A}_m} \sum_{i=1}^{n} \mathrm{Var}\big(f_a(Z_i)\big) \le \frac{M}{n} \quad \text{and} \quad B_m := \Big\| \max_{1\le i\le n} \sup_{a\in\mathscr{A}_m} |f_a(Z_i)| \Big\|_{\psi_1} \le \frac{\varepsilon}{nz} b.$$

Let us fix $x \in (0, (2\sigma_m/\kappa B_m)^2]$. We deduce from Theorem 9 that there exists a set $\Omega(x) \in \mathcal{A}$ such that $\mathbb{P}_s\big(\Omega^c(x)\big) \le 4\exp(-x)$ and on which

$$\sqrt{n}W(m) < 2\sqrt{MD_m} + 2\sqrt{Mx}.$$

Since m is built on $m^\star$, we have $\|\check{s}_m - s_m\|_\infty \le \varepsilon$ on $\Omega_\star$. Thus, on $\Omega_\star \cap \{\|\check{s}_m - s_m\| \ge z\}$,

$$\max_{I\in m} |\check{a}_I| / \sqrt{\mu(I)} \le \varepsilon/z,$$

so that $\|\check{s}_m - s_m\|$ and $W(m)$ coincide on that set. By setting $z = 2\sqrt{Mx/n}$, we obtain on $\Omega(x)$

$$\sqrt{n}\|\check{s}_m - s_m\|\mathbb{1}_{\Omega_\star} < 2\sqrt{MD_m} + 2\sqrt{Mx},$$

which proves (III.3.9) for all $x \in (0, (2\sigma_m/\kappa B_m)^2]$. Let us now fix $x > (2\sigma_m/\kappa B_m)^2$. We deduce this time from Theorem 9 that there exists a set $\Omega(x) \in \mathcal{A}$ such that $\mathbb{P}_s\big(\Omega^c(x)\big) \le 4\exp(-x)$ and on which

$$\sqrt{n}W(m) < 2\sqrt{MD_m} + \frac{\kappa\varepsilon b}{\sqrt{n}z} x.$$

We obtain (III.3.9) for such x by setting again $z = 2\sqrt{Mx/n}$.

Arguments similar to those used Section 2.2.3 in [Mas07] show that

$$\mathbb{P}_s(\Omega_\star^c) \le 2D_{m^\star} \exp\bigg(- C\frac{M^2}{b(bv + aM)} \ell_\star \bigg),$$

where $D_{m^\star} \le n/\ell_\star$, hence (III.3.10).

In the case where the X_i's are non-random, one can consider $\Omega_\star = \Omega$ if $a = 0$ or

$$\Omega_\star = \Big\{ \|\check{s}_{m^\star} - s_{m^\star}\|_\infty \le v/a \Big\}$$

otherwise, and conclude with the same arguments as in Sauvé [Sau09], Lemma 1.

III.5.5 Proof of Theorem 6

We will use the following lemma.

Lemma 3 *Assumptions* **(P)** *and* **(M)** *are supposed to be fulfilled. There exists $C(a, v)$ that only depends on a and v such that*

$$\mathbb{E}_s\big[\|s_{m^\star} - \check{s}_{m^\star}\|^4\big] \le C(a, v).$$

Moreover, $C(a, v)$ is an increasing function of a and an increasing function of v.

Proof. By Jensen's inequality,

$$\mathbb{E}_s\big[\|s_{m^\star} - \check{s}_{m^\star}\|^4\big] \le \int_0^1 \mathbb{E}_s|s_{m^\star}(x) - \check{s}_{m^\star}(x)|^4 \mathrm{d}x.$$

Since $s_{m^\star} = \mathbb{E}_s(\check{s}_{m^\star})$ and $\min_{I\in m^\star} n\mu(I) \ge 1$, we get

$$\mathbb{E}_s\big[\|s_{m^\star} - \check{s}_{m^\star}\|^4\big] \le \frac{D_{m^\star}}{n} \max_{I\in m^\star} \mathbb{E}_s\left[\left(\frac{1}{\sqrt{n\mu(I)}}\sum_{i=1}^n (w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s[w(Z_i)\mathbb{1}_I(X_i)])\right)^4\right].$$

By Assumption **(M)**, for all $I \in m^\star$ and all $\lambda \in (-1/a, 1/a)$,

$$\ln \mathbb{E}_s\left[\exp\left(\frac{\lambda}{\sqrt{n\mu(I)}}\sum_{i=1}^n \Big(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w(Z_i)\mathbb{1}_I(X_i)\big]\Big)\right)\right] \le \frac{v\lambda^2}{2(1-a|\lambda|)}.$$

Integrating this inequality, one obtains that there exists $C(a,v)$ that increases with a and v such that

$$\mathbb{E}_s\left[\left(\frac{1}{\sqrt{n\mu(I)}}\sum_{i=1}^n (w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s[w(Z_i)\mathbb{1}_I(X_i)])\right)^4\right] \le C(a,v). \tag{III.5.30}$$

Besides, $D_{m^\star} \le n/\ell_\star \le n$, hence Lemma 3. ■

Let us first introduce some notation. For all $t \in \mathcal{S}$, let

$$\nu_c(t) = \frac{1}{n}\sum_{i=1}^n \big(w(Z_i)t(X_i) - \mathbb{E}_s[w(Z_i)t(X_i)]\big)$$
$$R(t) = \frac{1}{n}\sum_{i=1}^n (\hat{w} - w)(Z_i)t(X_i).$$

Let us fix some partition $m \in \mathcal{M}^\star$, and set

$$\chi_c(m) = \sup_{\substack{t\in S_m\\ \|t\|=1}} |\nu_c(t)| \quad \text{and} \quad \chi_R(m) = \sup_{\substack{t\in S_m\\ \|t\|=1}} |R(t)|. \tag{III.5.31}$$

Notice that, from Cauchy-Schwarz inequality and (III.2.1),

$$\chi_c(m) = \|\check{s}_m - s_m\| \quad \text{and} \quad \chi_R(m) = \|\check{s}_m - \hat{s}_m\|. \tag{III.5.32}$$

From the definitions of $\hat{m}$ and $\hat{s}_m$, we get

$$\begin{aligned}\gamma(\tilde{s}) + \mathrm{pen}(\hat{m}) &\le \gamma(\hat{s}_m) + \mathrm{pen}(m)\\ &\le \gamma(s_m) + \mathrm{pen}(m).\end{aligned}$$

On the other hand, for all $t, u \in \mathcal{S}$,

$$\gamma(t) - \gamma(u) = \|s-t\|^2 - \|s-u\|^2 - 2(\nu_c + R)(t-u).$$

So

$$\|s - \tilde{s}\|^2 \le \|s - s_m\|^2 + \mathrm{pen}(m) + 2(\nu_c + R)(\tilde{s} - s_m) - \mathrm{pen}(\hat{m}). \tag{III.5.33}$$

In the sequel, for all partitions $m' \in \mathcal{M}^\star$, we denote by $m \cup m'$ the roughest partition built on m and m'. We recall that, for all positive θ and all nonnegative a and b,

$$2ab \le \frac{1}{\theta}a^2 + \theta b^2. \tag{III.5.34}$$

We deduce from the triangle inequality and the last inequality, applied several times, with $\theta = 1/2$ or $\theta = 1$, that

$$\begin{aligned} 2(\nu_c + R)(\tilde{s} - s_m) &\leq 2\|\tilde{s} - s_m\|\big(\chi_c(m \cup \hat{m}) + \chi_R(m \cup \hat{m})\big) \\ &\leq 2\big(\|s - \tilde{s}\| + \|s - s_m\|\big)\big(\chi_c(m \cup \hat{m}) + \chi_R(m \cup \hat{m})\big) \\ &\leq \frac{1}{2}\|s - \tilde{s}\|^2 + \|s - s_m\|^2 + 6\chi_c^2(m \cup \hat{m}) + 6\chi_R^2(m \cup \hat{m}). \end{aligned} \tag{III.5.35}$$

As $\mathcal{M}^\star$ only contains partitions built on $m^\star$, $S_{m\cup\hat{m}}$ is a subspace of $S_{m^\star}$, so

$$\chi_R^2(m \cup \hat{m}) \leq \chi_R^2(m^\star). \tag{III.5.36}$$

Combining Inequalities (III.5.33), (III.5.35) and (III.5.36) yields

$$\|s - \tilde{s}\|^2 \leq 4\|s - s_m\|^2 + 2\mathrm{pen}(m) + 12\chi_c^2(m \cup \hat{m}) - 2\mathrm{pen}(\hat{m}) + 12\chi_R^2(m^\star). \tag{III.5.37}$$

We shall first look for an upper-bound for the risk of $\tilde{s}$ on the set $\Omega_\star \cap A$. Let us fix some positive ζ. For all partitions $m' \in \mathcal{M}^\star$, let

$$x_{m'} = L_{m'} D_{m'} + \zeta,$$

$$\Omega_\zeta(m, m') = \left\{ n\chi_c^2(m \cup m')\mathbb{1}_{\Omega_\star} < M\Big(k_1(D_m + D_{m'}) + k_2 x_{m'} + 2k_3\sqrt{k_1 k_2 (D_m + D_{m'})x_{m'}}\Big) \right\}$$

and

$$\Omega_\zeta(m) = \bigcap_{m' \in \mathcal{M}^\star} \Omega_\zeta(m, m').$$

For all partitions $m' \in \mathcal{M}^\star$, Assumption **(P)** ensures that $m \cup m'$ is a partition built on $m^\star$, so we deduce from Proposition 14, Inequality (III.3.9), that

$$\mathbb{P}_s\big(\Omega_\zeta^c(m, m')\big) \leq k_4 \exp(-x_{m'}),$$

so that $\Omega_\zeta(m)$ is a set with probability

$$\mathbb{P}_s\big(\Omega_\zeta(m)\big) \geq 1 - k_4 \exp(-\zeta)\Sigma. \tag{III.5.38}$$

Using repeatedly that, for all nonnegative a and b, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and applying several times Inequality (III.5.34) with $\theta = 1$, we get on the set $\Omega_\zeta(m)$

$$\begin{aligned} &nM^{-1}\chi_c^2(m \cup \hat{m})\mathbb{1}_{\Omega_\star} \\ &\leq k_1(D_m + D_{\hat{m}}) + k_2(D_{\hat{m}}L_{\hat{m}} + \zeta) + 2k_3\sqrt{k_1 k_2 (D_m + D_{\hat{m}})(D_{\hat{m}}L_{\hat{m}} + \zeta)} \\ &\leq \Big(k_1 + k_2 L_{\hat{m}} + 2k_3\sqrt{k_1 k_2 L_{\hat{m}}}\Big) D_{\hat{m}} + k_1 D_m + k_2\zeta \\ &\quad + 2k_3\Big(\sqrt{k_1 D_{\hat{m}} k_2 \zeta} + \sqrt{k_2 D_{\hat{m}} L_{\hat{m}} k_1 D_m} + \sqrt{k_1 D_m k_2 \zeta}\Big) \\ &\leq \Big(k_1 + k_2 L_{\hat{m}} + 2k_3\sqrt{k_1 k_2 L_{\hat{m}}}\Big) D_{\hat{m}} + k_1 D_m + k_2\zeta \\ &\quad + k_3\Big(k_1 D_{\hat{m}} + k_2\zeta + k_2 D_{\hat{m}} L_{\hat{m}} + k_1 D_m + k_1 D_m + k_2\zeta\Big) \\ &\leq K_1(k_1, k_2, k_3, L_{\hat{m}}) D_{\hat{m}} + K_2(k_1, k_3) D_m + K_3(k_2, k_3)\zeta, \end{aligned}$$

where, for all partitions $m' \in \mathcal{M}^\star$,

$$\begin{aligned} K_1(k_1, k_2, k_3, L_{m'}) &= (1 + k_3)(k_1 + k_2 L_{m'}) + 2k_3\sqrt{k_1 k_2 L_{m'}} \\ K_2(k_1, k_3) &= k_1(1 + 2k_3) \\ K_3(k_2, k_3) &= k_2(1 + 2k_3). \end{aligned}$$

Since, on the set A, $M \leq k_7\widehat{M}$, we deduce from Inequality (III.5.37), the assumption on pen and the last inequality that, still on the set $\Omega_\zeta(m)$,

$$\begin{aligned}\|s-\tilde{s}\|^2\mathbb{1}_{\Omega_\star\cap A} \leq 4\|s-s_m\|^2 &\\ &+ 2\mathrm{pen}(m)\mathbb{1}_A + 12k_7K_2(k_1,k_3)\frac{\widehat{M}D_m}{n}\mathbb{1}_A\\ &+ 12MK_3(k_2,k_3)\frac{\zeta}{n} + 12\chi_R^2(m^\star)\mathbb{1}_{\Omega_\star\cap A}.\end{aligned}$$

Since $K_1(k_1,k_2,k_3,L_m) \geq k_1(1+k_3)$, we notice that

$$12k_7K_2(k_1,k_3)\frac{\widehat{M}D_m}{n}\mathbb{1}_A \leq K_4(k_3)\mathrm{pen}(m)\mathbb{1}_A,$$

where

$$K_4(k_3) = 2\frac{1+2k_3}{1+k_3}.$$

Consequently, still on the set $\Omega_\zeta(m)$, we have

$$\begin{aligned}\|s-\tilde{s}\|^2\mathbb{1}_{\Omega_\star\cap A} \leq 4\|s-s_m\|^2 + \big(2+K_4(k_3)\big)\mathrm{pen}(m)\mathbb{1}_A&\\ + 12MK_3(k_2,k_3)\frac{\zeta}{n} + 12\chi_R^2(m^\star)\mathbb{1}_{\Omega_\star\cap A}.&\end{aligned}$$

Finally, after integrating with respect to ζ, and taking the infimum over $m \in \mathcal{M}^\star$, we get the following upper-bound for the risk of $\tilde{s}$ on $\Omega_\star \cap A$:

$$\begin{aligned}\mathbb{E}_s\Big[\|s-\tilde{s}\|^2\mathbb{1}_{\Omega_\star\cap A}\Big] \leq \inf_{m\in\mathcal{M}^\star}\Big\{C_1\|s-s_m\|^2 + C_2(k_3)\mathbb{E}_s\big[\mathrm{pen}(m)\mathbb{1}_A\big]\Big\}&\\ + C_3(k_2,k_3,k_4)\frac{M\Sigma}{n} + 12\mathbb{E}_s\big[\chi_R^2(m^\star)\mathbb{1}_{\Omega_\star\cap A}\big].&\end{aligned} \tag{III.5.39}$$

Let us now bound the risk of $\tilde{s}$ on the complementary set of $\Omega_\star \cap A$. First, Pythagoras' equality leads to

$$\|s-\tilde{s}\|^2 = \|s-s_{\hat{m}}\|^2 + \|s_{\hat{m}}-\hat{s}_{\hat{m}}\|^2.$$

Besides, it follows from the triangle inequality and the expressions of χ_c and χ_R (*cf.* (III.5.32) and (III.5.31)) that for any $\theta > 0$,

$$\begin{aligned}\|s_{\hat{m}}-\hat{s}_{\hat{m}}\|^2 &\leq \big(\chi_c(\hat{m}) + \chi_R(\hat{m})\big)^2\\ &\leq (1+\theta^{-1})\chi_c^2(m^\star) + (1+\theta)\chi_R^2(m^\star).\end{aligned}$$

Therefore,

$$\|s-\tilde{s}\|^2 \leq \|s\|^2 + (1+\theta^{-1})\chi_c^2(m^\star) + (1+\theta)\chi_R^2(m^\star).$$

Let us denote by $p_\star$ the $\mathbb{P}_s-$measure of $\Omega_\star^c \cup A^c$. Applying Cauchy-Schwarz inequality, we get

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\mathbb{1}_{\Omega_\star^c\cup A^c}\big] \leq p_\star\|s\|^2 + (1+\theta^{-1})\sqrt{p_\star\mathbb{E}_s[\chi_c^4(m^\star)]} + (1+\theta)\mathbb{E}_s[\chi_R^2(m^\star)\mathbb{1}_{\Omega_\star^c\cup A^c}]. \tag{III.5.40}$$

It follows from Proposition 14 and the assumption on A that

$$p_\star \leq \frac{1}{n^2}\left(\frac{k_5}{\ell_\star}n^3\exp(-k_6\ell_\star) + C_A\right).$$

Since $p_\star \leq 1$ and as $\mathbb{E}_s[\chi_R^2(m^\star)] \leq C_\Delta/n$ under Assumption **(W)**, the first part of Theorem 6 follows from (III.5.39), Lemma 3 and (III.5.40) where we set $\theta = 11$. The second assertion in Theorem 6 is an easy consequence of the first one.

III.5.6 Proof of Proposition 15

According to Bernstein's inequality as stated in [Mas07] (Section 2.2.3), Assumption **(M)** is satisfied for instance when, for all integers $k \geq 2$,

$$\sum_{i=1}^{n} \mathbb{E}_s\big[|w(Z_i)|^k \mathbb{1}_I(X_i)\big] \leq \frac{k!}{2} nv|I|a^{k-2}.$$

As w is bounded, for all integers $k \geq 2$,

$$\sum_{i=1}^{n} \mathbb{E}_s\big[|w(Z_i)|^k \mathbb{1}_I(X_i)\big] \leq nM|I|\|w\|_\infty^{k-2},$$

so that Assumption **(M)** is satisfied with $v = M$ and $a = \|w\|_\infty/3$. Moreover, from (III.2.1), for all $I \in m^\star$,

$$\left(\frac{1}{n|I|}\sum_{i=1}^{n}(\hat{w}-w)(Z_i)\mathbb{1}_I(X_i)\right)^2 \leq \Delta^2\left(\frac{1}{n|I|}\sum_{i=1}^{n}(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s[w(Z_i)\mathbb{1}_I(X_i)]) + \|s\|_\infty\right)^2.$$

It thus follows from Cauchy-Schwarz inequality that Assumption **(W)** is fulfilled with

$$C_\Delta = 2C'_\Delta(\sqrt{C(a,v)} + \|s\|_\infty^2),$$

where $C(a,v)$ is defined as in (III.5.30).

Let us check that Conditions (III.3.11) and (III.3.13) are fulfilled. Let $\theta = 1/2$ and

$$A = \bigcap_{I\in m^\star}\left\{\left|\sum_{i=1}^{n}\Big(\hat{w}^2(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w^2(Z_i)\mathbb{1}_I(X_i)\big]\Big)\right| \leq \theta Mn|I|\right\}$$

On A, we immediately obtain the inequalities $M \leq (1-\theta)^{-1}\widehat{M}$ and $\widehat{M} \leq (1+\theta)M$. Let us now bound $\mathbb{P}_s(A^c)$. For all $I \in m^\star$, we introduce the sum of centered i.i.d. variables

$$S_I = \sum_{i=1}^{n}\Big(w^2(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w^2(Z_i)\mathbb{1}_I(X_i)\big]\Big)$$

and the sets

$$F_I = \Big\{|S_I| > (\theta/2)Mn|I|\Big\} \text{ and } G_I = \left\{\left|\sum_{i=1}^{n}(\hat{w}^2 - w^2)(Z_i)\mathbb{1}_I(X_i)\right| > (\theta/2)Mn|I|\right\}.$$

Moreover, we introduce

$$\Delta' = \max_{1\leq i\leq n}\frac{|\hat{w}^2 - w^2|}{w^2}(Z_i)$$

with the convention that $0/0 = 0$. For all $I \in m^\star$

$$\left|\sum_{i=1}^{n}(\hat{w}^2 - w^2)(Z_i)\mathbb{1}_I(X_i)\right| \leq \Delta'\Big(S_I + Mn|I|\Big),$$

so, for all positive c,

$$G_I \subset \big\{\Delta' > c\big\} \cup \Big\{S_I > (\theta/(2c) - 1)Mn|I|\Big\}.$$

Choosing c so that $\theta/(2c) - 1 = \theta/2$, i.e. $c = \theta/(\theta+2) = 1/5$, we get

$$A^c \subset \{\Delta' > \theta/(\theta+2)\} \cup \bigcup_{I \in m^\star} F_I.$$

Using the aforementionned version of Bernstein's inequality so as to bound $\mathbb{P}_s(F_I)$ and the inequality $D_{m^\star} \leq n/\ell_\star$, we obtain by summing over $I \in m^\star$ that

$$\sum_{I \in m^\star} \mathbb{P}_s(F_I) \leq \frac{2}{\ell_\star} n \exp\left(- C \frac{M \ell_\star}{\|w\|_\infty^2} \right)$$

where C is an aboslute constant. Moreover, since w satisfies Condition (III.2.1) and is nonnegative, we get, for all $I \in m^\star$,

$$C_w \langle s, \mathbb{1}_I \rangle \leq \mathbb{E}_s[w^2(Z_1)\mathbb{1}_I(X_1)] \leq \|w\|_\infty \langle s, \mathbb{1}_I \rangle,$$

hence

$$C_w \iota(s) \leq M \leq \|w\|_\infty \|s\|_\infty.$$

Since $\Delta' \leq \Delta^2 + 2\Delta$, it follows from the assumption on Δ that

$$\mathbb{P}_s(A^c) \leq C \times \frac{(C'_\Delta)^2 + C(C_w, \|w\|_\infty, \iota(s))}{n^2},$$

where C is an absolute constant and $C(C_w, \|w\|_\infty, \iota(s))$ is a decreasing function of $\iota(s)$ and C_w and an increasing function of $\|w\|_\infty$. Consequently, Conditions (III.3.11) and (III.3.13) in Theorem 6 are satisfied with $k_7 = 2$ and $k_8 = 3/2$ and we can set $b = \|w\|_\infty / \ln(2)$. Oracle-type inequality (III.3.15) then follows, where we can take $k_1, \ldots, k_5$ absolute constants and $k_6 = cC_w\iota(s)/\|w\|_\infty^2$ for some small enough absolute constant $c > 0$.

III.5.7 Proof of Theorem 8

First, we obtain as a straightforward consequence of Theorem 7 that

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C_6(1 + M + C_4) \inf_{m \in \mathcal{M}^\star} \left\{ \|s - s_m\|^2 + \frac{D_m}{n} \right\}.$$

Then, for all $m \in \mathcal{M}^\star$, it follows from the definition of s_m and the triangle inequality that, for all $t \in S_m$,

$$\|s - s_m\|^2 \leq 2(\|s - s_{m^\star}\|^2 + \|s_{m^\star} - t\|^2),$$

hence, by taking the infimum over $t \in S_m$ and $m \in \mathcal{M}^\star$,

$$\inf_{m \in \mathcal{M}^\star} \left\{ \|s - s_m\|^2 + \frac{D_m}{n} \right\} \leq 2\|s - s_{m^\star}\|^2 + 2 \inf_{m \in \mathcal{M}^\star} \left\{ \inf_{t \in S_m} \|s_{m^\star} - t\|^2 + \frac{D_m}{n} \right\}. \quad \text{(III.5.41)}$$

Let us first assume that $s \in \mathcal{V}(\alpha, R)$, with $0 < \alpha \leq 1$ and $R \geq 0$. For any subinterval I of $[0,1]$, we denote by s_I the mean value of s over I and define

$$V_\alpha(s, I) = \sup_{i \geq 1} \sup_{\substack{x_0 < \ldots < x_i \\ (x_0, \ldots, x_i) \in I^{i+1}}} \left(\sum_{j=1}^{i} |s(x_j) - s(x_{j-1})|^{1/\alpha} \right)^\alpha.$$

Remembering that the dyadic intervals of $[0,1]$ of length $2^{-J_\star}$ are the $\{I_{(J_\star,k)}\}_{k=0,\dots,2^{J_\star}-1}$ (*cf.* (III.4.20)), we have

$$\begin{aligned}\|s-s_{m^\star}\|^2 &\le \sum_{k=0}^{2^{J_\star}-1} |I_{(J_\star,k)}| \|(s-s_{I_{(J_\star,k)}})\mathbb{1}_{I_{(J_\star,k)}}\|_\infty^2 \\ &\le 2^{-J_\star} \sum_{k=0}^{2^{J_\star}-1} V_\alpha^2(s, I_{(J_\star,k)}).\end{aligned}$$

We then deduce from classical inequalities between ℓ_p-norms that

$$\|s-s_{m^\star}\|^2 \le 2^{J_\star((1-2\alpha)_+-1)} \left(\sum_{k=0}^{2^{J_\star}-1} V_\alpha^{1/\alpha}(s, I_{(J_\star,k)}) \right)^{2\alpha}.$$

As $V_\alpha(s,.)$ is subadditive and $V_\alpha(s) \le R$, we get

$$\|s-s_{m^\star}\|^2 \le 2^{J_\star((1-2\alpha)_+-1)} R^2. \tag{III.5.42}$$

Let us now bound $V_\alpha(s_{m^\star})$. Let i be a positive integer and $0 \le x_0 < \ldots < x_i \le 1$. For all $0 \le j \le i$, let $I_{(J_\star,k(j))}$ be the only dyadic interval of length $2^{-J_\star}$ that contains x_j. Since $1/\alpha \ge 1$, we deduce from Jensen's Inequality that

$$\begin{aligned}\sum_{j=1}^{i} |s_{m^\star}(x_j) - s_{m^\star}(x_{j-1})|^{1/\alpha} &= \sum_{j=1}^{i} \left| 2^{J_\star} \int_0^{1/2^{J_\star}} \left[s\big(x+k(j)2^{-J_\star}\big) - s\big(x+k(j-1)2^{-J_\star}\big) \right] \mathrm{d}x \right|^{1/\alpha} \\ &\le 2^{J_\star} \int_0^{1/2^{J_\star}} \sum_{j=1}^{i} \left| s\big(x+k(j)2^{-J_\star}\big) - s\big(x+k(j-1)2^{-J_\star}\big) \right|^{1/\alpha} \mathrm{d}x \\ &\le V_\alpha^{1/\alpha}(s),\end{aligned}$$

hence, by taking the supremum over all finite increasing sequences with values in $[0,1]$,

$$V_\alpha(s_{m^\star}) \le V_\alpha(s) \le R.$$

We can thus apply Proposition 3 of [Bir07] to $s_{m^\star}$. So, for all nonnegative integers j, there exist a partition m_j of $[0,1]$ into dyadic intervals and a function $t_j \in S_{m_j}$ such that

$$D_{m_j} \le c_1(\alpha) 2^j \text{ and } \|s_{m^\star} - t_j\|^2 \le c_2(\alpha) R^2 2^{-2\alpha j},$$

where $c_1(\alpha)$ and $c_2(\alpha)$ are positive reals whose precise values are given in [Bir07]. Moreover, since $s_{m^\star}$ is piecewise constant over $m^\star$, by construction, each $m_j \in \mathcal{M}^\star$. Therefore,

$$\begin{aligned}\inf_{m\in\mathcal{M}^\star} \left\{ \inf_{t\in S_m} \|s_{m^\star}-t\|^2 + \frac{D_m}{n} \right\} &\le \inf_{0\le j\le J_\star} \left\{ \|s_{m^\star}-t_j\|^2 + \frac{D_{m_j}}{n} \right\} \\ &\le C(\alpha) \inf_{0\le j\le J_\star} \left\{ R^2 2^{-2\alpha j} + \frac{2^j}{n} \right\}. \end{aligned} \tag{III.5.43}$$

Let J_0 be the greatest integer such that $2^{J_0}/n \le R^2 2^{-2\alpha J_0}$, *i.e.* such that $2^{J_0} \le (nR^2)^{1/(1+2\alpha)}$. If $1/n \le R^2 \le 2^{J_\star(1+2\alpha)}/n$, then $0 \le J_0 \le J_\star$ and $R^2 2^{-2\alpha J_0} + 2^{J_0}/n \le C(\alpha)(Rn^{-\alpha})^{2/(1+2\alpha)}$. Therefore, for $0 < \alpha \le 1/2$ and $1/n \le R^2 \le 2^{J_\star(1+2\alpha)}/n$, we deduce from Inequalities (III.5.41) to (III.5.43) that

$$\begin{aligned}\inf_{m\in\mathcal{M}^\star} \left\{ \|s-s_m\|^2 + \frac{D_m}{n} \right\} &\le C(\alpha)\big(R^2 2^{-2\alpha J_\star} + (Rn^{-\alpha})^{2/(1+2\alpha)}\big) \\ &\le C(\alpha)(Rn^{-\alpha})^{2/(1+2\alpha)}.\end{aligned}$$

For $1/2 < \alpha \leq 1$ and $1/n \leq R^2 \leq 2^{J_\star(1+2\alpha)}/n$, we deduce this time from Inequalities (III.5.41) to (III.5.43) that

$$\inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2+\frac{D_m}{n}\right\} \leq C(\alpha)\big(R^2 2^{-J_\star}+(Rn^{-\alpha})^{2/(1+2\alpha)}\big),$$

so that when $1/n \leq R^2 \leq 2^{J_\star(1+1/(2\alpha))}/n$, we still have

$$\inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2+\frac{D_m}{n}\right\} \leq C(\alpha)(Rn^{-\alpha})^{2/(1+2\alpha)}.$$

Let us now assume that $s \in \mathcal{B}(\alpha,p,R)$ for some $R \geq 0$, $p \geq 1$ and $1/p - 1/2 < \alpha < 1$. Let $(H_\lambda)_{\lambda\in\Lambda}$ be the Haar basis on $\mathbb{L}^2([0,1])$, indexed by $\Lambda = \bigcup_{j\geq -1}\Lambda(j)$ where

$$\Lambda(j) = \{(j,k), k = 0,\ldots,2^j-1\}.$$

Otherwise said, $H_{(-1,0)} = \mathbb{1}_{[0,1]}$ and

$$H_{(j,k)}(x) = 2^{j/2}H(2^j x - k), \text{ for } j \in \mathbb{N}, 0 \leq k \leq 2^j-1 \text{ and } x \in [0,1],$$

where $H = \mathbb{1}_{[0,1/2]} - \mathbb{1}_{(1/2,1]}$. Let $\Lambda_{J_\star-1} = \cup_{j=-1}^{J_\star-1}\Lambda(j)$. Developing s in the Haar basis yields

$$s = \sum_{\lambda\in\Lambda}\beta_\lambda H_\lambda \quad \text{and} \quad s_{m^\star} = \sum_{\lambda\in\Lambda_{J_\star-1}}\beta_\lambda H_\lambda.$$

Moreover, we recall that for all $p \geq 1$ and all $0 < \alpha < 1$,

$$\sup_{j\geq 0} 2^{j(\alpha+1/2-1/p)}\left(\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^p\right)^{1/p} \leq C(\alpha,p)|s|_{B^\alpha_{p,\infty}} \tag{III.5.44}$$

(*cf.* [DL92], Section 2.3, for instance). Using classical inequalities between ℓ_p-norms and Inequality (III.5.44), the error in approximating s by $s_{m^\star}$ can be upper bounded as follows:

$$\begin{aligned}
\|s-s_{m^\star}\|^2 &= \sum_{j\geq J_\star}\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^2 \\
&\leq \sum_{j\geq J_\star}\left(|\Lambda(j)|^{(1/2-1/p)_+}\left(\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^p\right)^{1/p}\right)^2 \\
&\leq \sum_{j\geq J_\star}2^{-2j\left(\alpha-(1/2-1/p)_-\right)}\left(2^{j(\alpha+1/2-1/p)}\left(\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^p\right)^{1/p}\right)^2. \\
&\leq C(\alpha,p)\sum_{j\geq J_\star}2^{-2j\left(\alpha-(1/2-1/p)_-\right)}|s|^2_{B^\alpha_{p,\infty}} \\
&\leq C(\alpha,p)2^{-2J_\star\left(\alpha-(1/2-1/p)_-\right)}R^2.
\end{aligned} \tag{III.5.45}$$

Let us now prove that $s_{m^\star}$ belongs to the space C^α_p equipped with the semi-norm $|.|_{C^\alpha_p}$, as defined in [DY90], Section 3. The function $(s_{m^\star})^\sharp_{\alpha,p}$ (*cf.* [DY90]) is in fact piecewise constant over $m^\star$, hence

$$|s_{m^\star}|^p_{C^\alpha_p} = 2^{-J_\star}\sum_{i=1}^{2^{J_\star}}\big((s_{m^\star})^\sharp_{\alpha,p}(i2^{-J_\star})\big)^p.$$

Besides, that semi-norm can be linked with $|s|_{B^\alpha_{p,\infty}}$. Let us denote by v the $\mathbb{R}^{2^{J_\star}}$-vector whose i-th coordinate is $v_i = s_{m^\star}(i2^{-J_\star})$, for $i = 1, \dots, 2^{J_\star}$, and consider the $\mathbb{R}^{2^{J_\star}}$-vector $v^{\sharp,\alpha,p}$ defined as in [Aka09], Section 6.2. It can easily be checked that, for $i = 1, \dots, 2^{J_\star}$,

$$(s_{m^\star})^\sharp_{\alpha,p}(i2^{-J_\star}) = 2^{\alpha J_\star} v_i^{\sharp,\alpha,p},$$

so that

$$|s_{m^\star}|^p_{C^\alpha_p} = 2^{J_\star(\alpha p-1)} \sum_{i=1}^{2^{J_\star}} \big(v_i^{\sharp,\alpha,p}\big)^p. \tag{III.5.46}$$

Let us equip $\mathbb{R}^{2^{J_\star}}$ with the scalar product

$$\langle a, b\rangle = \sum_{i=1}^{2^{J_\star}} a_i b_i,$$

and denote by $\{h_\lambda\}_{\lambda\in\Lambda_{J_\star-1}}$ the Haar basis of $\mathbb{R}^{2^{J_\star}}$ orthonormal for that scalar product. For all $\lambda \in \Lambda_{J_\star-1}$ and all $1 \le i \le 2^{J_\star}$, the i-th coordinate of h_λ satisfies

$$h_{\lambda\, i} = 2^{-J_\star/2} H_\lambda(i2^{-J_\star}).$$

Therefore, decomposing v in the Haar basis of $\mathbb{R}^{2^{J_\star}}$ gives

$$v = 2^{J_\star/2} \sum_{\lambda\in\Lambda_{J_\star-1}} \beta_\lambda h_\lambda.$$

From Proposition 5 of [Aka09], that still holds whatever $p \ge 1$, we get

$$\sum_{i=1}^{2^{J_\star}} \big(v_i^{\sharp,\alpha,p}\big)^p \le C(\alpha,p) 2^{-J_\star(\alpha p-1)} \sum_{j=0}^{J_\star-1} 2^{jp(\alpha+1/2-1/p)} \sum_{\lambda\in\Lambda(j)} |\beta_\lambda|^p. \tag{III.5.47}$$

Combining Inequalities (III.5.46), (III.5.47) and (III.5.44) then leads to

$$|s_{m^\star}|_{C^\alpha_p} \le C(\alpha,p)|s|_{B^\alpha_{p,\infty}} \le C(\alpha,p)R.$$

Since $s_{m^\star} \in C^\alpha_p$, it follows from Corollary 3.2. of [DY90] that, for all $1 \le D \le 2^{J_\star}$, there exist a partition m_D of $[0,1]$ into D dyadic intervals and a function $t_D \in S_{m_D}$ such that

$$\|s_{m^\star} - t_D\|^2 \le C(\alpha,p) R^2 D^{-2\alpha}. \tag{III.5.48}$$

Moreover, since $s_{m^\star}$ is piecewise constant over $m^\star$, the partitions m_D, $1 \le D \le 2^{J_\star}$, provided by the approximation algorithm described in [DY90] belong to $\mathcal{M}^\star$. From Inequalities (III.5.41), (III.5.45) and (III.5.48), we deduce as previously that, for $1/n \le R^2 \le 2^{J_\star(1+2\alpha)}/n$,

$$\begin{aligned}
\inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2 + \frac{D_m}{n}\right\} &\le 2\|s-s_{m^\star}\|^2 + 2\inf_{1\le D\le 2^{J_\star}}\left\{\|s_{m^\star}-t_D\|^2 + \frac{D}{n}\right\}\\
&\le C(\alpha,p)\big(2^{-2J_\star(\alpha-(1/2-1/p)_-)}R^2 + \inf_{1\le D\le 2^{J_\star}}\left\{R^2D^{-2\alpha} + D/n\right\}\big)\\
&\le C(\alpha,p)\big(\ln^{2\delta}(n)n^{-1}\big)^{2(\alpha-(1/2-1/p)_-)}R^2 + (Rn^{-\alpha})^{2/(1+2\alpha)}\big).
\end{aligned}$$

Therefore, for $p \ge 2$ and $1/n \le R^2 \le (\ln(n))^{-2\delta(1+2\alpha)} n^{2\alpha}$,

$$\inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2 + \frac{D_m}{n}\right\} \le C(\alpha,p)(Rn^{-\alpha})^{2/(1+2\alpha)}.$$

For $1 \le p < 2$, two cases have to be distinguished since $(\ln^{2\delta}(n)n^{-1})^{2(\alpha+1/2-1/p)}R^2 \le (Rn^{-\alpha})^{2/(1+2\alpha)}$ if and only if $R^2 \le n^{q(\alpha,p)-1}/(\ln(n))^{2\delta q(\alpha,p)}$.

Chapter IV

Conditional density estimation based on dependent data

This chapter presents a work in collaboration with Claire Lacour.

IV.1 Introduction

In this chapter, we are concerned with conditional density estimation. Such a model is expected to bring more information than the regression model, and also contains classical density estimation as a special case. Although statisticians often consider independent observations, it seems more natural to assume that the observed quantities are realizations of dependent random variables. There are many ways to quantify this dependence, and we will mainly work with mixing and Markovian data. Besides, our aim is to provide a nonparametric estimator that adapts to the unknown and possibly inhomogeneous smoothness of the function, as well as to its possible anisotropy. Otherwise said, the risk being measured in a $\mathbb{L}_q$-norm, our estimator should perform well for functions whose smoothness is measured in a $\mathbb{L}_p$-norm, where p is allowed to be smaller than q, and whose degree of smoothness may vary with the direction.

Many works have already been devoted to adaptive density estimation for stationary dependent observations. Tribouley and Viennet [TV98] study an estimator based on wavelet thresholding for absolutely regular data. Comte and Merlevède [CM02] consider absolutely regular as well as strongly mixing stationary processes, in discrete or continuous time, and provide adaptive estimators based on model selection via least-squares penalization. Density estimation has also been studied in the framework of Markovian observations. A major work in this context is due to Clémençon [Clé00a]. He gives optimal rates of convergence for the estimation of the stationary density of a Markov chain, and proposes a wavelet based estimator reaching this bound up to a logarithmic factor. Recent works consider weakly dependent data instead of mixing data. Gannaz and Wintenberger [GW09] study a wavelet thresholding method for a wide class of weakly dependent data. Lerasle [Ler09] estimates the density from τ-dependent observations, as defined by [DP05], via model selection and gives oracle inequalities in probability. Among the aforementionned works, only [CM02], [Clé00a] and [GW09] can cope with inhomogeneous smoothness, and these procedures are then optimal only up to a logarithmic factor. Even in density estimation based on independent data, few methods can do so. Let us cite block thresholding methods after [HKP98], for instance. Recently, three papers have proposed model selection procedures based on penalized minimum contrasts that consist in selecting from the data a best piecewise polynomial built on a partition into dyadic cubes or rectangles. Willett and Nowak [WN07] select best piecewise polynomials built on partitions into dyadic cubes via a penalized maximum likelihood contrast. Klemelä [Kle09] and Blanchard *et al.* [BSRM07] select best histograms based on partitions into dyadic rectangles via a penalized criterion based either on the $\mathbb{L}_2$ distance or on Kullback-Leibler divergence. Only the first two papers prove adaptivity results on inhomogeneous smoothness classes, both up to a logarithmic factor, [Kle09] being also able to adapt to anisotropy.

References about conditional density estimation are fewer, even for nonadaptive procedures based on independent data. Let us begin with some examples of nonadaptive estimators. For independent data, we can cite for instance Györfi and Kohler [GK07] for a histogram based procedure, or Faugeras [Fau07] for a copula-based kernel estimator. For dependent data, let us mention De Gooijer and Zerom [DGZ03] or Fan and Yim [FY04] for kernel methods. We refer to [Lac07] for a bibliography about nonadaptive estimation for the transition density of a Markov chain. Regarding now adaptive estimation of conditional density based on independent data, three recent papers give oracle inequalities and adaptivity results in the minimax sense for functions with anisotropic but homogeneous smoothness. Efromovich [Efr07; Efr08] uses a Fourier decomposition to build a blockwise-shrinkage Efromovich-Pinsker estimator, whereas Brunel *et al.* [BCL07] perform model selection based on a penalized least-squares criterion. Up to our knowledge, only two papers study adaptive estimators of the conditional density based on dependent data, and both are only concerned with Markovian observations: Clémençon [Clé00b] and Lacour [Lac07].

In this chapter, we lead a simultaneous study of conditional density and density estimation. The last problem, already widely studied, will serve as a benchmark to measure the performance of our procedure. Given a collection of partitions, we propose an estimator of the conditional density that selects from the data the best partition among that collection and then provides the best piecewise polynomial estimator built on that partition. To do so, we use a penalized least-squares criterion. To deal with the possible dependence of the observations, we mainly use β-mixing coefficients and their coupling properties. Thus, our dependence assumptions, while being satisfied by a wide class of Markov chains, are not restricted to Markovian assumptions. We first provide nonasymptotic oracle type inequalities fulfilled by any collection of partitions satisfying some mild structural conditions. We then consider the collections of partitions into dyadic cubes and dyadic rectangles, as in [WN07] or [Kle09; BSRM07]. In those two cases, we obtain oracle-type inequalities and adaptivity results in the minimax sense over a wide range of Besov smoothness classes, without logarithmic factor. Those classes contain functions with inhomogeneous smoothness, isotropic for the first collection, and possibly anisotropic for the second. For the collection of partitions into dyadic cubes, adaptivity relies on an approximation result due to Devore and Yu [DY90], not used in the aforementioned references. For the collection of partitions into dyadic rectangles, we prove an approximation result for functions with anisotropic smoothness inspired from [DY90], which is of independent interest. Moreover, determining in practice the penalized estimator based on that collection only requires a computational complexity linear in the size of the sample, up to a logarithmic factor.

The chapter is organized as follows. We begin by describing the framework and the estimation procedure, and present in Section IV.3 the dependence measures used in the sequel. In Section IV.4, estimation on one model is detailed. This study allows to understand the role of the dependence assumptions in our framework and what bound for the $\mathbb{L}_2$-risk we seek to obtain. The choice of a penalty yielding an oracle-type inequality is the topic of Section IV.5. Sections IV.6 and IV.7 are devoted to the collections of partitions into dyadic cubes and dyadic rectangles, and adaptivity results are proved for the associated penalized estimator. Most proofs are deferred to Section IV.8.

IV.2 General framework and estimation procedure

Let $\{Z_i\}_{i\in\mathbb{Z}} = \{(X_i, Y_i)\}_{i\in\mathbb{Z}}$ be a strictly stationary process defined on the measurable space $(\Omega, \mathcal{F}, \mathbb{P})$, where, for all $i \in \mathbb{Z}$, X_i and Y_i take values respectively in $[0,1]^{d_1}$ and $[0,1]^{d_2}$, with $d_1 \in \mathbb{N}$ and $d_2 \in \mathbb{N}^\star$. We set $[0,1]^{d_1} = \{0\}$ when $d_1 = 0$. Given some integer $n \geq 2$, our aim is to estimate, on the basis of the observation of $(Z_1, \ldots, Z_n)$, the marginal density s of Y_i conditionally to X_i. Let us mention two special cases of interest. If $d_1 = 0$, which amounts to no conditioning, then s is just the marginal density of Y_i. The expression "density estimation" will refer to that case. If $(X_i)_{i\in\mathbb{Z}}$ is a homogeneous Markov chain of order 1, and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then s is the transition density of the chain $(X_i)_{i\in\mathbb{Z}}$. When $d_1 \in \mathbb{N}^\star$, we assume that the conditioning variables $(X_i)_{i\in\mathbb{Z}}$ admit a bounded marginal density f with respect to the Lebesgue measure. When $d_1 = 0$, we set f equal to 1.

Let us introduce some standard notation. For any real-valued function t defined and bounded on some set $\mathcal{D}$, we set

$$\iota(t) = \inf_{x\in\mathcal{D}} |t(x)| \quad \text{and} \quad \|t\|_\infty = \sup_{x\in\mathcal{D}} |t(x)|.$$

For $k \in \mathbb{N}$, we denote by μ_k the Lebesgue measure on $[0,1]^k$ if k is positive, and the Dirac measure at 0 otherwise. We denote by $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ the set of all real-valued functions which are square integrable with respect to $\mu_{d_1} \otimes \mu_{d_2}$, and by $\langle ., . \rangle$ and $\|.\|$ the usual

scalar product and norm on $\mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$. Since f is bounded, we can also define on $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ the scalar product

$$\langle t, u \rangle_f = \int_{[0,1]^{d_1} \times [0,1]^{d_2}} t(x,y)u(x,y)f(x)\mu_{d_1}(\mathrm{d}x)\mu_{d_2}(\mathrm{d}y)$$

and the associated norm $\|.\|_f$. When $d_1 = 0$ (*i.e.* $[0,1]^{d_1} = \{0\}$), $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ is identified with $\mathbb{L}_2\left([0,1]^{d_2}\right)$ and $\|.\|_f$ coincides with $\|.\|$.

We consider the empirical criterion γ inspired from [BCL07] and defined on $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ by

$$\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{[0,1]^{d_2}} t^2(X_i, y)\mu_{d_2}(\mathrm{d}y) - 2t(X_i, Y_i)\right].$$

It satisfies, for all $t \in \mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$,

$$\mathbb{E}_s[\gamma(t) - \gamma(s)] = \|s - t\|_f^2,$$

so that s minimizes $t \mapsto \mathbb{E}_s[\gamma(t)]$. Given the aforementioned conventions, when $d_1 = 0$, *i.e.* for estimating the marginal density of $(Y_i)_{i\in\mathbb{Z}}$, we recover the usual least-squares criterion

$$\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{[0,1]^{d_2}} t^2(y)\mu_{d_2}(\mathrm{d}y) - 2t(Y_i)\right].$$

When $(X_i)_{i\in\mathbb{Z}}$ is a Markov chain and $Y_i = X_{i+1}$, we recover the contrast introduced in [Lac07] for transition density estimation,

$$\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{[0,1]^{d_2}} t^2(X_i, y)\mu_{d_2}(\mathrm{d}y) - 2t(X_i, X_{i+1})\right].$$

Let us now fix some nonnegative integer r. For a partition m of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles, we denote by S_m the space of all real-valued piecewise polynomial functions on $[0,1]^{d_1} \times [0,1]^{d_2}$ which are polynomial with coordinate degree $\leq r$ on each rectangle of m. We define a best estimator of s in the model S_m by setting

$$\hat{s}_m = \operatorname*{argmin}_{t\in S_m} \gamma(t).$$

We give ourselves a finite collection $\mathcal{M}$ of partitions of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles. Then, in view of choosing from the data the best estimator among the collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$, we consider the random selection procedure

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \{\gamma(\hat{s}_m) + \mathrm{pen}(m)\}$$

and the penalized estimator

$$\tilde{s} = \hat{s}_{\hat{m}},$$

where $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ is a so-called penalty function that remains to be chosen so that $\tilde{s}$ performs well.

In order to provide useful characterizations for $\hat{s}_m$ and $\hat{m}$ in practice, we need to introduce some adequate basis of each S_m, for $m \in \mathcal{M}$. We set $d = d_1 + d_2$. Assume that $d_1 \in \mathbb{N}^\star$. For K_1 rectangle in $[0,1]^{d_1}$, let $(\phi_{K_1,k_1})_{k_1\in\{0,\ldots,r\}^{d_1}}$ be a basis of the space of piecewise polynomial functions with support K_1 and coordinate degree $\leq r$, which is orthonormal for the norm $\|.\|$.

In the same way, we choose for each rectangle $K_2 \subset [0,1]^{d_2}$ a basis $(\psi_{K_2,k_2})_{k_2 \in \{0,\dots,r\}^{d_2}}$ of the space of piecewise polynomial functions with support K_2 and coordinate degree $\leq r$, which is orthonormal for the norm $\|.\|$. For K rectangle in $[0,1]^d$, we shall denote by K_1 and K_2 the rectangles in $[0,1]^{d_1}$ and $[0,1]^{d_2}$ such that $K = K_1 \times K_2$. For $k \in \{0,\dots,r\}^d$, we shall denote by k_1 and k_2 the multi-indices in $\{0,\dots,r\}^{d_1}$ and $\{0,\dots,r\}^{d_2}$ such that $k = (k_1, k_2)$. For any rectangle $K \in [0,1]^d$ and any multi-index $k \in \{0,\dots,r\}^d$, we then define $\Phi_{K,k}$ by

$$\Phi_{K,k}(x,y) = \phi_{K_1,k_1}(x)\psi_{K_2,k_2}(y)$$

for $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$. Thus, for a partition m of $[0,1]^d$ into rectangles, the family $(\Phi_{K,k})_{K \in m, k \in \{0,\dots,r\}^d}$ is a basis of S_m, orthonormal for the norm $\|.\|$. We will mention no index k when $r = 0$ and, when $d_1 = 0$, we will identify the rectangle K with K_2, the index k with k_2, ϕ_{K_1,k_1} with the constant 1 and $\Phi_{K,k}$ with ψ_{K_2,k_2}. We denote by

$$\hat{s}_m = \sum_{K \in m} \sum_{k \in \{0,\dots,r\}^d} \hat{a}_{K,k}\Phi_{K,k}$$

the decomposition of $\hat{s}_m$ in the basis $(\Phi_{K,k})_{K \in m, k \in \{0,\dots,r\}^d}$. For all $K \in m$, we define the matrices

$$A_K = \left(\hat{a}_{K,(k_1,k_2)}\right)_{(k_1,k_2) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_2}},$$

$$\Upsilon_K = \left(\frac{1}{n}\sum_{i=1}^n \phi_{K_1,k_1}(X_i)\psi_{K_2,k_2}(Y_i)\right)_{(k_1,k_2) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_2}},$$

and

$$G_{K_1} = \left(\frac{1}{n}\sum_{i=1}^n \phi_{K_1,k_1}(X_i)\phi_{K_1,l_1}(X_i)\right)_{(k_1,l_1) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_1}}.$$

Since ϕ_{K_1,k_1} and ϕ_{L_1,l_1} (resp. ψ_{K_2,k_2} and ψ_{L_2,l_2}) have disjoint supports when $K_1 \neq L_1$ (resp. $K_2 \neq L_2$) and $(\psi_{K_2,k_2})_{k_2 \in \{0,\dots,r\}^{d_2}}$ is orthonormal, we obtain after some computation that, for all $K \in m$, A_K is given by

$$G_{K_1}A_K = \Upsilon_K. \tag{IV.2.1}$$

Let us mention some special cases where we obtain explicit (and familiar) expressions for $\hat{s}_m$. In the case of density estimation, *i.e.* when $d_1 = 0$, we recover the projection estimator over S_m

$$\hat{s}_m = \sum_{K \in m} \sum_{k \in \{0,\dots,r\}^d} \left(\frac{1}{n}\sum_{i=1}^n \Phi_{K,k}(Y_i)\right)\Phi_{K,k}$$

and thus, when $r = 0$, the usual histogram

$$\hat{s}_m = \sum_{K \in m} \left(\frac{1}{n\mu_d(K)}\sum_{i=1}^n \mathbb{1}_K(Y_i)\right)\mathbb{1}_K.$$

For conditional density estimation with $d_1 \neq 0$ and when $r = 0$, we set, for all rectangle K,

$$\hat{s}_m\mathbb{1}_K = \frac{1}{\mu_{d_2}(K_2)\sum_{i=1}^n \mathbb{1}_{K_1}(X_i)}\sum_{i=1}^n \mathbb{1}_K(Z_i) \quad \text{if some } X_i \in K_1,$$

and $\hat{s}_m\mathbb{1}_K = 0$ otherwise.

IV.3 Measures of dependence

Let us introduce the notions of dependence used in the sequel. For two sub-σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, the α-mixing (or strong mixing) coefficient between $\mathcal{A}$ and $\mathcal{B}$ is defined by

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|\,,$$

the β-mixing (or absolute regularity) coefficient by

$$\beta(\mathcal{A},\mathcal{B}) = \mathbb{E}\left[\sup_{B\in\mathcal{B}} |\mathbb{P}(B|\mathcal{A}) - \mathbb{P}(B)|\right],$$

and the ρ-mixing (or maximal correlation) coefficient by

$$\rho(\mathcal{A},\mathcal{B}) = \sup_{X,Y} \frac{|\mathrm{Cov}(X,Y)|}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

where the supremum is taken with respect to all real-valued random variables X and Y that are respectively $\mathcal{A}$ and $\mathcal{B}$-measurable and square integrable. We recall that α, β and ρ-mixing are among the weakest forms of mixing conditions, in the sense that both β and ρ-mixing are implied by ϕ-mixing (uniform mixing) and imply α-mixing (see for instance [Dou94]). In particular,

$$0 \le 2\alpha(\mathcal{A},\mathcal{B}) \le \beta(\mathcal{A},\mathcal{B}) \quad \text{and} \quad 0 \le 4\alpha(\mathcal{A},\mathcal{B}) \le \rho(\mathcal{A},\mathcal{B}). \tag{IV.3.2}$$

Besides, in general, ρ-mixing does not imply β-mixing, and β-mixing does not imply ρ-mixing.

In the sequel, the letter θ stands for α, β or ρ. For all $j \in \mathbb{N}^\star$, let

$$\theta_j^{\boldsymbol{Z}} = \theta\left(\sigma(Z_i, i\le 0), \sigma(Z_i, i\ge j)\right).$$

The process $(Z_i)_{i\in\mathbb{Z}}$ is said to be θ-mixing when $\lim_{j\to+\infty}\theta_j^{\boldsymbol{Z}} = 0$. In particular, $(Z_i)_{i\in\mathbb{Z}}$ is geometrically θ-mixing with rate b, $b>0$, if there exists a positive constant a such that, for all $j\in\mathbb{N}^\star$,

$$\theta_j^{\boldsymbol{Z}} \le a\exp(-bj).$$

We shall also use the 2-mixing coefficients $\theta(\sigma(Z_0),\sigma(Z_j))$, that satisfy, for all $j\in\mathbb{N}^\star$,

$$\theta(\sigma(Z_0),\sigma(Z_j)) \le \theta_j^{\boldsymbol{Z}} \tag{IV.3.3}$$

and, if $(Z_i)_{i\in\mathbb{Z}}$ is a Markov chain,

$$\theta(\sigma(Z_0),\sigma(Z_j)) = \theta_j^{\boldsymbol{Z}}. \tag{IV.3.4}$$

Let us give sufficient conditions for $(Z_i)_{i\in\mathbb{Z}}$ to be θ-mixing. First, if $(X_i)_{i\in\mathbb{Z}}$ is a strictly stationary θ-mixing process, and $Y_i = X_{i+1}$ for all $i\in\mathbb{Z}$, then $(Z_i)_{i\in\mathbb{Z}}$ also is θ-mixing since, for all $j\ge 2$,

$$\theta_j^{\boldsymbol{Z}} = \theta_{j-1}^{\boldsymbol{X}}. \tag{IV.3.5}$$

In the sequel, we will mainly be concerned with mixing assumptions possibly involving ρ-mixing and β-mixing or ρ-mixing and α-mixing at the same time. Under adequate hypotheses, Markov chains (always assumed to be homogeneous of order 1) provide examples of such processes.

Example (M1): If $(Z_i)_{i\in\mathbb{Z}}$ is a Markov chain, then $(Z_i)_{i\in\mathbb{Z}}$ is geometrically ρ-mixing if and only if it is ρ-mixing (in fact, $(Z_i)_{i\in\mathbb{Z}}$ is geometrically ρ-mixing as soon as $\rho_j^{\boldsymbol{Z}} < 1$ for some

positive integer j) (*cf.* [Bra05], Theorem 3.3). Up to our knowledge, there is no such property for the β-mixing coefficients.

Example (M2): If $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary Harris ergodic Markov chain (aperiodic, irreducible, positive Harris recurrent), then $(Z_i)_{i\in\mathbb{Z}}$ is geometrically β-mixing if and only if it is geometrically ergodic (*cf.* [Bra05], Theorem 3.7).

Example (M3): If $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary Harris ergodic Markov chain that is also reversible and geometrically ergodic, then $(Z_i)_{i\in\mathbb{Z}}$ is both geometrically ρ-mixing and geometrically β-mixing (*cf.* [Jon04], Theorem 2).

Example (M4): If $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary, ergodic and aperiodic Markov chain satisfying the Doeblin condition, then $(Z_i)_{i\in\mathbb{Z}}$ is uniformly ergodic, hence both geometrically ρ-mixing and geometrically β-mixing (*cf.* [Bra05], 119–121, or [MT93], Section 16.2).

We refer to [DG83; Mok90; DT93; Dou94; AN98] for examples of stationary processes that are geometrically β-mixing or both geometrically β and ρ-mixing among commonly used time series such as nonlinear ARMA or nonlinear ARCH models.

IV.4 Upper-bounds for the risk on one model

In this section, we review dependence assumptions allowing to recover an upper-bound for the risk of each estimator $\hat{s}_m$ similar to the one obtained in the independent case, up to a constant or logarithmic factor. We recall that $d = d_1 + d_2$. We also use the following notation for each partition m of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles. We denote by $|m|$ the number of rectangles in m and by D_m the dimension of S_m, so that $D_m = (r+1)^d|m|$. The notation s_m stands for the orthogonal projection of s on S_m for the norm $\|.\|$. Besides, we say that the partition m is regular when all its rectangles have the same Lebesgue measure, *i.e.* when for all $K \in m$, $\mu_d(K) = 1/|m|$.

We consider the following dependence assumptions. Except for the last one, they are related to some rate of mixing. Some only involve 2-mixing coefficients or also impose conditions on the partition m.

Assumption ($D\alpha(m)$) *For all $K \in m$, $\mu_d(K) \geq 1/n$, and $(Z_i)_{i\in\mathbb{Z}}$ is geometrically α-mixing, with $a \geq 0$ and $b > 0$ such that, for all $j \in \mathbb{N}^\star$,*

$$\alpha_j^{\boldsymbol{Z}} \leq a\exp(-bj).$$

Assumption ($D2$-$\alpha(m)$) *The partition m is regular and the series $S_{2\text{-}\alpha} := \sum_{j\in\mathbb{N}^\star} \alpha(\sigma(Z_1), \sigma(Z_{j+1}))$ converges.*

Assumption ($D\rho$) *The series $S_\rho := \sum_{j\in\mathbb{N}} \rho_{2^j}^{\boldsymbol{Z}}$ converges.*

Assumption ($D2$-ρ) *The series* $S_{2\text{-}\rho} := \sum_{j\in\mathbb{N}^\star} \rho(\sigma(Z_1), \sigma(Z_{j+1}))$ *converges.*

Assumption (D_{cond}) *For all* $j \geq 2$, Z_j *is independent of* Z_1 *conditionally to* X_j.

Assumptions **($D2$-$\alpha(m)$)** and **($D2$-ρ)** are satisfied for instance if the involved 2-mixing coefficient decays as j^{-b}, with $b > 1$, whereas Assumption **($D\rho$)** is satisfied for instance when $\rho_j^{\boldsymbol{Z}}$ decays as $\ln^{-b}(j)$, with $b > 1$. In special cases, there are links between some of those dependence assumptions. For a regular partition m, it follows from (IV.3.3) and (IV.3.2) that

$$(D\alpha(m)) \Rightarrow (D2\text{-}\alpha(m)) \quad \text{and} \quad (D2\text{-}\rho) \Rightarrow (D2\text{-}\alpha(m)).$$

If $(Z_i)_{i\in\mathbb{Z}}$ is a Markov chain, then according to Example **($M1$)**, (IV.3.3) and (IV.3.2)

$$(D2\text{-}\rho) \Leftrightarrow (D\rho) \Rightarrow \begin{cases} (D\alpha(m)) \text{ if } \min_{K\in m} \mu_d(K) \geq 1/n \\ (D2\text{-}\alpha(m)) \text{ if } m \text{ is regular.} \end{cases}$$

On the other hand, Assumption **(D_{cond})** does not imply any of the others. For instance, if $(X_i)_{i\in\mathbb{Z}}$ is a Markov chain and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then Assumption **(D_{cond})** is satisfied, but $(X_i)_{i\in\mathbb{Z}}$, and therefore $(Z_i)_{i\in\mathbb{Z}}$, can be chosen non mixing. (*cf.* [DP05] for instance).

Let us first give some upper-bounds for the risk of $\hat{s}_m$ in the density estimation problem, which is already a widely studied problem under various dependence assumptions, and thus will serve as a benchmark. In that case, $(Z_i)_{i\in\mathbb{Z}}$ can be identified with $(Y_i)_{i\in\mathbb{Z}}$. We recall that when $(Y_i)_{i\in\mathbb{Z}}$ is a sequence of independent and identically distributed random variables,

$$\mathbb{E}_s\left[\|s - \hat{s}_m\|^2\right] \leq \|s - s_m\|^2 + \mathcal{N}(s)(2r+1)^d \frac{D_m}{n}, \tag{IV.4.6}$$

where $\mathcal{N}(s) = 1$ if the partition m is regular, and $\mathcal{N}(s) = \|s\|_\infty$ otherwise. A major interest of such a bound is that it allows to prove the optimality of $\hat{s}_m$ from the minimax point of view for a well-chosen partition m. Roughly speaking, when s has regularity σ measured in a $\mathbb{L}_p$-norm with $p \geq 2$, the bias term $\|s - s_m\|^2$ is at most of order $D_m^{-2\sigma/d}$ for any regular partition m. For a regular partition m realizing a good compromise between the bias and the variance terms, *i.e.* such that $D_m^{-2\sigma/d}$ and D_m/n are of the same order, the estimator $\hat{s}_m$ then reaches the optimal estimation rate $n^{-2\sigma/(2\sigma+d)}$. According to Proposition 17 below, that result still holds if $(Y_i)_{i\in\mathbb{Z}}$ only satisfies the α-mixing assumption **($D2$-$\alpha(m)$)**. But when s has regularity σ measured in a $\mathbb{L}_p$-norm with $p < 2$, one can only ensure that the bias term $\|s - s_m\|^2$ is at most of order $D_m^{-2\sigma/d}$ for some possibly *irregular* partitions m (see for instance Sections IV.6 and IV.7). In the independent case, a well chosen irregular partition m such that D_m is of order $n^{d/(2\sigma+d)}$ will still provide an estimator $\hat{s}_m$ reaching the optimal estimation rate $n^{-2\sigma/(2\sigma+d)}$. According to Proposition 17, that result still holds under Assumptions **($D\rho$)** or **($D2$-ρ)**, and also, up to a logarithmic factor, under Assumption **($D\alpha(m)$)**.

Proposition 17 *Assume that* $d_1 = 0$, *so that* s *is the marginal density of* $(Y_i)_{i\in\mathbb{Z}}$. *Let* m *be a partition of* $[0,1]^d$. *Let* $\mathcal{N}(s) = 1$ *if the partition* m *is regular, and* $\mathcal{N}(s) = \|s\|_\infty$ *otherwise. If one of the assumptions* **($D\alpha(m)$)**, **($D2$-$\alpha(m)$)**, **($D\rho$)**, **($D2$-ρ)** *is satisfied, then there exist* δ, C_1 *and* C_2 *nonnegative reals that do not depend on* n *such that*

$$\mathbb{E}_s\left[\|s - \hat{s}_m\|^2\right] \leq \|s - s_m\|^2 + C_1\mathcal{N}(s)(1 + \ln(n^\delta))\frac{D_m}{n} + C_2\frac{D_m}{n^2}. \tag{IV.4.7}$$

Suitable values for δ, C_1 *and* C_2 *are*

- $\delta = 1/b$, $C_1 = 9(2r+1)^d$ *and* $C_2 = 12(2r+1)^d a$ *under Assumption* **($D\alpha(m)$)** ;

- $\delta = 0$, $C_1 = 2(1+2S_{2\text{-}\alpha})(r+1)^d(2r+1)^d$ *and* $C_2 = 0$ *under Assumption* $(\boldsymbol{D2\text{-}\alpha(m)})$;
- $\delta = 0$, $C_1 = 250\prod_{j=0}^{\lfloor \log_2 n \rfloor}\left(1+\rho^{\boldsymbol{Y}}_{\lfloor 2^{j/3}\rfloor+1}\right)$ *and* $C_2 = 0$ *under Assumption* $(\boldsymbol{D\rho})$;
- $\delta = 0$, $C_1 = (1+2S_{2\text{-}\rho})(2r+1)^d$ *and* $C_2 = 0$ *under Assumption* $(\boldsymbol{D2\text{-}\rho})$.

Notice that Assumption $(\boldsymbol{D_{cond}})$ amounts in that case to $(Y_i)_{i\in\mathbb{Z}}$ being independent, case already covered by Assumption $(\boldsymbol{D2\text{-}\rho})$ with $S_{2\text{-}\rho} = 0$. Up to our knowledge, under Assumption $(\boldsymbol{D\alpha(m)})$, the bound (IV.4.7) does not appear in the literature, but it relies on a classical coupling property. Under Assumption $(\boldsymbol{D2\text{-}\alpha(m)})$, such a bound is proved for instance in [Rio00]. Assumption $(\boldsymbol{D\rho})$ is classical for proving Central Limit Theorems, and in that case, we use an inequality for the variance of partial sums first proved by [Pel82] (Lemma 3.4). Under Assumption $(\boldsymbol{D2\text{-}\rho})$, Inequality (IV.4.7) results immediately from the definition of the ρ-mixing coefficients.

Proof: Let us consider the orthnormal basis $(\Phi_{K,k})_{\substack{K\in m\\ k\in\{0,\dots,r\}^d}}$ defined from the Legendre polynomials as explained in Section IV.8.1. From Pythagoras' equality and the expressions of $\hat{s}_m$ and s_m in the basis $(\Phi_{K,k})_{\substack{K\in m\\ k\in\{0,\dots,r\}^d}}$, we get

$$\begin{aligned}\mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] &= \|s-s_m\|^2 + \mathbb{E}_s\left[\|\hat{s}_m - s_m\|^2\right]\\ &= \|s-s_m\|^2 + \frac{1}{n}\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d}\frac{1}{n}\mathrm{Var}_s\left(\sum_{i=1}^n \Phi_{K,k}(Y_i)\right).\end{aligned}$$

Under Assumption $(\boldsymbol{D\alpha(m)})$, we apply Inequality (IV.8.54) in Lemma 9 to each $\Phi_{K,k}$, $K\in m$, $k\in\{0,\dots,r\}^d$ and with $q_n = \lceil 3\ln(n^{1/b})\rceil$. Since $\min_{K\in m}\mu_d(K) \geq 1/n$ and $\|\Phi_{K,k}\|_\infty^2 \leq (2r+1)^d/\mu_d(K)$ (*cf.* Section IV.8.1), we get

$$n\mathbb{E}_s\left[\|\hat{s}_m - s_m\|^2\right] \leq 3(1+3\ln(n^{1/b}))\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d}\mathrm{Var}_s\big(\Phi_{K,k}(Y_1)\big) + 12(2r+1)^d a\frac{D_m}{n}.$$

Under Assumption $(\boldsymbol{D2\text{-}\alpha(m)})$, we apply Inequality (IV.8.55) and use the bound

$$\Phi_{K,k}^2 \leq (2r+1)^d|m|\mathbb{1}_K, \text{ for all } K\in m, k\in\{0,\dots,r\}^d, \tag{IV.4.8}$$

that results from the definition of those functions and the regularity of the partition m. We thus obtain

$$\begin{aligned}n\mathbb{E}_s\left[\|\hat{s}_m - s_m\|^2\right] &\leq 2(1+2S_{2\text{-}\alpha})\max_{K\in m}\left\|\sum_{k\in\{0,\dots,r\}^d}|\Phi_{K,k}|\right\|_\infty^2\\ &\leq 2(1+2S_{2\text{-}\alpha})(2r+1)^d(r+1)^{2d}|m|.\end{aligned}$$

Under Assumption $(\boldsymbol{D\rho})$, Lemma 8.15 in [Bra07] provides, for all $K\in m$ and $k\in\{0,\dots,r\}^d$,

$$\frac{1}{n}\mathrm{Var}_s\left(\sum_{i=1}^n\Phi_{K,k}(Y_i)\right) \leq C_1\mathrm{Var}_s\left(\Phi_{K,k}(Y_1)\right) \tag{IV.4.9}$$

with $C_1 = 250\prod_{j=0}^{\lfloor\log_2 n\rfloor}\left(1+\rho^{\boldsymbol{Y}}_{\lfloor 2^{j/3}\rfloor+1}\right)$. Under Assumption $(\boldsymbol{D2\text{-}\rho})$, we immediately deduce from the definition of the ρ-mixing coefficients and the stationarity of $(Y_i)_{i\in\mathbb{Z}}$ that, for all $K\in m$, $k\in\{0,\dots,r\}^d$ and $1\leq j\leq n-1$,

$$\left|\mathrm{Cov}_s\big(\Phi_{K,k}(Y_1),\Phi_{K,k}(Y_{j+1})\big)\right| \leq \rho(\sigma(Y_1),\sigma(Y_{j+1}))\mathrm{Var}_s\left(\Phi_{K,k}(Y_1)\right).$$

Thus,

$$\frac{1}{n}\mathrm{Var}_s\left(\sum_{i=1}^{n}\Phi_{K,k}(Y_i)\right) = \mathrm{Var}_s\big(\Phi_{K,k}(Y_1)\big) + 2\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\mathrm{Cov}_s\big(\Phi_{K,k}(Y_1),\Phi_{K,k}(Y_{j+1})\big)$$
$$\leq \left(1+2\sum_{j=1}^{n-1}\rho(\sigma(Y_1),\sigma(Y_{j+1}))\right)\mathrm{Var}_s\left(\Phi_{K,k}(Y_1)\right). \qquad \text{(IV.4.10)}$$

Last, for all $K \in m$ and $k \in \{0,\dots,r\}^d$, $\Phi_{K,k}^2 \leq (2r+1)^d \mathbb{1}_K/\mu_d(K)$, so

$$\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d}\mathrm{Var}_s\big(\Phi_{K,k}(Y_1)\big) \leq (2r+1)^d(r+1)^d\sum_{K\in m}\frac{\langle s,\mathbb{1}_K\rangle}{\mu_d(K)} \leq (2r+1)^d\mathcal{N}(s)D_m,$$

which completes the proof. ■

Remarks:

- For a regular partition m, the underlying property of the space S_m that allows to keep the variance term $\mathbb{E}_s\left[\|\hat{s}_m - s_m\|^2\right]$ of the same order as in the independent case under Assumption **(D2-α(m))** is

$$\|t\|_\infty^2 \leq C(r,d)D_m\|t\|^2, \text{ for all } t \in S_m, \qquad \text{(IV.4.11)}$$

  where $C(r,d)$ only depends on r and d (*cf.* (IV.4.8)). That property is also used in [Vie97] to obtain the estimation rate of $\hat{s}_m$ for a regular partition m (and more generally of projection estimators) under a β-mixing condition stronger than **(D2-α(m))**. But Inequality (IV.4.11) does not hold in general for an irregular partition.

- As shown by (IV.4.9) and (IV.4.10), a sufficient condition to ensure that $\mathbb{E}_s\left[\|\hat{s}_m - s_m\|^2\right]$ is of the same order as in the independent case is that for some constant C and all $t \in S_m$,

$$\mathrm{Var}\left(\sum_{i=1}^{n}t(Y_i)\right) \leq Cn\mathrm{Var}\left(t(Y_1)\right). \qquad \text{(IV.4.12)}$$

  Assumptions **(Dρ)** and **(D2-ρ)** are optimal for obtaining such an inequality in the following sense. Let us assume that $(Y_i)_{i\in\mathbb{N}}$ is a strictly stationary Harris ergodic and reversible Markov chain satisfying (IV.4.12) for all real-valued function t defined on $[0,1]^d$. Then the chain is variance bounding in the sense of [RR08], which implies that there is a spectral gap in $\mathbb{L}_2(s) := \{t : [0,1]^d \to \mathbb{R} \text{ s.t. } \langle t,s\rangle = 0 \text{ and } \|t\|_s < \infty\}$ (Theorem 14 in [RR08]). This leads to the geometrical ergodicity of the chain (Theorem 2.1 in [RR97]), which, given the reversibility assumption, implies that the chain is ρ-mixing (*cf.* Example **(M3)**). As a conclusion, a strictly stationary Harris ergodic and reversible Markov chain $(Y_i)_{i\in\mathbb{Z}}$ satisfies (IV.4.12) for all real-valued function t defined on $[0,1]^d$ if and only if it is ρ-mixing.

- Another assumption of dependence is used for instance by [Bos98] (Theorem 2.1) to prove that, asymptotically, the quadratic risk of kernel density estimators reaches the minimax rate (see also [CM02]).

  Assumption (Lip) *For all $j \in \mathbb{N}^\star$, (Y_1, Y_{j+1}) admits a density s_j with respect to the Lebesgue measure on $[0,1]^d \times [0,1]^d$, and $g_j(y_1,y_2) = s_j(y_1,y_2) - s(y_1)s(y_2)$ satisfies*

  $$|g_j(y_1,y_2) - g_j(y_1',y_2')| \leq \ell\|y-y'\|_{\mathbb{R}^{2d}}, \text{ for all } y=(y_1,y_2), y'=(y_1',y_2') \in [0,1]^d\times[0,1]^d,$$

for some nonnegative real ℓ. Moreover, there exist $a \geq 0$ and $b > 0$ such that, for all $j \in \mathbb{N}^\star$,

$$\alpha(\sigma(Y_1), \sigma(Y_{j+1})) \leq a j^{-b}.$$

Then, according to [Bos98] (Lemma 1.3),

$$\|g_j\|_\infty \leq \left(V_d^{-2} + \ell\sqrt{2}\right) \alpha(\sigma(Y_1), \sigma(Y_{j+1}))^{1/(2d+1)},$$

where V_d denotes the volume of the unit ball in $\mathbb{R}^d$. Therefore, for all measurable, real-valued and bounded functions t, u defined on $[0,1]^d$,

$$|\mathrm{Cov}_s\left(t(Y_1), u(Y_{j+1})\right)| \leq \left(V_d^{-2} + \ell\sqrt{2}\right) \alpha(\sigma(Y_1), \sigma(Y_{j+1}))^{1/(2d+1)} \left(\int_{[0,1]^d} |t|\right) \left(\int_{[0,1]^d} |u|\right). \tag{IV.4.13}$$

But, according to [Bra07] (Theorem 4.4), the usual ψ-mixing coefficient $\psi(\mathcal{A}, \mathcal{B})$ between two sub σ-fields $\mathcal{A}, \mathcal{B}$ of $\mathcal{F}$ satisfies

$$\psi(\mathcal{A}, \mathcal{B}) = \sup_{U,V} \frac{|\mathrm{Cov}(U, V)|}{\mathbb{E}[|U|]\mathbb{E}[|V|]}$$

where the supremum is taken over all real-valued and bounded random variables U and V that are respectively $\mathcal{A}$ and $\mathcal{B}$- measurable. Therefore, if $\iota(s) > 0$, then (IV.4.13) and the usual inequality between $\rho(\mathcal{A}, \mathcal{B})$ and $\psi(\mathcal{A}, \mathcal{B})$ (see for instance [Dou94]) imply

$$\rho(\sigma(Y_1), \sigma(Y_{j+1})) \leq \psi(\sigma(Y_1), \sigma(Y_{j+1})) \leq \left(V_d^{-2} + \ell\sqrt{2}\right) \iota^{-2}(s) \alpha(\sigma(Y_1), \sigma(Y_{j+1}))^{1/(2d+1)}.$$

Therefore, when s is bounded from below by a positive constant, Assumption **(Lip)** is much than stronger than Assumption $\boldsymbol{(D2\text{-}\rho)}$, which is enough for obtaining the optimal estimation rate from the minimax point of view.

Let us now consider conditional density estimation, with a nondeterministic conditioning variable. When $d_1 \in \mathbb{N}^\star$, a natural semi-norm to evaluate the risk of $\hat{s}_m$ is the random semi-norm $\|.\|_n$ defined, for all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$, by

$$\|t\|_n^2 = \frac{1}{n} \sum_{i=1}^n \int_{[0,1]^{d_2}} t^2(X_i, y) \mu_{d_2}(\mathrm{d}y).$$

We obtain the analogue of Proposition 18, provided $(X_i)_{i \in \mathbb{Z}}$ satisfies some mixing property.

Proposition 18 *Assume that $d_1 \in \mathbb{N}^\star$. Let m be a partition of $[0,1]^d$ built on a regular partition $m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes such that*

$$|m_1^\star| \leq \frac{n}{\ln^3(n)} \quad \text{and} \quad |m_1^\star|^2 \times |m_2^\star| \leq n^2.$$

Assume that s and f are bounded, and that f is also bounded from below by a positive constant. Assume that $(X_i)_{i \in \mathbb{Z}}$ is geometrically α-mixing, with $a_{\boldsymbol{X}} \geq 0$ and $b_{\boldsymbol{X}} > 0$ such that, for all $j \in \mathbb{N}$,

$$\alpha_j^{\boldsymbol{X}} \leq a_{\boldsymbol{X}} \exp(-b_{\boldsymbol{X}} j).$$

If one of the assumptions $\boldsymbol{(D\alpha(m))}$, $\boldsymbol{(D2\text{-}\alpha(m))}$, $\boldsymbol{(D\rho)}$, $\boldsymbol{(D2\text{-}\rho)}$, $\boldsymbol{(D_{cond})}$ is satisfied, then there exist nonnegative reals δ, C_1 and C_2 that do not depend on n such that

$$\mathbb{E}_s\left[\|s - \hat{s}_m\|_n^2\right] \leq 2\|s - s_m\|_f^2 + 11 C_1 (1 + \ln(n^\delta)) \frac{D_m}{n} + C_2 \frac{1}{n}. \tag{IV.4.14}$$

Suitable values for δ and C_1 are

- $\delta = 1/b$ *and* $C_1 = 9\|sf\|_\infty \iota^{-1}(f)$ *under Assumption* $(\boldsymbol{D\alpha(m)})$;
- $\delta = 0$ *and* $C_1 = 8(1+2S_{2\text{-}\alpha})(r+1)^d(2r+1)^d\iota^{-1}(f)$ *under Assumption* $(\boldsymbol{D2\text{-}\alpha(m)})$;
- $\delta = 0$ *and* $C_1 = 250\prod_{j=0}^{\lfloor \log_2 n \rfloor}\left(1+\rho^{\boldsymbol{Z}}_{\lfloor 2^{j/3}\rfloor+1}\right)\|s\|_\infty$ *under Assumption* $(\boldsymbol{D\rho})$;
- $\delta = 0$ *and* $C_1 = (1+2S_{2\text{-}\rho})\|s\|_\infty$ *under Assumption* $(\boldsymbol{D2\text{-}\rho})$;
- $\delta = 0$ *and* $C_1 = \|s\|_\infty$ *under Assumption* $(\boldsymbol{D_{cond}})$.

In those five cases, C_2 *only depends on* $r, d, a_{\boldsymbol{X}}, b_{\boldsymbol{X}}, \iota(f), \|f\|_\infty, \|s\|_\infty$.

A proof of that Proposition is given in Section IV.8.

Remark : The same proposition holds assuming that $(X_i)_{i\in\mathbb{Z}}$ is only arithmetically α-mixing with a fast enough mixing rate and under more restrictive conditions on $m^\star$.

IV.5 Choice of the penalty

Ideally, we would like to choose a penalty pen such that $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$, in the sense that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_f^2\right] \le C \min_{m\in\mathcal{M}} \mathbb{E}_s\left[\|s-\hat{s}_m\|_f^2\right] \tag{IV.5.15}$$

for some positive constant C. The following theorem suggests a form of penalty yielding an inequality akin to

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \le C \min_{m\in\mathcal{M}}\left\{\|s-s_m\|_f^2 + \frac{D_m}{n}\right\}. \tag{IV.5.16}$$

Yet, as recalled in the previous section, for each $m\in\mathcal{M}$, $\mathbb{E}_s\left[\|s-\hat{s}_m\|_f^2\right]$ is expected to be of order $\|s-s_m\|_f^2 + D_m/n$. So, Inequality (IV.5.16) is expected to be almost as good as Inequality (IV.5.15). In order to deal with a large collection $\mathcal{M}$ that may contain irregular partitions, we impose a minor structural condition on $\mathcal{M}$, assume that s and f are bounded and that $(Z_i)_{i\in\mathbb{Z}}$ at least satisfies some β-mixing assumption.

Assumption (P) *All the partitions in the collection* $\mathcal{M}$ *are built on the product partition* $m^\star = m_1^\star \times m_2^\star$, *where* $m_1^\star$ *and* $m_2^\star$ *are regular partitions of* $[0,1]^{d_1}$ *and* $[0,1]^{d_2}$ *into cubes such that*

$$|m_1^\star| \le \frac{n}{\ln^3(n)}\mathbb{I}_{d_1\neq 0}, \quad |m_2^\star| \le n \quad \textit{and} \quad |m_1^\star|\times|m_2^\star| \le \frac{n}{\ln^4(n)}.$$

Assumption (B) *For some positive constants* $\underline{L}, \overline{L}, \underline{\ell}, \overline{\ell}$,

$$\min\{\|sf\|_\infty, \|s\|_\infty\} \ge \underline{L}, \quad s \le \overline{L} \quad \textit{and} \quad \underline{\ell} \le f \le \overline{\ell}.$$

Assumption $(\boldsymbol{D\beta})$ *The process* $(Z_i)_{i\in\mathbb{Z}}$ *is geometrically* β*-mixing, with* $a \ge 0$ *and* $b > 0$ *such that, for all* $j\in\mathbb{N}^\star$,

$$\beta_j^{\boldsymbol{Z}} \le a\exp(-bj).$$

We then obtain the following model selection theorem, proved in Section IV.8.3.

Theorem 10 *Let $\mathcal{M}$ be a collection of partitions satisfying Assumption **(P)** and $\{L_m\}_{m\in\mathcal{M}}$ be a family of nonnegative reals such that*

$$\sum_{m\in\mathcal{M}} \exp(-L_m D_m) \leq 1. \tag{IV.5.17}$$

*Assume that $(Z_i)_{i\in\mathbb{Z}}$ satisfies Assumption **($D\beta$)** and s, f satisfy Assumption **(B)**. Let δ and ϑ be defined by*

- *$\delta = 1$ and $\vartheta = 1$, without further assumption on $\{Z_i\}_{i\in\mathbb{Z}}$;*
- *$\delta = 0$ and $\vartheta = 250\prod_{j=0}^{\lfloor \log_2 n\rfloor}\left(1+\rho^{\boldsymbol{Z}}_{\lfloor 2^{j/3}\rfloor+1}\right)$ under Assumption **($D\rho$)**;*
- *$\delta = 0$ and $\vartheta = (1+2S_{2\text{-}\rho})$ under Assumption **(D2-ρ)**;*
- *$\delta = 0$ and $\vartheta = 1$ under Assumption **(D_{cond})**.*

Let $\mathcal{N}_1(s,f)$ be defined by

$$\mathcal{N}_1(s,f) = \|s\|_\infty \text{ when } r = 0 \quad \text{and} \quad \mathcal{N}_1(s,f) = \|sf\|_\infty/\iota(f) \text{ otherwise.}$$

i) If the penalty satisfies, for all $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq \mathcal{C}\vartheta\mathcal{N}_1(s,f)(1+2L_m)\ln^\delta(n)\frac{D_m}{n},$$

for some large enough nonnegative constant $\mathcal{C}$, then

$$\begin{aligned}\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq & 3\min_{m\in\mathcal{M}}\left\{\|s-s_m\|_f^2 + \mathrm{pen}(m)\right\} \\ & + \vartheta C_1\left(3b^{-1}+1\right)^\delta \overline{L}\,\overline{\ell}\underline{\ell}^{-1}\frac{\ln^\delta(n)}{n} + \frac{C_2}{n\ln(n)}\end{aligned}$$

where C_1 is a nonnegative constant and C_2 is a nonnegative real that only depends on $a, b, \delta, r,\ d_1,\ d_2,\ \underline{L},\ \overline{L},\ \underline{\ell},\ \overline{\ell}$.

ii) In particular, if the penalty satisfies, for all $m \in \mathcal{M}$,

$$\mathrm{pen}(m) = \mathcal{C}\vartheta\mathcal{N}_2(s,f)(1+2L_m)\ln^\delta(n)\frac{D_m}{n}, \tag{IV.5.18}$$

where $\mathcal{N}_2(s,f)$ is an upper-bound for $\mathcal{N}_1(s,f)$ and $\mathcal{C}$ a large enough nonnegative constant, then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_3\left(1+\max_{m\in\mathcal{M}} L_m\right)\min_{m\in\mathcal{M}}\left\{\|s-s_m\|_f^2 + \ln^\delta(n)\frac{D_m}{n}\right\}. \tag{IV.5.19}$$

where C_3 is a positive constant that only depends on $\mathcal{C}$, ϑ, C_1, δ, a, b, r, d_1, d_2, $\underline{L}$, $\overline{L}$, $\underline{\ell}$, $\overline{\ell}$.

For density estimation, *i.e.* when $d_1 = 0$, $m^\star$ can be taken as a regular partition of $[0,1]^{d_2}$ into cubes such that

$$|m^\star| \leq \frac{n}{\ln^4(n)}$$

and pen can be chosen of the form

$$\mathrm{pen}(m) = \mathcal{C}\vartheta\mathcal{N}_2(s)(1+2L_m)\ln^\delta(n)\frac{D_m}{n}$$

where $\mathcal{N}_2(s)$ is an upper-bound for $\|s\|_\infty$ and $\mathcal{C}$ some large enough nonnegative constant. For conditional density estimation, with $d_1 \in \mathbb{N}^\star$, $m_1^\star$ and $m_2^\star$ can be chosen as regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes such that

$$|m_1^\star| \leq \frac{n^{d_1/d}}{\ln^4(n)} \quad \text{and} \quad |m_2^\star| \leq n^{d_2/d}$$

and pen can be chosen of the form

$$\mathrm{pen}(m) = \mathcal{C}\vartheta\mathcal{N}_2(s,f)(1+2L_m)\ln^\delta(n)\frac{D_m}{n}$$

where $\mathcal{N}_2(s,f)$ is an upper-bound for $\mathcal{N}_1(s,f)$ and $\mathcal{C}$ some large enough nonnegative constant. In practice, $\mathcal{N}_2(s)$ and $\mathcal{N}_2(s,f)$ can be replaced with estimators of $\mathcal{N}_1(s,f)$, as in [BM97] (Proposition 4) for instance.

Let us now comment on Inequality (IV.5.19), which is similar to (IV.5.16), up to the factors C_3, that does not depend on n, $(1+\max_{m\in\mathcal{M}} L_m)$ and $\ln^\delta(n)$. In fact, in most cases of dependence considered here, the logarithmic factor just vanishes. Besides, we describe in the next two sections interesting collections of partitions for which the factor $(1+\max_{m\in\mathcal{M}} L_m)$ can be bounded by a constant.

IV.6 Selection among partitions into dyadic cubes

We call dyadic cube of $[0,1]^d$ any product of d dyadic intervals of $[0,1]$ with same lengths, *i.e.* any set of the form $I_1 \times \ldots \times I_d$ where, for all $1 \leq l \leq d$,

$$I_l = [k_l 2^{-j}, (k_l+1)2^{-j}] \tag{IV.6.20}$$

with $j \in \mathbb{N}$ and $k_l \in \{0,\ldots,2^j-1\}$. For the sake of readability, we use in (IV.6.20) a slight abuse of notation : in order to build a real partition of $[0,1]^d$, we should use for instance left-open dyadic intervals, except for those beginning at 0. In this section, we are concerned with the collection of partitions of $[0,1]^d$ into dyadic cubes with sidelength $\geq 2^{-J_\star}$, where $J_\star \in \mathbb{N}$ will be chosen in Theorem 11 below. We denote by $\mathcal{M}^{cube}$ that collection of partitions. It should be noticed that a partition of $\mathcal{M}^{cube}$ may be composed of dyadic cubes with different Lebesgue measures. For such a collection, we obtain as a straightforward consequence of Theorem 10 that the estimator $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m\in\mathcal{M}^{cube}}$.

Theorem 11 *The notation are those of Theorem 10. Assumptions* ***(B)***, ***(Dβ)*** *are supposed to be fulfilled, and also, possibly, one of the Conditions* ***(Dρ)***, ***(D2-ρ)*** *or* ***(D$_{cond}$)***.
Let $\omega(n) = (\ln(n))^4$ *if* $d_1 = 0$ *and* $\omega(n) = (\ln(n))^{4d/d_1}$ *otherwise. Assume that* $n \geq \omega(n)$, *let*

$$J_\star = \max\left\{k \in \mathbb{N} \ s.t. \ 2^k \leq (n/\omega(n))^{1/d}\right\}$$

and let pen *be given on* $\mathcal{M}^{cube}$ *by*

$$\mathrm{pen}(m) = \mathcal{C}\vartheta\mathcal{N}_2(s,f)\ln^\delta(n)\frac{D_m}{n}$$

where $\mathcal{C}$ is some positive constant. If $\mathcal{C}$ is large enough, then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_4 \min_{m\in\mathcal{M}^{cube}}\left\{\|s-s_m\|_f^2 + \frac{\ln^\delta(n) D_m}{n}\right\} \tag{IV.6.21}$$

where C_4 is a positive real that only depends on $\mathcal{C}$, ϑ, C_1, δ, a, b, r, d_1, d_2, $\underline{L}$, $\overline{L}$, $\underline{\ell}$, $\overline{\ell}$.

Proof: We define $m_i^\star$, $i=1,2$, as the partition of $[0,1]^{d_i}$ into cubes with sidelength $2^{-J_\star}$. Thus, with our choice of $J_\star$, Assumption ***(P)*** is satisfied. Let $D\in\mathbb{N}^\star$. The partitions of $[0,1]^d$ into D dyadic cubes are in one-to-one correspondence with plane 2^d-ary trees with D leaves. According to [HP91], if $D=(2^d-1)k+1$ for some integer k, the number of such partitions is given by the generalized Catalan number

$$C(k,2^d) = \frac{1}{(2^d-1)k+1}\binom{2^d k}{k} \leq \frac{1}{D}\left(\frac{2^{d2^d/(2^d-1)}}{2^d-1}\right)^D.$$

Otherwise, there are no such partitions. Therefore, Condition (IV.5.17) is fulfilled for weights L_m all equal to the same constant, for instance

$$L_m = \left(\frac{d2^d}{2^d-1}+1\right)\ln(2), \text{ for all } m\in\mathcal{M}^{cube}.$$

Inequality (IV.6.21) is then a straightforward consequence of Theorem 10. ■

It should be noticed that, in most cases, Inequality (IV.6.21) is an oracle type inequality up to a factor that does not depend on n.

For such a collection, we also study the estimation rate of $\tilde{s}$ under various smoothness assumptions on s. For that purpose, let us introduce the following notation. Let $\sigma\in(0,r+1), p>0, R>0$. For a real-valued function $t\in\mathbb{L}_p([0,1]^d)$, let us denote by $\omega_{r+1}(g,.)_p$ the $r+1$-th modulus of smoothness as defined in [DL93] (Chapter 2), for instance. Let us also set

$$|t|_{\sigma,p,\infty} = \sup_{y>0} y^{-\sigma}\omega_{r+1}(t,y)_p \quad \text{and} \quad \|t\|_{\sigma,p,\infty} = \|t\|_p + |t|_{\sigma,p,\infty},$$

where $\|.\|_p$ stands for the usual norm (or quasi-norm when $p<1$) on $\mathbb{L}_p([0,1]^d)$. We consider the Besov balls

$$\mathcal{B}(\sigma,p,\infty,R) = \left\{t:[0,1]^d\to\mathbb{R} \text{ s.t. } \|t\|_{\sigma,p,\infty}\leq R\right\}.$$

Let

$$q(\sigma,d,p) = \frac{d+2\sigma}{\sigma}\left(\frac{\sigma}{d}+\frac{1}{2}-\frac{1}{p}\right) \quad \text{and} \quad \sigma_{d,p} = \frac{d}{2}\left(\frac{1}{p}-\frac{1}{2}\right)\left(1+\sqrt{\frac{2+3p}{2-p}}\right). \tag{IV.6.22}$$

It is worth mentioning that, for $0<p<2$ and $\sigma>1/p-1/2$, $\sigma_{d,p}/d\in(1/p-1/2,1/p)$, $q(\sigma,d,p)\in(0,1+2\sigma/d)$, and that $q(\sigma,d,p)>1$ if and only if $\sigma>\sigma_{d,p}$. We obtain the following results, proved in Section IV.8.5.

Theorem 12 *The notation are those of Theorems 10 and 11, and the assumptions those of Theorem 11. Assume that the parameters σ,p,R satisfy one of the two following sets of conditions*

- $p\geq 2$, $0<\sigma<\min\{r+1/p, r+1\}$ *and* $\ln^\delta(n)/n \leq R^2 \leq n^{2\sigma/d}\ln^\delta(n)/(\omega(n))^{(1+2\sigma/d)}$,
- $0<p<2$, $\sigma_{d,p}<\sigma<\min\{r+1/p, r+1\}$ *and* $\ln^\delta(n)/n \leq R^2 \leq n^{q(\sigma,d,p)-1}\ln^\delta(n)/(\omega(n))^{q(\sigma,d,p)}$,

then there exists some positive real $C(\sigma, r, d, p)$ that only depends on σ, r, d, p such that

$$\sup_{s\in\mathcal{B}(\sigma,p,\infty,R)} \mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_4 C(\sigma,r,d,p)\|f\|_\infty \left(R\left(\frac{n}{\ln^\delta(n)}\right)^{-\sigma/d}\right)^{2d/(d+2\sigma)}.$$

The estimation rate $(Rn^{-\sigma/d})^{2d/(d+2\sigma)}$ is expected to be the minimax rate over $\mathcal{B}(\sigma, p, \infty, R)$, in view of lower bounds for the minimax risk given for instance in [YB99] for density estimation based on independent data or in [Clé00a] for density and transition density of a Markov chain. Thus, Theorem 12 indicates that, for a wide range of smoothness assumptions on s, $\tilde{s}$ is also almost as good as the best estimator of s. The limiting condition $\sigma > \sigma_{d,p}$, equivalently expressed as $q(\sigma, d, p) > 1$, is similar to that described by [Kle09] (Theorem 1), but in our case σ is allowed to be larger than 1 since we consider piecewise polynomial estimation with arbitrary order. That limiting condition is satisfied in particular for $\sigma > d/p$. Thus, for functions with isotropic smoothness, our estimator achieves better performances than all the estimators mentioned in the introduction, due to a smaller or no logarithmic factor and the wider range of smoothness covered by Theorem 12.

IV.7 Selection among partitions into dyadic rectangles

In this section, we are concerned with a collection of partitions that contains $\mathcal{M}^{cube}$ but is also adapted to possible anisotropy of the function s. We call dyadic rectangle of $[0,1]^d$ any set of the form $I_1 \times \ldots \times I_d$ where, for all $1 \leq l \leq d$,

$$I_l = [k_l 2^{-j_l}, (k_l+1)2^{-j_l}]$$

with $j_l \in \mathbb{N}$ and $k_l \in \{0, \ldots, 2^{j_l}-1\}$, with the same abuse of notation as in (IV.6.20). Otherwise said, a dyadic rectangle of $[0,1]^d$ is defined as a product of d dyadic intervals of $[0,1]$ that may have different lengths. We consider the collection of partitions of $[0,1]^d$ into dyadic rectangles with sidelength $\geq 2^{-J_\star}$, where $J_\star$ is a nonnegative integer chosen according to Theorem 13 below. We denote by $\mathcal{M}^{rect}$ such a collection of partitions. Here also, let us underline that a partition of $\mathcal{M}^{rect}$ may be composed of rectangles with different Lebesgue measures. We first study the theoretical properties of the penalized estimator for that collection, and then provide an algorithm to implement it.

IV.7.1 Theoretical properties of the penalized estimator based on $\mathcal{M}^{rect}$

We deduce again from Theorem 10 that the penalized estimator based on that collection satisfies an oracle-type inequality, similar as in the isotropic case, for a suitable penalty.

Theorem 13 *The notation are those of Theorem 10. Assumptions* ***(B)***, ***(Dβ)*** *are supposed to be fulfilled, and also, possibly, one of the Conditions* ***(Dρ)***, ***(D2-ρ)***, ***(D_{cond})***.
Let $\omega(n) = (\ln(n))^4$ if $d_1 = 0$ and $\omega(n) = (\ln(n))^{4d/d_1}$ otherwise. Assume that $n \geq \omega(n)$, let

$$J_\star = \max\left\{k \in \mathbb{N} \text{ s.t. } 2^k \leq (n/\omega(n))^{1/d}\right\}$$

and let pen *be given on $\mathcal{M}^{rect}$ by*

$$\text{pen}(m) = \mathcal{C}\vartheta\mathcal{N}_2(s,f)\ln^\delta(n)\frac{D_m}{n}$$

where $\mathcal{C}$ is some positive constant. If $\mathcal{C}$ is large enough, then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_4 \min_{m\in\mathcal{M}^{rect}}\left\{\|s-s_m\|_f^2 + \frac{\ln^\delta(n)D_m}{n}\right\} \tag{IV.7.23}$$

where C_4 is a positive real that only depends on $\mathcal{C}$, ϑ, C_1, δ, a, b, r, d_1, d_2, $\underline{L}$, $\overline{L}$, $\underline{\ell}$, $\overline{\ell}$.

Proof: Let $D \in \mathbb{N}^\star$. Building a partition of $[0,1]^d$ into D dyadic rectangles amounts to choosing a vector $(l_1,\ldots,l_{D-1}) \in \{1,\ldots,d\}^{D-1}$ of cutting directions and growing a binary tree with root corresponding to $[0,1]^d$ and with D leaves. Since the number of binary trees with D leaves is given by the Catalan number

$$\frac{1}{D}\binom{2(D-1)}{D-1} \leq \frac{4^D}{D}$$

(see for instance [HP91]), the number of such partitions is at most $(4d)^D$. Therefore, Condition (IV.5.17) is fulfilled for weights L_m all equal to the same constant, and a possible choice is

$$L_m = \ln(8d), \text{ for all } m \in \mathcal{M}^{rect}.$$

Inequality (IV.7.23) is then a straightforward consequence of Theorem 10. ■

For the penalized estimator based on the collection $\mathcal{M}^{rect}$, we can indeed prove that it adapts to anisotropic smoothness by providing its estimation rate over anisotropic Besov balls. We denote by $(\mathbf{e}_1,\ldots,\mathbf{e}_d)$ the canonical basis of $\mathbb{R}^d$. For all $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_d) \in (0,r+1)^d$, all $p>0$, all $t \in \mathbb{L}_p([0,1]^d)$, all $h>0$ and all $1 \leq l \leq d$, setting $\mathcal{R} = [0,1]^{\mathrm{d}}$, we define

$$\mathcal{R}((r+1)h\mathbf{e}_l) = \{x \in [0,1]^d \text{ s.t. } x, x+h\mathbf{e}_l, \ldots, x+(r+1)h\mathbf{e}_l \in \mathcal{R}\},$$

$$\Delta_{h\mathbf{e}_l}^{r+1}t(x) = \sum_{k=0}^{r+1}\binom{r+1}{k}(-1)^{r+1-k}t(x+kh\mathbf{e}_l), \text{ for } x \in [0,1]^d \text{ such that } x+(r+1)h\mathbf{e}_l \in [0,1]^d,$$

$$\omega_{r+1}^{(l)}(t,y,\mathcal{R})_p = \sup_{0<h\leq y}\|\Delta_{h\mathbf{e}_l}^{r+1}t\mathbb{1}_{\mathcal{R}((r+1)h\mathbf{e}_l)}\|_p, \text{ for } y \geq 0,$$

$$|t|_{\boldsymbol{\sigma},p,p} = \sum_{l=1}^d\left(\int_0^\infty\left[y^{-\sigma_l}\omega_{r+1}^{(l)}(t,y,\mathcal{R})_p\right]^p\frac{\mathrm{d}y}{y}\right)^{1/p} \quad \text{and} \quad \|t\|_{\boldsymbol{\sigma},p,p} = \|t\|_p + |t|_{\boldsymbol{\sigma},p,p}.$$

For $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_d) \in (0,r+1)^d$, $R>0$, $p>0$, we consider the anisotropic Besov balls

$$\mathcal{B}(\boldsymbol{\sigma},p,p,R) = \{t : [0,1]^d \to \mathbb{R} \text{ s.t. } \|t\|_{\boldsymbol{\sigma},p,p} \leq R\},$$

we set $\underline{\boldsymbol{\sigma}} = \min_{1\leq l\leq d}\sigma_l$ and denote by $H(\boldsymbol{\sigma})$ the harmonic mean of $\sigma_1,\ldots,\sigma_d$, *i.e.*

$$\frac{1}{H(\boldsymbol{\sigma})} = \frac{1}{d}\sum_{l=1}^d\frac{1}{\sigma_l}.$$

In particular, if $\sigma_1 = \ldots = \sigma_d = \sigma$, then $H(\boldsymbol{\sigma}) = \sigma$ and, for all $p>q$, the isotropic Besov ball $\mathcal{B}(\sigma,p,\infty,R)$ (as defined in Section IV.6) is contained in $\mathcal{B}(\boldsymbol{\sigma},q,q,R')$ for some $R'>0$ (*cf.* [Tri06], Theorem 5.30). The estimation rate over those anisotropic Besov balls relies on the following approximation result, proved in Section IV.8.6. Theorem 14 below extends the results of DeVore and Yu [DY90], that are only devoted to functions with isotropic smoothness.

Theorem 14 *Let $R>0$, $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_d)\in(\mathbb{R}_*^+)^d$ such that $r+1>\max_{1\le l\le d}\sigma_l$ and $p>0$ such that*

$$H(\boldsymbol{\sigma})/d>\max\{1/p-1/2,0\}.$$

Assume that $s\in\mathcal{B}(\boldsymbol{\sigma},p,p,R)$. Then, for all $J\in\mathbb{N}$ and all $k\in\mathbb{N}$, there exists some partition m of $[0,1]^d$ that only contains dyadic rectangles with edge-length at least $2^{-J\underline{\boldsymbol{\sigma}}/\sigma_l}$ in the l-th direction, $l=1,\dots,d$, and such that

$$|m|\le C(d,p,\boldsymbol{\sigma})2^{kd}$$

and

$$\|s-s_m\|^2\le C(d,p,r,\boldsymbol{\sigma})R^2\left(2^{-2Jd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}+2^{-2kH(\boldsymbol{\sigma})}\right).$$

Let

$$q(\boldsymbol{\sigma},d,p)=\frac{\underline{\boldsymbol{\sigma}}}{H(\boldsymbol{\sigma})}\frac{d+2H(\boldsymbol{\sigma})}{H(\boldsymbol{\sigma})}\left(\frac{H(\boldsymbol{\sigma})}{d}+\frac12-\frac1p\right)$$

be the anisotropic counterpart of the real $q(\sigma,d,p)$ defined by (IV.6.22). Contrary to [Kle09], we have chosen a parameter $J_\star$ that does not depend on the unknown smoothness of s, hence the factor $\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})$ in the above definition. That factor, which is inferior or equal to 1 with equality only in the isotropic case, may be interpreted as an index measuring the lack of isotropy.

Theorem 15 *The notation are those of Theorems 10 and 13, and the assumptions those of Theorem 13. Let $0<p\le 2$ and $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_d)\in(\mathbb{R}_*^+)^d$ such that $\max_{1\le l\le d}\sigma_l<r+1$ and $H(\boldsymbol{\sigma})/d>1/p-1/2$. If $q(\boldsymbol{\sigma},d,p)>1$ and $\ln^\delta(n)/n\le R^2\le n^{q(\boldsymbol{\sigma},d,p)-1}\ln^\delta(n)/(\omega(n))^{q(\boldsymbol{\sigma},d,p)}$, then there exists some positive real $C(\boldsymbol{\sigma},r,d,p)$ that only depends on $\boldsymbol{\sigma},r,d,p$ such that*

$$\sup_{s\in\mathcal{B}(\boldsymbol{\sigma},p,p,R)}\mathbb{E}_s\left[\|s-\tilde s\|_n^2\right]\le C_4C(\boldsymbol{\sigma},r,d,p)\|f\|_\infty\left(R\left(\frac{n}{\ln^\delta(n)}\right)^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))}.$$

Proof: Since, for all $l=1,\dots,d$, $\underline{\boldsymbol{\sigma}}\le\sigma_l$, Theorem 14 applied with $J=J_\star$ only provides partitions that belong to $\mathcal{M}^{rect}$. Thus, with $\tau=H(\boldsymbol{\sigma})/d+1/2-1/p$, we obtain

$$\begin{aligned}
&\min_{m\in\mathcal{M}^{rect}}\left\{\|s-s_m\|^2+\frac{\ln^\delta(n)|m|}{n}\right\}\\
&\le C(d,p,r,\boldsymbol{\sigma})\left(R^2 2^{-2J_\star d\tau\underline{\boldsymbol{\sigma}}/H(\sigma)}+\inf_{k\in\mathbb{N}}\left\{R^2 2^{-2kH(\boldsymbol{\sigma})}+\frac{\ln^\delta(n)2^{kd}}{n}\right\}\right)\\
&\le C(d,p,r,\boldsymbol{\sigma})\left(R^2 2^{-2J_\star d\tau\underline{\boldsymbol{\sigma}}/H(\sigma)}+\left(R\left(\frac{n}{\ln^\delta(n)}\right)^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))}\right)
\end{aligned}$$

by choosing k as the greatest nonnegative integer such that $\ln^\delta(n)2^{kd}/n\le R^2 2^{-2kH(\boldsymbol{\sigma})}$ in order to reach approximately the infimum over $k\in\mathbb{N}$. We only retain the values of R for which the first term in the upper-bound is smaller than the second. ■

The case $p>2$ may de deduced from the above theorem for $p=2$ and the continuous embeddings between anisotropic Besov spaces (*cf.* [Tri06]). The rate $\left(Rn^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))}$ is expected to be the minimax one given the lower bounds proved for instance in [YB99] for density estimation based on independent data, or in [Lac07] for transition density estimation of a Markov chain. Let us underline that, among the references cited in the introduction, only [Kle09] can deal simultaneously with anisotropy and inhomogeneous smoothness. Theorem 15 improves on [Kle09] by allowing to approximately reach the minimax risk up to a factor

that does not depend on n in most cases and considering smoothness parameters possibly larger than 1.

Remarks :

- From continuous embeddings between Besov spaces, for all $q < p$, the anisotropic counterpart $\mathcal{B}(\boldsymbol{\sigma}, p, \infty, R)$ of $\mathcal{B}(\sigma, p, \infty, R)$ is contained in some Besov ball $\mathcal{B}(\boldsymbol{\sigma}, q, q, R(p, q))$. Thus, Theorem 15 can be extended to some Besov balls $\mathcal{B}(\boldsymbol{\sigma}, p, \infty, R)$.
- In the isotropic case, Theorem 15 also improves in some sense on Theorem 12. As a matter of fact, thanks to Theorem 14, the upper-limit for the σ_l's is always $r+1$, instead of $r + 1/p$ when $p \geq 1$ in Theorem 12.

IV.7.2 Computing the penalized estimator based on $\mathcal{M}^{rect}$

We keep the same notation as in Section IV.2. Thanks to Formula (IV.2.1), one can check that, for all $m \in \mathcal{M}^{rect}$,

$$\gamma(\hat{s}_m) = -\sum_{K \in m} \sum_{k \in \{0,\dots,r\}^d} (A_K)_{(k_1,k_2)} (\Upsilon_K)_{(k_1,k_2)}.$$

We shall consider a penalty pen of the form

$$\mathrm{pen}(m) = \mathcal{C}\vartheta \mathcal{N}_2(s, f) \ln^{\delta}(n) \frac{D_m}{n}, \tag{IV.7.24}$$

as in Theorem 13. With such a penalty, $\hat{m}$ is given by

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}^{rect}} \sum_{K \in m} \mathcal{L}(K) \tag{IV.7.25}$$

where, for all rectangle K,

$$\mathcal{L}(K) = \sum_{k \in \{0,\dots,r\}^d} \left(-(A_K)_{(k_1,k_2)} \sum_{i=1}^{n} \Phi_{K,k}(Z_i) + \mathcal{C}\vartheta \mathcal{N}_2(s, f) \ln^{\delta}(n) \right).$$

That characterization allows to determine $\hat{m}$ without having to compute all the estimators of the collection $\{\hat{s}_m\}_{m \in \mathcal{M}^{rect}}$. More precisely, we can adapt to our estimation framework either the algorithm proposed by [Don97; Kle09], or the one proposed by [BSRM07]. With our choice of $J_\star$ (*cf.* Theorem 13), the computational complexity is at most of order $n^{\log_2(2d)}$ in the first case, and at most of order $n \ln^{d+1}(n)$ in the second case, which may be more interesting for large values of d.

IV.8 Proofs

IV.8.1 Notation and preliminary lemma

In all the proofs, the letter C denotes a real that may change from line to line. The notation $C(\theta)$ means that the real C may depend on θ.

Let us first introduce some adequate bases of S_m, for m partition of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles. Let $(Q_j)_{j \in \mathbb{N}}$ be the orthogonal family of the Legendre polynomials in $\mathbb{L}_2([-1,1])$.

Assume that $d_1 \in \mathbb{N}^\star$. For $K_1 = \prod_{i=1}^{d_1}[u_i, v_i]$ rectangle of $[0,1]^{d_1}$, $k_1 = (k_1(1), \ldots, k_1(d_1)) \in \{0, \ldots, r\}^{d_1}$ and $x = (x_1, \ldots, x_{d_1}) \in [0,1]^{d_1}$, we set

$$\phi_{K_1,k_1}(x) = \frac{1}{\sqrt{\mu_{d_1}(K_1)}} \prod_{i=1}^{d_1} \sqrt{2k_1(i)+1}\, Q_{k_1(i)}\left(\frac{2x_i - u_i - v_i}{v_i - u_i}\right) \mathbb{1}_{K_1}(x).$$

For all $j \in \mathbb{N}$, we recall that Q_j also satisfies

$$\|Q_j\|_\infty = 1 \quad \text{and} \quad \|Q_j\|^2 = \frac{2}{(2j+1)}.$$

Therefore, for K_1 rectangle in $[0,1]^{d_1}$, $(\phi_{K_1,k_1})_{k_1 \in \{0,\ldots,r\}^{d_1}}$ is a basis of the space of piecewise polynomials functions with support K_1 and coordinate degree $\leq r$, which is orthonormal for the norm $\|.\|$ and satisfies

$$\|\phi_{K_1,k_1}\|_\infty^2 = \frac{\prod_{i=1}^{d_1}(2k_1(i)+1)}{\mu_{d_1}(K_1)}.$$

For K_2 rectangle in $[0,1]^{d_2}$ and $k_2 \in \{0, \ldots, r\}^{d_2}$, we define in the same way ψ_{K_2,k_2} on $[0,1]^{d_2}$. For K rectangle in $[0,1]^d$, we still denote by K_1 and K_2 the rectangles in $[0,1]^{d_1}$ and $[0,1]^{d_2}$ such that $K = K_1 \times K_2$. For $k \in \{0, \ldots, r\}^d$, we still denote by k_1 and k_2 the multi-indices in $\{0, \ldots, r\}^{d_1}$ and $\{0, \ldots, r\}^{d_2}$ such that $k = (k_1, k_2)$. For any rectangle $K \in [0,1]^d$ and any multi-index $k \in \{0, \ldots, r\}^d$, we define $\Phi_{K,k}$ by

$$\Phi_{K,k}(x,y) = \phi_{K_1,k_1}(x)\psi_{K_2,k_2}(y)$$

for $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$. Thus, for a partition m of $[0,1]^d$ into rectangles, the family $(\Phi_{K,k})_{K \in m, k \in \{0,\ldots,r\}^d}$ is a basis of S_m, orthonormal for the norm $\|.\|$. For each rectangle K_1 of $[0,1]^{d_1}$, let $(\phi^f_{K_1,k_1})_{k_1 \in \{0,\ldots,r\}_1^d}$ be the orthonormal family for the norm $\|.\|_f$ deduced from $(\phi_{K_1,k_1})_{k_1 \in \{0,\ldots,r\}_1^d}$ by applying the Gram-Schmidt orthonormalization algorithm. Setting

$$\Phi^f_{K,k}(x,y) = \phi^f_{K_1,k_1}(x)\psi_{K_2,k_2}(y)$$

for any rectangle $K \subset [0,1]^d$, $k \in \{0, \ldots, r\}^d$, and $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$, we obtain a basis $(\Phi^f_{K,k})_{K \in m, k \in \{0,\ldots,r\}^d}$ orthonormal for the norm $\|.\|_f$.

For all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$, let $\|.\|_n$ be the empirical semi-norm defined by

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n \int_{[0,1]^{d_2}} t^2(X_i, y)\mu_{d_2}(\mathrm{d}y),$$

and let $\langle ., . \rangle_n$ be the associated semi-scalar product. Notice that, when $d_1 = 0$, $\|.\|_n$ and $\langle ., . \rangle_n$ coincide with $\|.\|$ and $\langle ., . \rangle$. For all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$ and all $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$, let

$$\Gamma(t,z) = t(x,y) - \int_{[0,1]^{d_2}} t(x,u)s(x,u)\mu_{d_2}(\mathrm{d}u),$$

and let ν be the empirical process defined on $\mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$ by

$$\nu(t) = \frac{1}{n}\sum_{i=1}^n \Gamma(t, Z_i).$$

For all i, $\mathbb{E}_s[\Gamma(t, Z_i)|X_i] = 0$, so that ν is centered.

We will use several times the following lemma to bound some variance terms.

Lemma 4 *Let $q \in \mathbb{N}^\star$. For all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$,*

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma(t, Z_i)\right) \leq \vartheta q^{1+\delta} \mathrm{Var}_s\left(\Gamma(t, Z_1)\right) \tag{IV.8.26}$$

where

- $\delta = 1$ *and* $\vartheta = 1$, *without further assumption;*
- $\delta = 0$ *and* $\vartheta = 250 \prod_{j=0}^{\lfloor \log_2 n \rfloor} \left(1 + \rho_{\lfloor 2^{j/3} \rfloor + 1}^{\boldsymbol{Z}}\right)$ *under Assumption* $(\boldsymbol{D\rho})$;
- $\delta = 0$ *and* $\vartheta = 1 + 2S_{2\text{-}\rho}$ *under Assumption* $(\boldsymbol{D2\text{-}\rho})$;
- $\delta = 0$ *and* $\vartheta = 1$ *under Assumption* $(\boldsymbol{D_{cond}})$.

Besides,

$$\mathrm{Var}_s\left(\Gamma(t, Z_1)\right) \leq \mathbb{E}_s\left[t^2(Z_1)\right] \leq \min\{\|s\|_\infty \|t\|_f^2, \|sf\|_\infty \|t\|^2\}. \tag{IV.8.27}$$

Proof: Since $\Gamma(t, Z_i) = \varphi_t(Z_i)$, where φ_t is some real-valued and measurable function on $[0,1]^d$ that only depends on t, Inequality (IV.8.26) under Assumptions $(\boldsymbol{D\rho})$ or $(\boldsymbol{D2\text{-}\rho})$ follows from the arguments already used in the proof of Proposition 17. By stationarity,

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma(t, Z_i)\right) = q\mathrm{Var}_s\left(\Gamma(t, Z_1)\right) + 2\sum_{j=1}^{q-1}(q-j)\mathrm{Cov}_s\left(\Gamma(t, Z_1), \Gamma(t, Z_{j+1})\right).$$

For all $j \in \{1, \ldots, q-1\}$, Schwarz inequality and the stationarity of $(Z_i)_{i\in\mathbb{Z}}$ lead to

$$\mathrm{Cov}_s\left(\Gamma(t, Z_1), \Gamma(t, Z_{j+1})\right) \leq \mathrm{Var}_s\left(\Gamma(t, Z_1)\right),$$

whereas under Assumption $(\boldsymbol{D_{cond}})$

$$\mathrm{Cov}_s\left(\Gamma(t, Z_1), \Gamma(t, Z_{j+1})\right) = \mathbb{E}_s\left[\Gamma(t, Z_1)\mathbb{E}_s\left[\Gamma(t, Z_{j+1})|Z_1, X_{j+1}\right]\right] = 0,$$

hence Inequality (IV.8.26) in the other two cases. Besides,

$$\begin{aligned}\mathrm{Var}_s\left[\Gamma(t, Z_1)\right] &= \mathrm{Var}_s\left[t(Z_1)\right] + \mathbb{E}_s\left[\mathbb{E}_s\left[\left(\mathbb{E}_s[t(Z_1)] - \mathbb{E}_s[t(Z_1)|X_1]\right)^2 | X_1\right]\right] \\ &\leq \mathbb{E}_s[t^2(Z_1)] = \int_{[0,1]^{d_1}} \int_{[0,1]^{d_2}} t^2(x,y)s(x,y)f(x)\mu_{d_1}(\mathrm{d}x)\mu_{d_2}(\mathrm{d}y).\end{aligned}$$

■

IV.8.2 Proof of Proposition 18

Since $\hat{s}_m = \underset{t \in S_m}{\mathrm{argmin}}\, \gamma(t)$, we have $\gamma(\hat{s}_m) \leq \gamma(s_m)$. The contrast γ satisfies, for all $t, u \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$,

$$\gamma(t) - \gamma(u) = \|t - s\|_n^2 - \|u - s\|_n^2 - 2\nu(t - u),$$

hence

$$\|s - \hat{s}_m\|_n^2 \leq \|s - s_m\|_n^2 + 2\nu(\hat{s}_m - s_m).$$

Let

$$\chi_f(m) = \sup_{\substack{t\in S_m \\ \|t\|_f=1}} \nu(t),$$

and let θ be some positive constant, to be chosen later, then

$$\begin{aligned} 2\nu(\hat{s}_m - s_m) &\le 2\|\hat{s}_m - s_m\|_f \chi_f(m) \\ &\le \frac{1}{\theta}\|\hat{s}_m - s_m\|_f^2 + \theta\chi_f^2(m). \end{aligned}$$

Let us fix $\kappa > 0$, to be determined later, and define

$$\Omega_\kappa(m) = \left\{\text{For all } t \in S_m\backslash\{0\}, \|t\|_f^2 \le \kappa\|t\|_n^2\right\}. \tag{IV.8.28}$$

We deduce from the triangle inequality that, on $\Omega_\kappa(m)$,

$$\begin{aligned} 2\nu(\hat{s}_m - s_m) &\le \frac{\kappa}{\theta}\|\hat{s}_m - s_m\|_n^2 + \theta\chi_f^2(m) \\ &\le \frac{2\kappa}{\theta}\|s - \hat{s}_m\|_n^2 + \frac{2\kappa}{\theta}\|s - s_m\|_n^2 + \theta\chi_f^2(m) \end{aligned} \tag{IV.8.29}$$

Consequently, provided $\theta > 2\kappa$,

$$\left(1 - \frac{2\kappa}{\theta}\right)\|s - \hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\kappa(m)} \le \left(1 + \frac{2\kappa}{\theta}\right)\|s - s_m\|_n^2 + \theta\chi_f^2(m),$$

so that

$$\left(1 - \frac{2\kappa}{\theta}\right)\mathbb{E}_s\left[\|s - \hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\kappa(m)}\right] \le \left(1 + \frac{2\kappa}{\theta}\right)\|s - s_m\|_f^2 + \theta\mathbb{E}_s\left[\chi_f^2(m)\right].$$

Let us provide some upper-bounds for $\mathbb{E}_s\left[\chi_f^2(m)\right]$ under the dependence assumptions. Since ν is linear, we deduce from Cauchy-Schwarz inequality and its equality case that

$$\chi_f^2(m) = \sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d} \nu^2\left(\Phi_{K,k}^f\right),$$

so that

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] = \frac{1}{n}\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d} \mathrm{Var}\left(\sum_{i=1}^n \Gamma(\Phi_{K,k}^f, Z_i)\right). \tag{IV.8.30}$$

Besides, by linearity of ν and as $\|.\|^2 \le \|.\|_f^2/\iota(f)$, we obtain $\chi_f^2(m) \le \chi^2(m)/\iota(f)$ where $\chi(m) = \sup_{\substack{t\in S_m \\ \|t\|=1}} \nu(t)$. So we may also use

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] \le \frac{1}{n\iota(f)}\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d} \mathrm{Var}\left(\sum_{i=1}^n \Gamma(\Phi_{K,k}, Z_i)\right). \tag{IV.8.31}$$

Under Assumptions $(\boldsymbol{D\rho})$, $(\boldsymbol{D2\text{-}\rho})$ or $(\boldsymbol{D_{cond}})$, by applying Lemma 4 to each $\Phi_{K,k}^f$ and with $q = n$, we deduce from Equality (IV.8.30) that

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] \le \vartheta\|s\|_\infty D_m$$

where ϑ is defined as in Lemma 4. Under Assumption $(\boldsymbol{D\alpha(m)})$, Inequality (IV.8.31) and Inequality (IV.8.54) applied, for each $K \in m$ and each $k \in \{0,\dots,r\}^d$, with $g = \Gamma(\Phi_{K,k}, .)$ and $q = \lceil 3\ln(n^{1/b})\rceil$ yield

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] \le 9\left(1 + \ln(n^{1/b})\right)\|sf\|_\infty \iota^{-1}(f) D_m + 12(2r+1)^d a\iota^{-1}(f) D_m/n.$$

Under Assumption **(D2-$\alpha(m)$)**, Inequalities (IV.8.31) and (IV.8.55) lead to

$$
\begin{aligned}
n\mathbb{E}_s\left[\chi_f^2(m)\right] &\leq \frac{1}{n\iota(f)} \sum_{K\in m} \sum_{k\in\{0,\dots,r\}^d} \mathrm{Var}\left(\sum_{i=1}^n \Gamma(\Phi_{K,k}, Z_i)\right) \\
&\leq 2\iota^{-1}(f)(1+2S_{2\text{-}\alpha}) \left\| \sum_{K\in m} \sum_{k\in\{0,\dots,r\}^d} |\Gamma(\Phi_{K,k}, Z_1)| \right\|_\infty^2 \\
&\leq 2\iota^{-1}(f)(1+2S_{2\text{-}\alpha}) \frac{(2r+1)^d}{\min_{K\in m}\mu_d(K)} \left\| \sum_{k\in\{0,\dots,r\}^d} \sum_{K\in m} \left(\mathbb{1}_K(Z_1) + \mathbb{E}_s\left[\mathbb{1}_K(Z_1)|X_1\right]\right) \right\|_\infty^2 \\
&\leq 8\iota^{-1}(f)(1+2S_{2\text{-}\alpha})(2r+1)^d(r+1)^d D_m.
\end{aligned}
$$

In order to bound the risk of $\hat{s}_m$ on $\Omega_\kappa^c(m)$, we use the following two lemmas, proved just below.

Lemma 5 *Assume that $d_1 \in \mathbb{N}^\star$ and that s is bounded. Let m be a partition of $[0,1]^{d_1}\times[0,1]^{d_2}$ into rectangles built on a regular partition $m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes. Then*

$$\|s-\hat{s}_m\|^2 \leq 2\|s\|_\infty^2 + 2(r+1)^{d_2}(2r+1)^{d_2}|m_2^\star|.$$

Lemma 6 *Let $m^\star = m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes. Let $q\in\mathbb{N}^\star$, $\kappa>1$, $\Omega_\kappa(m^\star)$ be defined by* (IV.8.28)*. Then there exists an absolute constant C such that*

$$\mathbb{P}_s(\Omega_\kappa^c(m^\star)) \leq C(r+1)^{2d_1}|m_1^\star| \left(\exp\left(-\frac{\iota^2(f)(1-1/\kappa)^2 n}{100\|f\|_\infty C(r,d_1) q |m_1^\star|} \right) + \frac{\|f\|_\infty C(r,d_1)|m_1^\star|\alpha_q^{\boldsymbol{X}}}{\iota^2(f)(1-1/\kappa)^2} \right).$$

We now choose $\kappa = 7/6$, $\theta = 7$ and $q = \lceil 3\ln(n^{1/b})\rceil$. Since m is built on $m^\star = m_1^\star \times m_2^\star$, $\Omega_\kappa(m^\star) \subset \Omega_\kappa(m)$. Given the conditions on $m_1^\star$ and $m_2^\star$, we then obtain

$$
\begin{aligned}
\mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\kappa^c(m)}\right] &\leq 2\left(\|s\|_\infty^2 + (r+1)^{d_2}(2r+1)^{d_2}|m_2^\star|\right)\mathbb{P}_s(\Omega_\kappa^c(m^\star)) \\
&\leq C(r,d_1,a_{\boldsymbol{X}},b_{\boldsymbol{X}},s,f)/n,
\end{aligned}
$$

where $C(r,d_1,a_{\boldsymbol{X}},b_{\boldsymbol{X}},s,f)$ is a nonnegative real that only depends on r, d_1, $a_{\boldsymbol{X}}$, $b_{\boldsymbol{X}}$, $\|s\|_\infty$, $\iota(f)$ and $\|f\|_\infty$.

Let us end with the proofs of Lemmas 5 and 6.

Proof of Lemma 5: We shall use the notation

- $\|.\|_{\mathbb{R}^n}$ defined for $v=\{v_i\}_{1\leq i\leq n}\in\mathbb{R}^n$ by $\|v\|_{\mathbb{R}^n} = \sum_{i=1}^n v_i^2/n$;
- for $t\in\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$ and $y\in[0,1]^{d_2}$, $t^{\mathbf{X}}(y) = \{t(X_i,y)\}_{1\leq i\leq n}\in\mathbb{R}^n$;
- $\mathcal{V}_m^{\mathbf{X}}(y) = \{t^{\mathbf{X}}(y), t\in S_m\}$ and $\mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}$ the orthogonal projection of $\mathbb{R}^n$ on $\mathcal{V}_m^{\mathbf{X}}(y)$.

For all $y\in[0,1]^{d_2}$, let us also define the $\mathbb{R}^n$-vector

$$\hat{v}_m(y) = \left\{ \sum_{J\in m_2} \sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}(Y_i)\psi_{J,j}(y) \right\}_{1\leq i\leq n}.$$

As [Lac07] (Proposition 2.1), we can prove that $\hat{s}_m^{\mathbf{X}}(y) = \mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}\left(\hat{v}_m(y)\right)$. Using the triangle inequality and the shrinking property of $\mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}$, we get

$$\begin{aligned}\|s - \hat{s}_m\|_n^2 &= \int_{[0,1]^{d_2}} \|s^{\mathbf{X}}(y) - \hat{s}_m^{\mathbf{X}}(y)\|_{\mathbb{R}^n}^2 \mu_{d_2}(\mathrm{d}y) \\ &\leq 2\int_{[0,1]^{d_2}} \|s^{\mathbf{X}}(y)\|_{\mathbb{R}^n}^2 \mu_{d_2}(\mathrm{d}y) + 2\int_{[0,1]^{d_2}} \|\hat{v}_m(y)\|_{\mathbb{R}^n}^2 \mu_{d_2}(\mathrm{d}y).\end{aligned}$$

From the orthonormality of $\{\psi_{J,j}\}_{J\in m_2, j\in\{0,\dots,r\}^{d_2}}$, we deduce that

$$\int_{[0,1]^{d_2}} \|\hat{v}_m(y)\|_{\mathbb{R}^n}^2 \mu_{d_2}(\mathrm{d}y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{J\in m_2}\sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2(Y_i).$$

Besides, by grouping the $\psi_{J,j}$ having the same support, we get

$$\left\|\sum_{J\in m_2}\sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2\right\|_\infty \leq \max_{J\in m_2}\left\|\sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2\right\|_\infty \leq (2r+1)^{d_2}(r+1)^{d_2}/\min_{J\in m_2}\mu_{d_2}(J),$$

hence Lemma 5. ■

Proof of Lemma 6: The proof follows almost the same lines as the proof of Proposition 8 in [Lac07]. Let ν' be the centered empirical process defined for $u \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$ by

$$\nu'(u) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{[0,1]^{d_2}} u(X_i,y)\mu_{d_2}(\mathrm{d}y) - \int_{[0,1]^{d_1}\times[0,1]^{d_2}} u(x,y)f(x)\mu_{d_1}(\mathrm{d}x)\mu_{d_2}(\mathrm{d}y)\right).$$

Since $\|t\|_n^2 = \nu'(t^2) + \|t\|_f^2$ for all $t \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$, ν' is linear and $\kappa > 1$, we get

$$\Omega_\kappa^c(m^\star) \subset \left\{\sup_{t\in S_{m^\star}/\|t\|_f=1} |\nu'(t^2)| > 1 - 1/\kappa\right\}.$$

By construction of $(\Phi_{K,k})_{K\in m^\star, k\in\{0,\dots,r\}^d}$, for all $K, L \in m^\star$ and $k, l \in \{0,\dots,r\}^d$, and all $i \in \{1,\dots,n\}$,

$$\begin{aligned}\int_{[0,1]^{d_2}} \Phi_{K,k}(X_i,y)\Phi_{L,l}(X_i,y)\mu_{d_2}(\mathrm{d}y) &= \phi_{K_1,k_1}(X_i)\phi_{L_1,l_1}(X_i)\langle\psi_{K_2,k_2},\psi_{L_2,l_2}\rangle \\ &= \mathbb{1}_{K_1=L_1}\mathbb{1}_{(K_2,k_2)=(L_2,l_2)}\phi_{K_1,k_1}\phi_{L_1,l_1}(X_i). \qquad \text{(IV.8.32)}\end{aligned}$$

Let $t \in S_{m^\star}\backslash\{0\}$, and for $K_1 \in m_1^\star$ and $k_1 \in \{0,\dots,r\}^{d_1}$, let

$$a_{K_1,k_1} = \sqrt{\sum_{K_2\in m_2^\star}\sum_{k_2\in\{0,\dots,r\}^{d_2}} \langle t, \Phi_{K_1\times K_2,(k_1,k_2)}\rangle^2}/\|t\|.$$

It follows from (IV.8.32) and Schwarz inequality that

$$|\nu'(t^2)| \leq \|t\|^2 \sum_{K_1\in m_1}\sum_{k_1,l_1\in\{0,\dots,r\}^{d_1}} a_{K_1,k_1}a_{K_1,l_1}|\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})|$$

where ν'' is the centered empirical process defined on $\mathbb{L}_2([0,1]^{d_1})$ by

$$\nu''(u) = \frac{1}{n}\sum_{i=1}^{n}\left(u(X_i) - \int_{[0,1]^{d_1}} u(x)f(x)\mu_{d_1}(\mathrm{d}x)\right).$$

Consequently

$$\sup_{t\in S_m/\|t\|_f=1}|\nu'(t^2)| \le \iota^{-1}(f)\max_{a\in\mathcal{A}}\sum_{K_1\in m_1}\sum_{k_1,l_1\in\{0,\dots,r\}^{d_1}} a_{K_1,k_1}a_{K_1,l_1}|\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})|,$$

where $\mathcal{A}=\left\{a=(a_{K_1,k_1})_{K_1\in m_1,k_1\in\{0,\dots,r\}^{d_1}} \text{ s.t. } \sum_{K_1\in m_1}\sum_{k_1\in\{0,\dots,r\}^{d_1}} a^2_{K_1,k_1}=1\right\}$. Let us introduce $B=(B_{K_1,k_1,l_1})_{K_1\in m_1,k_1,l_1\in\{0,\dots,r\}^{d_1}}$ and $V=(V_{K_1,k_1,l_1})_{K_1\in m_1,k_1,l_1\in\{0,\dots,r\}^{d_1}}$ defined respectively by

$$B_{K_1,k_1,l_1}=\|\phi_{K_1,k_1}\phi_{K_1,l_1}\|_\infty \quad \text{and} \quad V_{K_1,k_1,l_1}=\|\phi_{K_1,k_1}\phi_{K_1,l_1}\|.$$

Let us set

$$\bar\rho(B)=\sup_{a\in\mathcal{A}}\sum_{K_1\in m_1}\sum_{k_1,l_1\in\{0,\dots,r\}^{d_1}}|a_{K_1,k_1}||a_{K_1,l_1}|B_{K_1,k_1,l_1},$$

define $\bar\rho(V)$ in the same way, and set $L(\phi)=\max\{\bar\rho^2(V),\bar\rho(B)\}$. Then, Schwarz inequality and the properties of the family $(\phi_{K_1,k_1})_{K_1\in m_1,k_1\in\{0,\dots,r\}^{d_1}}$ recalled in Section IV.8.1 provide

$$L(\phi)\le C(r,d_1)|m_1^\star|. \tag{IV.8.33}$$

Let

$$x=\frac{\iota^2(f)(1-1/\kappa)^2}{100q\|f\|_\infty L(\phi)}$$

and

$$\Delta=\bigcap_{K_1\in m_1^\star,k_1,l_1\in\{0,\dots,r\}^{d_1}}\left\{|\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})|<3\left(\sqrt{2q\|f\|_\infty x}V_{K_1,k_1,l_1}+\frac{2}{3}qB_{K_1,k_1,l_1}x\right)\right\}.$$

One can easily check that, on Δ, $\sup_{t\in S_{m^\star}/\|t\|_f=1}|\nu'(t^2)|\le 1-1/\kappa$, so that $\Omega_\kappa^c(m)\subset\Delta^c$. Lemma 6 then follows from Proposition 22 in the Appendix. ■

IV.8.3 Proof of Theorem 10

Let us fix $m\in\mathcal{M}$. For all $m'\in\mathcal{M}$, we will denote by $m\cup m'$ the roughest partition built on m and m'. We also fix $\kappa\ge 1$ and $\theta_1>0$, to be determined at the end of the proof. By definition of $\hat m$ and $\hat s_m$,

$$\begin{aligned}\gamma(\tilde s)+\mathrm{pen}(\hat m)&\le\gamma(\hat s_m)+\mathrm{pen}(m)\\&\le\gamma(s_m)+\mathrm{pen}(m).\end{aligned}\tag{IV.8.34}$$

Using the same arguments as in the proof of Proposition 18, we deduce from (IV.8.34) that

$$\|s-\tilde s\|_n^2\le\|s-s_m\|_n^2+\mathrm{pen}(m)+2\nu(\tilde s-s_m)-\mathrm{pen}(\hat m).$$

As $\tilde s-s_m\in S_{m\cup\hat m}\subset S_{m^\star}$, we obtain in the same way as Inequality (IV.8.29) that, on the set $\Omega_\kappa(m^\star)$ defined as in (IV.8.28),

$$2\nu(\tilde s-s_m)\le\frac{2\kappa}{\theta_1}\|s-\tilde s\|_n^2+\frac{2\kappa}{\theta_1}\|s-s_m\|_n^2+\theta_1\chi_f^2(m\cup\hat m).$$

Consequently, provided $\theta_1>2\kappa$,

$$\left(1-\frac{2\kappa}{\theta_1}\right)\|s-\tilde s\|_n^2\mathbb{1}_{\Omega_\kappa(m^\star)}\le\left(1+\frac{2\kappa}{\theta_1}\right)\|s-s_m\|_n^2\mathbb{1}_{\Omega_\kappa(m^\star)}+\mathrm{pen}(m)+\theta_1\chi_f^2(m\cup\hat m)-\mathrm{pen}(\hat m).\tag{IV.8.35}$$

Since the data are β-mixing, we can introduce blockwise independent data. More precisely, let q_n be some positive integer, to be determined later, and let (d_n, r_n) be the unique couple of nonnegative integers such that $n = d_n q_n + r_n$ and $0 \leq r_n < q_n$. For the sake of simplicity, we assume in the sequel that $r_n = 0$ and $d_n = 2p_n \in \mathbb{N}^\star$, but the other cases can be treated in a similar way. For $l = 0, \ldots, p_n - 1$, let us set

$$A_l = \{Z_i\}_{2lq_n+1 \leq i \leq (2l+1)q_n} \quad \text{and} \quad B_l = \{Z_i\}_{(2l+1)q_n+1 \leq i \leq (2l+2)q_n}.$$

As recalled for instance in [Vie97] (proof of Proposition 5.1), we can build, for $l = 0, \ldots, p_n - 1$,

$$A_l^\bullet = \{Z_i^\bullet\}_{2lq_n+1 \leq i \leq (2l+1)q_n} \quad \text{and} \quad B_l^\bullet = \{Z_i^\bullet\}_{(2l+1)q_n+1 \leq i \leq (2l+2)q_n}$$

such that, for all $l = 0, \ldots, p_n - 1$,

- A_l, $A_l^\bullet$, B_l and $B_l^\bullet$ have the same distribution;
- $\mathbb{P}_s(A_l \neq A_l^\bullet) \leq \beta_{q_n}^{\boldsymbol{Z}}$ and $\mathbb{P}_s(B_l \neq B_l^\bullet) \leq \beta_{q_n}^{\boldsymbol{Z}}$;
- $(A_l^\bullet)_{0 \leq l \leq p_n - 1}$ are independent random variables, and so are $(B_l^\bullet)_{0 \leq l \leq p_n - 1}$.

We set

$$\Omega_\bullet = \bigcap_{i=1}^{n} \{Z_i^\bullet = Z_i\}.$$

The proof of Theorem 10 heavily relies on the following concentration inequality satisfied by the random variables $\chi^2(m')$, for m' partition built on $m^\star$. The proof of that proposition is deferred to Section IV.8.4.

Proposition 19 *For all rectangle $K \subset [0,1]^{d_1} \times [0,1]^{d_2}$, let $\mu_f(K) = \int_K f(x)\mu_{d_1}(\mathrm{d}x)\mu_{d_2}(\mathrm{d}y)$. Let δ, ϑ and $\mathcal{N}_1(s,f)$ be defined as in Theorem 10, and let θ be some positive real. Under the assumptions of Theorem 10, there exists some set $\Omega_T \in \mathcal{F}$ such that*

$$\mathbb{P}_s(\Omega_T^c) \leq 4(r+1)^d |m^\star| \exp\left(-C(\theta, r, d, \vartheta)\mathcal{N}_1(s,f)\iota(f)\frac{n}{q_n^{2-\delta}|m^\star|}\right) \tag{IV.8.36}$$

and, for all partition m' built on $m^\star$ and all positive x,

$$\mathbb{P}_s\Big(n\chi_f^2(m')\mathbb{1}_{\Omega_\bullet \cap \Omega_T} \geq 2(1+\theta)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\left(D_{m'} + 2x\right)\Big) \leq 2\exp(-x). \tag{IV.8.37}$$

Let Ω_T be defined as in Proposition 19. We shall first bound the quadratic risk of $\tilde{s}$ on $\Omega_\kappa(m^\star) \cap \Omega_\bullet \cap \Omega_T$. Let us fix $\theta_2 > 0$, to be determined later. Let ξ be some positive real, and, for all $m' \in \mathcal{M}$, let $x_{m'} = L_{m'}D_{m'} + \xi$ and

$$\Omega(\xi, m, m') = \left\{n\chi_f^2(m \cup m')\mathbb{1}_{\Omega_\bullet \cap \Omega_T} \leq 2(1+\theta_2)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\left(D_m + D_{m'} + 2x_{m'}\right)\right\}.$$

According to Proposition 19 applied to each partition $m \cup m'$, for $m' \in \mathcal{M}$, the event $\Omega(\xi, m) = \bigcap_{m' \in \mathcal{M}} \Omega(\xi, m, m')$ happens with probability

$$\mathbb{P}_s(\Omega(\xi, m)) \geq 1 - 2\exp(-\xi).$$

Moreover, we deduce from Inequality (IV.8.35) that, on $\Omega(\xi, m)$,

$$\begin{aligned}\left(1 - \frac{2\kappa}{\theta_1}\right)\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_\kappa(m^\star) \cap \Omega_\bullet \cap \Omega_T} \leq \left(1 + \frac{2\kappa}{\theta_1}\right)\|s - s_m\|_n^2 + \mathrm{pen}(m) + 2\theta_1(1+\theta_2)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\frac{D_m}{n} \\ + 2\theta_1(1+\theta_2)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\frac{(1+2L_{\hat{m}})D_{\hat{m}}}{n} - \mathrm{pen}(\hat{m}) \\ + 4\theta_1(1+\theta_2)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\frac{\xi}{n}.\end{aligned}$$

Choosing pen such that, for all $m' \in \mathcal{M}$,

$$\mathrm{pen}(m') \geq 2\theta_1(1+\theta_2)^2\vartheta q_n^\delta \mathcal{N}_1(s,f)\frac{(1+2L_{m'})D_{m'}}{n},$$

we obtain, still on $\Omega(\xi,m)$,

$$\left(1-\frac{2\kappa}{\theta_1}\right)\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_\kappa(m^\star)\cap\Omega_\bullet\cap\Omega_T} \leq \left(1+\frac{2\kappa}{\theta_1}\right)\|s-s_m\|_n^2+2\mathrm{pen}(m)+4\theta_1(1+\theta_2)^2\vartheta q_n^\delta\mathcal{N}_1(s,f)\frac{\xi}{n}.$$

We recall that, for all real-valued random variable V,

$$\mathbb{E}[V] \leq \mathbb{E}[V^+] = \int_0^\infty \mathbb{P}(V^+ \geq \xi)\mathrm{d}\xi,$$

where V^+ stands for the positive part of V. Besides, for all deterministic function $t \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$, $\mathbb{E}_s[\|t\|_n^2] = \|t\|_f^2$. Using these two properties, we conclude that

$$\begin{aligned}\left(1-\frac{2\kappa}{\theta_1}\right)\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\mathbb{1}_{\Omega_\kappa(m^\star)\cap\Omega_\bullet\cap\Omega_T}\right] \leq \left(1+\frac{2\kappa}{\theta_1}\right)\|s-s_m\|_f^2 + 2\mathrm{pen}(m)\\ + 8\theta_1(1+\theta_2)^2\vartheta q_n^\delta\mathcal{N}_1(s,f)\frac{1}{n}. \qquad \text{(IV.8.38)}\end{aligned}$$

Let us now bound the quadratic risk of $\tilde{s}$ on $\Omega_\kappa^c(m^\star)\cup\Omega_\bullet^c\cup\Omega_T^c$. Lemma 5 can be extended to the case $d_1 = 0$. As a matter of fact, when $d_1 = 0$, we deduce from the triangle inequality and the convexity of the square function that, for all partition m,

$$\begin{aligned}\|s-\hat{s}_m\|^2 &\leq 2\|s\|^2 + 2\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d}\left[\frac{1}{n}\sum_{i=1}^n\Phi_{K,k}(Y_i)\right]^2\\ &\leq 2\|s\|^2+\frac{2}{n}\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d}\sum_{i=1}^n\Phi_{K,k}^2(Y_i)\\ &\leq 2\|s\|^2+2\max_{K\in m}\left\|\sum_{k\in\{0,\dots,r\}^d}\Phi_{K,k}^2\right\|_\infty.\end{aligned}$$

Consequently, we deduce from Lemma 5 extended to the case $d_1 = 0$ and applied to $\hat{m}$ that

$$\|s-\tilde{s}\|_n^2 \leq 2\|s\|_\infty^2 + 2(r+1)^{d_2}(2r+1)^{d_2}|m_2^\star|. \qquad \text{(IV.8.39)}$$

A straightforward upper-bound for the $\mathbb{P}_s$-measure of $\Omega_\kappa^c(m^\star)\cup\Omega_\bullet^c\cup\Omega_T^c$ is

$$\mathbb{P}_s(\Omega_\kappa^c(m^\star)\cup\Omega_\bullet^c\cup\Omega_T^c) \leq \mathbb{P}_s(\Omega_\kappa^c(m^\star)\cap\Omega_\bullet)+\mathbb{P}_s(\Omega_\bullet^c)+\mathbb{P}_s(\Omega_T^c).$$

One easily deduces from one of the properties of the $A_l^\bullet$'s and $B_l^\bullet$'s that

$$\mathbb{P}_s(\Omega_\bullet^c) \leq 2p_n\beta_{q_n}^{\boldsymbol{Z}} = \frac{n}{q_n}\beta_{q_n}^{\boldsymbol{Z}}. \qquad \text{(IV.8.40)}$$

If $d_1 = 0$ and $\kappa \geq 1$, then $\Omega_\kappa(m^\star) = \Omega$, so $\mathbb{P}_s(\Omega_\kappa^c(m^\star)\cap\Omega_\bullet) = 0$. Otherwise, in order to bound $\mathbb{P}_s(\Omega_\kappa^c(m^\star)\cap\Omega_\bullet)$, we follow the proof of Lemma 6. Thus there exists some constant $C(r,d_1)$ that only depends on d_1 and r such that

$$\mathbb{P}_s(\Omega_\kappa^c(m^\star)\cap\Omega_\bullet) \leq 4(r+1)^{2d_1}|m_1^\star|\exp\left(-C(r,d_1)\frac{\iota^2(f)(1-1/\kappa)^2}{\|f\|_\infty}\frac{n}{q_n|m_1^\star|}\right). \qquad \text{(IV.8.41)}$$

Combining Inequalities (IV.8.39) to (IV.8.41) and (IV.8.36) then provides for $\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2 \mathbb{1}_{\Omega_\kappa^c(m^\star)\cup\Omega_\bullet^c\cup\Omega_T^c}\right]$ the upper-bound

$$C(r,d_1,d_2,\overline{L})|m_2^\star|\left(\frac{n}{q_n}\beta_{q_n}^{\boldsymbol{Z}}+|m^\star|\exp\left(-C(\theta_2,\vartheta,r,d,\underline{L},\underline{\ell})\frac{n}{q_n^{2-\delta}|m^\star|}\right)\right.$$
$$\left.+|m_1^\star|\mathbb{1}_{d_1\geq 1}\exp\left(-C(\kappa,r,d_1,\underline{\ell},\overline{\ell})\frac{n}{q_n|m_1^\star|}\right)\right). \tag{IV.8.42}$$

Last, let us choose

$$\kappa=7/6,\quad \theta_1=7,\quad \theta_2=2/25\quad\text{and}\quad q_n=\lceil 3\ln(n^{1/b})\rceil.$$

Under Assumption (IV.5.17) on $m_1^\star$ and $m_2^\star$, we deduce from (IV.8.38) and (IV.8.42) that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right]\leq 3\left\{\|s-s_m\|_f^2+\mathrm{pen}(m)\right\}$$
$$+100\vartheta\mathcal{N}_1(s,f)(3\ln(n^{1/b})+1)^\delta\frac{1}{n}+\frac{C(a,b,\delta,r,d_1,d_2,\underline{L},\overline{L},\underline{\ell},\overline{\ell})}{n\ln(n)}.$$

Theorem 10 then follows by taking the minimum over $m\in\mathcal{M}$.

IV.8.4 Proof of Proposition 19

We define on $\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$ and for all $m'\in\mathcal{M}$

$$\nu^\bullet(t)=\frac{1}{n}\sum_{i=1}^n\Gamma(t,Z_i^\bullet)\quad\text{and}\quad\chi_f^\bullet(m')=\sup_{\substack{t\in S_{m'}\\ \|t\|_f=1}}\nu^\bullet(t)$$

$$\nu_{(1)}^\bullet(t)=\frac{1}{n}\sum_{l=0}^{p_n-1}\sum_{i=2lq_n+1}^{(2l+1)q_n}\Gamma(t,Z_i^\bullet)\quad\text{and}\quad\chi_{f,(1)}^\bullet(m')=\sup_{\substack{t\in S_{m'}\\ \|t\|_f=1}}\nu_{(1)}^\bullet(t)$$

$$\nu_{(2)}^\bullet(t)=\frac{1}{n}\sum_{l=0}^{p_n-1}\sum_{i=(2l+1)q_n+1}^{(2l+2)q_n}\Gamma(t,Z_i^\bullet)\quad\text{and}\quad\chi_{f,(2)}^\bullet(m')=\sup_{\substack{t\in S_{m'}\\ \|t\|_f=1}}\nu_{(2)}^\bullet(t).$$

Notice that $\nu=\nu^\bullet=\nu_{(1)}^\bullet+\nu_{(2)}^\bullet$ on $\Omega_\bullet$, hence

$$\chi_f^2(m)\mathbb{1}_{\Omega_\bullet}\leq 2\left(\left\{\chi_{f,(1)}^\bullet(m)\right\}^2+\left\{\chi_{f,(2)}^\bullet(m)\right\}^2\right)\mathbb{1}_{\Omega_\bullet}. \tag{IV.8.43}$$

We shall first prove concentration inequalities for $\chi_{(i)}^\bullet(m')$, $i\in\{1,2\}$, defined by

$$\chi_{(i)}^\bullet(m')=\sup_{\substack{t\in S_{m'}\\ \|t\|=1}}\nu_{(i)}^\bullet(t).$$

For $i\in\{1,2\}$, let

$$\Omega_{T(i)}=\bigcap_{K\in m^\star}\bigcap_{k\in\{0,\dots,r\}^d}\left\{|\nu_{(i)}^\bullet(\Phi_{K,k})|\leq\epsilon\sqrt{\mu_d(K)}\right\}$$

where

$$\epsilon=\frac{3\vartheta\theta^2}{2(\theta+3)(2r+1)^{2d}}\frac{\|sf\|_\infty}{q_n^{1-\delta}}.$$

On the set $\Omega_{T(1)}$, we have, for all $K \in m'$ and all $k \in \{0,\dots,r\}^d$,

$$|\nu^{\bullet}_{(1)}(\Phi_{K,k})| \leq \epsilon\sqrt{(r+1)^d \mu_d(K)}.$$

As a matter of fact, for $K \in m'$ and $k \in \{0,\dots,r\}^d$, decomposing $\Phi_{K,k}$ in the basis $(\Phi_{L,l})_{L\in m^\star, l\in\{0,\dots,r\}^d}$ gives

$$\Phi_{K,k} = \sum_L \sum_{l\in\{0,\dots,r\}^d} a_{L,l}\Phi_{L,l} \quad \text{with} \quad \sum_L \sum_{l\in\{0,\dots,r\}^d} a^2_{L,l} = 1$$

where the sum on L is only on disjoint rectangles $L \in m^\star$ whose union is K. Hence, it follows from the linearity of $\nu^{\bullet}_{(1)}$ and Schwarz inequality that, on the set $\Omega_{T(1)}$,

$$\begin{aligned}
|\nu^{\bullet}_{(1)}(\Phi_{K,k})| &\leq \sqrt{\sum_{L,l} |\nu^{\bullet}_{(1)}(\Phi_{L,l})|^2} \\
&\leq \epsilon\sqrt{\sum_{L,l} \mu_d(L)} \\
&\leq \epsilon\sqrt{(r+1)^d \mu_d(K)}.
\end{aligned}$$

We shall take advantage of two expressions for $\chi^{\bullet}_{(1)}(m')$. On the one hand, it follows from Schwarz inequality, its equality case, and the linearity of $\nu^{\bullet}_{(1)}$ that

$$\chi^{\bullet}_{(1)}(m') = \sqrt{\sum_{K\in m'} \sum_{k\in\{0,\dots,r\}^d} \left\{\nu^{\bullet}_{(1)}(\Phi_{K,k})\right\}^2} = \nu^{\bullet}_{(1)}(t^*) \tag{IV.8.44}$$

where t^* is the random element of $\{t \in S_{m'}\ s.t.\ \|t\| = 1\}$ defined by

$$t^* = \sum_{K\in m'} \sum_{k\in\{0,\dots,r\}^d} \frac{\nu^{\bullet}_{(1)}(\Phi_{K,k})}{\chi^{\bullet}_{(1)}(m')} \Phi_{K,k}.$$

On the other hand, according to the definition of $\chi^{\bullet}_{(1)}(m')$, we can write

$$\chi^{\bullet}_{(1)}(m') = \sup_{\substack{t\in S_{m'} \\ \|t\|=1}} \sum_{l=0}^{p_n-1} g_t(A^{\bullet}_l), \tag{IV.8.45}$$

where, for all $t \in S_{m'}$, g_t is defined on $([0,1]^{d_1} \times [0,1]^{d_2})^{q_n}$ by

$$g_t(z_1,\dots,z_{q_n}) = \frac{1}{n}\sum_{j=1}^{q_n} \Gamma(t, z_j).$$

Let us fix $x > 0$ and set

$$z = \sqrt{\vartheta q^{\delta}_n \frac{\|sf\|_\infty}{n}} x.$$

We introduce a countable and dense subset $\mathcal{A}_{m'}$ of $\{t \in S_{m'}/\|t\| = 1, \|t\|_\infty \leq \epsilon(2r+1)^{2d}/z\}$, define

$$W(m') = \sup_{t\in\mathcal{A}_{m'}} \sum_{l=0}^{p_n-1} g_t(A^{\bullet}_l),$$

and set

$$\sigma_{m'}^2 = \sup_{t\in\mathcal{A}_{m'}} \sum_{l=0}^{p_n-1} \mathrm{Var}_s\left(g_t(A_l^\bullet)\right) \quad \text{and} \quad b_{m'} = \sup_{t\in\mathcal{A}_{m'}} \|g_t\|_\infty.$$

Since $n = 2p_nq_n$, we deduce from Lemma 4 that

$$\sigma_m^2 \leq \vartheta\|sf\|_\infty q_n^\delta \frac{1}{2n}.$$

As, for all $t \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$ and $z \in [0,1]^{d_1}\times[0,1]^{d_2}$, $|\Gamma(t,z)| \leq 2\|t\|_\infty$, we get

$$b_{m'} \leq \frac{2\epsilon(2r+1)^{2d}q_n}{nz}.$$

Thanks to Talagrand's inequality (as stated for instance in [Mas07], Inequality (5.50)), there exists a set $\Omega(x) \in \Omega$ such that $\mathbb{P}_s(\Omega(x)) \geq 1-\exp(-x)$ and on which

$$W(m') \leq (1+\theta)\left(\mathbb{E}_s[W(m')] + \sqrt{\vartheta\|sf\|_\infty q_n^\delta \frac{x}{n}}\right).$$

Let us now come back to $\chi_{(1)}^\bullet(m')$. Given expression (IV.8.45) for $\chi_{(1)}^\bullet(m')$ and the definition of $\mathcal{A}_{m'}$, we have $W(m') \leq \chi_{(1)}^\bullet(m')$. Moreover, given expression (IV.8.44) for $\chi_{(1)}^\bullet(m')$ and the definition of $\mathcal{A}_{m'}$, $W(m')$ and $\chi_{(1)}^\bullet(m')$ coincide on $\Omega_{T(1)}\cap\{\chi_{(1)}^\bullet(m') \geq z\}$. As a matter of fact, since $\Phi_{K,k}$ has support K, we get

$$\|t^*\|_\infty \leq \frac{1}{\chi_{(1)}^\bullet(m')} \max_{K\in m'} \max_{k\in\{0,\dots,r\}^d} |\nu_{(1)}^\bullet(\Phi_{K,k})| \sum_{k\in\{0,\dots,r\}^d} \|\Phi_{K,k}\|_\infty$$

so that $t^* \in \mathcal{A}_{m'}$ when the event $\Omega_{T(1)}\cap\{\chi_{(1)}^\bullet(m') \geq z\}$ happens. Besides, we deduce from expression (IV.8.44), the convexity of the square function, the independence of the $A_l^\bullet$'s and Lemma 4 that

$$\begin{aligned}
\mathbb{E}_s^2[\chi_{(1)}^\bullet(m')] &\leq \sum_{K\in m'}\sum_{k\in\{0,\dots,r\}^d} \mathrm{Var}_s[\nu_{(1)}^\bullet(\Phi_{K,k})] \\
&\leq \frac{p_n}{n^2}\sum_{K\in m'}\sum_{k\in\{0,\dots,r\}^d} \mathrm{Var}_s\left(\sum_{i=1}^{q_n}\Gamma(\Phi_{K,k},Z_i)\right) \\
&\leq \vartheta q_n^\delta\|sf\|_\infty\frac{D_{m'}}{2n}.
\end{aligned}$$

Therefore, we conclude that, on the set $\Omega(x)$,

$$\chi_{(i)}^\bullet(m')\mathbb{1}_{\Omega_{T(i)}} \leq (1+\theta)\sqrt{\vartheta q_n^\delta\frac{\|sf\|_\infty}{2}}\left(\sqrt{\frac{D_{m'}}{n}}+\sqrt{2\frac{x}{n}}\right).$$

Since $\nu_{(i)}^\bullet$, $i=1,2$, is linear and $\|.\|^2 \leq \|.\|_f^2/\iota(f)$,

$$\left\{\chi_{f,(i)}^\bullet(m')\right\}^2 \leq \frac{1}{\iota(f)}\left\{\chi_{(i)}^\bullet(m')\right\}^2.$$

With the same arguments, we obtain similar concentration properties for $\chi_{(2)}^\bullet(m')\mathbb{1}_{\Omega_{T(2)}}$. Given Inequality (IV.8.43), we finally obtain (IV.8.37) with $\Omega_T = \Omega_{T(1)}\cap\Omega_{T(2)}$. In order to bound $\mathbb{P}_s(\Omega_T^c)$, it is enough to bound, for all $K \in m^\star$ and $k \in \{0,\dots,r\}^d$,

$$\mathbb{P}_s\left(\left|\sum_{l=0}^{p_n-1} U_{(K,k),l}\right| > \frac{n\epsilon}{\sqrt{|m^\star|}}\right),$$

where $U_{(K,k),l} = \sum_{i=2lq_n+1}^{(2l+1)q_n} \Gamma(\Phi_{K,k}, Z_i^\bullet)$. For that purpose, we apply Bernstein's inequality, as stated for instance in [Mas07], to the variables $(U_{(K,k),l})_{0\le l\le p_n-1}$, which are independent by construction of $(A_l^\bullet)_{0\le l\le p_n-1}$, centered and satisfy

$$\max_{0\le l\le p_n-1} |U_{(K,k),l}| \le 2\sqrt{(2r+1)^d |m^\star| q_n}$$

and, thanks to Lemma 4,

$$\sum_{l=0}^{p_n-1} \mathrm{Var}_s(U_{(K,k),l}) \le \frac{\vartheta q_n^\delta \|sf\|_\infty}{2} n.$$

Assume now that $r = 0$. The functions $(\Phi_K^f)_{K\in m'}$ introduced in Section IV.8.1 as a basis of $S_{m'}$ orthonormal for the norm $\|.\|_f$ can simply be expressed as

$$\Phi_K^f = \mathbb{1}_K / \sqrt{\mu_f(K)}, \text{ for all } K \in m'.$$

In that case, we obtain directly a concentration inequality for $\chi_{f,(1)}^\bullet(m')$ by using the basis $(\Phi_K^f)_{K\in m'}$ and the same arguments as previously, which explains why the penalty does not depend on f. ■

IV.8.5 Proof of Theorem 12

Let us fix $p > 0$, σ such that $d.\max\{1/p - 1/2, 0\} < \sigma < \min\{r + 1/p, r + 1\}$, $R > 0$, and assume that $s \in \mathcal{B}(\sigma, p, \infty, R)$. We shall use the following lemma, proved at the end of that subsection.

Lemma 7 *Let $p > 0$, σ such that $d.\max\{1/p - 1/2, 0\} < \sigma < \min\{r + 1/p, r + 1\}$, $R > 0$. Let $(x)_+$ denote the positive part of a real x. For all $s \in \mathcal{B}(\sigma, p, \infty, R)$, there exists a function $T_{J_\star}(s) \in S_{m^\star}$ such that*

$$\|T_{J_\star}(s)\|_{\sigma,p,\infty} \le C(r, d, p)R$$

and

$$\|s - T_{J_\star}(s)\|^2 \le C(\sigma, r, d, p)R^2 2^{-2dJ_\star(\sigma/d - (1/p-1/2)_+)}.$$

Let $T_{J_\star}(s)$ be given by Lemma 7. It follows from the triangle inequality that, for all partition $m \in \mathcal{M}^{cube}$,

$$\|s - s_m\|^2 \le 2\left(\|s - T_{J_\star}(s)\|^2 + \inf_{t\in S_m} \|T_{J_\star}(s) - t\|^2\right)$$

hence,

$$\min_{m\in\mathcal{M}^{cube}} \left\{\|s - s_m\|_f^2 + \frac{\ln^\delta(n) D_m}{n}\right\}$$
$$\le 2\|f\|_\infty (r+1)^d \left(\|s - T_{J_\star}(s)\|^2 + \min_{1\le D\le 2^{dJ_\star}} \left\{\min_{\substack{m\in\mathcal{M}^{cube}\\ |m|\le D}} \inf_{t\in S_m} \|T_{J_\star}(s) - t\|^2 + \frac{\ln^\delta(n) D}{n}\right\}\right).$$

Let us fix $1 \le D \le 2^{dJ_\star}$. Since $T_{J_\star}(s) \in \mathcal{B}(\sigma, p, \infty, C(r, d, p)R)$, we deduce from Corollary 3.2 in [DY90] that there exist some partition $m_D(s)$ of $[0, 1]^d$ into at most D dyadic cubes and some function $\bar{s}_D$, piecewise polynomial with coordinate degree $\le r + 1$ over that partition, such that

$$\|T_{J_\star}(s) - \bar{s}_D\|^2 \le C(\sigma, r, p, d)R^2 D^{-2\sigma/d}.$$

Besides, as $T_{J_\star}(s) \in S_{m^\star}$, the partition $m_D(s)$ generated by the algorithm described in [DY90] (Section 2) only contains dyadic cubes with sidelength $\geq 2^{-J_\star}$. Otherwise said, $m_D(s) \in \mathcal{M}^{cube}$, hence

$$\min_{\substack{m\in\mathcal{M}^{cube}\\|m|\leq D}} \inf_{t\in S_m} \|T_{J_\star}(s)-t\|^2 \leq \|T_{J_\star}(s)-\bar{s}_D\|^2 \leq C(\sigma,r,p,d)R^2D^{-2\sigma/d}.$$

Let us now choose $D^\star$ as the greatest integer D such that $D\ln^\delta(n)/n \leq R^2D^{-2\sigma/d}$. The assumptions over R ensure that $1 \leq D^\star \leq 2^{dJ_\star}$, and such a choice of D leads to

$$\min_{1\leq D\leq 2^{dJ_\star}} \left\{ \min_{\substack{m\in\mathcal{M}^{cube}\\|m|\leq D}} \inf_{t\in S_m} \|T_{J_\star}(s)-t\|^2 + \frac{\ln^\delta(n)D}{n} \right\} \leq C(\sigma,d)\left(R\left(\frac{n}{\ln^\delta(n)}\right)^{-\sigma/d}\right)^{2d/(d+2\sigma)}.$$

We then deduce from Lemma 7 that

$$\begin{aligned}&\min_{m\in\mathcal{M}^{cube}} \left\{ \|s-s_m\|_f^2 + \frac{\ln^\delta(n)D_m}{n} \right\}\\ &\leq 2\|f\|_\infty(r+1)^dC(\sigma,r,d,p)\left(R^2 2^{-2dJ_\star(\sigma/d-(1/p-1/2)_+)} + \left(R\left(\frac{n}{\ln^\delta(n)}\right)^{-\sigma/d}\right)^{2d/(d+2\sigma)}\right)\end{aligned}$$

and then only retain the values of R for which the first term in the upper-bound is smaller than the second, hence Inequality (IV.6.23).

We end with the proof of Lemma 7. We shall use the same notation as in [DP88], that we will not redefine formally here, and apply their results with r replaced with $r+1$ and α replaced with σ. Let us first assume that $p \leq 2$. For all $k \in \mathbb{N}$, let $T_k(s)$ be defined as in [DP88] (4.20), by taking $A = 1$ in [DP88] (3.5). For all $k \in \mathbb{N}$, let $t_k(s) = T_k(s) - T_{k-1}(s)$, with $T_{-1}(s) = 0$ and let $\mathbf{D}_k([0,1]^d)$ be the set of all dyadic cubes of $[0,1]^d$ with sidelength $\geq 2^{-k}$. In $\mathbb{L}_p([0,1]^d)$, we have the equality

$$s - T_{J_\star}(s) = \sum_{k\geq J_\star+1} t_k(s).$$

By construction, each $T_k(s)$ is polynomial with coordinate degree $\leq r$ on each dyadic cube in $\mathbf{D}_k([0,1]^d)$, and thus, so is $t_k(s)$. For all $I \in \mathbf{D}_k([0,1]^d)$, Markov Inequality for polynomials (see for instance [DL93], Theorem 2.7, Chapter 4) provides

$$\|t_k(s)\mathbb{1}_I\| \leq C(r,d,p)2^{dk(1/p-1/2)}\|t_k(s)\mathbb{1}_I\|_p.$$

So, by using the usual inequality between ℓ_p-norm and ℓ_2-norm, we get

$$\|t_k(s)\|^2 \leq C(r,d,p)2^{2dk(1/p-1/2)}\|t_k(s)\|_p^2.$$

Using the B-spline decomposition of $t_k(s)$ given in [DP88] (5.6)

$$t_k(s) = \sum_{\nu\in\Lambda(k)} \alpha_{\nu,k}(s)N_{\nu,k},$$

and their Lemma 4.2 to bound $\|t_k(s)\|_p^2$, we get

$$\|t_k(s)\| \leq C(r,d,p)2^{-dk(\sigma/d+1/2-1/p)}2^{\sigma k}\left(\sum_{\nu\in\Lambda(k)} |\alpha_{\nu,k}(s)|^p 2^{-kd}\right)^{1/p}.$$

Let $N_4(s)$ be defined as in [DP88], Corollary 5.3. From the triangle inequality, we deduce that

$$\|s - T_{J_\star}(s)\| \leq \sum_{k \geq J_\star + 1} \|t_k(s)\| \leq C(\sigma, r, d, p) N_4(s) 2^{-dJ_\star(\sigma/d + 1/2 - 1/p)}.$$

Last, since $0 < \sigma < r + 1$, the proof of Corollary 5.3 in [DP88] provides

$$N_4(s) \leq C(r, d, p) \|s\|_{\sigma,p,p},$$

hence the second inequality in Lemma 7. According to [DP88] (4.21),

$$\|T_{J_\star}(s)\|_p \leq C(r, d) \|s\|_p.$$

For all $k \in \mathbb{N}$, let Σ_k be the space of splines with coordinate degree $\leq r$ associated to the collection $\mathbf{D}_k([0,1]^d)$ defined as in [DP88], Section 4. For all $k \in \mathbb{N}$ and all $f \in \mathbb{L}_p([0,1]^d$, let $s_k(f)_p$ be the approximation error of f by Σ_k in $\mathbb{L}_p$-norm, as defined by [DP88] (4.27). Given the definition of $T_{J_\star}(s)$, we obtain that $s_k(T_{J_\star}(s))_p = 0$ if $k \geq J_\star$. For $k < J_\star$, we apply the quasi-triangle inequality (*cf.* [DP88], Section 2) and Corollary 4.7 of [DP88] to get

$$s_k(T_{J_\star}(s))_p \leq C(r, d, p)(s_{J_\star}(s)_p + s_k(s)_p) \leq C(r, d, p) s_k(s)_p.$$

Let N_1 be defined as [DP88], Theorem 5.1, we then deduce that

$$N_1(T_{J_\star}(s)) \leq N_1(s).$$

According to Theorem 5.1 in [DP88], N_1 is equivalent to $\|.\|_{\sigma,p,p}$ when $\sigma < \min\{r + 1/p, r + 1\}$, which concludes the proof for $p \leq 2$. The case $p \geq 2$ follows from the case $p = 2$ since $\|.\|_{\sigma,2,2} \leq C(\sigma, d, p)\|.\|_{\sigma,p,p}$ when $p \geq 2$.

IV.8.6 Proof of Theorem 14

We consider a special subcollection of dyadic rectangles adapted to an anistropic smoothness measured by the parameter $\boldsymbol{\sigma}$. Such rectangles, or similar ones, have been used by [Lei03] and [Hoc02] for instance to study wavelet approximation in anisotropic Besov spaces. Let $\underline{\boldsymbol{\sigma}}$ be defined by $\underline{\boldsymbol{\sigma}} = \min_{1 \leq l \leq d} \sigma_l$. For $j \in \mathbb{N}$, we define $\mathcal{D}_j^{\boldsymbol{\sigma}}$ as the set of all dyadic rectangles $I_1 \times \ldots \times I_d \subset [0,1]^d$ such that, for all $1 \leq l \leq d$,

$$I_l = \left[k_l 2^{-\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor}, (k_l + 1) 2^{-\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor} \right],$$

with $k_l \in \{0, \ldots, 2^{\lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor} - 1\}$. Let $j \in \mathbb{N}$ and $K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$, we call children of K all the dyadic rectangles of $\mathcal{D}_{j+1}^{\boldsymbol{\sigma}}$ that are included in K, and also refer to K as the parent of its children. It should be noticed that the children of K form a partition of K into

$$\prod_{l=1}^{d} 2^{\lfloor (j+1)\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor - \lfloor j\underline{\boldsymbol{\sigma}}/\sigma_l \rfloor} \leq 2^d 2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \tag{IV.8.46}$$

dyadic rectangles from $\mathcal{D}_{j+1}^{\boldsymbol{\sigma}}$. We set $\mathcal{D}^{\boldsymbol{\sigma}} = \cup_{j \in \mathbb{N}} \mathcal{D}_j^{\boldsymbol{\sigma}}$.

We shall use the following approximation algorithm, adapted from [DY90]. Let $t \in \mathbb{L}_2([0,1]^d)$. For any rectangle $K \subset [0,1]^d$, let

$$\mathcal{E}_2(t, K) = \inf_{P \in \mathscr{P}_r} \|(t - P)\mathbb{1}_K\|,$$

where $\mathscr{P}_r$ stands for the set of all polynomial functions on $[0,1]^d$ with coordinate degree $\leq r$. We fix some threshold $\epsilon > 0$. At the beginning of the algorithm, the set $\mathcal{I}^1(t, \epsilon)$ contains $[0,1]^d$.

If $\mathcal{E}_2(t,[0,1]^d) < \epsilon$, then the algorithm stops. Else, $[0,1]^d$ is replaced with his children in $\mathcal{I}^1(t,\epsilon)$, hence a new partition $\mathcal{I}^2(t,\epsilon)$. In the same way, the k-th step begins with a partition $\mathcal{I}^k(t,\epsilon)$ of $[0,1]^d$ into dyadic rectangles that belong to $\mathcal{D}^{\boldsymbol{\sigma}}$. If $\max_{K\in\mathcal{I}^k(t,\epsilon)} \mathcal{E}_2(t,K) < \epsilon$, then the algorithm stops. Else, a dyadic rectangle $K \in \mathcal{I}^k(t,\epsilon)$ such that $\mathcal{E}_2(t,K) \geq \epsilon$ is chosen and replaced with his children in $\mathcal{I}^k(t,\epsilon)$, hence a new partition $\mathcal{I}^{k+1}(t,\epsilon)$. Since $t \in \mathbb{L}_2([0,1]^d)$, $\mathcal{E}_2(t,K)$ tends to 0 when $\mu_d(K)$ tends to 0, so the algorithm finally stops. The final partition $\mathcal{I}(t,\epsilon)$ only contains dyadic rectangles that belong to $\mathcal{D}^{\boldsymbol{\sigma}}$ and such that $\max_{K\in\mathcal{I}(t,\epsilon)} \mathcal{E}_2(t,K) < \epsilon$. For all $K \in \mathcal{I}(t,\epsilon)$, let $P_K(t)$ be the polynomial function on K with coordinate degree $\leq r$ such that $\|(t-P_K(t))\mathbb{1}_K\| = \mathcal{E}_2(t,K)$. We define $A(t,\epsilon)$ by

$$A(t,\epsilon) = \sum_{K\in\mathcal{I}(t,\epsilon)} P_K(t),$$

and then have

$$\|t - A(t,\epsilon)\|^2 = \sum_{K\in\mathcal{I}(t,\epsilon)} \|(t-P_K(t))\mathbb{1}_K\|^2 < |\mathcal{I}(t,\epsilon)|\epsilon^2. \tag{IV.8.47}$$

For that algorithm, a first approximation result can be stated as follows.

Proposition 20 *Let $k \in \mathbb{N}$, $R > 0$, $0 < p \leq 2$, $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_d) \in (\mathbb{R}_*^+)^d$ and $t \in \mathbb{L}_2([0,1]^d)$. Assume that*

$$H(\boldsymbol{\sigma})/d > 1/p - 1/2$$

and that

$$\sup_{j\in\mathbb{N}} 2^{jd(\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))(H(\boldsymbol{\sigma})/d+1/2-1/p)} \left(\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_2^p(t,K)\right)^{1/p} \leq R. \tag{IV.8.48}$$

Then, there exists some partition m of $[0,1]^d$ that only contains dyadic rectangles from $\mathcal{D}^{\boldsymbol{\sigma}}$ and such that

$$|m| \leq C_1(d,p,\boldsymbol{\sigma})2^{kd}$$

and

$$\|t - t_m\|^2 \leq C_2(d,p,\boldsymbol{\sigma})R^2 2^{-2kH(\boldsymbol{\sigma})},$$

where t_m is the orthogonal projection of t on S_m. Besides, if for some $J \in \mathbb{N}$, t is polynomial with coordinate degree $\leq r$ over each rectangle of $\mathcal{D}_J^{\boldsymbol{\sigma}}$, then m only contains dyadic rectangles from $\cup_{j=0}^J \mathcal{D}_j^{\boldsymbol{\sigma}}$.

The following proof is inspired from [DPW08] (Lemma 2.1).

Proof: For $k = 0$, we can just choose m as the trivial partition of $[0,1]^d$ since

$$\|t - t_m\|^2 = \mathcal{E}_2^2(t,[0,1]^d) \leq R^2.$$

Let us now fix $k \geq 1$, set

$$\tau = H(\boldsymbol{\sigma})/d + 1/2 - 1/p \quad \text{and} \quad \lambda = 2^{(1+(1+\tau p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))d/p},$$

and choose

$$\epsilon = \lambda R 2^{-kd(\tau+1/p)}.$$

If $\mathcal{I}(t,\epsilon)$ is trivial, then (IV.8.47) provides

$$\|t - A(t,\epsilon)\|^2 \leq |\mathcal{I}(t,\epsilon)|\epsilon^2 \leq \lambda^2 R^2 2^{-2kH(\boldsymbol{\sigma})}.$$

Let us now assume that $\mathcal{I}(t,\epsilon)$ is not trivial and fix $j \geq 1$ such that $\mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}$ is not empty. If $K \in \mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}$, then K is a child of a dyadic rectangle $\tilde{K} \in \mathcal{D}_{j-1}^{\boldsymbol{\sigma}}$ such that

$$\epsilon \leq \mathcal{E}_2(t,\tilde{K}),$$

hence

$$\epsilon^p \leq 2^{-(j-1)dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} 2^{(j-1)dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \mathcal{E}_2^p(t,\tilde{K}).$$

By grouping all the rectangles $K \in \mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}$ having the same parent in $\mathcal{D}_{j-1}^{\boldsymbol{\sigma}}$, and taking into account Remark (IV.8.46), we obtain

$$|\mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}| \epsilon^p \leq 2^{d(1+(1+p\tau)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} R^p.$$

Replacing ϵ by its value, we deduce that

$$|\mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}| \leq 2^{kd(1+p\tau)} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}. \tag{IV.8.49}$$

Besides, for all $j \geq 1$,

$$|\mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}| \leq |\mathcal{D}_j^{\boldsymbol{\sigma}}| \leq 2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}.$$

Let us denote by J the greatest integer $j \geq 1$ such that

$$2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd(1+p\tau)} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})},$$

i.e. such that

$$2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd}.$$

Since $\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}) \leq 1$, the last inequality is satisfied by $k \geq 1$ for instance, so that J is well-defined. Besides, J is characterized by

$$2^{Jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \leq 2^{kd} < 2^{(J+1)d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}.$$

Therefore,

$$\begin{aligned}
|\mathcal{I}(t,\epsilon)| &= \sum_{j\geq 1} |\mathcal{I}(t,\epsilon) \cap \mathcal{D}_j^{\boldsymbol{\sigma}}| \\
&\leq \sum_{j=1}^{J} 2^{jd\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{kd(1+p\tau)} \sum_{j\geq J+1} 2^{-jdp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} \\
&\leq C_1(d,p,\boldsymbol{\sigma}) 2^{kd}
\end{aligned}$$

where

$$C_1(d,p,\boldsymbol{\sigma}) = \frac{2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}}{2^{d\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} - 1} + \frac{1}{1 - 2^{-dp\tau\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})}}.$$

Moreover, we deduce from (IV.8.47) that

$$\|t - A(t,\epsilon)\|^2 \leq |\mathcal{I}(t,\epsilon)| \epsilon^2 \leq C_2(d,p,\boldsymbol{\sigma}) R^2 2^{-2kH(\boldsymbol{\sigma})},$$

where

$$C_2(d,p,\boldsymbol{\sigma}) = C_1(d,p,\boldsymbol{\sigma}) \lambda^2.$$

We then set $m = \mathcal{I}(t,\epsilon)$, so that $A(t,\epsilon) = t_m$.

The last assertion in Proposition 20 is a straightforward consequence of the approximation algorithm. ■

We also need the following lemma, relating Assumption (IV.8.48) to Besov type smoothness.

Lemma 8 *Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in (0, r+1)^d$ and $0 < p \leq 2$ such that $H(\boldsymbol{\sigma})/d > 1/p - 1/2$. Then $B_p^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d)) \subset L_2([0,1]^d)$. Besides, for all $J \in \mathbb{N}$ and all $t \in B_p^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d))$, there exists t_J, polynomial with coordinate degree $\leq r$ over each dyadic rectangle of $\mathcal{D}_J^{\boldsymbol{\sigma}}$, such that*

$$\|t - t_J\|^2 \leq C(d,p,r,\boldsymbol{\sigma}) 2^{-2Jd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} |t|^2_{\boldsymbol{\sigma},p,p}$$

and

$$\sup_{j\in\mathbb{N}} 2^{jd(\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))(H(\boldsymbol{\sigma})/d+1/2-1/p)} \left(\sum_{K\in\mathcal{D}_j^{\boldsymbol{\sigma}}} \mathcal{E}_2^p(t_J, K) \right)^{1/p} \leq C(d,p,r,\boldsymbol{\sigma}) |t|_{\boldsymbol{\sigma},p,p}.$$

Proof: According to [Tri06] (Theorem 5.30, Embeddings (1.299) and Equality (1.2)), if $0 < p \leq 2$ and $H(\boldsymbol{\sigma})/d > 1/p - 1/2$, then

$$B_p^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d)) \subset B_2^0(\mathbb{L}_2([0,1]^d)) = \mathbb{L}_2([0,1]^d).$$

Let us fix $j \in \mathbb{N}$ and $K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$. For all $k \geq j$, we denote by $\mathcal{C}_k(K)$ the set of all rectangles from $\mathcal{D}_k^{\boldsymbol{\sigma}}$ that are included in K. Hence, $\mathcal{C}_j(K)$ is reduced to $\{K\}$, $\mathcal{C}_{j+1}(K)$ is the set of all the children of K, *etc.*... For any rectangle $I \subset [0,1]^d$, we denote by

$$\mathcal{E}_p(t, I) = \inf_{Q\in\mathscr{P}_r} \|(t-Q)\mathbb{1}_I\|_p,$$

and by $Q_I(t)$ a polynomial function on I with coordinate degree $\leq r$ such that

$$\|(t - Q_I(t))\mathbb{1}_I\|_p = \mathcal{E}_p(t, I).$$

For all $k \geq j$, we set

$$S_k(t, K) = \sum_{I\in\mathcal{C}_k(K)} Q_I(t)\mathbb{1}_I$$

and

$$e_k(t, K) = \inf_{P\in\Pi_{k,r}} \|(t-P)\mathbb{1}_K\|_p,$$

where $\Pi_{k,r}$ is the set of all functions that are polynomial with coordinate degree $\leq r$ over each rectangle in $\mathcal{D}_k^{\boldsymbol{\sigma}}$. It should be noticed that

$$e_k(t, K) = \|(t - S_k(t, K))\mathbb{1}_K\|_p = \left(\sum_{I\in\mathcal{C}_k(K)} \mathcal{E}_p^p(t, I) \right)^{1/p}.$$

Besides, for all $y \geq 0$ and $1 \leq l \leq d$, we define $\omega_{r+1}^{(l)}(t, y, K)_p$ as $\omega_{r+1}^{(l)}(t, y, [0,1]^d)_p$, just replacing $[0,1]^d$ with K, and

$$|t|_{\boldsymbol{\sigma},p,p,K} = \sum_{l=1}^d \left(\int_0^\infty \left[y^{-\sigma_l} \omega_{r+1}^{(l)}(t, y, K)_p \right]^p \frac{\mathrm{d}y}{y} \right)^{1/p}.$$

In particular, $|t|_{\boldsymbol{\sigma},p,p,[0,1]^d}$ coincides with $|t|_{\boldsymbol{\sigma},p,p}$. The sequence $(S_k(t,K))_{k\geq j}$ converges to $t\mathbb{1}_K$ in $\mathbb{L}_p$. As a matter of fact, we can prove as Theorem 3.5 in [DS84] when $p \geq 1$, and as Theorem 3.1 in [Hoc02] when $0 < p < 1$, that, for all dyadic rectangles $I \in \mathcal{D}_k^{\boldsymbol{\sigma}}$,

$$\mathcal{E}_p(t, I) \leq C(d,p,\boldsymbol{\sigma}) 2^{-k\underline{\boldsymbol{\sigma}}} |t|_{\boldsymbol{\sigma},p,p,I}.$$

Therefore,

$$\|(t - S_k(t,K))\mathbb{1}_K\|_p^p \leq \sum_{I\in\mathcal{D}_k^{\boldsymbol{\sigma}}} \mathcal{E}_p^p(t, I) \leq C(d,p,\boldsymbol{\sigma}) 2^{-pk\underline{\boldsymbol{\sigma}}} \sum_{I\in\mathcal{D}_k^{\boldsymbol{\sigma}}} |t|^p_{\boldsymbol{\sigma},p,p,I}.$$

Besides , $\mathcal{D}_k^{\boldsymbol{\sigma}}$ is a partition of $[0,1]^d$, so by using the equivalence between the modulus of smoothness $\omega_r^{(l)}$ and the averaged modulus of smoothness, as explained for instance in [Hoc02] (Section 2.2), we obtain

$$\sum_{I\in\mathcal{D}_k^{\boldsymbol{\sigma}}} |t|^p_{\boldsymbol{\sigma},p,p,I} \leq C(d,p,\boldsymbol{\sigma})|t|^p_{\boldsymbol{\sigma},p,p}, \tag{IV.8.50}$$

hence

$$e_k(t,K) \leq C(d,p,\boldsymbol{\sigma})2^{-k\underline{\boldsymbol{\sigma}}}|t|_{\boldsymbol{\sigma},p,p} \xrightarrow[k\to+\infty]{} 0. \tag{IV.8.51}$$

Let us now fix $k \geq j$. Using Markov Inequality for polynomials on dyadic rectangles (see for instance [Hoc02], Lemma 5.1), we obtain

$$\begin{aligned}\|S_{k+1}(t,K) - S_k(t,K)\|^2 &= \sum_{I\in\mathcal{C}_{k+1}(K)} \|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|^2 \\ &\leq C(d,p,r) \sum_{I\in\mathcal{C}_{k+1}(K)} \mu_d(I)^{-2(1/p-1/2)}\|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p^2 \\ &\leq C(d,p,r)2^{2(k+1)d(1/p-1/2)\underline{\boldsymbol{\sigma}}/H(\sigma)} \sum_{I\in\mathcal{C}_{k+1}(K)} \|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p^2.\end{aligned} \tag{IV.8.52}$$

Let us also fix $I \in \mathcal{C}_{k+1}(K)$. Then

$$(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I = (Q_I(t) - Q_{\tilde{I}}(t))\mathbb{1}_I$$

where $\tilde{I} \in \mathcal{C}_k(K)$ is the parent of I. Let $\kappa(p) = 2^{1/p}$ if $p < 1$, and $\kappa(p) = 1$ otherwise. From the quasi-triangle inequality, we then get

$$\begin{aligned}\|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p &\leq \kappa(p)\left(\|(t - Q_I(t))\mathbb{1}_I\|_p + \|(t - Q_{\tilde{I}}(t))\mathbb{1}_I\|_p\right) \\ &\leq \kappa(p)\left(\mathcal{E}_p(t,I) + \mathcal{E}_p(t,\tilde{I})\right),\end{aligned}$$

hence

$$\|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p^2 \leq 2\kappa^2(p)\left(\mathcal{E}_p^2(t,I) + \mathcal{E}_p^2(t,\tilde{I})\right).$$

By grouping all the rectangles $I \in \mathcal{C}_{k+1}(K)$ that have the same parent, we obtain

$$\sum_{I\in\mathcal{C}_{k+1}(K)} \|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p^2 \leq 2\kappa^2(p)\left(\sum_{I\in\mathcal{C}_{k+1}(K)} \mathcal{E}_p^2(t,I) + 2^{d(1+\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))}\sum_{\tilde{I}\in\mathcal{C}_k(K)} \mathcal{E}_p^2(t,\tilde{I})\right).$$

Since $p \leq 2$, the classical inequality between ℓ_p and ℓ_2-norms provides

$$\sum_{I\in\mathcal{C}_{k+1}(K)} \|(S_{k+1}(t,K) - S_k(t,K))\mathbb{1}_I\|_p^2 \leq 2\kappa^2(p)\left(e_{k+1}^2(t,K) + 2^{d(1+\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma}))}e_k^2(t,K)\right).$$

Then, it follows from (IV.8.52) that for $J \geq j$,

$$\sum_{k=j}^{J} \|S_{k+1}(t,K) - S_k(t,K)\| \leq C(d,p,r,\boldsymbol{\sigma})\sum_{k=j}^{J+1} 2^{kd(1/p-1/2)\underline{\boldsymbol{\sigma}}/H(\sigma)}e_k(t,K).$$

From (IV.8.51), we then deduce that

$$\sum_{k=j}^{J} \|S_{k+1}(t,K) - S_k(t,K)\| \leq C(d,p,r,\boldsymbol{\sigma})\left(\sum_{k=j}^{J+1} 2^{-kd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\sigma)}\right)|t|_{\boldsymbol{\sigma},p,p,K}.$$

Since $H(\boldsymbol{\sigma})/d + 1/2 - 1/p > 0$, $(S_k(t,K))_{k\geq j}$ also converges in $\mathbb{L}_2$ to $t\mathbb{1}_K$. From the definition of $\mathcal{E}_2(t,K)$ and the triangle inequality, it follows that

$$\begin{aligned}\mathcal{E}_2(t,K) &\leq \|(t - Q_K(t))\mathbb{1}_K\| \\ &\leq \sum_{k\geq j} \|S_{k+1}(t,K) - S_k(t,K)\| \\ &\leq C(d,p,r,\boldsymbol{\sigma})|t|_{\boldsymbol{\sigma},p,p,K} 2^{-jd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\sigma)}. \end{aligned} \tag{IV.8.53}$$

For all $J \in \mathbb{N}$, let us set

$$t_J = S_J(t,[0,1]^d) = \sum_{K\in\mathcal{D}_J^{\boldsymbol{\sigma}}} Q_K(t)\mathbb{1}_K.$$

Since $\|t - t_J\|^2 = \sum_{K\in\mathcal{D}_J^{\boldsymbol{\sigma}}} \|(t - Q_K(t))\mathbb{1}_K\|^2$, the first inequality stated in Lemma 8 results from (IV.8.53), the usual inequality between ℓ_p and ℓ_2-norms for $p \leq 2$ and (IV.8.50). Let us fix anew $j \in \mathbb{N}, K \in \mathcal{D}_j^{\boldsymbol{\sigma}}$ and $k \geq j$. Since t_J is polynomial with coordinate degree $\leq r$ on each rectangle from $\cup_{j\geq J}\mathcal{D}_j^{\boldsymbol{\sigma}}$, $e_k(t_J,K) = 0$ for $k \geq J$. Let us now assume that $k < J$. Then, the quasi-triangle inequality, the definition of t_J and the inclusion $\Pi_{k,r} \subset \Pi_{J,r}$ provide successively

$$\begin{aligned} e_k(t_J,K) &\leq \kappa(p)\left(\|(t - t_J)\mathbb{1}_K\|_p + e_k(t,K)\right) \\ &\leq \kappa(p)\left(e_J(t,K) + e_k(t,K)\right) \\ &\leq 2\kappa(p)e_k(t,K). \end{aligned}$$

Therefore, we obtain as (IV.8.53) that

$$\mathcal{E}_2(t_J,K) \leq C(d,p,r,\boldsymbol{\sigma})|t|_{\boldsymbol{\sigma},p,p,K} 2^{-jd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\sigma)}.$$

Since $\mathcal{D}_j^{\boldsymbol{\sigma}}$ is a partition of $[0,1]^d$, Inequality (IV.8.50) allows to complete the proof. ■

We can now prove Theorem 14. If $p > 2$, then Theorem 5.30 in [Tri06] provides the continuous embedding

$$B_p^{\boldsymbol{\sigma}}(\mathbb{L}_p([0,1]^d)) \hookrightarrow B_2^{\boldsymbol{\sigma}}(\mathbb{L}_2([0,1]^d)),$$

so it is enough to prove Theorem 14 for $p \leq 2$. Let $k \in \mathbb{N}$ and s_J be given by Lemma 8. According to Proposition 20, there exists a partition m into dyadic rectangles from $\cup_{j=0}^J \mathcal{D}_j^{\boldsymbol{\sigma}}$ such that

$$|m| \leq C(d,p,\boldsymbol{\sigma})2^{kd}$$

and

$$\inf_{t\in S_m} \|s_J - t\|^2 \leq C(d,p,r,\boldsymbol{\sigma})R^2 2^{-2kH(\boldsymbol{\sigma})}.$$

So, we deduce from the triangle inequality and the upper-bound for $\|s - s_J\|^2$ given in Lemma 8 that for all $t \in S_m$

$$\begin{aligned}\|s - t\|^2 &\leq 2\left(\|s - s_J\|^2 + \|s_J - t\|^2\right) \\ &\leq C(d,p,r,\boldsymbol{\sigma})R^2\left(2^{-2Jd(H(\boldsymbol{\sigma})/d+1/2-1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{-2kH(\boldsymbol{\sigma})}\right),\end{aligned}$$

which completes the proof.

Appendix

Proposition 21 (Coupling property) *Let $(W_1, \dots, W_n)$ be real-valued and bounded random variables. We can redefine the vector $(W_1, \dots, W_n)$ on a richer probability space together with an independent sequence of random variables $(W_1^\bullet, \dots, W_n^\bullet)$ with the following properties, for each $1 \leq k \leq n$,*

- *$W_k^\bullet$ has the same distribution as W_k;*
- *$W_k^\bullet$ is independent of $\sigma(W_1, \dots, W_{k-1})$;*
- *$\mathbb{E}\left[|W_k - W_k^\bullet|\right] \leq 4\|W_1\|_\infty \alpha(\sigma(W_1, \dots, W_{k-1}), \sigma(W_k))$.*

We refer to [Rio00] (Lemma 5.2 and proof of Theorem 6.1) (beware of the definition of $\alpha(.,.)$) or [Pel02] (Theorem 2) for a proof.

Lemma 9 (Variance of partial sums) *Let $(W_i)_{i\in\mathbb{Z}}$ be a strictly stationary process, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathcal{W} \subset \mathbb{R}^p$, $p \in \mathbb{N}^\star$. Let $n \in \mathbb{N}^\star$ and $q \in \mathbb{N}^\star$. Then*

i) for all bounded function $g : \mathcal{W} \to \mathbb{R}$,

$$\frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n} g(W_i)\right) \leq 3q\mathrm{Var}\left(g(W_1)\right) + 12n\|g\|_\infty^2 \alpha_q^{\boldsymbol{W}}. \tag{IV.8.54}$$

ii) for all finite family $(g_I)_{I\in\mathcal{I}}$ of real-valued and bounded functions defined on $\mathcal{W}$

$$\frac{1}{n}\sum_{I\in\mathcal{I}} \mathrm{Var}\left(\sum_{i=1}^{n} g_I(W_i)\right) \leq 2(1 + 2S_{2\text{-}\alpha}) \left\| \sum_{I\in\mathcal{I}} |g_I(W_1)| \right\|_\infty^2 \tag{IV.8.55}$$

where $S_{2\text{-}\alpha} := \sum_{j\in\mathbb{N}^\star} \alpha\big(\sigma(W_1), \sigma(W_{j+1})\big)$.

Proof:

Proof of Inequality (IV.8.54)*:* Let (d, r) be the unique couple of nonnegative integers such that $n = dq + r$ and $0 \leq r < q$. Let us assume, for instance, that $d = 2p + 1$, $p \in \mathbb{N}$. The case when d is even can be treated in a similar way. For $l \in \mathbb{N}$, let us set

$$A_l = (W_i)_{2lq+1\leq i\leq (2l+1)q} \quad \text{and} \quad B_l = (W_i)_{(2l+1)q+1\leq i\leq (2l+2)q}.$$

For all $v = (v_1, \dots, v_q) \in \mathcal{W}^q$, let us set $S_g(v) = \sum_{i=1}^q g(v_i)$. Since $(W_i)_{i\in\mathbb{Z}}$ is strictly stationary, so are $(A_l)_{l\in\mathbb{N}}$ and $(B_l)_{l\in\mathbb{N}}$, and $(B_l)_{l\in\mathbb{N}}$ is distributed as $(A_l)_{l\in\mathbb{N}}$. Therefore, using also the convexity of the square function and the stationarity of $(W_i)_{i\in\mathbb{Z}}$, we deduce that

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} g(W_i)\right) &= \mathrm{Var}\left(\sum_{l=0}^{p} S_g(A_l) + \sum_{l=0}^{p-1} S_g(B_l) + \sum_{i=(2p+1)q+1}^{n} g(W_i)\right) \\
&\leq 3\left[\mathrm{Var}\left(\sum_{l=0}^{p} S_g(A_l)\right) + \mathrm{Var}\left(\sum_{l=0}^{p-1} S_g(B_l)\right) + \mathrm{Var}\left(\sum_{i=1}^{r} g(W_i)\right)\right] \\
&\leq 3\left[d\mathrm{Var}\left(S_g(A_0)\right) + \mathrm{Var}\left(\sum_{i=1}^{r} g(W_i)\right) + 2d\sum_{l=1}^{p} \mathrm{Cov}\left(S_g(A_0), S_g(A_l)\right)\right]
\end{aligned}$$

and

$$\mathrm{Var}\left(S_g(A_0)\right) \leq q\mathbb{E}\left[\sum_{i=1}^{q}\left(g(W_i)-\mathbb{E}[g(W_i)]\right)^2\right] = q^2\mathrm{Var}\left(g(W_1)\right),$$

hence

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(W_i)\right) \leq 3nq\mathrm{Var}\left(g(W_1)\right) + 6d\sum_{l=1}^{p}\mathrm{Cov}\left(S_g(A_0), S_g(A_l)\right).$$

According to Proposition 21, for all $l = 1, \ldots, p$, we can build a random variable $S_{g,l}^{\bullet}$ distributed as $S_g(A_l)$, independent of $S_g(A_0)$ and such that $\mathbb{E}\left[\left|S_g(A_l) - S_{g,l}^{\bullet}\right|\right] \leq 4\|g\|_\infty q\alpha_q^{\boldsymbol{W}}$. Therefore, we get, for all $l \in \{1, \ldots, p\}$,

$$\begin{aligned}\left|\mathrm{Cov}\left(S_g(A_0), S_g(A_l)\right)\right| &= \left|\mathbb{E}\left[S_g(A_0)\left(S_g(A_l) - S_{g,l}^{\bullet}\right)\right]\right| \\ &\leq 4q^2\|g\|_\infty^2\alpha_q^{\boldsymbol{W}}\end{aligned}$$

hence Inequality (IV.8.54).

Proof of Inequality (IV.8.55)*:* The argument is used by [Rio00] in the proof of Theorem 1.3. We reproduce it here for the sake of completeness. Let $(\varepsilon_I)_{I\in\mathcal{I}}$ be independent random variables such that, for all $I \in \mathcal{I}$, $\mathbb{P}(\varepsilon_I = 1) = \mathbb{P}(\varepsilon_I = -1) = 1/2$. Assume moreover that $(\varepsilon_I)_{I\in\mathcal{I}}$ is independent of $(W_i)_{i\in\mathbb{Z}}$. The random variables $(\sum_{i=1}^n \varepsilon_I g_I(W_i))_{I\in\mathcal{I}}$ are centered, uncorrelated and such that, for all $I \in \mathcal{I}$, $\mathrm{Var}\left(\sum_{i=1}^n \varepsilon_I g_I(W_i)\right) = \mathrm{Var}\left(\sum_{i=1}^n g_I(W_i)\right)$, so

$$\mathrm{Var}\left(\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}\varepsilon_I g_I(W_i)\right) = \sum_{I\in\mathcal{I}}\mathrm{Var}\left(\sum_{i=1}^{n} g_I(W_i)\right).$$

Let us denote by U some random variable uniformly distributed over $[0,1]$. According to [Rio00] Corollary 1.1 and Section 1.4.4, for all real-valued strictly stationary process $(X_k)_{k\in\mathbb{Z}}$, there exist measurable functions α^{-1} and $(Q_k)_{k\in\mathbb{Z}}$ such that $2\int_0^1 \alpha^{-1}(x)\mathrm{d}x \leq 1 + 2\sum_{j\in\mathbb{N}^\star}\alpha(\sigma(X_1), \sigma(X_{j+1}))$, $Q_k^2(U)$ is distributed as X_k^2 and

$$\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) \leq 4\sum_{k=1}^{n}\mathbb{E}\left[\alpha^{-1}(U)Q_k^2(U)\right],$$

hence

$$\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) \leq 2\left(1 + 2\sum_{j\in\mathbb{N}^\star}\alpha(\sigma(X_1), \sigma(X_{j+1}))\right)\sum_{k=1}^{n}\|X_k\|_\infty^2.$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{I\in\mathcal{I}}\varepsilon_I g_I(W_i)\Big|(\varepsilon_I)_{I\in\mathcal{I}}\right) &\leq 2n(1+2S_{2\text{-}\alpha})\left\|\sum_{I\in\mathcal{I}}\varepsilon_I g_I(W_1)\right\|_\infty^2 \\ &\leq 2n(1+2S_{2\text{-}\alpha})\left\|\sum_{I\in\mathcal{I}}|g_I(W_1)|\right\|_\infty^2\end{aligned}$$

which completes the proof. ■

Proposition 22 (Bernstein-type inequality) *Let $(W_i)_{i\in\mathbb{Z}}$ be a strictly stationary process, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathcal{W} \subset \mathbb{R}^p$, $p \in \mathbb{N}^\star$. Let $n \in \mathbb{N}^\star$, $q \in \mathbb{N}^\star$*

and g be a real-valued and bounded function defined on $\mathcal{W}$. Let $\sigma_g^2 = \mathrm{Var}(g(W_1))$. Then, for all $x > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\left(g(W_i) - \mathbb{E}[g(W_i)]\right)\right| \geq 3\left(\sqrt{2qn\sigma_g^2 x} + 2q\|g\|_\infty x/3\right)\right)$$
$$\leq 2\exp\left(-x\right) + 4n\|g\|_\infty \alpha_q^{\boldsymbol{W}} / \max\left\{2q\|g\|_\infty x/3, \sqrt{2qn\sigma_g^2 x}\right\}. \qquad \text{(IV.8.56)}$$

The proof uses the same arguments as Theorem 6.1 in [Rio00] and relies on the Bernstein inequality for independent random variables as stated in [Mas07] (Section 2.2.3), for instance.

Perspectives

Selecting a best partition into dyadic intervals or dyadic rectangles seems to provide in theory a highly interesting estimation procedure in many frameworks. Besides, the simulations realized in Chapter II tend to confirm those qualities in practice. These are promising results that could be pushed further by considering several questions linked with the choice of the penalty. We could for instance try to provide a more refined one, of the form

$$\mathrm{pen}(m) = \frac{c}{n} \sum_{I \in m} p(I)$$

for the problems treated in Chapters III and IV, where c is a positive constant and $\{p(I)\}_I$ is an adequate collection of penalties associated with the dyadic rectangles. When all the $p(I)$'s are equal, we recover the penalty proposed in the previous works. From a practical point of view, that would only amount to change the definition of the weights assigned to each dyadic rectangle in the shortest-path algorithm used to compute the penalized estimator. Thus, the computational complexity would still be linear in the sample size. Another piece of work consists in realizing an extensive simulation study, so as to propose an adequate choice of the penalty constant, that would thus make the procedure really usable in practice. In particular, regarding the density estimation framework with possibly dependent data, where a factor depending on the unknown dependence structure appears in the penalty, it would be interesting to study whether the data-driven choice of the constant proposed in Chapter II still performs well. Last, beginning with the classical density estimation framework, several reference procedures, and among them spatially adaptive ones, should be implemented together with our estimator so as to provide a comparative study for various sample sizes and various measures of estimation accuracy.

The estimation problems we have studied also lead to new ones, that require a different treatment. Thus, copula density estimation can be viewed as an estimation problem with ill-observed data that does not fit into the framework defined in Chapter III, but can indeed be tackled via model selection based on an adequate contrast. Regarding estimation based on dependent data, it would be interesting to provide spatially adaptive estimators under weaker conditions of dependence, such as α-mixing conditions or weak dependence conditions as defined for instance in [DL99; DP05]. Yet, the lack of smoothness of piecewise polynomial functions seems to be a real drawback, even to study the quadratic risk of one histogram built on an irregular partition. Therefore, in such cases, the use of the collection of wavelet-based models inspired from the compression algorithm of Birgé and Massart [BM00; Mas07] should be considered, and extended to multivariate functions with possible anisotropic smoothness.

Bibliographie

[Ada08] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13 :no. 34, 1000–1034, 2008.

[Aka09] N. Akakpo. Estimating a discrete distribution via histogram selection. *ESAIM Probab. Statist.*, To appear, 2009.

[AN98] P. Ango Nze. Critères d'ergodicité géométrique ou arithmétique de modèles linéaires perturbés à représentation markovienne. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(3) :371–376, 1998.

[AV95] M. Aerts and N. Veraverbeke. Bootstrapping a nonparametric polytomous regression model. *Math. Methods Statist.*, 4(2) :189–200, 1995.

[Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.

[Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[BB09] Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2) :239–284, 2009.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[BBM00] J. V. Braun, R. K. Braun, and H.-G. Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2) :301–314, 2000.

[BBM08] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2) :489–531, 2008.

[BC05] E. Brunel and F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3) :441–475, 2005.

[BC06] E. Brunel and F. Comte. Adaptive nonparametric regression estimation in presence of right censoring. *Math. Methods Statist.*, 15(3) :233–255, 2006.

[BC08] E. Brunel and F. Comte. Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory Methods*, 37(8-10) :1284–1305, 2008.

[BC09] E. Brunel and F. Comte. Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.*, 3 :1–24, 2009.

[BCG09] E. Brunel, F. Comte, and A. Guilloux. Nonparametric density estimation in presence of bias and censoring. *TEST*, 18(1) :166–194, 2009.

[BCL07] E. Brunel, F. Comte, and C. Lacour. Adaptive Estimation of the Conditional Density in Presence of Censoring. *Sankhyā*, 69(4) :734–763, 2007.

[BCV01] Y. Baraud, F. Comte, and G. Viennet. Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, 5 :33–49 (electronic), 2001.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[Bir04] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6) :1039–1051, 2004.

[Bir06a] L. Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3) :273–325, 2006.

[Bir06b] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4) :497–537, 2006.

[Bir07] L. Birgé. Model selection for Poisson processes. In *Asymptotics : particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 32–64. Inst. Math. Statist., Beachwood, OH, 2007.

[BLM99] D. Bitouzé, B. Laurent, and P. Massart. A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(6) :735–763, 1999.

[BM93] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2) :113–150, 1993.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM98a] L. Birgé and P. Massart. Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329–375, 1998.

[BM98b] J.V. Braun and H.-G. Müller. Statistical methods for DNA sequence segmentation. *Statistical Science*, pages 142–162, 1998.

[BM00] L. Birgé and P. Massart. An adaptive compression algorithm in Besov spaces. *Constr. Approx.*, 16(1) :1–36, 2000.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[Bos98] D. Bosq. *Nonparametric statistics for stochastic processes*, volume 110 of *Lecture Notes in Statistics*. Springer-Verlag, New York, second edition, 1998. Estimation and prediction.

[Bra05] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2 :107–144 (electronic), 2005. Update of, and a supplement to, the 1986 original.

[Bra07] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 1.* Kendrick Press, Heber City, UT, 2007.

[BS67] M. Š. Birman and M. Z. Solomjak. Piecewise polynomial approximations of functions of classes $W_p{}^{\alpha}$. *Mat. Sb. (N.S.)*, 73 (115) :331–355, 1967.

[BS88] C. Bennett and R. Sharpley. *Interpolation of operators*, volume 129 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, 1988.

[BSRM07] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2) :209–241, 2007.

[Cai99] T. T. Cai. Adaptive wavelet estimation : a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3) :898–924, 1999.

[Cai02] T. T. Cai. On block thresholding in wavelet regression : adaptivity, block size, and threshold level. *Statist. Sinica*, 12(4) :1241–1273, 2002.

[Cas00] G. Castellan. Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8) :729–732, 2000.

[Cas03] G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8) :2052–2060, 2003.

[CC05] E. Chicken and T. T. Cai. Block thresholding for density estimation : local and global adaptivity. *J. Multivariate Anal.*, 95(1) :76–106, 2005.

[Che07] C. Chesneau. Wavelet block thresholding for samples with random design : a minimax approach under the L^p risk. *Electron. J. Stat.*, 1 :331–346 (electronic), 2007.

[Che08] C. Chesneau. Wavelet estimation via block thresholding : a minimax study under L^p risk. *Statist. Sinica*, 18(3) :1007–1024, 2008.

[Che09] C. Chesneau. Wavelet block thresholding for density estimation in the presence of bias. *Journal of the Korean Statistical Society*, To appear, 2009.

[Clé00a] S. Clémençon. *Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov*. PhD thesis, Doctoral thesis, Université Paris VII, 2000.

[Clé00b] S. J. M. Clémençon. Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.*, 9(4) :323–357, 2000.

[CLRS01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[CM02] F. Comte and F. Merlevède. Adaptive estimation of the stationary density of discrete and continuous time mixing processes. *ESAIM Probab. Statist.*, 6 :211–238 (electronic), 2002.

[CS01] T. T. Cai and B. W. Silverman. Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā Ser. B*, 63(2) :127–148, 2001. Special issue on wavelets.

[Csű04] M. Csűrös. Algorithms for finding maximal-scoring segment sets (extended abstract). In *Algorithms in bioinformatics*, volume 3240 of *Lecture Notes in Comput. Sci.*, pages 62–73. Springer, Berlin, 2004.

[CZ09] T. T. Cai and H. H. Zhou. A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.*, 37(2) :569–595, 2009.

[DeV98] R. A. DeVore. Nonlinear approximation. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[DG83] P. Doukhan and M. Ghindès. Estimation de la transition de probabilité d'une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d'ordre 1. *Stochastic Process. Appl.*, 15(3) :271–293, 1983.

[DGZ03] J. G. De Gooijer and D. Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2) :159–176, 2003.

[DJ94a] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.*, 319(12) :1317–1322, 1994.

[DJ94b] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[DJ95] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224, 1995.

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3) :879–921, 1998.

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369, 1995. With discussion and a reply by the authors.

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2) :508–539, 1996.

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27 :642–669, 1956.

[DL92] R. A. DeVore and B. J. Lucier. Wavelets. In *Acta numerica, 1992*, Acta Numer., pages 1–56. Cambridge Univ. Press, Cambridge, 1992.

[DL93] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[DL99] Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84(2) :313–342, 1999.

[DLT09] C. Durot, E. Lebarbier, and AS Tocquet. Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli*, 15(2) :475–507, 2009.

[Don93] D. L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.*, 1(1) :100–115, 1993.

[Don96] D. L. Donoho. Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Anal.*, 3(4) :388–392, 1996.

[Don97] D. L. Donoho. CART and best-ortho-basis : a connection. *Ann. Statist.*, 25(5) :1870–1911, 1997.

[Dou94] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples.

[DP88] R. A. DeVore and V. A. Popov. Interpolation of approximation spaces. In *Constructive theory of functions (Varna, 1987)*, pages 110–119. Publ. House Bulgar. Acad. Sci., Sofia, 1988.

[DP05] J. Dedecker and C. Prieur. New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2) :203–236, 2005.

[DPW08] R. A. DeVore, G. Petrova, and P. Wojtaszczyk. Anisotropic smoothness spaces via level sets. *Comm. Pure Appl. Math.*, 61(9) :1264–1297, 2008.

[DR02] S. Döhler and L. Rüschendorf. Adaptive estimation of hazard functions. *Probab. Math. Statist.*, 22(2, Acta Univ. Wratislav. No. 2470) :355–379, 2002.

[DS84] R. A. DeVore and R. C. Sharpley. Maximal functions measuring smoothness. *Mem. Amer. Math. Soc.*, 47(293) :viii+115, 1984.

[DT93] P. Doukhan and A. B. Tsybakov. Nonparametric recurrent estimation in nonlinear ARX models. *Problemy Peredachi Informatsii*, 29(4) :24–34, 1993.

[DY90] R. A. DeVore and X. M. Yu. Degree of adaptive approximation. *Math. Comp.*, 55(192) :625–635, 1990.

[Efr07] S. Efromovich. Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6) :2504–2535, 2007.

[Efr08] S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, pages 1–27, 2008.

[Eng94] J. Engel. A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.*, 49(2) :242–254, 1994.

[Eng97] J. Engel. The multiresolution histogram. *Metrika*, 46(1) :41–57, 1997.

[Fau07] O. P. Faugeras. A product type non-parametric estimator of the conditional density by quantile transform and copula representation. www. *Arxiv preprint math.ST/0709.3192 v1*, 2007.

[FC90] Y.-X. Fu and R. N. Curnow. Maximum likelihood estimation of multiple change points. *Biometrika*, 77(3) :563–573, 1990.

[FY04] J. Fan and T. H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4) :819–834, 2004.

[GK07] L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5) :1872–1879, 2007.

[GL08] S. Gey and E. Lebarbier. Using cart to detect multiple change-points in the mean for large samples. Technical report, Technical report, Preprint SSB, 2008.

[GN97] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2) :135–170, 1997.

[GN05] S. Gey and E. Nédélec. Model selection for CART regression trees. *IEEE Trans. Inform. Theory*, 51(2) :658–670, 2005.

[GW09] I. Gannaz and O. Wintenberger. Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, To appear, 2009.

[HKP98] P. Hall, G. Kerkyacharian, and D. Picard. Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26(3) :922–942, 1998.

[HKP99] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, 9(1) :33–49, 1999.

[HKPT98] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[HNB03] M. Hoebeke, P. Nicolas, and P. Bessieres. MuGeN : simultaneous exploration of multiple genomes and computer analysis results, 2003.

[Hoc02] R. Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2) :179–208, 2002.

[HP91] P. Hilton and J. Pedersen. Generalizations of Eulerian numbers and their q-analogues. *Nieuw Arch. Wisk. (4)*, 9(3) :271–298, 1991.

[Jon04] G. L. Jones. On the Markov chain central limit theorem. *Probab. Surv.*, 1 :299–320 (electronic), 2004.

[Jud97] A. Juditsky. Wavelet estimators : adapting to unknown smoothness. *Math. Methods Statist.*, 6(1) :1–25, 1997.

[Kle09] J. Klemelä. Multivariate histograms with data-dependent partitions. *Statist. Sinica*, 19(1) :159–176, 2009.

[KLP01] G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*, 121(2) :137–170, 2001.

[KN04] E. D. Kolaczyk and R. D. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *Ann. Statist.*, 32(2) :500–527, 2004.

[Lac07] C. Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5) :571–597, 2007.

[Leb02] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Doctoral thesis, Université Paris XI, UFR Orsay, 2002.

[Lei03] C. Leisner. Nonlinear wavelet approximation in anisotropic Besov spaces. *Indiana Univ. Math. J.*, 52(2) :437–455, 2003.

[Lep91] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4) :645–659, 1991.

[Ler09] M. Lerasle. Adaptive density estimation of stationary β-mixing and τ-mixing processes. *Math. Methods Statist.*, 18(1) :59–83, 2009.

[LGM96] G. G. Lorentz, M. van Golitschek, and Y. Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. Advanced problems.

[Li07] L. Li. On the minimax optimality of wavelet estimators with censored data. *J. Statist. Plann. Inference*, 137(4) :1138–1150, 2007.

[Li08] L. Li. On the block thresholding wavelet estimators with censored data. *J. Multivariate Anal.*, 99(8) :1518–1543, 2008.

[LMS97] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness : an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3) :929–947, 1997.

[LN07] E. Lebarbier and E. Nédélec. Change-points detection for discrete sequences via model selection. *SSB preprint, Research report*, 9, 2007.

[Mal98] S. Mallat. *A wavelet tour of signal processing.* Academic Press Inc., San Diego, CA, 1998.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3) :1269–1283, 1990.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics.* Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mok90] A. Mokkadem. Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.*, 26(2) :219–260, 1990.

[MT93] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[MvdG97] E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1) :387–413, 1997.

[NBM+02] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum, and P. Bessières. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic acids research*, 30(6) :1418, 2002.

[Neu00] M. H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statist. Sinica*, 10(2) :399–431, 2000.

[NvS97] M. H. Neumann and R. von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, 25(1) :38–76, 1997.

[Pee76] J. Peetre. *New thoughts on Besov spaces.* Mathematics Department, Duke University, Durham, N.C., 1976. Duke University Mathematics Series, No. 1.

[Pel82] M. Peligrad. Invariance principles for mixing sequences of random variables. *Ann. Probab.*, 10(4) :968–981, 1982.

[Pel02] M. Peligrad. Some remarks on coupling of dependent random variables. *Statist. Probab. Lett.*, 60(2) :201–209, 2002.

[RB03] P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1) :103–153, 2003.

[RB06] P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4) :633–661, 2006.

[Ren99] O. Renaud. Density estimation with wavelets : variability, invariance, and discriminant power. *Thèse de doctorat, Ecole Polytechnique Fédérale de Lauzane*, 1999.

[Rio00] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2000.

[RR97] G. O. Roberts and J. S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.*, 2 :no. 2, 13–25 (electronic), 1997.

[RR08] G. O. Roberts and J. S. Rosenthal. Variance bounding Markov chains. *Ann. Appl. Probab.*, 18(3) :1201–1214, 2008.

[Sau02] M. Sauvé. *Sélection de modèles en régression non gaussienne. Application à la sélection de variables et aux tests de survie accélérés.* PhD thesis, Doctoral thesis, Université Paris XI, UFR Orsay, 2002.

[Sau09] M. Sauvé. Histogram selection in non Gaussian regression. *ESAIM Probab. Stat.*, 13 :70–86, 2009.

[SRS05] W. Szpankowski, W. Ren, and L. Szpankowski. An optimal DNA segmentation based on the MDL principle. *International Journal of Bioinformatics Research and Applications*, 1(1) :3–17, 2005.

[SW86] G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[Tal96] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3) :505–563, 1996.

[Tri83] H. Triebel. *Theory of function spaces*, volume 78 of *Monographs in Mathematics.* Birkhäuser Verlag, Basel, 1983.

[Tri06] H. Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics.* Birkhäuser Verlag, Basel, 2006.

[TV98] K. Tribouley and G. Viennet. $\mathbb{L}_p$ adaptive density estimation in a β mixing framework. *Ann. Inst. H. Poincaré Probab. Statist.*, 34(2) :179–208, 1998.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes. With application to statistics.* Springer Series in Statistics. Springer-Verlag, New York, 1996.

[Vie97] G. Viennet. Inequalities for absolutely regular sequences : application to density estimation. *Probab. Theory Related Fields*, 107(4) :467–492, 1997.

[WN07] R. M. Willett and Robert D. Nowak. Multiscale Poisson intensity and density estimation. *IEEE Trans. Inform. Theory*, 53(9) :3171–3187, 2007.

[YB99] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5) :1564–1599, 1999.

Estimation adaptative par sélection de partitions en rectangles dyadiques

Résumé

Dans cette thèse, nous étudions divers problèmes d'estimation par sélection d'estimateurs constants ou polynomiaux par morceaux sur des partitions en intervalles ou rectangles dyadiques, en utilisant un critère de type moindres carrés pénalisé adéquat. Nos travaux portent sur trois sujets différents. Nous nous intéressons tout d'abord à l'estimation d'une loi de probabilité discrète, ainsi qu'à une application à la détection de ruptures multiples. Puis, nous proposons un cadre unifié pour l'estimation fonctionnelle basée sur des données éventuellement censurées. Enfin, nous étudions simultanément l'estimation de densité multivariée et de densité conditionnelle pour des données dépendantes. Le choix de la collection de partitions en intervalles ou rectangles dyadiques s'avère intéressant aussi bien en théorie qu'en pratique. En effet, notre estimateur pénalisé vérifie dans chacun des cadres une inégalité de type oracle non-asymptotique, pour une pénalité bien choisie. Il atteint également la vitesse minimax à constante près sur de nombreuses classes de fonctions, dont la régularité est éventuellement à la fois non homogène et non isotrope. Cette propriété, qui à notre connaissance n'a été démontrée pour aucun autre estimateur, repose sur des résultats d'approximation dont les preuves sont inspirées d'un article de DeVore et Yu. Par ailleurs, le calcul de notre estimateur dans un cadre univarié est basé sur un algorithme de plus court chemin dont la complexité est seulement linéaire en la taille de l'échantillon.

Mots-clefs : sélection de modèle, histogramme, inégalité d'oracle, adaptation au sens minimax; approximation non-linéaire, régularité non-homogène, espace de Besov, fonctions à α-variations bornées; détection de ruptures, données censurées, données dépendantes.

Adaptive estimation by selecting a best partition into dyadic rectangles

Abstract

In this thesis, we study several estimation problems by selection of a best piecewise constant or piecewise polynomial estimator built on a partition into dyadic intervals or rectangles, using an adequate least-squares type criterion. Our works are devoted to three topics. First, we are concerned with discrete distribution estimation and provide an application to multiple change-point detection. Then, we propose a unified approach to functional estimation problems based on possibly censored data. Last, we lead a simultaneous study of multivariate density and conditional density estimation based on dependent data. The choice of the collection of partitions into dyadic intervals or rectangles reveals highly interesting in theory and in practice. As a matter of fact, our penalized estimator satisfies in each framework a nonasymptotic oracle-type inequality for a well-chosen penalty. It also reaches the minimax rate, up to a constant, over a wide range of classes of functions that may have inhomogeneous and anisotropic smoothness. Such a property, that, up to our knowledge, has never been proved for any other estimator, follows from approximation results whose proofs are inspired from a paper by DeVore and Yu. Besides, in a univariate framework, our estimator can be determined via a shortest-path algorithm whose computational complexity is only linear in the sample size.

Keywords : model selection, histogram, oracle inequality, adaptive estimation in the minimax sense; nonlinear approximation, inhomogeneous smoothness, Besov space, functions with bounded α-variation; multiple change-point detection, censored data, dependent data.