# Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie

Gaëlle Chagny

**Université Paris Descartes**
Laboratoire MAP5 UMR CNRS 8145
École doctorale 386 : Sciences Mathématiques de Paris Centre

**THÈSE**
présentée pour obtenir le grade de

Docteur de l'Université Paris Descartes

*Spécialité : Mathématiques*

Présentée par

**Gaëlle Chagny**

---

# Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie.

---

*Soutenue le 5 juillet 2013 devant la commission d'examen composée de :*

| | | |
|---|---|---|
| Fabienne Comte | Université Paris Descartes | Directrice de thèse |
| Jérôme Dedecker | Université Paris Descartes | Examinateur |
| Cécile Durot | Université Paris-Ouest | Examinatrice |
| Christophe Giraud | Université Paris-Sud et École Polytechnique | Examinateur |
| Agathe Guilloux | Université Pierre et Marie Curie | Examinatrice |
| Dominique Picard | Université Paris Diderot | Examinatrice |
| Patricia Reynaud-Bouret | Université Nice Sophia-Antipolis | Rapporteure |

**Rapporteurs :**

| | |
|---|---|
| Oleg Lepski | Université Aix-Marseille |
| Patricia Reynaud-Bouret | Université Nice Sophia-Antipolis |

# Merci !

J'ai toujours pensé pendant ma thèse que les pages de remerciements seraient les plus faciles à écrire. Je me rends compte aujourd'hui à quel point l'exercice est difficile, tant la liste des personnes ayant contribué à cette thèse est longue. Je m'excuse par avance auprès de ceux que je vais oublier de citer, et auprès de ceux qui mériteraient bien plus que le simple (mais sincère et chaleureux !) "Merci" qui suit : si je détaillais plus, vous m'auriez tous encore dit que je ne sais pas faire court !

Merci tout d'abord à Fabienne, ma directrice de thèse. Fabienne, ce n'est pas seulement parce que c'est la tradition que je commence par TOI (TU as vu, j'essaye enfin de TE tutoyer) : sans ton encadrement parfait, je n'aurais sans doute jamais été au bout de cette aventure. Merci d'avoir accepté, il y a plus de trois ans, de me proposer un sujet de stage et de thèse, alors que j'hésitais encore entre statistique et probabilités, enseignement et recherche. Merci de m'avoir fait confiance depuis ce jour-là. Merci pour tout le temps (précieux dans ton agenda si chargé) que tu as passé avec moi, pour partager ton immense connaissance en statistique, pour répondre à mes longues listes de questions, ou pour discuter de la suite. Merci pour ta sérénité, ton optimisme et ton soutien qui m'ont toujours permis de reprendre courage dans les moments de doute. Je réalise depuis longtemps maintenant quelle chance j'ai eu d'avoir une directrice si disponible, dont la porte du bureau est toujours ouverte, et qui répond aux e-mails si rapidement. J'espère que nous pourrons continuer à travailler ensemble par la suite.

Merci à mes rapporteurs, Patricia Reynaud-Bouret et Oleg Lepski. Patricia, merci pour l'intérêt que tu as porté à mon travail depuis le CIRM l'an passé. Merci pour tous les conseils que tu m'as donnés pour améliorer ce manuscrit. J'espère que nous aurons encore l'occasion de discuter. Oleg Lepski, votre méthode est à l'origine de nombreux résultats de cette thèse : merci pour la bienveillance avec laquelle vous les avez considérés.

Merci à Cécile Durot, Agathe Guilloux, Dominique Picard, Jérôme Dedecker et Christophe Giraud d'avoir si gentiment et rapidement accepté de participer au jury. C'est indéniablement grâce au merveilleux cours de M1 (puis de M2) de Cécile Durot que j'ai choisi de faire des statistiques : merci pour votre rigueur et votre enthousiasme. Agathe, merci beaucoup d'avoir accepté de m'aider en analyse de survie : tes conseils m'ont permis de "débloquer" les résultats du Chapitre 7. Dominique Picard est l'auteure, avec Gérard Kerkyacharian, de l'article qui a motivé le sujet de cette thèse : c'est un honneur pour moi que vous soyez présente aujourd'hui. Jérôme, merci pour ta gentillesse et ta disponibilité, que ce soit pour discuter de fonctions-quantiles empiriques, pour apprendre à utiliser certains logiciels étranges, ou pour régler in extremis des problèmes administratifs. Je suis très admirative des nombreuses thématiques de recherche de Christophe Giraud : merci d'avoir accepté ce cadeau (empoisonné !) que constitue mon manuscrit.

Merci aux membres permanents du MAP5, de l'UFR Math-Info, et du département STID de l'IUT pour leur accueil. Merci tout d'abord à l'équipe de Statistique : en particulier à Valentine Genon-Catalot (pour les cours accélérés de rédaction !), à Adeline Samson (pour les nombreuses fois où je t'ai dérangée avec des problèmes aussi graves (!) que la relecture de flyers, la notation des TP pour lesquels tu m'as fait confiance, la compilation d'un fichier TeX récalcitrant, et bien d'autres choses encore...), à Olivier Bouaziz (pour les discussions en passant chercher un thé, et ton point de vue sur la densité relative : peut-être qu'on pourra travailler ensemble une autre fois ?), à Antoine Chambaz (pour tes conseils toujours précieux sur des slides ou un CV), à Jean-Claude Fort (pour ton avis tranché sur les probas de petites boules), à Flora Alarcon, Servane Gey et Avner Bar-Hen (pour votre gentillesse).

Merci également à Annie Raoult pour prendre soin des doctorants de son labo, et à Christine Graffigne, pour m'avoir permis de faire des enseignements variés. J'en profite pour exprimer ma

gratitude envers Mireille Chaleyat-Maurel, qui m'a confié trois de ses TD, à Nicole Rigal, Georges Koepfler et Elisabeth Ottenwaelter et Florence Muri, pour avoir soutenu mes premières expériences en tant qu'enseignante. Merci à Charlotte Laclau pour les bons moments passés à l'IUT et les rires de préparation et corrections de TP. Merci bien sûr aussi aux étudiants, qui, bien qu'ayant sans doute "essuyé les plâtres" de mes débuts, m'ont si souvent apporté le café et m'ont confirmé dans mon désir d'enseigner.

Merci à Marie-Hélène Gbaguidi, Isabelle Valéro, Christophe Castellani et Michel Guillemot : votre aide est toujours efficace pour les démarches administratives complexes. Merci à Vincent Delos d'une part, et Thierry Raedersdorff et son équipe d'autre part : même si comme tu le dis, je fais "partie des meubles" à P5, j'ai toujours du mal à décoder votre langue d'informaticiens !

Merci à tous ceux qui m'ont fait profiter de leurs connaissances en statistique depuis le master. Je dois à Pascal Massart mon intérêt pour la sélection de modèles : merci pour ce cours si vivant ! Je vous suis également très reconnaissante pour les nombreux conseils, que vous et Elisabeth Gassiat m'avez donnés en M2. Un grand merci aussi à Gilles Stoltz (pour les max de gaussiennes), à Erwan Le Pennec (ne m'en veux pas de ne pas avoir assez su tirer profit des TP d'ondelettes), à Stéphane Boucheron, à Michael Chichignoud, à Élodie Brunel. Merci à Emeline Schmisser, Sandra Plancade, et Claire Lacour : vos parcours à toutes les trois sont des modèles pour moi. Claire, merci d'avoir affronté le RER plusieurs fois pour venir discuter de borne inf ou d'autres choses !

Merci à Angelina Roche pour le travail commun qui éclaire cette fin de thèse. Angelina, ta persévérance sur les calculs et tes larges connaissances sur les données fonctionnelles m'impressionnent. Merci d'avoir eu la patience de laisser un peu de côté notre collaboration pendant la préparation de la soutenance. Merci également à Thomas Laloë et Rémi Servien de s'être intéressés à mes histoires de déformation.

Merci aussi à tous les enseignants qui m'ont donné envie d'enseigner à mon tour et que j'ai fatigués avec mes milliers de questions (et c'est un euphémisme) tout au long de ma scolarité et de mes études, de la psycho à P7, au maths. Parmi ceux que j'ai le plus embêté en maths à P5, merci à Nicole Rigal (pour m'avoir toujours poussée à tout comprendre), Bernard Locker, Thierry Cabanal, Hermine Biermé (je t'ai dérangée avec le concours 3A en plein congé maternité). A Cachan et Orsay respectivement, merci à Vincent Beck et Jean-François Le Gall, avec qui j'ai sans doute battu le record de questions. Merci également à ceux qui m'ont encadrée pour de petits projets de recherche : je pense à Jean-Stéphane Dhersin, Claude Zuily et Alano Ancona. Enfin, un immense merci à Pierre Calka. Pierre, merci pour ton soutien constant depuis des années maintenant, même quand tu es victime de mon "indécision chronique". Tu resteras aussi le prof qui effaçait le tableau plus vite que son ombre !

Merci à tous les "TJD" du MAP5, avec qui j'ai partagé un bureau et/ou des pauses au LoungEX ou sur la terrasse : par ordre d'"ancienneté" , Benjamin, Mahendra, Jean-Pascal, Pierre, Robin et Wilson (le trio des ATER de choc !), Baptiste (tu ne me fais plus sursauter en ouvrant la porte !), Anne-Cécile, Laureen, Mariella, Julien, Mélina, Samuel, Nicolas C. et M., Imen, Djénéba, The-Minh, Ardo, Fanny, Kévin (merci pour les chaussures de mariage !), Léon, Georges, Rebecca, Charlotte D., Christelle, Claire, Gwennaelle, Nina et Jean. Merci spécialement à Diarra Fall et Christophe Denis, pour votre aide pré-soutenance. Christophe, je suis admirative et envieuse de ton calme en toute circonstance ! Je garde aussi d'excellents souvenirs des séjours en conférences que nous avons partagés (j'espère que le récent voyage en avion ne t'a pas découragé de repartir avec moi !). Un merci très chaleureux à Laurent Navoret (pour le café de 9h), à Rémy Abergel et Anne-Sophie Macé (pour votre gentillesse, et votre disponibilité pour répondre à des problèmes informatiques), à Arthur Leclaire (pour ton amitié, ton rôle de relecteur d'e-mails, et de spécialiste des questions d'analyse, de probas

et d'abeilles, et pour tout le chocolat qu'on a mangé ensemble : la version 19 du Guide d'accueil intègrera-t-elle un plan pour accéder aux réserves de chocolat ?).

Merci également à tous les (jeunes) statisticiens croisés au détour d'une conférence, de balades (avec plus ou moins d'escalade...) au CIRM, de parties de ping-pong et des footing à Fréjus ou dans un séminaire parisien : dans le désordre merci à Aurore, Caroline B., Alice, Florian, Sylvain(s), Maud, Céline, Saskia, Thierry, Marc-Antoine, Thomas, Nelo, Benjamin, Cécile, Sarah, Sébastien, Cyrielle, Vincent, et Célia. Un merci tout particulier à Laure : tu es la voisine de chambre parfaite pour tous les colloques ! Merci aussi pour tes relectures et tes réponses rapides à toutes les questions (débiles).

Merci maintenant aux amis pour tous les indispensables moments et week-ends hors-thèse ! Merci à Aurore, Flo et Virginie : vous restez les amies de lycée (de maternelles !) sur qui je sais que je peux compter. Merci à tous ceux qui ont pris sur leur temps (et leurs congés !) pour venir assister à ma soutenance : je pense en particulier à Romain et Yves (bonjour à Edwige et au petit Théo !). Merci à la bande des joyeux cachanais, Irène, Jimmy et Sandra, Popoff, Kron, Romain et Agathe, Antoine, Cécile et Arthur pour les soirées et les vacances entre les Alpes, la Corse et l'Ile de Ré. Je vous dois des connaissances bien plus élargies et poussées que ce qu'il y a dans cette thèse, en pâtisserie (si si), en escalade (bientôt le 6c), en dinosaures, en jeux de société (coinche ou dominion ? !), en $\pi^2/12$, et Vice Et Versa. Romain, Antoine et Cécile, nos pique-niques entre le CDF-P5-P7 et l'Agro auront rendu les journées de travail plus courtes et agréables. Antoine, même si nous ne partageons pas la même définition d'un estimateur à noyau, merci pour ton humour. Et puis c'est toujours rassurant de passer un moment avec toi, depuis que je sais que tu es hyper-réactif pour me récupérer par terre en plein cours de distribution et me porter sur 300m ! Cécile, merci pour ton aide en toute circonstance (mariage, pot de thèse,...), pour tous les moments sympas avec vous et votre craquante petite Jeanne !

Avant de conclure par des remerciements plus personnels, quelques "merci" matériels. Mes plus vifs remerciements à une certaine compagnie de transports, qui, bien qu'ayant tenté de faire échouer de nombreuses fois cette thèse, m'a finalement toujours amené à bon port (ou plutôt en bonne gare) à Vernon ou Paris. Merci au PSG, à la PS3 et à Pedro Miguel pour avoir si souvent occupé mon mari pendant les soirées et week-ends consacrés aux maths. Merci à mon levain (mes amis auraient été déçus si je ne le citais pas !), qui dort paisiblement au réfrigérateur en attendant le prochain pain. Et merci à mes chaussures de randonnée qui ont bénéficié d'un congé spécial inespéré en ce printemps.

Merci à ma famille ! Marion, ma grande petite sœur, c'est toujours rassurant de savoir que je peux compter sur toi en cas de défaillance SNCF, ou plus simplement pour rire et se raconter nos petites vies. Merci également de m'avoir rappelé que tout de même, 300 pages en 3 ans, c'est une bien faible moyenne (c'est sûr que comparer à 60 patients à la demi-journée, je ne pouvais pas rivaliser...) ! Papa, maman, merci pour tout ce que je vous dois. Merci d'avoir toujours respecté mes choix (même les plus farfelus) et mon entêtement, et d'avoir toujours tout fait pour me faciliter la vie : merci Maman d'avoir taillé mes crayons de couleurs si longtemps, merci pour les pains d'épices du dimanche ! Je te dois également d'être ma plus fidèle correctrice d'orthographe, même dans les documents mathématiques incompréhensibles selon toi (dont l'intro de cette thèse). Papa, même si tu ne veux plus entendre parler de maths aujourd'hui, tu restes pour moi le meilleur prof de maths, avant même d'être celui qui sait le mieux s'orienter hors chemins en forêt de Fontainebleau ! Merci à tous les 3 d'être toujours là, et de m'aimer. Et promis, un jour j'apprendrai la signification du mot "repos" !

Enfin, le "merci" le plus tendre et amoureux à Baptiste. Seul toi me sup-porte au quotidien, dans les (nombreux) moments d'hyper-activité de défis mathématiques, culinaires ou sportifs et dans les (au moins aussi nombreux) moments de doute et de tristesse. Merci de m'apporter l'optimisme et ton infaillible confiance en la vie qui me font parfois défaut. Les mots ne suffisent pas à exprimer à quel point je te suis reconnaissante, et redevable de ta patience. Mieux vaut donc faire court, sans doute : merci pour tout, pour tout l'amour, que je ne saurais jamais si bien te rendre.

## Organisation de la thèse

La présente thèse rassemble quelques travaux sur des problèmes d'estimation non-paramétrique, étudiés à l'aide de méthodes adaptatives. Tous sont motivés par la construction d'estimateurs ayant de "bonnes" propriétés théoriques (au sens de l'adaptation), tout en ayant une expression simple et en étant numériquement stables. Ceci conduit à l'étude de méthodes dites de déformation, introduites dans chacun des modèles étudiés, et pour les différentes classes d'estimateurs considérés (estimateurs par projection ou à noyaux). Dans tous ces cas, la théorie sous-jacente aux résultats prouvés est inspirée de la sélection de modèles, et à ce titre, plusieurs preuves suivent le même schéma et font appel aux mêmes outils théoriques, ce qui peut engendrer quelques répétitions. Précisément, l'organisation de la thèse est la suivante.

- Les **Chapitres 1** et **2** sont introductifs.
  - Le **Chap. 1** est une introduction générale : le contexte et les méthodes d'estimation utilisées sont rappelés ; chacune des parties de la thèse est présentée. En particulier, la méthode de déformation est motivée sur l'exemple de la régression additive.
  - Le **Chap. 2** présente les principaux outils requis pour les preuves.
- La **Partie A** de cette thèse est consacrée à la construction d'estimateurs bâtis à l'aide de données déformées :
  - la méthode est détaillée au **Chap. 3** pour la régression additive en bases déformées,
  - elle est étendue au **Chap. 4** à d'autres cadres d'estimation d'une fonction d'une variable réelle, et à la construction d'estimateurs à noyaux,
  - le **Chap. 5** tente d'illustrer la pertinence de la méthode pour estimer la densité conditionnelle.
- La **Partie B** traite un problème à deux échantillons, l'estimation de la densité relative :
  - le cas de données complètes est l'objet du **Chap. 6**,
  - le **Chap. 7** envisage la possibilité d'échantillons censurés.

Chaque chapitre peut globalement être lu indépendamment des autres. Cependant, nous avons tenté de mettre en évidence les liens existants entre eux : tout d'abord, le lien entre les deux parties est expliqué dans l'**Introduction** (**Chap. 1**). Ensuite, des notations consistantes sont utilisées. Chaque chapitre est également précédé d'un court résumé permettant de le situer dans son contexte. Enfin, les cadres étudiés nous ont conduits à mettre en œuvre différentes procédures inspirées de la sélection de modèles (méthode de Birgé-Massart), ou de la sélection de fenêtres (méthode de Goldenshluger-Lepski). Nous avons également profité de ce travail pour souligner différents liens entre ces méthodes, que ce soit dans l'**Introduction**, dans le **Chap. 3**, ou en appendice des **Chap. 4** et **5**.

**Un résumé figure à la toute fin du manuscrit.**

## Organization of the thesis

The present thesis collects works on nonparametric estimation problems, studied with adaptive methods. Their common motivation is to build estimators which enjoy both interesting theoretical properties and a simple practical computation and numerical stability. This leads to the study of what we call a "warping device", which is introduced in each of the studied frameworks, and for any type of the considered estimators (projection or kernel). Whatever the example, the theory we use is inspired by model selection. Thus, several proofs are similar, and require identical tools. This may induce some redundancy. The organization is the following.

- **Chap. 1** and **2** are introductory chapters.
  - **Chap. 1** is a general presentation: the statistical framework and the estimation methods are introduced. Each part of the dissertation is presented. The warping device is motivated in the additive regression setting.
  - **Chap. 2** presents the main tools of the proofs.
- **Part A** deals with the building of warped-estimators:
  - the method is detailed in **Chap. 3** for warped-bases regression,
  - it is extended in **Chap. 4** to other estimation problems of one-variable functions, and to kernel estimates,
  - **Chap. 5** illustrates the relevance of the warping device to recover a conditional density.
- **Part B** aims at solving a two-sample problem, the estimation of the relative density:
  - the estimation from complete data-samples is the subject of **Chap. 6**,
  - the problem of censored-data is tackled in **Chap. 7**.

Each chapter can mainly be read separately from the others. Nevertheless, we have tried to underline the links between them. First, the link between the two parts is explained in the **Introduction** (**Chap. 1**). Next, consistent notations are used. Last, each chapter begin with a short abstract, to put it in context. Moreover, the settings we study lead us to apply different strategies inspired by model selection (Birgé-Massart method) and by bandwidth selection (Goldenshluger-Lepski method). We take advantage of this work to highlight the links between the methods, in the **Introduction**, in **Chap. 3** or in the appendix of **Chap. 4** and **Chap. 5**.

**A short abstract can be found at the end of the thesis.**

# Table des matières

# Chapitre 1

# Introduction.

## Sommaire

Dans cette thèse, nous nous intéressons à quelques problèmes d'estimation fonctionnelle pouvant être étudiés à partir de données transformées ou déformées (en un sens que nous définirons). Ce travail s'inscrit plus généralement dans le cadre de la construction d'estimateurs non-paramétriques, possédant des propriétés d'adaptation quant à la régularité de la fonction à retrouver. Ces notions sont présentées en début d'introduction, avec les principales méthodes d'estimation que nous utilisons. Pour cela, nous nous plaçons dans le modèle élémentaire de l'estimation d'une densité (Section 1.1), modèle qui ne sera pas repris dans la suite. La problématique qui motive l'introduction de données transformées est ensuite proposée à partir du modèle de régression (Section 1.2.1). Les résultats déjà établis sur ce sujet sont également présentés. Les principaux modèles statistiques auxquels la méthode est appliquée font l'objet de la Section 1.2.2. Nous présentons les résultats obtenus. La seconde partie de la thèse, consacrée à l'estimation de la densité relative, une fonction récemment définie dans les problèmes à deux échantillons, est introduite à la Section 1.3 : nous présentons la problématique, ainsi que les résultats obtenus au regard des procédures existantes. Nous décrivons le lien entre ce nouveau problème et la méthode de déformation (Section 1.3.3). Enfin, nous concluons avec quelques perspectives de recherche dans le prolongement des travaux de thèse.

## 1.1 Contexte : estimation non-paramétrique adaptative

### 1.1.1 Cadre statistique

Les problèmes d'inférence statistique auxquels nous nous intéressons relèvent d'une approche qui présente deux caractéristiques principales :

- elle est *non-paramétrique* : les quantités que nous cherchons à retrouver à partir d'observations sont des fonctions, sur lesquelles on ne fait pas d'hypothèses de forme a priori, et qui ne sont pas de prime abord paramétrées (les données ne sont ainsi pas excessivement modélisées). Ceci n'exclut toutefois pas certaines hypothèses plus générales, comme l'appartenance de la fonction cible à un espace fonctionnel (de dimension infinie) par exemple.
- elle est *non-asymptotique* : notre objectif est principalement d'obtenir des résultats théoriques avec un nombre d'observations fixe et ne tendant pas vers l'infini (des vitesses de convergence pourront aussi être déduites dans un deuxième temps). Cependant, si la taille des échantillons n'est pas nécessairement supposée très grande, faire de l'estimation non-paramétrique avec un nombre très faible de données relève le plus souvent du défi.

Nous considérons donc dans la suite l'estimation d'une fonction $s$ à valeurs réelles, sur un borélien $A$ de $\mathbb{R}$ ou $\mathbb{R}^2$, à partir d'un échantillon $(X_1, \dots, X_n)_{n\in\mathbb{N}\setminus\{0\}}$ d'une certaine variable (ou d'un certain vecteur) aléatoire $X$, sur un espace probabilisable $(\Omega, \mathcal{F})$. La loi $\mathbb{P}_s$ de la variable $X$ dépend de la fonction inconnue $s$.

Toute fonction mesurable des observations $(X_i)_{i\in\{1,\dots,n\}}$ étant considérée comme un estimateur, un critère de qualité est requis pour évaluer ses performances. Nous ferons toujours l'hypothèse que la fonction $s$ appartient à $L^2(A)$, l'ensemble des fonctions de carré intégrable

par rapport à la mesure de Lebesgue sur $A$ et nous définissons la fonction de *perte* $L^2$ par

$$\ell(s,\hat{s}) = \|s-\hat{s}\|^2 = \int_A (s(x)-\hat{s}(x))^2 dx.$$

Le *risque quadratique intégré* (en anglais, M.I.S.E. : *Mean Integrated Squared Error*) est l'espérance de cette dernière quantité, sous la loi $\mathbb{P}_s$ : $\mathbb{E}_s[\ell(s,\hat{s})]$, ce que nous noterons abusivement $\mathbb{E}[\ell(s,\hat{s})]$. Le choix du risque $L^2$ est essentiellement motivé par la stratégie par projection que nous allons mettre en œuvre [(1)]. Nous serons également amenés à considérer des normes $L^2$ pondérées de type $\ell_w(s,\hat{s}) = \int_A (s(x)-\hat{s}(x))^2 w(x)dx$ pour une certaine fonction $w$.

Une exigence naturelle lorsque l'on bâtit un estimateur est que son risque soit le plus "petit possible". On peut bien sûr espérer qu'il tende vers 0 quand le nombre d'observations tend vers l'infini, et plus précisément quantifier la décroissance en fonction de l'indice de régularité de la fonction estimée. On dira qu'un estimateur $\hat{s}$ *atteint la vitesse* $\psi_n$ sur une classe de fonctions $\mathcal{S}$, si, pour $s \in \mathcal{S}$,

$$\mathbb{E}\left[\|s-\hat{s}\|^2\right] \leq C\psi_n,$$

pour une constante strictement positive $C$, et pour une suite $(\psi_n)_{n\in\mathbb{N}\setminus\{0\}}$ qui décroît vers 0 quand $n$ tend vers l'infini. Dans les problèmes à deux échantillons qui font l'objet de la Partie B de la thèse, $(\psi_n)_n$ sera remplacée par une suite $(\psi_{n,n_0})_{n,n_0}$ dépendant des tailles $n$, et $n_0$ des deux échantillons. Rappelons également que la suite $(\psi_n)_n$ est appelée *vitesse optimale au sens du risque minimax* pour l'estimation de la fonction $s$ sur la classe $\mathcal{S}$, si aucun estimateur de $s$ n'atteint une vitesse "plus rapide" pour estimer les fonctions de $\mathcal{S}$, c'est-à-dire, si

$$\inf_{\substack{\hat{s}\\ \text{estimateur}}} \sup_{s\in\mathcal{S}} \mathbb{E}\left[\|s-\hat{s}\|^2\right] \geq C'\psi_n,$$

pour une constante $C' > 0$. Le membre de gauche de cette inégalité est appelé *risque minimax*. Une autre approche consiste à exiger que l'estimateur ait un risque aussi petit, à constante près, et à terme de reste près, que le meilleur des risques possible dans une collection d'estimateurs. Nous nous plaçons plutôt du côté de cette seconde possibilité, et cherchons dans cette thèse des estimateurs présentant des propriétés d'*optimalité au sens de l'oracle*, ce que nous définirons précisément plus loin (voir Section 1.1.3).

### 1.1.2 Estimation par projection et estimateurs à noyaux

Deux grandes familles d'estimateurs non-paramétriques seront considérées dans la suite de cette thèse : les estimateurs obtenus par minimisation de contraste et les estimateurs fondés sur des noyaux. Nous les décrivons donc dans la section qui suit, pour le modèle simple d'estimation d'une densité. Nous tenterons d'établir des parallèles entre les deux approches. Cette section s'inspire fortement de Massart (2007), Tsybakov (2009) et Comte (2012). Soit donc $(X_i)_{i\in\{1,\dots,n\}}$ un $n-$échantillon de densité $s \in L^2(A)$, inconnue (avec $A$ un intervalle de $\mathbb{R}$).

(1). Notons que d'autres fonctions de perte sont souvent considérées en statistique, comme la perte ponctuelle $\ell(s,\hat{s}) = |\hat{s}(x_0) - s(x_0)|$, la perte associée à la norme $L^p$, pour un indice $p \geq 1$ quelconque, la perte définie par la distance de Kullback ou celle de Hellinger. Nous ne les utiliserons pas dans cette thèse.

### Estimation par projection

**Principe.** La méthode d'estimation par projection repose fortement sur l'hypothèse selon laquelle la fonction $s$ appartient à $L^2(A)$, muni du produit scalaire usuel $\langle .,. \rangle$ et de la norme qui en découle $\|.\|$. Considérant une base hilbertienne $(\varphi_j)_{j\in J}$ (famille orthonormée et totale) de cet espace, on peut alors développer la fonction cible

$$s = \sum_{j\in J} a_j \varphi_j, \quad a_j = \langle s, \varphi_j \rangle,$$

l'égalité ayant lieu dans $L^2$ (convergence de la série en ce sens). Une stratégie raisonnable d'estimation consiste alors à estimer certains coefficients du développement précédent. Ceci revient dans un premier temps à estimer la projection de la fonction $s$ sur certains sous-espaces vectoriels de $L^2(A)$, que l'on choisira de dimensions finies (ce qui revient à se ramener d'abord à un problème paramétrique), puis à choisir dans un second temps ce sous-espace. Précisément, on se donne une collection finie $\mathcal{M}_n$ d'indices (dont le cardinal dépend de la taille $n$ de l'échantillon d'observations), et pour chaque $m \in \mathcal{M}_n$ ($\mathcal{M}_n \subset \mathbb{N}$ ou $\mathbb{N}^2$), on considère $S_m$ un sous-espace de $L^2(A)$, appelé *sous-espace d'approximation* ou *modèle* (*model* ou *sieve* en anglais, voir Birgé & Massart 1998). Dans notre travail, $S_m$ sera toujours un sous-espace vectoriel de dimension $D_m > 0$ finie. Notons $(\varphi_j)_{j\in J_m}$ une base qui l'engendre (avec $J_m \subset J$). On bâtit pour chaque indice $m$ un estimateur de $\Pi_{S_m}(s) = \sum_{j\in J_m} a_j \varphi_j$ la projection de $s$ sur $S_m$. Sachant que celle-ci est définie de telle sorte que $\Pi_{S_m}(s) = \arg\min_{t\in S_m} \|s-t\|^2$, l'idée est de remplacer la norme inconnue par un équivalent empirique appelé *contraste* (voir Birgé & Massart 1993, p.117 pour une définition précise ou encore Birgé & Massart 1998, p.318). Dans le cas de l'estimation de densité, on peut choisir [(2)]

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t(X_i), \quad t \in L^2(A), \tag{1.1}$$

puisque l'on a alors $\mathbb{E}[\gamma_n(t)] = \|t\|^2 - 2\langle t, s\rangle = \|t-s\|^2 - \|s\|^2$. Minimiser $\gamma_n$ sur $S_m$ est donc cohérent pour proposer un M-estimateur de $s$ :

$$\hat{s}_m = \arg\min_{t\in S_m} \gamma_n(t). \tag{1.2}$$

Ceci définit $\hat{s}_m$ de manière unique : $\hat{s}_m = \sum_{j\in J_m} \hat{a}_j \varphi_j$, avec $\hat{a}_j = n^{-1}\sum_{i=1}^n \varphi_j(X_i)$. Remarquons que $\hat{s}_m$ estime sans biais la projection $\Pi_{S_m}(s)$ de $s$ sur le modèle $S_m$ : $\mathbb{E}[\hat{s}_m] = \Pi_{S_m}(s)$.

**Modèles.** La question qui se pose est celle du choix de la base hilbertienne $(\varphi_j)_{j\in J}$ et des modèles $S_m$, $m \in \mathcal{M}_n$. Ceux que nous utiliserons dans cette thèse sont des sous-espaces vectoriels de dimension finie de $L^2(A)\cap L^\infty(A)$ ($L^\infty(A)$ étant l'ensemble des fonctions bornées presque partout sur $A$, muni de la norme sup $\|.\|_\infty$). Nous les choisissons de la forme $S_m = \text{Vect}\{\varphi_j\ ,\ j \in J_m\}$ comme indiqué ci-dessus. Ces modèles seront supposés vérifier les propriétés suivantes :

(2). Il existe également des contrastes fondés sur la vraisemblance des observations, mais nous ne les utiliserons pas ici. On pourra par exemple se référer à Cohen & Le Pennec (2012), pour le cadre d'estimation d'une densité conditionnelle, cadre que nous abordons également au Chapitre 5.

($\mathcal{M}_1$) La dimension $D_m$ de $S_m$ est majorée par le nombre d'observations $n$.

($\mathcal{M}_2$) Les modèles sont emboîtés : si $m \leq m'$ alors $S_m \subset S_{m'}$ (si $\mathcal{M}_n \subset \mathbb{N}^2$, la relation "$\leq$" est prise terme à terme).

($\mathcal{M}_3$) Il existe une constante $\phi_0 > 0$, telle que, pour $t \in S_m$, $\|t\|_\infty^2 \leq \phi_0^2 D_m \|t\|^2$.

Ces trois hypothèses sont classiques : on pourra se référer par exemple à Birgé & Massart (1998), Baraud (2002) ou Brunel & Comte (2005). La Propriété ($\mathcal{M}_1$) est raisonnable, si l'on garde en mémoire que $D_m$ est le nombre de coefficients à estimer : s'il est supérieur au nombre d'observations, on ne peut espérer obtenir un estimateur de qualité. La Propriété ($\mathcal{M}_2$) est relativement forte. Elle peut dans certains cas être relâchée et remplacée par l'hypothèse de l'existence d'un espace englobant tous les modèles de la collection (voir par exemple l'hypothèse $\mathcal{H}_2$ de Brunel & Comte 2005). La Propriété ($\mathcal{M}_3$) est équivalente à l'inégalité suivante (voir le Lemme 1 de Birgé & Massart 1998) :

$$\left\| \sum_{j \in J_m} \varphi_j \right\|_\infty^2 \leq \phi_0^2 D_m.$$

Notons qu'une propriété plus forte de localisation (comme la propriété ($H_{Loc}$) de Baraud 2002) est parfois ajoutée. Il existe également des hypothèses permettant de limiter de manière plus ou moins importante la complexité de la collection de modèles, au sens du nombre de modèles ayant une même dimension $D$ fixée : on se réfèrera par exemple au travail de Akakpo (2009), qui obtient des résultats pour des collections de modèles de grande complexité. Dans ce travail, les collections considérées comporteront toujours au plus un modèle par dimension.

Les modèles étant emboîtés, nous pouvons considérer le réagencement suivant : $J = \mathbb{N}\backslash\{0\}$, $J_m = \{1, \ldots, D_m\}$, et donc $S_m = \text{Vect}\{\varphi_1, \ldots, \varphi_{D_m}\}$. Notons également que ces propriétés et notations s'étendent au cas de sous-espaces vectoriels de $L^2(\mathbb{R}^2)$ (voir le Chapitre 5, ou encore Lacour 2007 ou Plancade 2013 parmi d'autres).

Il existe dans la littérature des espaces "de référence" classiquement utilisés, vérifiant les trois propriétés ci-dessus : modèles fondés sur la base de Fourier, sur les bases de polynômes par morceaux avec partition régulière dyadique, ou les modèles fondés sur des ondelettes à support compact. Le choix est principalement guidé par les propriétés d'approximation que peuvent avoir les sous-espaces, c'est-à-dire sur la possibilité de majorer convenablement $\|s - \Pi_{S_m}(s)\|^2$ (le lecteur peut se référer à la Section 2.2.2 pour des précisions). Nous renvoyons aussi à Birgé & Massart (1998), Baraud (2002) ou Brunel & Comte (2005) pour des descriptions générales de ces trois types de modèles, à Efromovich (1999) ou Fourdrinier & Pergamenshchikov (2007) pour des détails sur l'utilisation en statistique de la base de Fourier et à Härdle *et al.* (1998) pour celle des ondelettes (entre autres). La base trigonométrique sera la plus fréquemment utilisée dans cette thèse. Les espaces correspondants sont précisément décrits à la Section 2.2.1, avec les propriétés qui motivent leur choix. Notons que d'autres espaces sont également utilisés dans la littérature (nous donnons à titre d'exemple une ou deux références par modèle, sans bien sûr prétendre être exhaustifs) : histogrammes fondés

sur des partitions dyadiques (Akakpo & Durot, 2010), bases fondées sur la fonction sinus cardinal (Lacour, 2008), sur des splines (Koo & Lee, 1998; Schmisser, 2012). Des travaux récents autorisent également des modèles composés de fonctions de nature différente (ce que l'on appelle un *dictionnaire*) ce qui conduit à des problèmes théoriques de nature différente : citons par exemple les travaux de Massart & Meynet (2011) dans le modèle de bruit blanc gaussien, ou ceux de Pham Ngoc & Rivoirard (2013) en déconvolution sur la sphère.

**Risque d'un estimateur par projection, et choix non adaptatif de modèles.** Le choix de la base étant fixé, nous disposons de la collection $(\hat{s}_m)_{m\in\mathcal{M}_n}$ d'estimateurs de la densité $s$ (définis par (1.2)). On cherche à retrouver celui ayant le risque le plus petit. La fonction réalisant la meilleure performance est appelée *oracle* et définie comme $\hat{s}_{m^*}$, avec $m^* = \arg\min_{m\in\mathcal{M}_n} \mathbb{E}[\|\hat{s}_m - s\|^2]$ (cette notion est introduite par Donoho & Johnstone 1994). Elle est cependant inaccessible puisque $s$ est précisément inconnue.

On considère donc le risque de chacun des estimateurs $\hat{s}_m$, pour envisager ensuite quel choix d'indice $m$ peut ensuite être fait. L'erreur quadratique se décompose en deux termes, par le Théorème de Pythagore

$$\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] = \|s - \Pi_{S_m}(s)\|^2 + \mathbb{E}\left[\|\hat{s}_m - \Pi_{S_m}(s)\|^2\right]. \tag{1.3}$$

Le premier terme est le terme de biais, ou erreur d'approximation, que l'on commet en estimant $\Pi_{S_m}(s)$ plutôt que $s$. Le second est le terme de variance ou erreur d'estimation, provenant de l'erreur commise en remplaçant $a_j$ par une version empirique $\hat{a}_j$. Ces deux termes évoluent de façons opposées en fonction de la dimension : le biais décroît quand la dimension $D_m$ augmente (plus le modèle est gros, mieux $s$ est approchée par sa projection $\Pi_{S_m}(s)$), alors que le terme de variance augmente avec $D_m$ (le nombre de coefficients estimés augmente avec la taille du modèle). Il convient donc, pour minimiser l'erreur quadratique, d'équilibrer ces deux termes, ce que l'on appelle le *compromis biais-variance*. Précisément, avec la propriété $(\mathcal{M}_3)$,

$$\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] = \sum_{j=1}^{D_m} \mathrm{Var}(\hat{a}_j) = \frac{1}{n}\sum_{j=1}^{D_m} \mathrm{Var}\left(\varphi_j(X_1)\right) \leq \frac{1}{n}\sum_{j=1}^{D_m} \mathbb{E}\left[\varphi_j^2(X_1)\right] \leq \phi_0^2 \frac{D_m}{n}. \tag{1.4}$$

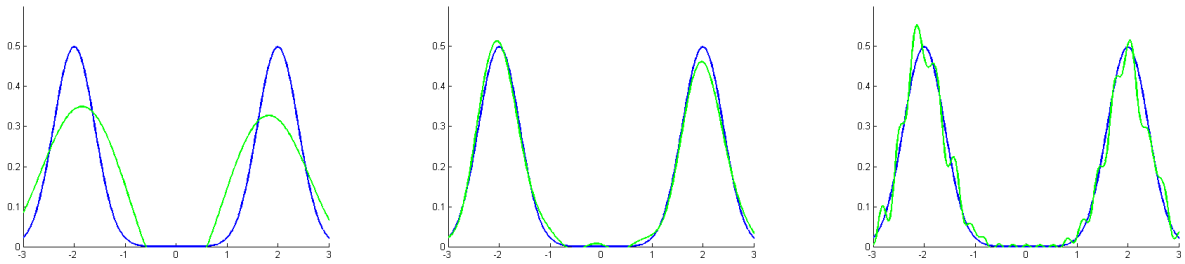
On obtient donc

$$\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] \leq \|s - \Pi_{S_m}(s)\|^2 + \phi_0^2 \frac{D_m}{n}. \tag{1.5}$$

**Remarque 1.1.** Si l'on souhaite s'affranchir de la propriété $(\mathcal{M}_3)$, on peut obtenir une majoration du même ordre de grandeur en supposant en contrepartie $s$ bornée ($\|s\|_\infty$ désignant son supremum essentiel) : en effet, en reprenant le calcul (1.4),

$$\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] \leq \frac{1}{n}\sum_{j=1}^{D_m} \mathbb{E}\left[\varphi_j^2(X_1)\right] \leq \frac{\|s\|_{L^\infty(A)}}{n} \int_A \sum_{j=1}^{D_m} \varphi_j^2(x)dx = \|s\|_\infty \frac{D_m}{n},$$

par orthonormalité de la base. On obtient donc à nouveau une majoration de type (1.5), où $\phi_0^2$ est remplacé par $\|s\|_\infty$. Dans l'objectif de construire une procédure adaptative, la majoration initiale (1.5) sera toutefois préférée, car elle ne dépend d'aucune quantité inconnue.



$D_m = 5$ $D_m = 14$ $D_m = 41$

FIGURE 1.1 – Exemple d'estimateurs par projection d'une densité définie comme un mélange de deux gaussiennes, avec $n = 500$ observations, pour trois choix différents de dimensions $D_m$ de modèles trigonométriques. En trait foncé : la densité estimée. En trait clair : l'estimateur.

La Figure Figure 1.1 illustre le compromis à effectuer : si le modèle choisi a une dimension trop faible, le biais est très important, mais au contraire, si le modèle a une dimension trop importante, sa complexité entraîne une variabilité élevée.

Si la fonction $s$ a pour indice de régularité $\alpha$ dans une certaine classe fonctionnelle, le terme de biais a typiquement pour ordre de grandeur $D_m^{-2\alpha}$ (des résultats d'approximations sont rappelés au Chapitre 2, Section 2.2.2). Si cet indice $\alpha$ est connu, on peut choisir le modèle $S_{\tilde{m}(\alpha)}$ tel que $\tilde{m}(\alpha) = \arg\min_{m \in \mathcal{M}_n}\{D_m^{-2\alpha} + D_m/n\}$ (en omettant les constantes). On obtient alors un estimateur $\hat{s}_{\tilde{m}(\alpha)}$ qui atteint la vitesse $n^{-2\alpha/(2\alpha+1)}$, au sens défini ci-dessus (Section 1.1.1), vitesse qu'on sait être optimale au sens minimax (c.f. par exemple Donoho *et al.* (1996)). La construction de $\hat{s}_{\tilde{m}}$ dépend cependant de la régularité $\alpha$ qui n'a aucune raison d'être connue sachant que $s$ elle-même ne l'est pas. L'enjeu de la sélection de modèles est donc de faire un choix d'estimateur uniquement fondé sur les données. Ce sera l'objet de la Section 1.1.3.

### Estimation à noyaux

**Principe.** Les méthodes à noyaux constituent l'autre grande famille de méthodes permettant d'estimer des fonctions, comme ici la densité $s$. On supposera pour simplifier la présentation que $A = \mathbb{R}$ (quitte à remplacer $s$ par $s\mathbf{1}_A$, où $\mathbf{1}_A$ est la fonction indicatrice de $A$). Un *noyau* est une fontion $K : \mathbb{R} \to \mathbb{R}$, intégrable sur $\mathbb{R}$, d'intégrale égale à 1. Pour tout réel $h > 0$, on note $K_h$ la fonction définie par $K_h : \ x \ \mapsto \ K(x/h)/h$. La propriété qui est à la base de l'intérêt des noyaux pour l'estimation est la suivante : la famille $(K_h)_{h>0}$ forme une *approximation de l'unité* pour le produit de convolution (voir Briane & Pagès 2006 par exemple). En particulier, la convolée $K_h \star s := \int_{\mathbb{R}} K_h(x - x')s(x')dx'$ converge vers $s$, quand $h$ tend vers 0 en norme $L^2$ et d'autant plus rapidement que $s$ est régulière (voir la Section 2.2.2 pour plus de détails). On peut donc approcher $s$ par $K_h \star s$. Ceci présente un intérêt du point de vue de l'estimation puisque l'on note que $K_h \star s(x) = \mathbb{E}[K_h(x - X_1)]$. L'estimateur

à noyau de la densité $s$ pour le paramètre $h > 0$ fixé est donc

$$\hat{s}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad x \in \mathbb{R}. \tag{1.6}$$

La notation est abusive, sachant que l'on a déjà noté $\hat{s}_m$ l'estimateur par projection sur un modèle $S_m$. Cependant, comme l'on note systématiquement $m$ (ou $m'$ plus bas) l'indice des modèles et $h$ (ou $h'$) le paramètre de l'estimateur à noyau, il n'y a pas de confusion possible. L'estimateur (1.6) a d'abord été introduit par Rosenblatt (1956), pour le noyau rectangulaire, défini par $K = \mathbf{1}_{]-1;1[}/2$ : sa construction est alors guidée par le fait que la densité $s$ est la dérivée de la fonction de répartition de $X$ (voir Tsybakov 2009 p.2 à ce sujet). Cette définition est généralisée par Parzen (1962) à un noyau quelconque.

Suivant van der Vaart (1998) (Chap. 24, p.342), on peut également voir la définition de l'estimateur (1.6) comme un moyen de répartir la masse 1 que doit avoir une densité au total entre toutes les observations : un poids $1/(nh)$ est d'abord attribué à chaque donnée $X_i$, puis le restant de la masse est attribué à un voisinage de chacun des $X_i$, dépendant de $h$ et de $K$. Le paramètre de lissage $h$ est la *fenêtre* ("*bandwidth*" en anglais) : son choix est crucial pour la qualité de la reconstruction. L'interprétation de van der Vaart (1998) de l'estimateur à noyau justifie déjà que l'on suppose $h > 1/n$ (comme on avait imposé $D_m \leq n$ au paragraphe précédent).

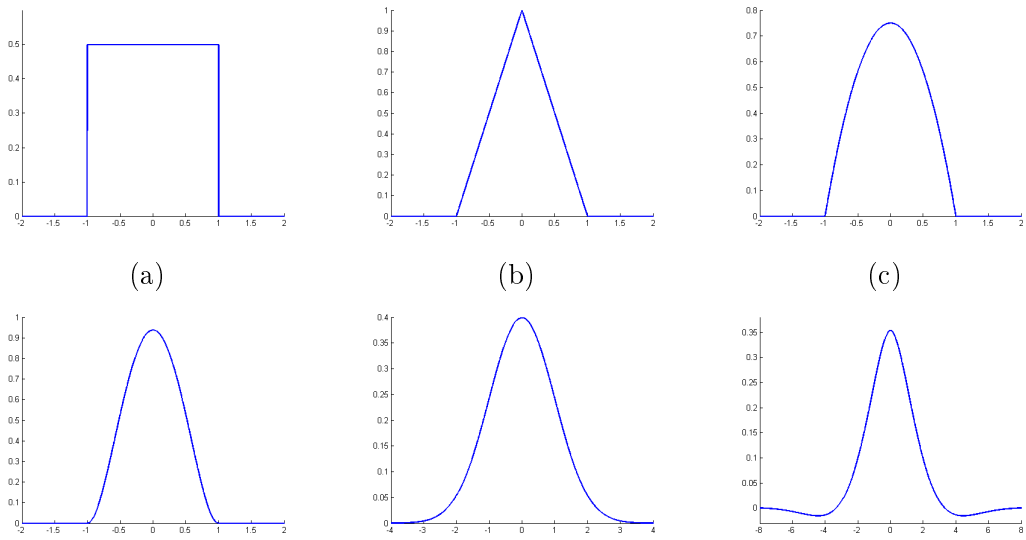
**Remarque 1.2.** Nous pouvons déjà établir un parallèle entre la construction d'un estimateur à noyau, et celle d'un estimateur par projection . Dans le cadre de l'estimation de la densité $s$, nous pouvons voir les estimateurs (1.2) et (1.6) comme provenant tous les deux du même constat : pour toute fonction $t$ de carré intégrable, $\mathbb{E}[t(X_i)] = \langle t, s\rangle$. En prenant $t = K_h(x - .)$, on obtient une justification de l'estimateur à noyau (1.6). En choisissant $t = \varphi_j$, on légitime le calcul des coefficients de l'estimateur par projection (1.2). Cette remarque, extrêmement simple dans le cas de l'estimation de densité, permettra également d'introduire de manière similaire nos estimateurs à la Section 1.2.1.

**Noyaux usuels.** De la même façon qu'il existe des familles de modèles classiques, il existe des noyaux fréquemment utilisés. Tsybakov (2009) (p.3) définit par exemple six noyaux classiques, représentés à la Figure Figure 1.2.

A titre d'exemple, la Figure Figure 1.3 représente les fonctions $K_h$ pour $h$ de plus en plus proche de 0, dans le cas où $K$ est le noyau rectangulaire ou le noyau gaussien.

Nous utiliserons pour implémenter les estimateurs bâtis dans cette thèse le noyau gaussien. Pour la partie théorique, nous ferons uniquement l'hypothèse que le noyau utilisé est dérivable, et nous supposerons également qu'il est d'*ordre* suffisant, notion introduite à la Section 2.2.2.

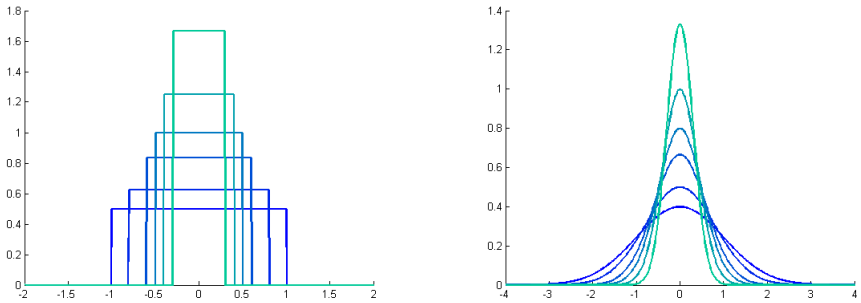
**Risque d'un estimateur à noyau.** En considérant une collection finie $\mathcal{H}_n \subset \mathbb{R}_+^*$ de fenêtres, on dispose, comme en projection, d'une collection d'estimateurs $(\hat{s}_h)_{h\in\mathcal{H}_n}$ pour la densité $s$. Parmi eux, existe également un oracle $\hat{s}_{h^*}$, inconnu, puisque $h^* = \arg\min_{h\in\mathcal{H}_n} \mathbb{E}[\|\hat{s}_h -$



(a) (b) (c)

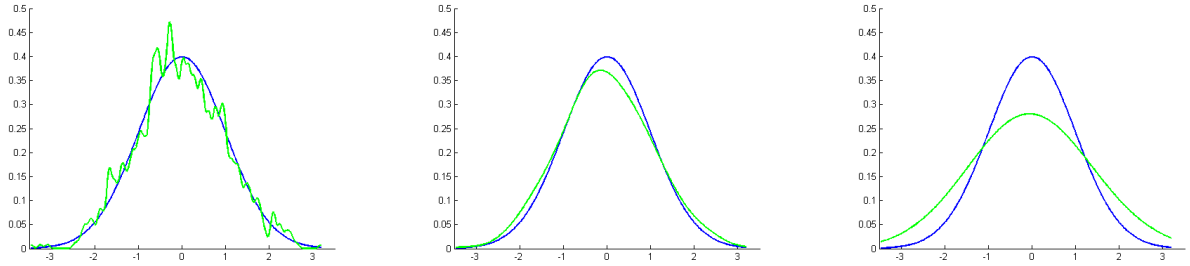
(d) (e) (f)

FIGURE 1.2 – Noyaux classiquement utilisés : (a) noyau rectangulaire, (b) noyau triangulaire, (c) noyau d'Epanechnikov, (d) noyau "biweight", (e) noyau gaussien, (f) noyau de Silverman.



(a) (e)

FIGURE 1.3 – Approximation de l'unité : fonctions $(K_h)_h$ où $K$ est (a) le noyau rectangulaire, (e) le noyau gaussien.



$h = 0.005$ $h = 0.04$ $h = 1$

FIGURE 1.4 – Exemple d'estimateurs à noyaux de la densité de la loi gaussienne centrée réduite, avec $n = 500$ observations, pour trois choix différents de fenêtres $h$. En trait foncé : la densité estimée. En trait clair : l'estimateur.

$s\|^2]$, et l'objectif est toujours d'étudier le risque d'un estimateur avec fenêtre fixée pour comprendre comment la sélectionner ensuite.

Sachant que $\mathbb{E}[\hat{s}_h] = K_h \star s$, on obtient également une décomposition biais-variance analogue, à (1.3) :

$$\mathbb{E}\left[\|\hat{s}_h - s\|^2\right] = \|s - K_h \star s\|^2 + \mathbb{E}\left[\|\hat{s}_h - K_h \star s\|^2\right].$$

Cette fois, le terme de biais est d'autant plus petit que la fenêtre $h$ est petite (puisque $K_h \star s$ tend vers $s$ en norme $L^2$ quand $h$ tend vers 0). Le terme de variance est à l'opposé croissant quand $h$ tend vers 0 (si $h$ est trop petit, l'estimateur "colle" trop aux observations) : avec les mêmes arguments qu'en projection, et en utilisant que $\|K_h\|^2 = \|K\|^2/h$,

$$\mathbb{E}\left[\|\hat{s}_h - K_h \star s\|^2\right] = \int_{\mathbb{R}} \mathrm{Var}(\hat{s}_h(x))dx \leq \frac{1}{n}\int_{\mathbb{R}} \mathbb{E}[K_h^2(x - X_1)] \leq \frac{\|K\|^2}{nh}.$$

On obtient cette fois

$$\mathbb{E}\left[\|\hat{s}_h - s\|^2\right] \leq \|s - K_h \star s\|^2 + \|K\|^2 \frac{1}{nh}. \tag{1.7}$$

Ici encore, un arbitrage s'impose pour choisir la fenêtre $h$ permettant d'obtenir un risque petit : une fenêtre trop proche de 0 fait exploser le terme de variance, une fenêtre trop grande fait exploser le biais : voir les Figures Figure 1.4 et Figure 1.5.

Mettant en parallèle la majoration (1.7) avec (1.5), on peut faire la comparaison suivante : *la fenêtre $h$ joue pour les estimateurs à noyaux le rôle de $D_m^{-1}$ pour les estimateurs par projection.* Cette analogie sera poursuivie tout au long de ce travail, en particulier à la Section 1.1.3 ci-dessous.

On peut poursuivre comme en projection : si la fonction $s$ a pour indice de régularité $\alpha$ dans une certaine classe fonctionnelle, le terme de biais a pour ordre de grandeur $h^{-2\alpha}$. Si cet indice $\alpha$ est connu, on peut choisir le modèle $S_{\tilde{h}(\alpha)}$ minimisant le membre de droite de

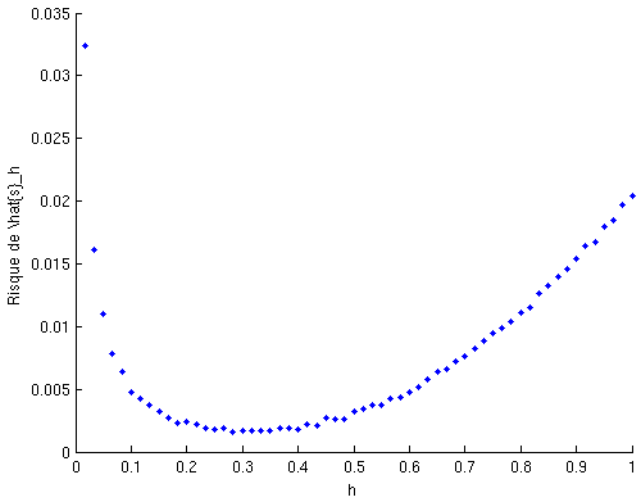



FIGURE 1.5 – Risque des estimateurs à noyaux de la densité de la loi gaussienne centrée réduite, avec $n = 500$ observations, en fonction de la fenêtre considérée.

l'Inégalité (1.7) pour obtenir à nouveau un estimateur $\hat{s}_{\tilde{h}(\alpha)}$ qui atteint la vitesse $n^{-2\alpha/(2\alpha+1)}$. L'enjeu de la sélection de fenêtres est donc analogue à celui de la sélection de modèles : faire un choix d'estimateur dans la collection uniquement fondé sur les données.

### 1.1.3 Problème d'adaptation

Pour simplifier la présentation, nous considérons toujours l'exemple de l'estimation de la densité $s$ à partir d'un $n-$échantillon $(X_i)_{i\in\{1,\dots,n\}}$. Nos travaux utilisent bien sûr les procédures présentées dans des cadres plus variés.

#### Problématique et principales méthodes d'estimation adaptative

L'estimation non-paramétrique *adaptative*, qui est l'objectif principal de cette thèse, vise à estimer une fonction inconnue $s$ en faisant le moins d'hypothèses possibles sur ses caractéristiques. En particulier, on peut souhaiter définir des estimateurs, qui, bien que construits uniquement sur la base des observations (*data driven* en anglais), atteignent les mêmes vitesses que celles pouvant être atteintes en supposant la régularité de $s$ connue. On peut aussi chercher des estimateurs présentant des critères d'optimalité au sein d'une famille donnée. Typiquement, dans l'exemple de l'estimation de la densité ci-dessus, les estimateurs $\hat{s}_{m(\alpha)}$ ou $\hat{s}_{h(\alpha)}$ sont respectivement les meilleurs dans les familles $(\hat{s}_m)_{m\in\mathcal{M}_n}$ et $(\hat{s}_h)_{h\in\mathcal{H}_n}$, mais ne sont pas adaptatifs, car leur construction nécessite la connaissance de l'indice de régularité $\alpha$ de $s$.

De nombreuses méthodes adaptatives se sont développées depuis les années 1990. Les principales sont les procédures de sélection de modèles par pénalisation, les méthodes dites de

Lepski, le seuillage de coefficients en base d'ondelettes, ainsi que les méthodes d'agrégation d'estimateurs. Nous présentons les deux premières, sur lesquelles sont fondés nos travaux. Pour les méthodes de seuillage de coefficients d'ondelettes, le lecteur peut se référer aux travaux fondateurs de Donoho & Johnstone (1994) ou Donoho *et al.* (1995), au livre de Härdle *et al.* (1998) pour une introduction ou encore à l'introduction (section 1.7) de la thèse d'Akakpo (2009) pour de nombreuses références. Pour les procédures d'agrégation, renvoyons aux travaux de Nemirovski (2000) ou Rigollet & Tsybakov (2007) par exemple. Récemment, des procédures très générales de sélection d'estimateurs (dans une collection pouvant être non dénombrable) ont vu le jour, englobant notamment, comme cas particulier, la sélection de modèles et l'agrégation : voir Baraud *et al.* (2012).

**Approche oracle**

Nous sommes volontairement restés imprécis ci-dessus quant à la définition précise de l'*adaptation* et de l'*optimalité* d'un estimateur. Il existe de nombreuses définitions possibles pour ces deux termes. Nous avons introduit, à la Section 1.1.1, l'optimalité au sens minimax [3], et évoqué brièvement l'optimalité au sens de l'oracle, approche sur laquelle nous nous basons. Celle-ci est liée à ce que Barron *et al.* (1999) définissent comme l'"*adaptation à la fonction cible*" (p.361 et suivantes).

Nous supposons donnée une collection $(\hat{s}_b)_{b\in\mathcal{B}}$ dépendant d'un paramètre $b$, élément d'un ensemble $\mathcal{B}$. La lettre $b$ désigne l'indice $m$ du modèle dans le cas d'estimateurs par projection, la fenêtre (paramètre de lissage) $h$ pour les estimateurs à noyaux. La Figure Figure 1.6 présente par exemple une collection d'estimateurs à noyaux de la densité gaussienne standard, comme ceux construits en Section 1.1.2.

Notre but est de définir un choix [4] $\hat{b}$ du paramètre de telle sorte que l'estimateur $\hat{s}_{\hat{b}}$ imite l'oracle (notion définie en Section 1.1.2). Son risque doit-être aussi proche que possible de celui de l'oracle, c'est-à-dire satisfaire une *inégalité oracle* :

$$\mathbb{E}\left[\|\hat{s}_{\hat{b}} - s\|^2\right] \leq C \inf_{b\in\mathcal{B}} \mathbb{E}\left[\|\hat{s}_b - s\|^2\right] + R_n. \tag{1.8}$$

La lettre $C$ désigne une constante strictement positive, et $R_n$ est un terme de reste, qui doit être négligeable devant $\inf_{b\in\mathcal{B}} \mathbb{E}[\|\hat{s}_b - s\|^2]$. L'inégalité est d'autant meilleure que la constante $C$ est proche de 1.

Nous obtiendrons, dans cette thèse, des *inégalités de type-oracle* pour des estimateurs à noyaux ou en projection au sens suivant :

$$\mathbb{E}\left[\|\hat{s}_{\hat{b}} - s\|^2\right] \leq C \inf_{b\in\mathcal{B}} \left\{\|\mathbb{E}\left[\hat{s}_b\right] - s\|^2 + Q(b)\right\} + R_n, \tag{1.9}$$

(3). On peut aussi adopter une approche de type *maxiset* pour mesurer les performances d'une procédure d'estimation : cela consiste à déterminer l'ensemble de fonctions le plus gros possible sur lequel un estimateur atteint une vitesse donnée. Kerkyacharian & Picard (2004) en rappellent une définition précise, par exemple.

(4). Il s'agit donc de sélectionner un estimateur dans une collection. Notre approche adaptative est donc différente de l'approche par agrégation d'estimateurs, ou de l'approche par seuillage de coefficients en base d'ondelettes.

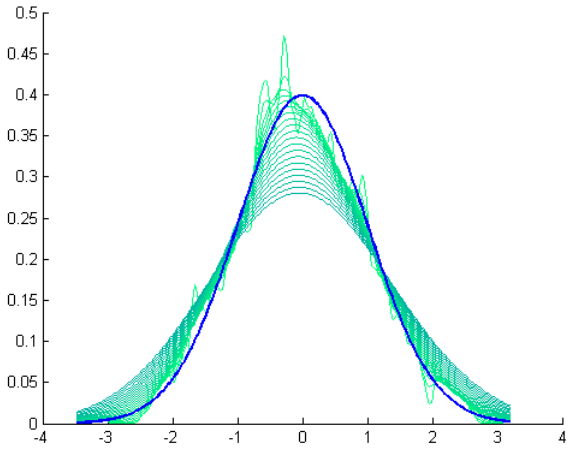



FIGURE 1.6 – Collection d'estimateurs à noyaux de la densité gaussienne centrée réduite. En trait foncé : la densité estimée. En traits clairs : les estimateurs.

où $Q(b)$ désigne l'ordre de grandeur du majorant du terme de variance, soit, pour l'estimation d'une densité,

$$Q(b) = Q(m) = \phi_0^2 \frac{D_m}{n} \quad \text{pour les estimateurs par projection (voir (1.5))},$$
$$Q(b) = Q(h) = \frac{\|K\|^2}{nh} \quad \text{pour les estimateurs à noyaux (voir (1.7))}.$$

Rappelons également que dans ce cadre $\mathbb{E}[\hat{s}_b] = \Pi_{S_m}(s)$ pour un estimateur par projection ($\Pi_{S_m}(s)$ étant la projection de $s$ sur $S_m$), et $\mathbb{E}[\hat{s}_b] = K_h \star s$ pour les noyaux.

**Remarque 1.3. Comparaison des Inégalités (1.8) et (1.9).** La différence entre ces deux inégalités réside dans leurs membres de droite : le risque $\mathbb{E}[\|\hat{s}_b - s\|^2]$ de $\hat{s}_b$ intervenant dans (1.8) a été remplacé par son majorant $\|\mathbb{E}[\hat{s}_b] - s\|^2 + Q(b)$ dans (1.9). Cependant, on peut montrer dans de nombreux cas (dont ceux traités dans cette thèse), que la différence est minime. A titre d'exemple, prouvons-le dans le cas des estimateurs par projection de la densité. Le passage de (1.8) à (1.9) provient de la majoration (1.5), que nous rappelons,

$$\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] \leq \|s - \Pi_{S_m}(s)\|^2 + \phi_0^2 \frac{D_m}{n}.$$

La décomposition du risque en un terme de biais et un terme de variance est exacte en norme $L^2$ (voir (1.3)). Les seules majorations proviennent donc du calcul de la variance (1.4) : deux inégalités sont successivement appliquées.

– La première est la majoration de la variance $\mathrm{Var}(\varphi_j(X_1))$ par $\mathbb{E}[\varphi_j^2(X_1)]$. Le terme que l'on néglige à cette occasion est exactement (voir (1.4)) :

$$\frac{1}{n}\sum_{j=1}^{D_m} \left(\mathbb{E}\left[\varphi_j(X_1)\right]\right)^2 = \frac{1}{n}\sum_{j=1}^{D_m} \left(\langle s, \varphi_j\rangle\right)^2 = \frac{\|\Pi_{S_m}(s)\|^2}{n} \leq \frac{\|s\|^2}{n},$$

par les propriétés du projeté orthogonal $\Pi_{S_m}(s)$ de $s$ sur le sous-espace $S_m$. Ainsi, le terme "oublié" lors de cette majoration est réellement négligeable par rapport à l'ordre de grandeur $D_m/n$.

- La seconde concerne la base $(\varphi_j)_j$, qui est choisie par l'utilisateur : on majore tout d'abord la somme des carrés des fonction $\varphi_j$ par sa norme infinie, puis on utilise la propriété de connexion de normes $(\mathcal{M}_3)$ des modèles : $\| \sum_{j=1}^{D_m} \varphi_j^2 \|_\infty \leq \phi_0^2 D_m/n$. Dans le cas de modèles trigonométriques comme ceux utilisés dans cette thèse, cette propriété est une égalité avec $\phi_0^2 = 1$.

Remarquons que la deuxième étape de la majoration peut-être différente, si on considère la version de la majoration (1.5), énoncée à la Remarque 1.1, où $\phi_0^2$ est remplacée par $\|s\|_\infty$ : dans ce cas, on peut argumenter que la différence entre les deux inégalités considérées est toujours minime, puisque, partant de $\mathbb{E}[\|\hat{s}_m - s\|^2] \leq n^{-1} \sum_{j=1}^{D_m} (\langle s, \varphi_j \rangle)^2$, l'on a l'encadrement

$$\inf_{x \in A} s(x) \frac{D_m}{n} \leq \frac{1}{n} \sum_{j=1}^{D_m} (\langle s, \varphi_j \rangle)^2 \leq \|s\|_\infty \frac{D_m}{n}.$$

Dans le cas d'estimateurs à noyaux de la densité, la seule majoration du risque entraînant le passage de (1.8) à (1.9) est analogue au premier item ($\mathrm{Var}(Z) \leq \mathbb{E}[Z^2]$, avec $Z = K_h(x - X_1)$). On peut également montrer que le terme omis est négligeable (cf. Tsybakov 2009).

Un estimateur sera dit *optimal au sens de l'oracle*, ou *adaptatif*, s'il vérifie une inégalité de type (1.8) ou (1.9), tout en étant fondé uniquement sur les observations. Rappelons que ces notions d'adaptation et d'optimalité font partie de celles définies par Barron *et al.* (1999), et utilisées par exemple par Reynaud-Bouret & Rivoirard (2010) (parmi d'autres auteurs). Le compromis biais-variance est automatiquement réalisé, puisque le membre de droite de (1.9) représente le meilleur compromis possible entre le terme de biais $\|\mathbb{E}[\hat{s}_b] - s\|^2$ et la variance $V(b)$. En faisant l'hypothèse que $s$ appartient à un certain espace fonctionnel, on peut dans un second temps calculer le minimum intervenant dans le membre de droite de (1.9) pour déduire la vitesse de convergence du risque, la meilleure que l'on puisse obtenir dans la collection $(\hat{s}_b)_b$. L'estimateur s'adaptera bien à la régularité de $s$, au sens où nous n'aurons pas utilisé l'appartenance de la cible à l'espace fonctionnel pour le bâtir. Le terme $R_n$ est alors négligeable s'il a un ordre de grandeur moins important que celui de la vitesse : $R_n \approx 1/n$ ou $R_n \approx \ln^p(n)/n$ ($p$ un réel positif) généralement.

L'obtention d'inégalités de la forme (1.9) dans les cadres faisant l'objet de la thèse nécessitera des contrôles et calculs relativement techniques et longs. Par souci de clarté, nous n'avons pas cherché à avoir les constantes les plus fines possibles dans les majorations. Ainsi, la constante $C$ des inégalités de type-oracle que nous démontrons n'est pas nécessairement proche de 1.

Nous décrivons dans les deux paragraphes qui suivent, deux méthodes permettant d'obtenir des inégalités de type (1.9), sans faire d'hypothèses de régularité sur la fonction $s$ : la sélection de modèles par pénalisation pour choisir un estimateur par projection et une méthode inspirée des travaux de Lepski pour sélectionner un estimateur à noyau. Enfin, on

présente une stratégie mixte permettant de sélectionner un estimateur par projection à l'aide d'une méthode inspirée de celle de Lepski.

**Sélection de modèles**

Trouvant son origine dans les travaux de Akaike (1973) et Mallows (1973), la théorie de la sélection de modèles a été formalisée par Birgé & Massart (1997) et Barron *et al.* (1999) (on peut aussi se référer au livre de Massart 2007).

Considérons la famille d'estimateurs $(\hat{s}_m)_{m\in\mathcal{M}_n}$ de la densité $s$ définis par (1.2) à la Section 1.1.2. La collection $\mathcal{M}_n$ est supposée finie. Nous souhaitons définir un critère permettant de sélectionner parmi eux un estimateur dont le risque est aussi proche que possible du meilleur des risques que l'on peut obtenir dans la collection. Comme l'on souhaite précisément obtenir une inégalité (1.9) de type oracle, il est cohérent de sélectionner un indice $m$ minimisant une version empirique de la somme $\|s - \Pi_{S_m}(s)\|^2 + Q(m)$ où $Q(m) = \phi_0^2 D_m/n$ est l'ordre de grandeur du majorant de la variance. Il s'agit donc d'estimer le terme de biais, qui se réécrit

$$\|s - \Pi_{S_m}(s)\|^2 = \|s\|^2 - \|\Pi_{S_m}(s)\|^2,$$

par définition de la projection orthogonale $\Pi_{S_m}(s)$ de $s$ sur $S_m$. Par définition du contraste (voir (1.1)), on constate que $\gamma_n(\hat{s}_m) = -\|\hat{s}_m\|^2$, et donc minimiser cette quantité en $m$ revient à minimiser le biais. Ainsi, on définit

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\}, \tag{1.10}$$

où $\mathrm{pen}(m) = \kappa\phi_0^2 D_m/n$ (pour une constante numérique $\kappa > 0$) estime le terme de variance.

**Remarque 1.4.** Une autre manière de concevoir la méthode consiste à considérer, qu'au vu de la définition des estimateurs $\hat{s}_m$, $m \in \mathcal{M}_n$ comme minimisant le contraste $\gamma_n$ sur chaque modèle $S_m$, il serait naturel de définir $\hat{m} = \arg\min_{m\in\mathcal{M}_n} \gamma_n(\hat{s}_m)$. Mais, si $S_{m'} \subset S_m$, alors $\hat{s}_{m'} \in S_m$ et par construction $\gamma_n(\hat{s}_m) \leq \gamma_n(\hat{s}_{m'})$. Le contraste, vu comme version empirique du risque, décroît quand la dimension augmente et donc sous-estime le vrai risque. Il faut donc pénaliser les modèles en fonction de leur dimension, ce qui légitime l'introduction de la fonction "pen", appelée *pénalité*. L'estimateur $\hat{s}_{\hat{m}}$ est appelé *estimateur par minimum de contraste pénalisé*.

Nous avons justifié heuristiquement la définition (1.10). On peut démontrer que l'estimateur $\hat{s}_{\hat{m}}$ est bien optimal au sens de l'oracle. L'argument principal est le contrôle des fluctuations de processus empiriques (principalement pour montrer que le terme de reste est bien négligeable). Ces outils sont rappelés au Chapitre 2, Section 2.1.1. Nous renvoyons le lecteur au livre de Massart (2007) pour des résultats bien plus généraux, et donnons dans la suite de cette thèse des preuves de résultats analogues dans les cadres que nous étudions (voir par exemple le Théorème 3.1).

L'implémentation d'un estimateur comme $\hat{s}_{\hat{m}}$ pose le choix de la constante intervenant dans la pénalité. Les démonstrations en fournissent une valeur théorique, souvent pessimiste

car résultant de majorations nombreuses. La calibration de pénalités minimales/optimales fait l'objet de nombreux travaux depuis celui de Birgé & Massart (2007). En pratique, des méthodes de type "heuristique de pente" sont maintenant développées. Nous ferons appel au package CA.P.U.S.HE (Matlab) pour l'utiliser (Baudry *et al.* 2012).

**Sélection de fenêtres**

Considérons la collection $(\hat{s}_h)_{h\in\mathcal{H}_n}$ d'estimateurs de la densité $s$ définis par (1.6). La méthode dite de Lepski est une méthode bien connue permettant de choisir le paramètre de lissage ou fenêtre d'un estimateur à noyau pour assurer des propriétés d'adaptation. Elle est fondée, au départ, sur la comparaison d'estimateurs avec fenêtre fixée deux à deux entre eux, c'est-à-dire sur l'étude de quantités impliquant $\hat{s}_h - \hat{s}_{h'}$, $h, h' \in \mathcal{H}_n$.

La méthode trouve son origine dans les travaux de Lepskiĭ (1991, 1992a,b). Dans ces premiers articles, une procédure générale est déjà donnée, relative tout autant aux estimateurs à noyaux que par polynômes locaux, et concernant les différents risques possibles ($L^p$, ponctuels...). Elle est améliorée ensuite par Lepski *et al.* (1997) (modèle de bruit blanc gaussien), et Goldenshluger & Nemirovski (1997) (régression gaussienne). Les premières versions reposent sur l'estimation du paramètre de régularité $\alpha$ de la fonction estimée (estimation $\hat{\alpha}$, requièrant le rayon $L$ de la boule correspondante en estimation ponctuelle au moins), et les résultats sont des résultats d'adaptation au sens minimax. Les premières inégalités oracles obtenues avec ce type de méthodes le sont par Kerkyacharian *et al.* (2001). C'est dans leur travail aussi qu'apparait pour la première fois un estimateur artificiel, intermédiaire, auquel on compare un estimateur à fenêtre fixe. Ces idées sont reprises par Goldenshluger & Lepski (2008, 2009) de manière très générale, tant du point de vue des modèles statistiques que des estimateurs étudiés.

La procédure que nous présentons ici est inspirée du récent article de Goldenshluger & Lepski (2011a), qui propose une sélection de fenêtres d'estimateurs pour la densité d'un vecteur aléatoire en norme $L^p$. Des résultats d'adaptation non-asymptotiques y sont obtenus. Ce que nous proposons peut-être considéré comme une version simplifiée de la méthode de Lepski en ce sens que

1. nous nous restreignons à considérer le risque $L^2$ (et pas l'ensemble des risques $L^p$, ou le risque ponctuel),
2. nous considérons une collection finie $\mathcal{H}_n$ de fenêtres possibles, et non pas un intervalle, et ce, par souci de faisabilité de la méthode (un ensemble fini est raisonnable d'un point de vue pratique).

Ces restrictions nous permettent en particulier d'utiliser les outils typiques de la sélection de modèles dans les démonstrations (concentration de processus empiriques). Notons que cette méthodologie est également suivie par Comte & Lacour (2013) ou Doumic *et al.* (2012).

L'idée peut-être expliquée comme suit. Comme en sélection de modèles, nous cherchons à définir $\hat{s}_{\hat{h}}$ qui satisfait une inégalité de type (1.9), c'est-à-dire dont le risque est aussi proche

que possible que $\min_{h\in\mathcal{H}_n}\{\|s-s\star K_h\|^2+\|K\|^2(nh)^{-1}\}$. Nous remplaçons les quantités intervenant dans cette somme biais-variance par des contreparties empiriques. Nous commençons par poser $V(h) = \kappa'\|K\|^2/(nh)$, pour une constante $\kappa' > 0$, pour approcher le terme de variance [5]. La spécificité de la méthode réside dans l'estimation du biais : posons

$$A(h) = \max_{h'\in\mathcal{H}_n}\left(\left\|\hat{s}_{h'} - \hat{s}_{h,h'}\right\|^2 - V(h')\right)_+, \quad h \in \mathcal{H}_n, \tag{1.11}$$

où $\hat{s}_{h,h'} = K_h \star \hat{s}_{h'}$ est l'estimateur auxiliaire évoqué plus haut. Une heuristique conduisant à cette définition est la suivante : pour approcher le biais $\|s - s \star K_h\|^2$, on remplace $s$ par un estimateur avec fenêtre fixe, $\hat{s}_{h'}$. Ceci conduit à $\|\hat{s}_{h'} - K_h \star \hat{s}_{h'}\|^2 = \|\hat{s}_{h'} - \hat{s}_{h,h'}\|^2$. Mais, contrairement au biais, cette quantité "contient" de l'aléa et donc de la variabilité : pour corriger ceci, il est nécessaire de retrancher la "part de variance" $V(h')$ correspondante. Enfin, comme il n'y a pas de raison de choisir une fenêtre $h' \in \mathcal{H}_n$ plutôt qu'une autre, on balaye l'ensemble de la collection. Ceci n'est bien sûr qu'une justification "avec les mains", et l'on peut démontrer rigoureusement que $A(h)$ a bien l'ordre de grandeur du biais au sens où $A(h) \leq C\|s - s \star K_h\|^2$ (voir par exemple le Lemme 3.2, Chap. 3). La preuve que nous donnons repose sur les propriétés de la convolution et sur les mêmes outils que la sélection de modèles : le contrôle de $(\|\hat{s}_{h'} - s\star K_{h'}\|^2 - V(h'))_+$ fait intervenir l'Inégalité de Talagrand (Proposition 2.2, Chap. 2).

De ces définitions découle le choix suivant de fenêtre :

$$\hat{h} = \arg\min_{h\in\mathcal{H}_n}\{A(h) + V(h)\}.$$

On prouve alors que $\hat{s}_{\hat{h}}$ vérifie une inégalité de type oracle (1.9). Cette méthode adaptative, utilisée dans la thèse, sera appelée méthode inspirée de Goldenshluger et Lepski ou plus rapidement *méthode de Goldenshluger-Lepski*.

Tout comme la minimisation d'un critère pénalisé où la pénalité dépend d'une constante $\kappa$ à calibrer, se pose en pratique la question du choix de la constante $\kappa'$ impliquée dans la définition de $V(h)$. Nous calibrons dans les illustrations la constante par une série préalable de simulations (cf. Appendice 1, Section 3.6, Chapitre 3 pour des détails).

**Stratégie mixte**

De la façon dont nous les avons présentées, les stratégies de sélection de modèles ou de fenêtres précédentes présentent de nombreux points communs. En particulier, toutes deux visent à sélectionner un indice $\hat{b}$ tel que $\hat{s}_{\hat{b}}$ ait un risque proche de $\min_{b\in\mathcal{B}}(\|s-\mathbb{E}[\hat{s}_b]\|^2+Q(b))$, pour reprendre les notations introduites en début de section. Ceci se fait dans les deux cas en définissant un équivalent empirique de la variance $Q(b)$ et du biais $\|s - \mathbb{E}[\hat{s}_b]\|^2$. La version empirique du biais est donnée par l'utilisation du contraste en sélection de modèles, et par le terme $A(h)$ défini par (1.11) pour la méthode de Goldenshluger-Lepski.

(5). La définition exacte est en fait $V(h) = \kappa'(1 + \|K\|_{L^1}^2)\|K\|^2/(nh)$, la quantité $(1 + \|K\|_{L^1}^2)$ provenant des preuves. Notre but étant seulement de présenter de façon simple la méthodologie, nous omettons cette constante ici.

Nous proposons alors une procédure de sélection mixte : sélection d'un estimateur par projection inspirée de la méthode de sélection de fenêtres présentée. La méthode que nous suggérons est basée sur la définition d'une quantité $A(m)$ équivalente à (1.11) pour les estimateurs par projections. Comme dans la stratégie de Goldenshluger-Lepski, l'idée est d'estimer le biais en comparant des estimateurs entre eux. En procédant comme dans l'heuristique du paragraphe précédent, on peut tout d'abord penser remplacer $\|s - \Pi_{S_m}(s)\|^2$ par $\|\hat{s}_{m'} - \Pi_{S_m}(\hat{s}_{m'})\|^2$. Or, $\hat{s}_{m'}$ estime $\Pi_{S_{m'}}(s)$. Donc, $\Pi_{S_m}(\hat{s}_{m'})$ estime $\Pi_{S_m}(\Pi_{S'_m}(s))$, soit encore, comme les modèles sont emboîtés (hypothèse $\mathcal{M}_2$, Section 1.1.2), $\Pi_{S_{m\wedge m'}}(s)$. Donc on substitue plutôt $\|\hat{s}_{m'} - \hat{s}_{m\wedge m'}\|^2$. Un raisonnement identique à celui de l'heuristique ci-dessus conduit alors à poser

$$A(m) = \max_{m' \in \mathcal{M}_n} \left( \|\hat{s}_{m'} - \hat{s}_{m\wedge m'}\|^2 - V(m') \right)_+, \quad m \in \mathcal{M}_n,$$

où $V(m') = \kappa' \phi_0^2 D_{m'}/n$ estime la variance. On choisit alors

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{A(m) + V(m)\}.$$

Nous montrons toujours que $\hat{s}_{\hat{m}}$ satisfait (1.9) (voir par exemple la preuve du Théorème 3.1, où cette méthodologie est utilisée). Une telle sélection de modèles inspirée de Goldenshluger & Lepski (2011a) est aussi utilisée par Comte & Johannes (2012). Notons qu'une stratégie fondée sur des idées analogues, mais s'inspirant de versions antérieures de la méthode de Lepski (et donc n'impliquant par d'intermédiaire de la forme $\hat{s}_{m\wedge m'}$) a été étudiée par Birgé (2001). Enfin, en retour de cette dernière méthode, et en tenant compte de l'analogie entre $h$ et $D_m^{-1}$, on peut également penser faire de la sélection de fenêtre avec un critère impliquant $\hat{s}_{h\vee h'}$. Nous envisagerons cette possibilité à l'Appendice 3 du Chapitre 4 (Section 4.9).

Pour conclure la présentation des méthodes adaptatives utilisées dans la thèse, le Tableau 1.1 récapitule les trois méthodes de sélection présentées. Notons que des comparaisons de ces méthodes seront effectuées tout au long de la thèse, à l'occasion des différents exemples rencontrés : comparaison de la méthode de pénalisation avec la sélection de modèles inspirée de Goldenshluger & Lepski (2011a) tout au long du Chapitre 3 (cas de l'estimation de fonctions d'une seule variable) et dans l'Appendice 1 (Section 5.6) du Chapitre 5 (cas de l'estimation de fonctions de deux variables) ou comparaison de différentes méthodes de sélection de fenêtres (à la façon de Goldenshluger & Lepski 2011a), à l'Appendice 3 (Section 4.9) du Chapitre 4.

| | Sélection de modèles par pénalisation | Sélection de modèles inspirée de G.L. | Sélection de fenêtres inspirée de G.L. |
|---|---|---|---|
| quantité à estimer | $\|s - \Pi_{S_m}(s)\|^2 + \phi_0^2 \frac{D_m}{n}$ | $\|s - \Pi_{S_m}(s)\|^2 + \phi_0^2 \frac{D_m}{n}$ | $\|s - s \star K_h\|^2 + \|K\|^2 \frac{1}{nh}$ |
| biais estimé | $\gamma_n(\hat{s}_m)$ | $A(m) = \max_{m' \in \mathcal{M}_n} \left( \|\hat{s}_{m'} - \hat{s}_{m \wedge m'}\| - V(m') \right)_+$ | $A(h) = \max_{h' \in \mathcal{H}_n} \left( \|\hat{s}_{h'} - K_h \star \hat{s}_{h'}\| - V(h') \right)$ |
| variance estimée | $\mathrm{pen}(m) = \kappa \phi_0^2 \frac{D_m}{n}$ | $V(m) = \kappa' \phi_0^2 \frac{D_m}{n}$ | $V(h) = \kappa' \|K\|^2 \frac{1}{nh}$ |
| Indice sélectionné | $\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\}$ | $\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{A(m) + V(m)\}$ | $\hat{h} = \arg\min_{h \in \mathcal{H}_n} \{A(h) + V(h)\}$ |

TABLE 1.1 – Récapitulatif des méthodes de sélection de modèles ou de fenêtres utilisées dans la thèse. Notation : "G.L". pour Goldenshluger-Lepski

## 1.2 Partie A : Méthode de déformation pour l'estimation adaptative

L'objectif de cette section est de donner une vue d'ensemble des problématiques qui font l'objet de la Partie A, des méthodes statistiques développées pour les résoudre, ainsi que des principaux résultats obtenus au regard de la littérature existante.

### 1.2.1 Motivation : modèle de régression

#### Régression additive

Nous considérons dans cette section le modèle le plus classique étudié dans cette thèse, celui de la régression additive, pour justifier dans un premier temps l'introduction d'observations déformées. L'objectif est la reconstruction d'un signal $s$ à partir d'observations bruitées, issues de l'échantillonnage de $s$ : les données sont constituées de couples de variables aléatoires réelles indépendants et identiquement distribués (*i.i.d.* dans la suite) $(X_i, Y_i)_{i=1,\dots,n}$. Les variables $X_i$, $i = 1, \dots, n$, appelées design, représentent des variables explicatives, et les $Y_i$, $i = 1, \dots, n$ sont les variables à expliquer. La relation entre les $X_i$ et les $Y_i$ est modélisée par la fonction $s$, à un bruit additif près :

$$Y_i = s(X_i) + \varepsilon_i, \quad i = 1, \dots, n. \tag{1.12}$$

Les variables aléatoires $\varepsilon_i$ sont *i.i.d.*, centrées et indépendantes des $X_i$, ce qui signifie que l'on fait l'hypothèse selon laquelle la relation entre la réponse $Y_i$ et le prédicteur $X_i$ est totalement expliquée par $s$ : $s(X) = \mathbb{E}[Y|X]$. La loi des variables $\varepsilon_i$ caractérise donc uniquement les variations de $Y_i$ autour de cette espérance conditionnelle. Le modèle (1.12) est dit *homoscédastique*, puisque les $\varepsilon_i$ ont même loi et donc même variance $\sigma^2$ (supposée finie). Nous supposerons que les variables $X_i$ admettent une densité notée $f_X$, et nous noterons $F_X$ leur fonction de répartition.

Ce modèle a été étudié de manière intensive depuis la seconde moitié du XXème siècle, et il existe une vaste littérature consacrée à l'estimation adaptative de la fonction de régression, avec des critères d'optimalité fondés tout autant sur l'approche oracle que sur l'approche minimax. Les méthodes historiques sont fondées sur les noyaux et ont été initiées par Nadaraya (1964) et Watson (1964) qui ont bâti l'estimateur qui porte désormais leur nom. Partant du constat selon lequel $\mathbb{E}[Y_i K_h(x - X_i)] = K_h \star (sf_X)$ (avec $K_h$ défini à partir d'un noyau $K$ comme à la Section 1.1.2), ils proposèrent un estimateur de $s$ construit comme le quotient d'un estimateur du produit $sf_X$ sur un estimateur de la densité $f_X$. L'estimateur de Nadaraya-Watson est donc le suivant :

$$\tilde{s}^{NW} : x \mapsto \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}. \tag{1.13}$$

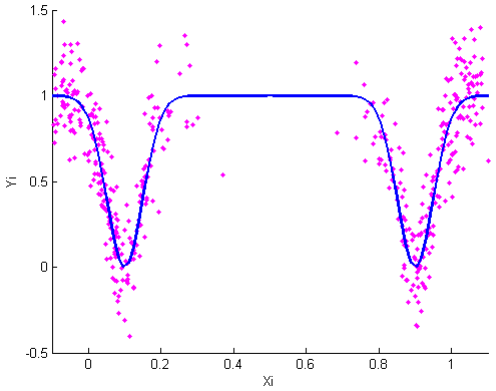



FIGURE 1.7 – Exemple de données $(X_i, Y_i)_i$ suivant le modèle de régression additive (1.12). En trait plein : fonction de régression $s$.

Il fait toujours aujourd'hui l'objet d'études et de propositions d'extension, basées notamment sur des techniques de type polynômes locaux, comme indiqué dans la bibliographie introductive propre au Chapitre 3.

D'un point de vue théorique, l'estimateur (1.13) initial est construit comme le quotient des estimateurs de deux fonctions différentes. Même si le choix de la fenêtre $h$ est généralement le même pour les deux, il n'y a pas de raisons *a priori*, pour obtenir des résultats adaptatifs, de sélectionner la même fenêtre pour l'estimateur à noyau du numérateur, et pour celui du dénominateur. De plus, la mise en place de deux procédures de sélection conduit à des difficultés d'ordre théorique. Un quotient d'estimateur est par nature difficile à gérer (voir à ce sujet Penskaya 1995). Les estimateurs fondés sur des polynômes locaux ont de même généralement des expressions complexes rendant l'adaptation difficile : le caractère récent des travaux de Chichignoud (2010) sur l'estimateur bayésien fondé sur un critère de type pseudo-vraisemblance (avec sélection de fenêtre par méthode de Lepski) en est une illustration. D'un point de vue pratique également, la présence d'un dénominateur peut entraîner une instabilité numérique lorsque ce dernier est proche de zéro, ce qui peut être le cas par exemple lorsque le design présente des "trous" : la Figure Figure 1.7 donne un exemple de telles données. Des travaux fondés sur le risque ponctuel d'estimateurs définis à partir de polynômes locaux permettent toutefois de prendre en compte des données très inhomogènes : Gaïffas (2007) obtient des résultats d'optimalité au sens minimax avec une méthode de Lepski.

Dans une autre direction, de nombreuses méthodes existent pour estimer la fonction $s$ en tirant parti de son développement dans certaines bases fonctionnelles. Nous renvoyons de même à l'introduction du Chapitre 3. Les stratégies fondées sur la minimisation d'un contraste et inspirées des méthodes de sélection de modèles reposent généralement sur des critères de type moindres-carrés pénalisés (Polyak & Tsybakov 1990, Baraud 2002 par exemple) : l'expression à minimiser sur un modèle $S_m$ comme ceux décrits à la Section 1.1.2, par rapport

à une fonction $t$ est

$$\gamma_n^{MC}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2,$$

soit encore

$$\gamma_n^{MCbis}(t) = \|t\|_n^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i t(X_i), \quad \text{où} \quad \|t\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} t^2(X_i). \tag{1.14}$$

Le calcul de l'estimateur des moindres-carrés $\hat{s}_m^{MC} := \arg\min_{t\in S_m} \gamma_n^{MCbis}(t)$ aboutit à une expression explicite seulement sous l'hypothèse d'inversibilité de la matrice $G = (g_{j,k})_{j,k\in\{1,\dots,D_m\}}$, définie par

$$g_{j,k} := \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i)\varphi_k(X_i), \quad j,k \in \{1,\dots,D_m\}.$$

On obtient alors $\hat{s}_m^{MC} = \sum_{j=1}^{D_m} \hat{a}_j\varphi_j$ où les coefficients $\hat{a} = (\hat{a}_j)_{j=1,\dots,D_m}$ sont définis par la relation $\hat{a} = G^{-1}b$, avec $b_j = n^{-1}\sum_{i=1}^{n} Y_i\varphi_j(X_i)$. L'inversion la matrice $G$ (matrice de Gram pour la norme empirique) peut donc également entraîner, si l'estimateur est implémenté sans correction particulière, une instabilité, et la majoration du risque de l'estimateur de type moindres-carrés $\hat{s}_m^{MC}$ requiert l'introduction d'événements de grande probabilité portant sur le contrôle des valeurs propres de la matrice $G$. Des résultats adaptatifs de type inégalité oracle peuvent toutefois être obtenus : voir par exemple Baraud (2002) pour un contrôle du risque $L^2$, ou Birgé (2004) pour un risque fondé sur la distance de Hellinger.

**Déformation**

La prise en compte de données très irrégulières, ainsi que la volonté de bâtir des estimateurs ayant une expression simple, ne faisant intervenir ni quotient ni inversion de matrices tout en ayant des propriétés d'adaptativité, nous ont amenés à considérer la possibilité de bâtir des estimateurs de la fonction de régression $s$ dans le modèle (1.12) à partir de données déformées par la fonction de répartition $F_X$ du design, de la forme $(F_X(X), Y)$.

Notons $A$ le support de la variable $X$ : nous supposons qu'il s'agit d'un intervalle, qui représente aussi l'intervalle d'estimation. La fonction de répartition $F_X$ est supposée bijective sur $A$, ce qui est par exemple le cas dès que la densité $f_X$ de $X$ ne s'annule pas sur $A$. L'inverse étant noté $F_X^{-1} : (0;1) \to A$, nous partons du constat suivant : pour toute fonction $t \in L^2((0;1))$,

$$\mathbb{E}\left[Yt \circ F_X(X)\right] = \langle t, s \circ F_X^{-1}\rangle, \tag{1.15}$$

où $\langle .,. \rangle$ est le produit scalaire de $L^2((0;1))$. En effet,

$$\mathbb{E}\left[Yt \circ F_X(X)\right] = \mathbb{E}\left[s(X)t \circ F_X(X)\right] = \int_A s(x)t \circ F_X(x) f_X(x)dx = \int_0^1 s \circ F_X^{-1}(u)t(u)du,$$

en posant le changement de variables $u = F_X(x)$.

La relation (1.15) permet d'envisager une procédure d'estimation en deux étapes : dans un premier temps, on peut reconstruire la fonction auxiliaire

$$g = s \circ F_X^{-1}, \quad \text{définie sur } (0;1), \tag{1.16}$$

à l'aide d'un estimateur $\hat{g}$, puis dans un second temps, définir un estimateur de la fonction cible $s$ en posant $\hat{s} = \hat{g} \circ F_X$ si la loi du design, et donc aussi la fonction $F_X$, sont supposées connues, ou plus généralement

$$\hat{s} = \hat{g} \circ \hat{F}_n,$$

où $\hat{F}_n = n^{-1} \sum_{i=1}^n \mathbf{1}_{[X_i,\infty[}$ est la fonction de répartition empirique de l'échantillon $(X_i)_{i=1,\dots,n}$ sinon. Une telle procédure peut être mise en place pour bâtir des estimateurs appartenant aux deux grandes familles décrites à la Section 1.1.2 : projection ou noyaux, en suivant la Remarque 1.2. La fonction $F_X$ sera appelée *fonction de déformation*.

Pour construire des estimateurs de type projection, prenons $t = \varphi_j$ dans l'égalité (1.15), en notant $(\varphi_j)_{j\in\mathbb{N}\setminus\{0\}}$ une base hilbertienne de $L^2((0;1))$. De cette façon, on obtient $\mathbb{E}\left[Y\varphi_j \circ F_X(X)\right] = a_j$, où la suite $(a_j)_j$ désigne les coefficients du développement de la fonction $g$ dans la base $(\varphi_j)_j$. Un estimateur par projection sur $S_m = \mathrm{Vect}\{\varphi_1, \dots, \varphi_{D_m}\}$ au sens défini à la Section 1.1.2 est donc

$$\hat{g}_m = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{F}} \varphi_j, \quad \text{où} \quad \hat{a}_j^{\hat{F}} = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_j \circ \hat{F}_n(X_i),$$

et l'estimateur correspondant pour la fonction $s$, défini par $\hat{s}_m = \hat{g}_m \circ \hat{F}_n$, s'écrit donc

$$\hat{s}_m = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{F}} \varphi_j \circ \hat{F}_n. \tag{1.17}$$

Cet estimateur admet donc un développement non pas dans une base orthonormée classique, mais dans une base dite *déformée* : $(\varphi_j \circ \hat{F}_n)_j$. Les coefficients du développement sont exprimés comme de simple moyennes empiriques, ne faisant intervenir aucune inversion de matrice (par opposition à l'estimateur des moindres-carrés $\tilde{s}_m^{MC}$ présenté à la Section 1.2.1).

Pour construire un estimateur à noyau, nous choisissons cette fois $t = K_h(x - .)$ dans l'égalité (1.15), $K_h$ étant toujours défini comme à la Section 1.1.2). Nous obtenons $\mathbb{E}\left[YK_h(x - F_X(X))\right] = K_h \star (g\mathbf{1}_{(0;1)})$. L'estimateur résultant pour $g$ est donc

$$\hat{g}_h(u) = \frac{1}{n} \sum_{i=1}^n K_h\left(u - \hat{F}_n(X_i)\right), \quad u \in (0;1),$$

et l'estimateur correspondant pour la fonction $s$, défini par $\hat{s}_h = \hat{g}_h \circ \hat{F}_n$, s'écrit donc

$$\hat{s}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h\left(\hat{F}_n(x) - \hat{F}_n(X_i)\right), \quad x \in \mathbb{R}. \tag{1.18}$$

Nous parlerons donc cette fois de *noyau déformé*. L'expression de l'estimateur $\hat{s}_h$ ne fait pas intervenir de quotient, contrairement à l'estimateur $\tilde{s}_h^{NW}$ défini par (1.13).

**Résultats existants**

Les estimateurs fondés sur les données transformées $(F_X(X_i), Y_i)_i$ ou $(\hat{F}_n(X_i), Y_i)$ ont déjà été introduits dans la littérature.

L'estimateur (1.18) a, à notre connaissance, été défini par Yang (1981) puis repris par Stute (1984) : le premier démontra la convergence en moyenne quadratique, et le second la normalité asymptotique. Stute (1986a,b) a également étendu la stratégie de déformation à l'estimation à noyaux de la fonction de répartition conditionnelle : la consistance, ainsi qu'un principe d'invariance de type Donsker furent prouvés. Dans les travaux de Mehra *et al.* (2000), une loi du logarithme itéré est énoncée pour l'estimateur à noyau déformé de la densité conditionnelle.

Des estimateurs de type (1.17), fondés sur le développement de la fonction à reconstruire dans des bases déformées, ont également été étudiés, pour mettre en place des procédures de seuillage de coefficients en base d'ondelettes : un travail fondateur dans ce cadre est celui de Kerkyacharian & Picard (2004). Les auteurs ont calculé pour l'estimateur dont les coefficients déformés sont seuillés, des vitesses de convergence du risque $L^p$ lorsque la fonction cible appartient à un espace de régularité "à poids" (introduit dans l'article) et lorsque le design vérifie des propriétés de type Muckenhoupt (voir hypothèse $(\mathcal{H}_p)$ p.1062), le bruit étant supposé gaussien : les vitesses obtenues sont les vitesses classiques en estimation non-paramétrique, à un facteur logarithmique près, dont on sait qu'il est inévitable pour le contrôle $L^p$ de procédures de type ondelettes. Le problème du biais est étudié de manière très approfondie. La méthode de Kerkyacharian & Picard (2004) est ensuite reprise par Pham Ngoc (2009) pour proposer des estimateurs bayésiens dans le modèle de régression, par Chesneau (2007) pour le modèle de bruit blanc gaussien (ces deux derniers auteurs supposant $F_X$ connue), et par Kulik & Raimondo (2009) pour estimer la fonction de régression $s$ dans des contextes où les erreurs sont dépendantes (ces derniers prouvèrent également des bornes inférieures pour le risque $L^p$ de leurs estimateurs). Chesneau & Willer (2012) s'intéressent à des ondelettes déformées pour prendre en compte des données dépendantes.

### 1.2.2 Vue d'ensemble des résultats de la Partie A

**Contribution**

La contribution de notre travail à l'étude de méthodes de déformation (**Partie A** essentiellement) peut être résumée par les points suivants.

1. Premièrement, nous mettons en place des méthodes inspirées de la sélection de modèles (telle qu'elle est décrite dans Massart 2007) et de la sélection de fenêtres (inspirée de Goldenshluger & Lepski 2011a) pour les estimateurs en bases déformées (1.17) ou à noyaux déformés (1.18) de la fonction de régression, dans le modèle à design aléatoire (et de fonction de répartition inconnue), avec un bruit de loi quelconque (loi absolument continue pour la mesure de Lebesgue, et admettant un moment d'ordre 2). L'utilisation des bases déformées pour la régression est le sujet du **Chapitre 3**, l'estimation à noyaux déformés étant abordée au **Chapitre 4**.

2. Deuxièment, nous étendons la méthode de déformation (version bases et version noyaux) présentée dans le cadre de la régression à la Section 1.2.1 à d'autres cadres statistiques variés, incluant des modèles utilisés en analyse de survie : ceci nécessite dans certains cas l'utilisation de fonctions de déformation plus complexes que la fonction de répartition $F_X$. Des procédures de sélection de modèles ou de fenêtres sont également mise en place. C'est l'objet du **Chapitre 4**. Un exemple original de déformation est aussi traité au **Chapitre 7** de la Partie B.
3. Troisièment, nous démontrons, au **Chapitre 5** la pertinence de la déformation ainsi que des méthodes de sélection associées pour estimer la densité conditionnelle d'une variable aléatoire sachant un prédicteur réel.

Les modèles statistiques correspondant aux points 2 et 3 précédents sont présentés ci-dessous. Pour chacun, les résultats obtenus sont introduits à la section qui suit.

**Exemples étudiés**

Nous présentons dans ce paragraphe les modèles statistiques étudiés à l'aide d'une méthode de déformation dans la Partie A de la thèse, à l'exception du modèle de **régression additive à design aléatoire** décrit à la Section 1.2.1 et que nous noterons **Exemple 1.**, et du modèle d'estimation de la **densité relative avec censure** qui sera introduit à la Section 1.3.1. Des références bibliographiques seront précisées dans les Chapitres 4 (Exemples 2-4) et 5 (Exemple 5).

**Exemple 2. Régression multiplicative (modèle hétéroscédastique).** Les observations sont constituées de couples *i.i.d.* $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ tels que

$$Y_i = \sigma(X_i)\varepsilon_i, \tag{1.19}$$

où $(\varepsilon_i)_i$ est un échantillon de variables aléatoires centrées réduites, admettant un moment d'ordre 4. La fonction d'intérêt est la fonction de volatilité $\sigma^2$. Comme noté par Chichignoud (2012), en posant $Y_i' = Y_i^2$ et $\eta_i = \varepsilon_i^2 - 1$, le modèle (1.19) devient un modèle de régression hétéroscédastique :

$$Y_i' = \sigma^2(X_i) + \sigma^2(X_i)\eta_i.$$

Ce modèle est fortement connecté aux modèles plus complexes de type auto-régressifs ($X_i = Y_{i-1}'$), qui ont fait l'objet de nombreuses études, aussi bien d'estimateurs fondés sur des noyaux que d'estimateurs par projection, en partie pour leurs importantes applications en finance ou économétrie.

**Exemple 3. Estimation de la fonction de répartition à partir de données censurées par intervalle, Cas 1.** Dans ce modèle, on s'intéresse à la fonction de répartition $F_Z$ d'une variable aléatoire positive $Z$ qui n'est pas observée. Les données sont constituées de couples *i.i.d.* $(X_i, \mathbf{1}_{Z_i\leq X_i})$, où les $Z_i$, $i = 1,\dots,n$, sont distribuées comme $Z$ et où les $X_i$ sont également des variables positives, indépendantes des $Z_i$. Ce problème peut être décrit par la

régression de $\mathbf{1}_{Z_1 \leq X_1}$ sur $X_1$, puisque

$$F_Z(x) = \mathbb{E}\left[\mathbf{1}_{Z_1 \leq x} | X_1 = x\right], \quad x \in \mathbb{R}.$$

Un tel modèle provient de problématiques liées au domaine bio-médical, et fournit des données qui sont englobées sous le terme générique de *données de survie* : typiquement, $Z_i$ est l'instant où un individu $i$ est contaminé par un virus. Cet instant n'est jamais observé : on dispose généralement uniquement de l'instant $X_i$ où le patient se soumet à un test de dépistage, ainsi que du résultat de celui-ci : la variable binaire $\mathbf{1}_{Z_i \leq X_i}$ indique si oui ou non le patient est contaminé au temps $X_i$, c'est-à-dire si $Z_i$ appartient à l'intervalle de temps $[0; X_i]$ ou $]X_i, \infty[$ : la seule information est donc l'état "courant" du patient (d'où le nom de *current status data* donné aussi à ce type d'observations). Comme la plupart des problèmes issus de l'analyse de survie, de telles données apparaissent aussi dans des applications éloignées des études médicales (voir van de Geer 1993, Example 3.3(a)).

**Exemple 4. Estimation du taux risque instantané à partir de données censurées à droite** Comme dans l'exemple précédent, la variable d'intérêt est un temps de survie non observé. Les données sont des couples *i.i.d.*, de la forme $(Z_i \wedge C_i, \mathbf{1}_{Z_i \leq C_i})$, avec $Z_i$ et $C_i$ des variables positives et indépendantes. La fonction cible est le taux de risque instantané (*hazard rate*) défini par

$$s = \frac{f_Z}{1 - \bar{F}_Z}, \tag{1.20}$$

où $f_Z$ et $F_Z$ sont respectivement une densité et la fonction de répartition des variables $Z_i$.

De telles données sont une nouvelle fois observées dans des études médicales de suivis de $n$ patients : $Z_i$ représente la survie du patient $i$ après un traitement. Celui-ci peut disparaître de l'étude (décéder pour une cause indépendante du traitement) à un instant aléatoire $C_i$ appelé instant de censure, avant que son temps de survie ne soit observé. La fonction $s$ définie en (1.20), prise en un instant $x$, représente la probabilité que le patient meure juste après $x$, sachant qu'il était vivant jusqu'au temps $x$ : $s(x) \approx \mathbb{P}(Z_1 \in [x, x + dx] | Z_1 \geq x) dx$. Des problématiques analogues peuvent intervenir en fiabilité : des exemples sont donnés par Rivoirard & Stoltz (2012) (Chap. 14).

**Exemple 5. Estimation de la densité conditionnelle** Le dernier exemple que nous avons étudié est celui de l'estimation de la densité conditionnelle d'une variable $Y_1$ sachant un prédicteur $X_1$, à partir d'un $n-$échantillons de variables $(X_i, Y_i)_{i=1,\ldots,n}$ : notant $f_{(X,Y)}$ une densité du couple, et $f_X$ la marginale de $X$, celle-ci s'écrit,

$$\pi(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}, \quad (x, y) \text{ tel que } f_X(x) \neq 0. \tag{1.21}$$

**Intérêt de la déformation dans les Exemples 2 à 5.** Deux courtes remarques justifient maintenant la mise en place d'une procédure d'estimation adaptative par déformation dans ces quatre problèmes d'estimation.

- Les Exemples 2 et 3 peuvent être considérés comme des problèmes de régression. La motivation d'une stratégie de déformation est donc celle présentée ci-dessus pour l'Exemple 1 : construction d'estimateurs minimisant un contraste qui ne soit pas un contraste classique des moindre-carrés [6] , pour obtenir un estimateur explicite, sans inversion de matrice.
- Les fonctions d'intérêt dans les Exemples 4 et 5 ont par nature la forme d'un quotient. On se propose de bâtir, en utilisant une déformation des estimateurs qui ne sont pas construits comme quotient de deux estimateurs, pour simplifier la mise en place de procédures de sélection de fenêtres ou de modèles dans un second temps.

**Principaux résultats**

Les problèmes d'estimation présentés ci-dessus (Exemples 1 à 5) peuvent être résumés de la façon générale suivante : l'objectif est d'estimer une fonction $s$ à valeurs réelles, sur la base d'un $n-$échantillon de couples de variables aléatoires réelles $(X_i, Y_i)_{i\in\{1,\dots,n\}}$, où chaque $X_i$ admet une densité $f_X$ portée par un intervalle de $A$ de $\mathbb{R}$. La fonction $s$ peut être la fonction de régression (Ex.1), la fonction de volatilité (Ex.2), une fonction de répartition (Ex.3), le taux de hasard (Ex.4), et la densité conditionnelle (Ex.5).

Considérons les Exemples 1 à 4 pour l'instant (estimation de fonctions d'une variable réelle), qui font l'objet des Chapitres 3 et 4. La stratégie d'estimation que nous proposons étend celle présentée à la Section 1.2.1 pour l'Ex.1. Nous montrons que l'on peut, dans chacun des cas, exhiber une fonction de déformation $\Phi : A \to \Phi(A)$, bijective sous certaines hypothèses, et permettant de proposer une procédure d'estimation pour la cible $s$ à partir des données transformées $(\Phi(X_i), Y_i)_{i=1,\dots,n}$, au sens où l'on peut montrer

$$\mathbb{E}\left[\theta(Y_1)t\circ\Phi(X_1)\right] = \langle t, s\circ\Phi^{-1}\rangle, \quad t\in L^2(\Phi(A)), \tag{1.22}$$

pour $\theta(Y) := Y$ dans les Exemples 1, 3 et 4, et $\theta(Y) = Y^2$ dans l'Exemple 2. La fonction $\Phi$ est la fonction de répartition de $X$ dans les Exemples 1 à 3, et une primitive de la fonction de survie de $X$ dans l'Exemple 4 : $\Phi(x) = \int_0^x (1 - F_X(t))dt$. La relation (1.22) est prouvée sous la forme d'un énoncé un peu plus général (au sens où il ne se limite pas aux quatre exemples proposés) à la Proposition 4.1 du Chapitre 4.

L'égalité clé (1.22) permet de proposer une procédure d'estimation par projection ou à noyaux pour la fonction auxiliaire $g = s\circ\Phi^{-1}$ (de manière similaire à ce qui a été développé à la Remarque 1.2, ou à la Section 1.2.1 à partir de la relation (1.15)). Puis, à partir d'un estimateur $\hat{g}$ de $g$, la fonction estimant la cible $s$ est $\hat{s} = \hat{g}\circ\hat{\Phi}$, où $\hat{\Phi}$ est une contrepartie empirique pour $\Phi$. Des collections d'estimateurs pour la fonction cible $s$ sont donc déduites,

(6). au sens où le contraste ne fera pas intervenir une norme empirique, contrairement au contraste des moindres carrés (1.14), mais une "vraie" norme $L^2$ (voir la définition (3.3) au Chapitre 3 ou plus généralement la définition (4.37) du Chapitre 4 par exemple).

de la forme

$$\hat{s}_m(x) = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{\Phi}} \varphi_j \left( \hat{\Phi}(x) \right), \quad \text{avec} \quad \hat{a}_j^{\hat{\Phi}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_j \left( \hat{\Phi}(X_i) \right),$$
$$\hat{s}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h \left( \hat{\Phi}(x) - \hat{\Phi}(X_i) \right).$$

Les estimateurs par projection peuvent également être obtenus par minimisation d'un contraste. Cette construction est étendue à l'Exemple 5 (estimation de la densité conditionnelle) au Chapitre 5. Les performances théoriques et pratiques des estimateurs bâtis sont considérées.

Le risque quadratique intégré, pondéré par la dérivée $\phi = \Phi'$ de la fonction de déformation est étudié : nous obtenons une majoration semblable à une décomposition biais-variance, à termes de reste près (Proposition 3.1 ou 4.3), sous des hypothèses concernant la taille des collections des modèles ou des fenêtres :

$$\mathbb{E}\left[ \|\hat{s}_m - s\|_\phi^2 \right] \leq c \left( \|s - \Pi_{S_m}(g) \circ \Phi\|_\phi^2 + \frac{D_m}{n} + \frac{\ln(n)}{n} \right),$$
$$\mathbb{E}\left[ \|\hat{s}_h - s\|_\phi^2 \right] \leq c \left( \|s - (K_h \star g\mathbf{1}_{\Phi(A)}) \circ \Phi\|_\phi^2 + \frac{1}{nh} + \frac{\ln(n)}{n} \right),$$

pour une constante $c$ indépendante de $n$ (et de $m$ ou $h$) et où $\Pi_{S_m}(g)$ est la projection de $g$ sur un modèle $S_m$, dans les Exemples 1 à 4. De tels résultats nécessitent des décompositions du risque en de nombreux termes, dont les majorations respectives font appel à des astuces de calculs différentes. Le principal enjeu consiste à se ramener à contrôler l'écart entre la déformation $\Phi$ et son estimateur $\hat{\Phi}$ : la substitution de $\hat{\Phi}$ à $\Phi$, bien que naturelle, requiert des calculs non triviaux.

Nous mettons ensuite en place des procédures de sélection d'un modèle $\hat{m}$ ou d'une fenêtre $\hat{h}$, inspirées à la fois des travaux de Massart (2007) et Goldenshluger & Lepski (2011a) (voir Section 1.1.3). Les estimateurs obtenus sont entièrement déterminés par les données. Nous prouvons l'optimalité de ces méthodes au sens de l'oracle. Dans le cas des noyaux par exemple, on obtient, pour les Exemples 2 à 4,

$$\mathbb{E}\left[ \|\hat{s}_{\hat{h}} - s\|_\phi^2 \right] \leq c \min_{h \in \mathcal{H}_n} \left\{ \|s - (K_h \star g\mathbf{1}_{\Phi(A)}) \circ \Phi\|_\phi^2 + \frac{1}{nh} \right\} + c' \frac{\ln(n)}{n}.$$

Si les résultats semblent similaires dans les différents exemples, et si les démonstrations font appel aux mêmes outils (inégalités de concentration par exemple), chacune d'entre elles requiert cependant des calculs spécifiques. Ceci justifie que nous traitons séparément chaque cas : voir les Théorèmes 3.1 et 3.2 (pour l'Exemple 1, version projection, Chap. 3), Theorème 4.1 (pour les Exemples 1 à 4, version noyaux, Chap. 4), et Théorème 5.1 (Exemple 5, version projection, Chap. 5). Les vitesses classiques de convergence pour le risque sont également retrouvées. Les résultats sont exprimés pour des régularités concernant la fonction auxiliaire $g$ : appartenance à des boules d'espaces de Besov ou de Sobolev lorsque l'on s'intéresse aux estimateurs par projection, appartenance à des boules d'espaces de Hölder ou de Nikol'skiĭ

pour les noyaux (voir le Chapitre 2 pour une brève présentation de ces espaces). Dans le cas de l'Exemple 5, des régularités anisotropiques sont autorisées.

La comparaison des procédures d'adaptation (pénalisation ou méthode inspirée de Goldenshluger-Lepski - voir Section 1.1.3) est considérée : nous montrons l'équivalence de ces méthodes d'un point de vue théorique pour l'estimation d'une fonction d'une variable (Chapitre 3). Nous discutons au Chapitre 5 le choix de la méthode de Lepski adoptée pour estimer la densité conditionnelle (Section 5.6). Une autre procédure possible de choix de fenêtres est aussi discutée au Chapitre 4 (Section 4.9).

Enfin, d'un point de vue pratique, la déformation permet une grande simplicité quant à l'estimation : la procédure semble compétitive par rapport à d'autres méthodes existantes. Chaque résultat est illustré par des simulations (qualité visuelle de reconstruction et calculs de risques) : voir les Sections 3.4, 4.5 et 5.4.

Ces résultats font l'objet d'articles publiés ou soumis.
- Les résultats principaux du **Chap. 3** sont publiés dans la revue *Esaim P&S.*, Vol. **17**, p.328-358 (2013), sous le nom *Penalization versus Goldenshluger-Lepski strategies in regression estimation with warped bases* (accepté pour publication en novembre 2011).
- Les travaux du **Chap. 4** font l'objet d'un article *Adaptive warped kernel estimators*, soumis. `http://hal.archives-ouvertes.fr/hal-00715184` (version longue).
- Ceux du **Chap. 5** sont résumés dans l'article *Warped basis for conditional density estimation*, soumis pour publication, et actuellement en révision. `http://hal.archives-ouvertes.fr/hal-00641560/`.

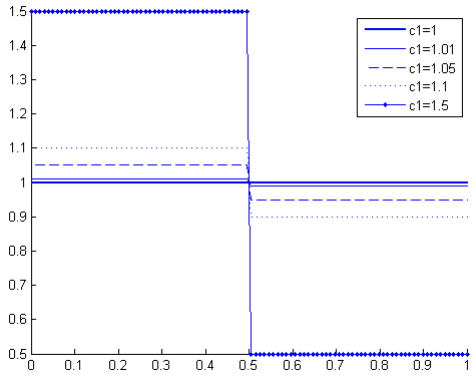
## 1.3 Partie B : Comparaison de la loi de deux échantillons, estimation de la densité relative

La Partie B de la thèse est motivée par la comparaison de deux échantillons de variables possiblement censurées : l'approche est fondée sur l'estimation de la densité relative.

### 1.3.1 Problématique

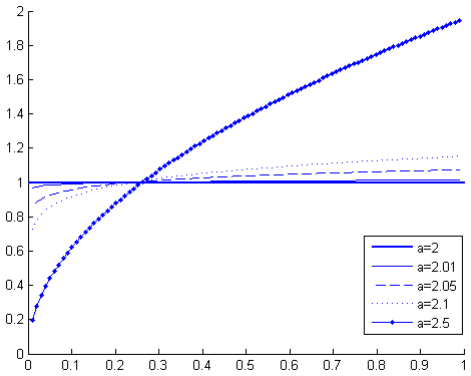
La comparaison d'un groupe d'individus à une population de référence au travers de la comparaison de deux fonctions de répartition $F$ et $F_0$ de variables réelles $X$ et $X_0$ est un objectif important en statistique et les applications dans des domaines variés comme la recherche médicale (comparaison d'un groupe de malades à un groupe sain) ou les sciences sociales sont nombreuses. Cette question rentre dans le cadre des problèmes dits *à deux échantillons* : les observations sont constituées d'échantillons indépendants de $X$ et $X_0$, notés $(X_i)_{i=1,\dots,n}$ et $(X_{0,i_0})_{i_0=1,\dots,n_0}$. Les tailles $n$ et $n_0$ ne sont pas nécessairement identiques. Les variables $X_i$ et/ou $X_{0,i_0}$ pourront dans un second temps être supposées censurées à droite.

Ecart à la loi uniforme $\mathcal{U}_{(0;1)}$




$X_0 \sim \mathcal{U}_{(0;1)}$
$X \sim f = c_1 \mathbf{1}_{(0;1/2)} + (2 - c_1)\mathbf{1}_{(1/2;1)}$

Ecart à la loi bêta $\mathcal{B}(2;5)$




$X_0 \sim \mathcal{B}(2;5)$
$X \sim \mathcal{B}(a;5)$

FIGURE 1.8 – Exemples de densités relatives $r$

Le but est de proposer une réponse au problème de comparaison par l'estimation non-paramétrique et adaptative d'un outil récent, la *densité relative de $X$ par rapport à $X_0$*. Ces deux variables sont supposées admettre des densités $f$ et $f_0$ respectivement, portées par des intervalles $A$ et $A_0$. On suppose que $f_0$ ne s'annule pas sur son support $A_0$. La densité relative $r$ est définie alors comme une densité de la variable $F_0(X)$, c'est-à-dire

$$r(x) = \frac{f \circ F_0^{-1}}{f_0 \circ F_0^{-1}}(x), \quad x \in F_0(A). \tag{1.23}$$

L'estimation de cette fonction est un problème relativement peu étudié, à notre connaissance. Les méthodes classiques permettant de répondre au problème initial posé, à savoir la comparaison de $F$ et $F_0$, sont des tests statistiques, comme le test de Kolmogorov-Smirnov (fondé sur la fonction de répartition empirique), ou ceux de Wilcoxon ou Mann-Whitney (fondés sur les rangs des variables des échantillons). Un autre outil classique est la courbe ROC, définie comme la courbe représentative de la répartition de la variable $1 - F_0(X)$, et connue pour capturer aussi le compromis entre la sensitivité et la spécificité d'un test. De façon très proche, un intérêt s'est peu à peu développé pour l'estimation de la fonction de répartition de la variable $F_0(X)$, appelée fonction de répartition relative (voir par exemple Handcock & Morris 1999 pour des références). La densité relative définie en (1.23) est la dérivée, quand elle existe, de cette fonction de répartition. Elle présente l'intérêt d'être un outil graphique relativement simple pour comparer $F$ à $F_0$ : si $F = F_0$, $r$ est la densité de la loi uniforme ; plus $F$ est éloignée de $F_0$, plus $r$ est différente de cette distribution. La Figure Figure 1.8 représente des exemples de densité relative, lorsque la loi de $X$ devient peu à peu différente de celle de $X_0$.

La fonction $r$ présente également l'avantage de donner des informations plus détaillées que la fonction de répartition relative ou la courbe ROC pour comparer deux tests diagnostics et

établir des règles de classification. On se réfèrera à Molanes-López & Cao (2008b) pour plus de détails.

Les estimateurs de la densité relative $r$ dans la littérature sont des estimateurs à noyaux, inspirés des méthodes d'estimation de la densité classique, et étudiés d'un point de vue asymptotique, en supposant que les tailles $n$ et $n_0$ des échantillons sont proportionnelles. Pour le cas de données complètes, nous renvoyons à Ćwik & Mielniczuk (1993), Handcock & Janssen (2002), et Molanes-López & Cao (2008a). Le cas d'observations censurées (au sens de l'Exemple 4 ci-dessus) est traité par Cao *et al.* (2000, 2001) ou encore Molanes-López & Cao (2008b). Le lecteur peut également se référer aux introductions des Chapitres 6 et 7, pour une bibliographie plus détaillée. Il n'existe pas de résultats adaptatifs dans ces cadres.

### 1.3.2 Principaux résultats

Nous nous proposons de bâtir des estimateurs par projection de la densité relative (1.23), vérifiant des propriétés d'adaptivité, dans le cas d'observations complètes (Chapitre 6) et dans le cas d'observations censurées à droite (Chapitre 7, travail en cours).

**Chapitre 6.** Nous supposons observer entièrement les deux échantillons $(X_i)_{i\in\{1,\dots,n\}}$ et $(X_{0,i_0})_{i_0\in\{1,\dots,n\}}$ et nous faisons l'hypothèse que les deux variables $X$ et $X_0$ ont même support. L'intervalle d'estimation est l'intervalle de définition de $r$, à savoir $F_0(A) = (0;1)$.

Nous modifions le contraste de densité (1.1) pour l'adapter au cadre des deux échantillons, en introduisant $\hat{F}_{0,n_0}$, la fonction de répartition empirique de l'échantillon $(X_{0,i_0})_{i_0}$. Le critère à minimiser s'écrit donc sous la forme

$$\gamma_n(t, \hat{F}_{0,n_0}) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t \circ \hat{F}_{0,n_0}(X_i), \quad t \in L^2((0;1)). \tag{1.24}$$

Sa construction précise est expliquée en Section 6.2.2. Pour chaque modèle trigonométrique $S_m$ ($m$ élément d'une collection $\mathcal{M}_{n,n_0}$), nous définissons l'estimateur par $\hat{r}_m(.,\hat{F}_{0,n_0}) = \arg\inf_{t\in S_m} \gamma_n(t, \hat{F}_{0,n_0})$.

Dans le cas "jouet" où la fonction $F_0$ est supposée connue, le risque quadratique intégré de $\hat{r}_m(.,F_0)$ est égal à la somme de deux termes, un terme de variance de l'ordre de $D_m/n$, et un terme de biais $\|\Pi_{S_m} r - r\|^2$. La difficulté ici provient de l'introduction de la contrepartie empirique $\hat{F}_{0,n_0}$ de $F_0$. Le résultat obtenu est alors de la forme suivante, pour une constante $C$,

$$\mathbb{E}\left[\left\|\hat{r}_m(.,\hat{F}_{0,n_0}) - r\right\|^2\right] \leq C\left(\left(\frac{D_m}{n} + \frac{D_m}{n_0}\right) + \|r_m - r\|^2 + \frac{1}{n} + \frac{1}{n_0}\right), \tag{1.25}$$

$r_m$ étant la projection de $r$ sur $S_m$. La vitesse que l'on en déduit (si $r$ est de régularité $\alpha$) est non classique, car dépendant des deux tailles d'échantillons :

$$\mathbb{E}[\|\hat{r}_m(.,\hat{F}_{0,n_0}) - r\|^2] = O\left(\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}\right).$$

Le modèle est ensuite sélectionné par la méthode présentée en Section 1.1.3, le terme de "pénalité" étant de l'ordre de $D_m/n + D_m/n_0$. Le résultat principal est une inégalité de type oracle énoncée au Théorème 6.1, sous des hypothèses très faibles (ne supposant pas en particulier de lien entre $n$ et $n_0$). Des simulations viennent illustrer la méthode (Section 6.4). Nous confirmons en particulier que les tailles des deux échantillons ne jouent pas des rôles symétriques : l'augmentation de $n_0$, le nombre d'observations de référence améliore la qualité de l'estimation de façon plus significative qu'une augmentation identique de $n$.

**Chapitre 7.** Au vu des applications potentielles dans lesquelles surviennent des problèmes à deux échantillons, il était naturel de chercher à adapter la méthode du Chapitre 6 au cas de données censurées à droite. Ce travail est en cours. Les observations sont alors constituées des deux échantillons indépendants suivants : un premier échantillon $(Z_{0,i_0}, \delta_{0,i_0})_{i_0=1,\dots,n_0}$ de variables aléatoires $Z_{0,i_0} = X_{0,i_0} \wedge C_{0,i_0}$, et $\delta_{0,i_0} = \mathbf{1}_{X_{0,i_0} \leq C_{0,i_0}}$, et un second échantillon $(Z_i, \delta_i)_{i=1,\dots,n_0}$ de variables aléatoires $Z_i = X_i \wedge C_i$, et $\delta_i = \mathbf{1}_{X_i \leq C_i}$. Rappelons que les variables $X_i$ et $X_{0,i_0}$ sont les variables d'intérêt, supposées ici à supports positifs $(0;\tau_X)$ et $(0;\tau_{X_0})$ respectivement. Les fonctions $f$ et $f_0$ (resp. $F$ et $F_0$) désignent toujours leurs densités. Les censures $C_i$ et $C_{0,i_0}$ ont pour supports respectifs $(0;\tau_C)$ et $(0;\tau_{C_0})$.

Les méthodes classiques d'analyse de survie (estimateur de Kaplan-Meier, correction de censure de type Koul *et al.* (1981)) nous conduisent d'abord à modifier le contraste (1.24). Nous nous plaçons sur $L^2((0;\tau))$ pour $\tau < \tau_X \wedge \tau_{X_0} \wedge \tau_C \wedge \tau_{C_0}$, et nous définissons :

$$\gamma_n^c(t, \check{F}_{0,n_0}, \check{\bar{G}}_n) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \frac{\delta_i}{\check{\bar{G}}_n(Z_i)} t \circ \check{F}_{0,n_0}(Z_i) \mathbf{1}_{Z_i \leq \tau},$$

où $\check{\bar{G}}_n$ (resp. $\check{F}_{0,n_0}$) est l'estimateur de Kaplan-Meier modifié de la survie $\bar{G} = 1 - G$ de la variables $C_i$ (resp. de la répartition $F_0$ des variables $X_{0,i_0}$). La définition de ces quantités est rappelée en Section 2.1.3. Nous obtenons par minimisation du contraste une collection d'estimateurs pour $r$, qui satisfont une borne identique à (1.25). La procédure de sélection conduit encore à la vitesse attendue (Théorème 7.1).

Nous proposons aussi une seconde méthode, fondée sur le principe de déformation (introduit à la Section 1.2.1 ci-dessus), permettant d'appéhender le cas où seul l'échantillon lié à la variable $X$ est censurée. Celle-ci repose sur l'estimation préalable de la fonction auxiliaire $g = r \circ \Phi^{-1}$, où $\Phi$ est la fonction de déformation suivante

$$\Phi(x) = \int_0^x \bar{G} \circ F_0^{-1}(u) du, \quad x \in (0;1), \tag{1.26}$$

dont l'introduction est expliquée en Section 7.3.1. Le contraste prend alors la forme suivante, dans le cas "jouet" où $\Phi$ est connue,

$$\gamma_n^{(cc)}(t, \Phi) = \|t\|^2_{L^2((0;B(\tau_Z)))} - \frac{2}{n}\sum_{i=1}^{n} \delta_i t \circ \Phi \circ F_0(Z_i).$$

Sa minimisation conduit à des estimateurs par projections pour $g$, dépendant d'une version empirique $\hat{\Phi}$ de $\Phi$, estimateurs que l'on déforme ensuite pour estimer $r$. Les performances théoriques sont étudiées (majoration non-asymptotique du risque, et sélection de modèles).

### 1.3.3 Liens entre les deux parties de la thèse

Hormis la dernière stratégie d'estimation de la densité relative dans le cas d'observations censurée (fondée sur la fonction (1.26)), les méthodes proposées ci-dessus pour reconstruire cette fonction ne relèvent pas à proprement parler d'une stratégie de déformation, au sens introduit à la Section 1.2.1 : nous ne définissons au Chapitre 6 ni fonction de déformation, ni fonction auxiliaire à estimer.

Cependant, nous pouvons remarquer que les résultats développés dans la Partie B relèvent des mêmes techniques et méthodes que celles mises en place dans la Partie A pour gérer la déformation. Le développement des estimateurs par projection dans la base orthonormée $(\varphi_j)_j$ font intervenir des quantités de type $\varphi_j(\hat{\Phi}(X_i))$, où $\hat{\Phi}$ est la contrepartie empirique de la fonction de déformation dans la Partie A, la fonction de répartition empirique de l'échantillon $(X_{0,i_0})_{i_0}$ au Chap. 6, l'estimateur de Kaplan-Meier de $F_0$ au Chap. 7.

Outre bien sûr les méthodes spécifiques à l'estimation adaptative via la sélection de modèles (estimation par projection, concentration de processus empiriques...), la majoration de telles quantités reposent sur l'introduction de la "vraie" fonction $\Phi$, et donc le contrôle d'écarts de la forme : $\varphi_j(\hat{\Phi}(X_i)) - \varphi_j(\Phi(X_i))$. Après développement de Taylor pour la fonction $\varphi_j$, l'enjeu est donc de majorer :

$$\sup_x \left| \hat{\Phi}(x) - \Phi(x) \right|. \tag{1.27}$$

Ceci fait principalement appel à deux ensembles de résultats.

- Dans certains cas, $\hat{\Phi}$ est égale à (ou tout au moins dépend de) la répartition empirique d'un échantillon de variables *i.i.d.* : les arguments sont alors basés sur des inégalités déduites de Dvoretzky *et al.* (1956) (voir Section 2.1.2).
- Dans les autres cas, $\hat{\Phi}$ est égale à (ou tout au moins dépend de) l'estimateur de Kaplan-Meier d'une fonction de survie de données censurées à droite : tout repose alors sur des inégalités analogues à celles de Dvoretzky *et al.* (1956), et démontrées par Bitouzé *et al.* (1999) (voir Section 2.1.3).

Par conséquent, bien que différentes de par la variété des cadres étudiés, les preuves détaillées dans ce travail présentent une certaine unité.

## 1.4 Perspectives de recherche

Les travaux présentés dans cette thèse tentent d'apporter un ensemble de contributions au problème de la construction d'estimateurs non-paramétriques satisfaisant des propriétés d'adaptation, tout en étant facilement implémentables et numériquement stables, dans des cadres statistiques variés. Nous prouvons, dans chacun des exemples étudiés, des propriétés théoriques non-asympotiques pour les méthodes mises en place, et nous montrons également que celles-ci sont pertinentes en pratique et font au moins jeu égal avec les procédures existantes. Les résultats obtenus nous permettent d'envisager différentes pistes de recherche.

La Partie A généralise la méthode dite de déformation pour bâtir des estimateurs adaptatifs à des cadres nouveaux (extension en analyse de survie notamment), nous démontrons des inégalités de type oracle pour des estimateurs fondés sur des données transformées. La déformation est étudiée pour l'instant pour transformer des variables réelles uniquement (ce qui n'a toutefois pas été un obstacle pour reconstruire une fonction de deux variables comme la densité conditionnelle, cf. Chap. 5). Il pourrait cependant être intéressant de considérer la question de la déformation d'un vecteur aléatoire de $\mathbb{R}^d$ ($d > 1$). Un exemple important pour comprendre les enjeux peut être à nouveau celui de la régression à design aléatoire : comment peut-on mettre en place une méthode analogue à celle étudiée lorsque la variable $X_i$ du modèle de régression appartient à $\mathbb{R}^d$ et non plus $\mathbb{R}$, la fonction $s$ à reconstruire étant alors une fonction de $d$ variables ? Ceci nécessiterait en particulier un contrôle non-asymptotique de l'écart entre la fonction de répartition, et son équivalent empirique en dimension supérieure à 1, analogue aux bornes de Dvoretzky *et al.* 1956, présentées à la Section 2.1.2.

Par ailleurs, l'étude du modèle de régression multiplicative (1.19) permet de penser que la méthode de déformation est robuste à un bruit hétéroscédastique. Cependant, des questions demeurent concernant l'estimation simultanée, par transformation de données, de $s$ et $\sigma^2$ dans un modèle de type $Y_i = s(X_i) + \sigma^2(X_i)\varepsilon_i$. Nous pourrions nous interroger également sur la robustesse de la méthode à la dépendance. Il serait par exemple intéressant de voir si nous pouvons l'utiliser dans un modèle autorégressif de la forme $X_{i+1} = s(X_i) + \varepsilon_{i+1}$, les observations n'étant plus constituées que des $X_i$.

La Partie B est centrée sur l'estimation adaptative d'une fonction récemment étudiée dans les problèmes à deux échantillons, la densité relative. Le Chapitre 6 propose une procédure optimale, au sens de l'oracle. Une vitesse de convergence du risque, impliquant les tailles des deux échantillons est exhibée (Corollaire 6.1). Nous ne pouvons pour l'instant nous permettre d'affirmer l'optimalité de cette vitesse au sens minimax. Une perspective immédiate consiste donc à étudier la borne inférieure pour le risque minimax associé sur des boules d'espaces de Besov. D'autre part, comme nous l'avons indiqué, le travail encore en cours sur l'estimation à partir de données censurées à droite (Chapitre 7) mérite d'être finalisé : il sera naturel de proposer une comparaison des deux estimateurs bâtis dans ce chapitre d'un point de vue pratique. Outre des simulations, et considérant l'étendue des applications possibles des problèmes à deux échantillons, le test de nos méthodes sur des jeux de données "réels" est un enjeu majeur. Ces mêmes applications font de l'estimation adaptative de la densité relative un problème riche qui amène de nombreuses autres pistes de recherche. En particulier, la prise en compte de covariables pour les variables d'intérêt $X$ et $X_0$ serait un pari intéressant : la question récurrente dans les problèmes médicaux de la comparaison d'un groupe d'individus sains à un groupe d'individus malades motive par exemple une problématique conditionnelle à des quantités comme l'âge. La première question est alors la définition même de l'objet sur lequel travailler : comment introduire fonction de répartition conditionnelle et/ou densité conditionnelle dans l'écriture de la densité relative ? Sur ce sujet, l'enjeu est donc tout autant la modélisation que l'estimation. Il pourrait être alors judicieux de commencer l'étude par

l'estimation adaptative d'une fonction de répartition relative conditionnelle ou courbe ROC relative conditionnelle.

Enfin, nous étudions et comparons dans cette thèse différentes possibilités pour sélectionner un estimateur dans une collection donnée : méthodes inspirées à la fois de la sélection de modèles développée par Barron *et al.* (1999), et des travaux de Goldenshluger & Lepski (2011a). L'enjeu méthodologique de cette comparaison nous semble important et pourrait être approfondi pour d'autres problèmes d'estimation. Nous avons par exemple récemment amorcé un travail, en collaboration avec Angelina Roche (I3M, Univ. Montpellier II) pour estimer de manière adaptative la fonction de répartition d'une variable réelle, conditionnellement à une covariable fonctionnelle, à l'aide d'estimateurs à noyaux définis par Ferraty *et al.* (2006) ou Ferraty *et al.* (2010). Le choix de la fenêtre, au coeur du problème, ne peut être mis en œuvre avec une application directe des travaux de Goldenshluger & Lepski (2011a), puisque le biais, ne s'exprimant pas comme convolée du noyau avec la fonction estimée, ne permet pas alors de définir des estimateurs auxiliaires similaires à ceux impliqués dans la méthode. Il est donc nécessaire de modifier la sélection pour prendre en compte la spécificité du cadre étudié.

# Chapitre 2

# Quelques résultats utiles de probabilités et d'analyse.

**Sommaire**

**Résumé.** L'objectif de ce chapitre est d'introduire rapidement quelques outils mathématiques, aussi bien des résultats de nature probabiliste que des résultats d'analyse, liés principalement à la théorie de l'approximation. Par souci de clarté, nous choisissons le plus souvent de présenter les résultats sous la forme précise où ils seront utilisés dans les chapitres qui suivent, quitte à perdre de la généralité, et à imposer des hypothèses parfois plus fortes que celles réellement nécessaires. Nous donnons parfois des éléments de démonstration des énoncés, et toujours des références précises.

## 2.1 Outils probabilistes

### 2.1.1 Inégalités de concentration

Les résultats principaux que nous obtenons dans cette thèse sont des inégalités de type oracle pour le risque des estimateurs, prouvant leurs propriétés d'adaptivité. Pour démontrer ces bornes non-asymptotiques, les arguments reposent sur la concentration de processus empiriques (ou de leurs suprema) autour de leur espérance, sous la forme de bornes en probabilités, ou d'inégalités intégrées. On utilisera particulièrement deux inégalités classiques en sélection de modèles : l'Inégalité de Bernstein et une forme intégrée de l'Inégalité de Talagrand. Ces outils ne sont bien sûr pas spécifiques des méthodes de déformation.

#### Inégalité de Bernstein

L'Inégalité de Bernstein exprime comment une moyenne de variables aléatoires indépendantes se concentre autour de sa moyenne. On énonce un résultat concernant uniquement les variables aléatoires bornées, ce qui est suffisant pour les applications que nous visons.

**Proposition 2.1.** *Soit $(Z_i)_{i=1,\dots,n}$ des variables aléatoires indépendantes. Supposons qu'il existe deux nombres positifs $v$ et $b$ tels que*

$$\sum_{i=1}^{n} \mathbb{E}\left[(Z_i)^2\right] \leq v \text{ et, } \forall i \in \{1,\dots,n\},\ Z_i \leq b.$$

*Alors, pour tout réel strictement positif $u$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq \sqrt{2\frac{v}{n}u} + \frac{b}{3}u\right) \leq \exp(-nu).$$

Cette version est prouvée par exemple dans Massart (2007) (voir la Proposition 2.9 p.24, et les commentaires qui la suivent p.25 pour des hypothèses identiques à celles imposées ci-dessus). Sa démonstration est fondée sur la méthode de Cramér-Chernoff.

### Inégalité de Talagrand, version intégrée

L'Inégalité de Bernstein permet de contrôler la concentration d'un processus empirique autour de son espérance, mais ne permet pas de borner les déviations de suprema de processus empiriques. On utilisera dans ce cas une des inégalités de Talagrand, qui peut être vue comme une extension de l'Inégalité de Bernstein à des fonctionnelles plus générales de variables aléatoires indépendantes. Notons que Goldenshluger & Lepski (2011b) évoquent à ce propos des *bornes uniformes en probabilité ou en espérance* (*uniform probability and moment upper bounds*), et démontrent des résultats très généraux pour des fonctionnelles sous-additives de somme de variables aléatoires indépendantes.

Le point de départ de la preuve de la forme intégrée de l'inégalité que nous utilisons est la version de Klein & Rio (2005), partant des résultats de Talagrand (1996).

**Théorème 2.1.** *Soient $\mathcal{S}$ un ensemble au plus dénombrable de fonctions mesurables définies sur $\mathcal{X}$ espace métrique polonais et à valeurs dans $[-1;1]^n$, $(\xi_i)_{i\in\{1,\dots,n\}}$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$. On suppose que pour tout vecteur $s=(s^1,\dots,s^n)\in\mathcal{S}$, pour tout $i\in\{1,\dots,n\}$, $\mathbb{E}[s^i(\xi_i)]=0$. On définit :*

$$Z=\sup_{s\in\mathcal{S}s=(s^1,\dots,s^n)}\sum_{i=1}^n s^i(\xi_i),\ \textit{et}\ V_n=\sup_{s\in\mathcal{S}}\ \textit{Var}\ \left(\sum_{i=1}^n s^i(\xi_i)\right).$$

*Alors, pour tout $\lambda>0$,*

$$\log\mathbb{E}\left[\exp(\lambda Z)\right]\leq\lambda\mathbb{E}(Z)+\frac{\lambda}{2}\left(2\mathbb{E}[Z]+V_n\right)\left\{\exp\left(\frac{e^{2\lambda}-1}{2}\right)-1\right\}.$$

*En particulier, en notant $\underline{v}=2\mathbb{E}[Z]+V_n$, on obtient, pour tout $x>0$,*

$$\begin{aligned}\mathbb{P}\left(Z\geq\mathbb{E}(Z)+x\right) &\leq \exp\left(-\frac{x}{4}\log\left(1+2\log\left(1+\frac{x}{\underline{v}}\right)\right)\right),\\ &\leq \exp\left(-\frac{x^2}{2\underline{v}+3x}\right).\end{aligned}$$

Des techniques fondées sur l'entropie permettent de prouver la première assertion du théorème. La seconde en découle immédiatement, en utilisant la méthode de Chernoff. Ceci entraîne le résultat suivant dont on donne une démonstration ci-dessous.

**Proposition 2.2.** *Soient $\xi_1,\dots,\xi_n$ des variables aléatoires indépendantes. Soit $\nu_n(r)=\frac{1}{n}\sum_{i=1}^n r(\xi_i)-\mathbb{E}[r(\xi_i)]$, pour $r$ appartenant à une classe dénombrable $\mathcal{R}$ de fonctions mesurables à valeurs réelles. Alors, pour $\delta>0$, il existe des constantes $c_l$, $l=1,2,3$, telles que*

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{r\in\mathcal{R}}\left(\nu_n\left(r\right)\right)^2-c(\delta)H^2\right)_+\right] &\leq c_1\left\{\frac{v}{n}\exp\left(-c_2\delta\frac{nH^2}{v}\right)\right.\\ &\quad\left.+\frac{M_1^2}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\frac{nH}{M_1}\right)\right\},\end{aligned}\tag{2.1}$$

*avec,* $C(\delta) = (\sqrt{1+\delta} - 1) \wedge 1$, $c(\delta) = 2(1+2\delta)$ *et*

$$\sup_{r\in\mathcal{R}} \|r\|_{L^\infty((0;1))} \leq M_1,\ \mathbb{E}\left[\sup_{r\in\mathcal{R}} |\nu_n(r)|\right] \leq H,\ et\ \sup_{r\in\mathcal{R}} Var\left(r\left(\xi_1\right)\right) \leq v.$$

**Remarque 2.1.** On a généralement besoin d'utiliser cette inégalité pour une classe $\mathcal{R}$ de fonctions non dénombrable. L'égalité suivante permet de le faire sous certaines conditions de densité : notons $(B, d)$ espace métrique séparable. Soient $A$ partie dénombrable dense de $B$ et $\nu : B \to \mathbb{R}$ une application continue. Alors,

$$\sup_{f\in B} \nu(f) = \sup_{f\in A} \nu(f). \tag{2.2}$$

En effet,

- d'une part, $A \subset B$, donc $\sup_{f\in A} \nu(f) \leq \sup_{f\in B} \nu(f)$,
- d'autre part, soit $f_B \in B$ et $\varepsilon > 0$. Par continuité de $\nu$, soit $\eta > 0$ tel que pour tout $f \in B$,
$$d\left(f, f_B\right) \leq \eta \ \Rightarrow\ |\nu(f) - \nu\left(f_B\right)| \leq \varepsilon.$$

De plus, par densité de $A$ dans $B$, soit $f_A \in A$ tel que $d(f_A, f_B) \leq \eta$. On a donc $|\nu\left(f_A\right) - \nu\left(f_B\right)| \leq \varepsilon$. En particulier,

$$\nu\left(f_B\right) \leq \nu\left(f_A\right) + \varepsilon \leq \sup_{f\in A} \nu(f) + \varepsilon,$$

et ainsi $\sup_{f\in B} \nu(f) \leq \sup_{f\in A} \nu(f) + \varepsilon$. Comme ceci est valable quel que soit $\varepsilon > 0$, l'égalité (2.2) est prouvée.

**Preuve de la Proposition 2.2.** Posons $\mathcal{S} = \{s_r = (s_r^1, \ldots, s_r^n),\ r \in \mathcal{R}\}$, où, pour $r \in \mathcal{R}$,

$$s_r^i : x \mapsto s_r^i(x) = \frac{r}{2M_1}(x) - \mathbb{E}\left[\frac{r}{2M_1}(\xi_i)\right].$$

On a, pour $r \in \mathcal{R}$ et $i \in \{1, \ldots, n\}$, $\mathbb{E}[s_r^i(\xi_i)] = 0$, $s_r^i$ à valeurs dans $[-1; 1]$ (par définition de $M_1$) et $Z$ se réécrit $Z = \sup_{r\in\mathcal{R}}(n/2M_1)\nu_n(r)$.
Par le Théorème 2.1, quel que soit $x > 0$,

$$\mathbb{P}\left(\frac{n}{2M_1}\sup_{r\in\mathcal{R}} \nu_n(r) \geq \frac{n}{2M_1}\mathbb{E}\left[\sup_{r\in\mathcal{R}} \nu_n(r)\right] + x\right) \leq \exp\left(-\frac{x^2}{2\underline{v} + 3x}\right),$$

c'est à dire, pour tout $y > 0$, en prenant $x = ny/2M_1$, et en utilisant la définition de $H$,

$$\begin{aligned}\mathbb{P}\left(\sup_{r\in\mathcal{R}} \nu_n(r) \geq H + y\right) &\leq \mathbb{P}\left(\sup_{r\in\mathcal{R}} \nu_n(r) \geq \mathbb{E}\left[\sup_{r\in\mathcal{R}} \nu_n(r)\right] + y\right), \\ &\leq \exp\left(-\frac{n^2y^2}{8M_1^2\underline{v} + 6M_1ny}\right).\end{aligned}$$

On majore ensuite $\underline{v} = 2\mathbb{E}[Z] + V_n$ : d'une part, $\mathbb{E}(Z) \leq (nH)/(2M_1)$, d'autre part,

$$\begin{aligned} V_n &= \sup_{r\in\mathcal{R}} \text{Var} \left( \sum_{i=1}^{n} \frac{r(\xi_i)}{2M_1} \right), \\ &= \frac{n}{4M_1^2} \sup_{r\in\mathcal{R}} \frac{1}{n} \text{Var} \left( \sum_{i=1}^{n} r(\xi_i) \right), \\ &\leq \frac{nv}{4M_1^2}. \end{aligned}$$

On a donc $\underline{v} \leq nH/M_1 + vn/4M_1^2$. Il vient ainsi,

$$\mathbb{P}\left( \sup_{r\in\mathcal{R}} \nu_n(r) \geq H + y \right) \leq \exp\left( -\frac{ny^2}{2(4M_1H + v) + 6M_1 y} \right).$$

On utilise ensuite

$$\begin{aligned} \mathbb{P}\left( \sup_{r\in\mathcal{R}} |\nu_n(r)| \geq H + y \right) &\leq \mathbb{P}\left( \sup_{r\in\mathcal{R}} \nu_n(r) \geq H + y \right) + \mathbb{P}\left( \sup_{r\in\mathcal{R}} -\nu_n(r) \geq H + y \right), \\ &= \mathbb{P}\left( \sup_{r\in\mathcal{R}} \nu_n(r) \geq H + y \right) + \mathbb{P}\left( \sup_{r\in\mathcal{R}} \nu_n(-r) \geq H + y \right), \\ &\leq 2 \times \exp\left( -\frac{ny^2}{2(4M_1H + v) + 6M_1 y} \right). \end{aligned}$$

On applique ceci avec $y = \lambda + \eta H$. Suivons maintenant la démarche observée par Birgé & Massart (1998) (Corollaire 2 p.354).

$$\begin{aligned} \frac{y^2}{2(v + 4M_1H) + 6M_1 y} &= \frac{\lambda^2 + \eta^2 H^2 + 2n\lambda H}{2v + 8HM_1 + 6M_1\lambda + 6M_1\eta H}, \\ &\geq \frac{\lambda^2 + 2n\lambda H}{2v + 6M_1\lambda + M_1 H(8 + 6\eta)} := \frac{a+b}{c+d+e}. \end{aligned}$$

Or quels que soient $a, b, c, d, e > 0$,

$$\frac{a+b}{c+d+e} \geq \frac{1}{3}\left( \frac{a}{c} \wedge \frac{a}{d} \wedge \frac{b}{e} \right),$$

puisque,

$$\begin{aligned} \frac{a+b}{c+d+e} &\geq \frac{a+b}{3(c \vee d \vee e)}, \\ &= \frac{1}{3}\left( \frac{a+b}{c} \wedge \frac{a+b}{d} \wedge \frac{a+b}{e} \right), \\ &\geq \frac{1}{3}\left( \frac{a}{c} \wedge \frac{a}{d} \wedge \frac{b}{e} \right). \end{aligned}$$

Ainsi,

$$\frac{y^2}{2(v + 4M_1H) + 6M_1 y} \geq \frac{1}{6}\left[ \frac{\lambda^2}{v} \wedge \frac{2\lambda}{M_1}\left( \frac{1}{6} \wedge \frac{\eta}{4 + 3\eta} \right) \right].$$

Ensuite, on a aussi

$$\frac{1}{6} \wedge \frac{\eta}{4+3\eta} \geq \frac{\eta \wedge 1}{7}.$$

En effet, on a d'une part $1/6 \geq (\eta \wedge 1)/7$, et d'autre part,

$$\frac{\eta}{4+3\eta} - \frac{\eta \wedge 1}{7} = \begin{cases} \dfrac{3(1-\eta)}{7(4+3\eta)} \geq 0 \text{ pour } \eta \leq 1, \\ \\ \dfrac{4(\eta-1)}{7(4+3\eta)} \geq 0 \text{ pour } \eta \geq 1. \end{cases}$$

Il vient donc

$$\frac{y^2}{2(v+4M_1H)+6M_1y} \geq \frac{1}{6}\left[\frac{\lambda^2}{v} \wedge \frac{2(\eta\wedge 1)}{7}\frac{\lambda}{M_1}\right],$$

et par suite,

$$\mathbb{P}\left(\sup_{r\in\mathcal{R}}|\nu_n(r)| \geq \lambda + (\eta+1)H\right) \leq 2\exp\left(-\frac{n}{6}\left\{\frac{\lambda^2}{v} \wedge \frac{2(\eta\wedge 1)}{7}\frac{\lambda}{M_1}\right\}\right). \tag{2.3}$$

Il reste maintenant à intégrer cette inégalité. Notons $X = (\sup_{r\in\mathcal{R}}(\nu_n(r))^2 - 2(1+2\varepsilon)H^2)_+$. On commence par

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty \mathbb{P}\left(\sup_{r\in\mathcal{R}}(\nu_n(r))^2 \geq 2(1+2\varepsilon)H^2 + t\right)dt, \\ &= \int_0^\infty \mathbb{P}\left(\sup_{r\in\mathcal{R}}|\nu_n(r)| \geq \sqrt{2(1+\varepsilon)H^2 + 2(\varepsilon H^2 + t/2)}\right)dt, \\ &\leq \int_0^\infty \mathbb{P}\left(\sup_{r\in\mathcal{R}}|\nu_n(r)| \geq \sqrt{1+\varepsilon}H + \sqrt{\varepsilon H^2 + t/2}\right)dt. \end{aligned}$$

En appliquant (2.3) avec $\eta = \sqrt{1+\varepsilon} - 1$ et $\lambda = \sqrt{\varepsilon H^2 + t/2}$, on en déduit, en notant $K_1 = 1/6$ pour simplifier,

$$\begin{aligned} \mathbb{E}[X] &\leq \int_0^\infty 2\exp\left(-K_1 n\left\{\frac{\varepsilon H^2 + t/2}{v} \wedge \frac{2(\eta\wedge 1)}{7}\frac{\sqrt{\varepsilon H^2 + t/2}}{M_1}\right\}\right)dt, \\ &\leq 2\int_0^\infty \exp\left(-K_1 n\left\{\frac{\varepsilon H^2 + t/2}{v}\right\}\right)dt \\ &\quad +2\int_0^\infty \exp\left(\left\{-K_1 n\frac{2(\eta\wedge 1)}{7}\frac{\sqrt{\varepsilon H^2 + t/2}}{M_1}\right\}\right)dt. \end{aligned}$$

On utilise dans la seconde intégrale $\sqrt{\varepsilon H^2 + t/2} \geq (\sqrt{\varepsilon}H + \sqrt{t/2})/\sqrt{2}$ pour en déduire :

$$\begin{aligned} \mathbb{E}[X] &\leq 2\int_0^\infty \exp\left(-\frac{K_1 n}{v}(\varepsilon H^2 + t/2)\right)dt \\ &\quad +2\int_0^\infty \exp\left(-\frac{2K_1 n(\eta\wedge 1)}{7M_1\sqrt{2}}(\sqrt{\varepsilon}H + \sqrt{t/2})\right)dt, \\ &= 2\exp\left(-\frac{K_1 n}{v}\varepsilon H^2\right)\int_0^\infty \exp\left(-\frac{K_1 n}{2v}t\right)dt \\ &\quad +2\exp\left(-\frac{\sqrt{2}K_1 n(\eta\wedge 1)}{7M_1}(\sqrt{\varepsilon}H)\right)\int_0^\infty \exp\left(-\frac{K_1 n(\eta\wedge 1)}{7M_1}\sqrt{t}\right)dt, \end{aligned}$$

d'où, en calculant les intégrales,

$$\begin{aligned}\mathbb{E}[X] &\leq 2\left\{\exp\left(-\frac{K_1 n}{v}\varepsilon H^2\right)\frac{2v}{K_1 n} + \exp\left(-\frac{\sqrt{2}K_1 n(\eta\wedge 1)}{7M_1}(\sqrt{\varepsilon}H)\right)\frac{2\times 49 M_1^2}{K_1^2 n^2(\eta\wedge 1)^2}\right\},\\ &= \frac{4}{K_1}\left\{\frac{v}{n}\exp\left(-K_1\varepsilon\frac{nH^2}{v}\right) + \frac{49}{K_1(\eta\wedge 1)^2}\frac{M_1^2}{n^2}\exp\left(-\frac{\sqrt{2}K_1(\eta\wedge 1)\sqrt{\varepsilon}}{7}\frac{nH}{M_1}\right)\right\},\end{aligned}$$

ce qui est l'inégalité cherchée.

□

Une preuve analogue pourra être trouvée dans Lacour (2008) (Lemma 5 p.812) par exemple, avec des constantes légèrement différentes.

### 2.1.2 Quelques propriétés de la fonction de répartitition empirique d'un échantillon de variables aléatoires réelles

Les méthodes de déformation que nous nous proposons d'étudier sont fondées sur l'utilisation de données transformées. La transformation est parfois égale à la fonction de répartition des observations, ou met tout au moins celle-ci en jeu. Nous sommes donc amenés à utiliser son équivalent empirique, et quelques propriétés de celui-ci. Pour une revue plus générale concernant la fonction de répartition empirique et d'autres estimateurs possible de la fonction de répartition d'un échantillon de variables réelles, on pourra se référer au travail de Servien (2009).

#### Notations et définitions

Considérons $(X_1,\dots,X_n)$ un échantillon d'une variable aléatoire réelle $X$, de densité $f_X$, et de fonction de répartition $F_X$, supposée inversible. On note $\hat{F}_n$ la fonction de répartition empirique associée, définie par

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i\leq x},\quad x\in\mathbb{R}. \tag{2.4}$$

On note également $U_i = F_X(X_i)$, $\hat{U}_n$ la fonction de répartition empirique associée à l'échantillon $(U_i)_{i=1,\dots,n}$, et $id$ la fonction $u\mapsto u$. Rappelons que $U_i$ suit la loi uniforme sur l'intervalle $(0;1)$, et que la variable aléatoire $\|\hat{U}_n - id\|_{L^\infty((0;1))}$ a même loi que $\|\hat{F}_n - F_X\|_{L^\infty(\mathbb{R})}$.

#### Inégalités exponentielles pour la fonction de répartition empirique

Le contrôle en probabilité et en espérance de la norme $\|\hat{F}_n - F_X\|_{L^\infty(\mathbb{R})}$ est l'une des clés des majorations de risques des estimateurs par projection développés en base déformée ou des estimateurs à noyaux déformés. L'Inégalité de Dvoretzky-Kiefer-Wolfowitz (Dvoretzky *et al.*, 1956) sera donc très fréquemment utilisée, sous l'une des formes suivantes.

**Proposition 2.3.** *(Inégalité Dvoretzky-Kiefer-Wolfowitz) Il existe une constante $C > 0$, telle que, pour tout entier $n \geq 1$ et tout réel $\lambda > 0$,*

$$\mathbb{P}\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))} \geq \lambda\right) \leq C\exp\left(-2n\lambda^2\right).$$

Massart (1990) a montré que la constante $C$ de l'inégalité ci-dessus peut-être choisie égale à 2. Par intégration, on en déduit les bornes suivantes.

**Proposition 2.4.** *Pour tout entier $p > 0$, il existe une constante $C_p > 0$ telle que*

$$\mathbb{E}\left[\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))}\right] \leq \frac{C_p}{n^{p/2}}.$$

**Preuve de la Proposition 2.4.** On écrit l'espérance sous forme d'intégrale, puis on applique la Proposition 2.3 :

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))}\right] &= p\int_0^\infty x^{p-1}\mathbb{P}\left(\sup_{u\in(0;1)}\left|\hat{U}_n(u) - u\right| \geq x\right)dx, \\
&\leq cp\int_0^\infty x^{p-1}\exp(-2nx^2)dx = \frac{cp}{2(2n)^{p/2}}\int_0^\infty u^{\frac{p}{2}-1}\exp(-u)du, \\
&= \frac{cp\Gamma(p/2)}{2(2n)^{p/2}} = \frac{C_p}{n^{p/2}}.
\end{aligned}$$

□

**Corollaire 2.1.** *Pour tout $\kappa > 0$, pour tout entier $p \geq 2$, il existe une constante $C$ telle que*

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))} - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right] \leq Cn^{-2^{\frac{2-p}{p}}\kappa^{2/p}}. \tag{2.5}$$

*De plus,*

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))} - \kappa\frac{\ln(n)}{n}\right)^2\right] \leq Cn^{-2-2\kappa}. \tag{2.6}$$

L'Inégalité (2.6) est une version légèrement plus précise de l'Inegalité (2.5) dans le cas $p = 2$.

**Preuve du Corollaire 2.1.** Commençons par l'Inégalité (2.5) :

$$\begin{aligned}
&\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))} - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right] \\
&\quad= \int_0^\infty \mathbb{P}\left(\left(\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))} - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+ > t\right)dt, \\
&\quad\leq \int_0^\infty \mathbb{P}\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))} > \left(t + \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)^{1/p}\right)dt,
\end{aligned}$$

Par concavité de la fonction $u : x \mapsto x^{1/p}$, l'inégalité suivante est vraie : $(a+b)^{1/p} \geq 2^{1/p-1}(a^{1/p} + b^{1/p})$, pour $a, b > 0$. Par conséquent, en utilisant ceci, puis la Proposition 2.4,

$$
\begin{aligned}
&\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^p_{L^\infty((0;1))} - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right] \\
&\quad\leq \int_0^\infty \mathbb{P}\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))} > 2^{(1-p)/p}\left(t^{1/p} + \frac{\kappa^{1/p}\ln^{1/2}(n)}{n^{1/2}}\right)\right) dt \\
&\quad\leq C\int_0^\infty \exp\left\{-2n\times 2^{(2-2p)/p}\left(t^{1/p} + \frac{\kappa^{1/p}\ln^{1/2}(n)}{n^{1/2}}\right)^2\right\} dt, \\
&\quad\leq C\exp\left\{-2^{(2-p)/p}\kappa^{2/p}\ln(n)\right\}\int_0^\infty \exp\left\{-2^{(2-p)/p}nt^{2/p}\right\} dt, \qquad (2.7) \\
&\quad= Cn^{-2^{\frac{2-p}{p}}\kappa^{2/p}}\int_0^\infty \exp\left\{-2^{(2-p)/p}nt^{2/p}\right\} dt,
\end{aligned}
$$

où l'on a utilisé pour l'inégalité (2.7) que $(a+b)^2 \geq a^2 + b^2$, pour $a, b > 0$. Comme l'intégrale restante peut grossièrement être majorée par 1, cela conduit au résultat cherché. Pour l'Inégalité (2.6), la majoration est similaire :

$$
\begin{aligned}
&\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))} - \kappa\frac{\ln(n)}{n}\right)^2\right] \\
&\quad\leq \int_0^\infty 2t\mathbb{P}\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))} > \sqrt{\kappa\frac{\ln(n)}{n} + t}\right) dt, \\
&\quad\leq C\int_0^\infty 2t\exp\left(-2n\left(\kappa\frac{\ln(n)}{n} + t\right)\right) dt, \\
&\quad= (C/2)n^{-2-2\kappa}\int_0^\infty u\exp(-u)du = (C/2)n^{-2-2\kappa}.
\end{aligned}
$$

□

**Remarque 2.2.** La borne supérieure de l'Inégalité (2.5) peut-être légèrement améliorée en $Cn^{-2^{\frac{2-p}{p}}\kappa^{2/p}-p/2}$, ce que l'on obtient lorsque l'on calcule l'intégrale intervenant dans (2.7) (au lieu de la majorer par 1) :

$$
\int_0^\infty \exp\left\{-2^{(2-p)/p}nt^{2/p}\right\} dt = \frac{p}{2\times 2^{(2-p)/2}}\int_0^\infty u^{p/2-1}\exp(-u)du\, n^{-p/2}.
$$

### 2.1.3 Quelques propriétés de l'estimateur de Kaplan-Meier de la fonction de survie d'un échantillon de variables aléatoires réelles censurées à droite

Les applications que nous visons de la méthode de déformation nous amènent à considérer également le cas de données incomplètes, car censurées de manière aléatoire à droite. Dans ce cas, la fonction de répartition empirique des variables d'intérêt ne peut-être considérée comme un estimateur et on a alors recours à l'estimateur de Kaplan-Meier de la fonction de survie, pour lequel on peut démontrer des inégalités exponentielles analogues.

**Notations et définitions**

Considérons toujours une variable aléatoire réelle d'intérêt $X$, de fonction de répartition $F_X$, de densité $f_X$, à support positif $(0;\tau_X)$ ($\tau_X \in (0;\infty]$). Soit également une autre variable positive $C$ (de répartition $G$ et de densité $f_C$ portée par $(0;\tau_C)$), indépendante de $X$, et posons $Z = X \wedge C$. Soit enfin $\delta = \mathbf{1}_{X\leq C}$ l'indicateur de non-censure associé (on pourra se référer en particulier au Chapitre 4, Section 4.2 pour plus de détails sur ce modèle, ou encore à l'Introduction, Section 1.2.2).

Supposons que les observations sont constituées d'un $n$-échantillon de paires $(Z_i, \delta_i)$ distribuées selon la loi de $(Z, \delta)$. La fonction $\hat{F}_n$, définie par (2.4) ne peut être utilisée comme estimateur la répartition $F_X$, puisque sa définition implique des données non observées (les $X_i$ censurés ne sont pas disponibles). De plus, garder uniquement les variables non-censurées conduit à une procédure biaisée. L'estimateur standard utilisé dans ce cadre pour $F_X$ (et aussi pour $F_C$) est celui de Kaplan & Meier (1958). Sa construction s'appuie sur la statistique d'ordre associée aux variables $Z_i$, $i = 1,\dots,n$, notée $(Z_{(i)})_{i=1,\dots,n}$ et définie de telle sorte que l'on a $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ et $\{Z_{(1)},\dots,Z_{(n)}\} = \{Z_1,\dots,Z_n\}$. On note aussi $\delta_{(i)}$ la statistique d'ordre induit par celle des $Z_{(i)}$. La formule suivante définit l'estimateur de Kaplan-Meier de la fonction de survie $\bar{F}_X = 1 - F_X$.

$$\breve{\bar{F}}_n^{(1)}(x) = \begin{cases} \displaystyle\prod_{\substack{i=1\\ Z_{(i)}\leq x}}^{n} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}}, \text{ si } x \leq Z_{(n)}, \\ 0, \text{ sinon.} \end{cases}$$

Pour comprendre sa construction, on pourra se référer au Chapitre 14 de Rivoirard & Stoltz (2012) ou à l'introduction de Bouaziz (2009). Une version différente présentant l'avantage de ne pas s'annuler sera préférable dans notre cadre : la modification est donnée par Lo *et al.* (1989).

$$\breve{\bar{F}}_n(x) = \begin{cases} \displaystyle\prod_{\substack{i=1\\ Z_{(i)}\leq x}}^{n} \left(\frac{n-i+1}{n-i+2}\right)^{\delta_{(i)}}, \text{ si } x \leq Z_{(n)}, \\ \breve{\bar{F}}_n(Z_{(n)}), \text{ sinon.} \end{cases} \tag{2.8}$$

L'avantage de ce dernier estimateur par rapport au précédent réside dans la minoration : $\breve{\bar{F}}_n(x) \geq (n+1)^{-1}$ quel que soit $x$. On a aussi $\sup_{0\leq x\leq\tau} |\breve{\bar{F}}_n - \breve{\bar{F}}_n^{(1)}| = O(n^{-1})$ p.s. pour $\tau < \tau_C \wedge \tau_X$.

L'estimateur de Kaplan-Meier (resp. modifié) de la répartition de $X$ est $\breve{F}_n^{(1)} = 1 - \breve{\bar{F}}_n^{(1)}$ (resp. $\breve{F}_n = 1 - \breve{\bar{F}}_n$). Le problème étant presque symétrique en $X$ et en $C$, on peut définir de façon analogue un estimateur $\breve{G}_n^{(1)}$ (ou $\breve{G}_n$) pour la répartition $G$ de la censure en remplaçant simplement $\delta_i$ par $1 - \delta_i$.

**Expression sous forme de somme de l'estimateur de Kaplan-Meier**

L'estimateur de Kaplan-Meier est une fonction constante par morceaux, continue à droite et limitée à gauche, comme la fonction de répartition empirique. Les sauts ont lieu uniquement aux instants de décès. On peut donc écrire l'estimateur sous forme d'une somme de la forme

$$\breve{\breve{F}}_n^{(1)}(x) = \sum_{i=1}^{n} W_{i,n} \mathbf{1}_{Z_i \leq x}, \quad x \in \mathbb{R},$$

avec des poids $W_{i,n}$ nuls quand $\delta_i = 0$. Précisément, les poids peuvent être calculés, et l'on obtient l'expression suivante prouvée par exemple par le Lemme 1.1 de Bouaziz (2009), reprenant les résultats des travaux de Satten & Datta (2001) :

$$\breve{\breve{F}}_n^{(1)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\breve{\breve{G}}_n^{(1)}(X_i-)} \mathbf{1}_{X_i \leq x}. \tag{2.9}$$

Comme $\breve{\breve{G}}_n^{(1)}$ est continue à gauche, on a $\breve{\breve{G}}_n^{(1)}(X_i-) = \breve{\breve{G}}_n^{(1)}(X_i)$. La preuve de l'égalité (2.9) peut être adaptée à l'estimateur de Kaplan-Meier modifié. On obtient ainsi le Lemme suivant :

**Lemme 2.1.** *L'estimateur de Kaplan-Meier modifié peut s'écrire*

$$\breve{\breve{F}}_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \frac{\delta_i}{\breve{\breve{G}}_n(X_i)} \mathbf{1}_{X_i \leq x}, \quad x \in \mathbb{R}. \tag{2.10}$$

L'expression sous forme de somme est plus facilement manipulable que la définition sous forme de produit.

**Preuve du Lemme 2.1.** On suit le schéma de preuve donné par Bouaziz (2009) que l'on adapte à l'estimateur de Kaplan-Meier modifié. On calcule la valeur du saut à la $i-$ème observation $Z_{(i)}$. Celle-ci vaut, par définition, $\breve{\breve{F}}_n(Z_{(i)}) - \breve{\breve{F}}_n(Z_{(i-1)})$. En utilisant la définition (2.8),

$$\begin{aligned}
\breve{F}_n(Z_{(i)}) - \breve{F}_n(Z_{(i-1)}) &= \prod_{j=1}^{i-1} \left( \frac{n-j+1}{n-j+2} \right)^{\delta_{(j)}} - \prod_{j=1}^{i-1} \left( \frac{n-j+1}{n-j+2} \right)^{\delta_{(j)}}, \\
&= \prod_{j=1}^{i-1} \left( \frac{n-j+1}{n-j+2} \right)^{\delta_{(j)}} \left( 1 - \left( \frac{n-i+1}{n-i+2} \right)^{\delta_{(i)}} \right),
\end{aligned}$$

or,

$$1 - \left( \frac{n-i+1}{n-i+2} \right)^{\delta_{(i)}} = \begin{cases} 0 & \text{si } \delta_{(i)} = 0, \\ \dfrac{1}{n-i+2} & \text{sinon .} \end{cases}$$

On obtient donc

$$\breve{F}_n(Z_{(i)}) - \breve{F}_n(Z_{(i-1)}) = \prod_{j=1}^{i-1} \left( \frac{n-j+1}{n-j+2} \right)^{\delta_{(j)}} \frac{\delta_{(i)}}{n-i+2}. \tag{2.11}$$

D'autre part, $1-\check{G}_n(Z_{(i)}) = \prod_{j=1}^{i-1}((n-j+1)/(n-j+2))^{1-\delta_{(j)}}$, et on peut calculer

$$\begin{aligned}\prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j+2}\right)^{\delta_{(j)}} \times \left(1-\check{G}_n(Z_{(i)})\right) &= \prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j+2}\right)^{\delta_{(j)}} \times \prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j+2}\right)^{1-\delta_{(j)}}, \\ &= \prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j+2}\right) = \frac{n-i+2}{n+1}.\end{aligned}$$

En insérant ceci dans (2.11), on obtient le résultat cherché.

□

**Inégalités exponentielles pour l'estimateur de Kaplan-Meier**

On va utiliser des propriétés de déviations du Kaplan-Meier, analogue à l'Inégalité de Dvoretzky-Kiefer-Wolfowitz pour la répartition empirique (voir Proposition 2.3 ci-dessus). Rappelons tout d'abord le résultat obtenu par Bitouzé *et al.* (1999) (découlant du Théorème 1 p.737).

**Proposition 2.5.** *L'estimateur de Kaplan-Meier de $F_X$, construit à partir d'un échantillon $(Z_i,\delta_i)_{i\leq n}$, vérifie l'inégalité suivante, pour tout $\lambda>0$*

$$\forall x>0,\ \ \mathbb{P}\left(\left\|\bar{G}\left(\check{F}_n^{(1)}-F_X\right)\right\|_{L^\infty(\mathbb{R})}\geq x\right)\leq 2.5\exp\left(-2nx^2+Cx\sqrt{n}\right).$$

*pour une constante $C>0$.*

Notons que Wellner (2007) a ensuite proposé des bornes pour la constante $C$ de la proposition. On en déduit la forme intégrée suivante, dont la preuve est analogue à la preuve de la Proposition 2.4 (voir aussi Brunel & Comte 2006, Lemme 3.1) :

**Proposition 2.6.** *Quel que soit $p$ entier naturel non nul, il existe une constante $\bar{c}_p$, telle que*

$$\mathbb{E}\left[\left\|\bar{G}\left(\check{F}_n^{(1)}-F\right)\right\|_{L^\infty(\mathbb{R})}^p\right]\leq \bar{c}_p n^{-p/2}.$$

Ces résultats sont également valables pour l'estimateur de Kaplan-Meier modifié $\check{F}_n$ (se référer à l'égalité (3.2) et à la preuve du Lemme 3.1 de Brunel & Comte 2006).

Enfin, l'inégalité suivante est un outil important pour les preuves des résultats adaptatifs, similaire au Corollaire 2.1.

**Corollaire 2.2.** *Quel que soit $p$ entier naturel non nul, il existe une constante $\tilde{c}_p$, telle que*

$$\mathbb{E}\left[\left(\left\|\bar{G}\left(\check{F}_n-F\right)\right\|_{L^\infty(\mathbb{R})}^p-\kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right]\leq \tilde{c}_p n^{-2^{(2-2p)/p}\kappa^{2/p}}.$$

### 2.1.4 Autre résultat utile

On rappelle enfin l'inégalité suivante, dûe à Rosenthal (1970), et permettant de contrôler les moments d'ordre $k > 2$ d'une sommes de variables aléatoires indépendantes.

**Proposition 2.7.** *(Inégalité de Rosenthal) Soit $(\xi_i)_{i=1,\dots,n_0}$ des variables aléatoires rélles indépendantes, centrées, admettant un moment d'ordre $k > 2$. Alors, il existe une constante $\bar{\bar{c}}_k$, telle que*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n_0}\xi_i\right|^k\right] \leq \bar{\bar{c}}_k \left\{\sum_{i=1}^{n_0}\mathbb{E}\left[|\xi_i|^k\right] + \left(\sum_{i=1}^{n_0}\mathbb{E}\left[\xi_i^2\right]\right)^{k/2}\right\}.$$

Une démonstration simple de ce résultat est donnée par Härdle *et al.* (1998) (Appendix C, Theorem C.2 p.244). Pour une preuve aboutissant à des constantes exactes, on pourra se référer aux travaux de Ibragimov & Sharakhmetov (2001) par exemple.

## 2.2 Outils d'Analyse

L'objectif de cette thèse est de bâtir des estimateurs s'adaptant à la régularité (inconnue) de la fonction à reconstruire. Cette régularité est généralement matérialisée par un paramètre "$\alpha$" faisant référence à une certaine classe fonctionnelle. Différents espaces seront considérés, en fonction des propriétés d'approximation cherchées : approximation d'une fonction par sa projection orthogonale sur un sous-espace dans le cas d'estimateurs par projection, ou approximation d'une fonction par sa convolée avec une approximation de l'unité dans le cas d'estimateurs à noyaux. Dans cette section, nous présentons les différentes classes utilisées, ainsi que les principaux résultats nécessaires au contrôle du biais des estimateurs.

### 2.2.1 Définition des espaces fonctionnels considérés

Nous rappelons brièvement ici quelques définitions d'espaces classiquement utilisés en estimation non paramétrique pour l'estimation de fonctions d'une variable ou de deux variables réelles, ce qui est suffisant pour les cadres étudiés. Dans le cas des fonctions de deux variables, l'indice de régularité pourra être différent dans chacune des deux directions (par rapport à chaque variable). On parlera d'espaces anisotropiques. Pour plus de clarté, nous donnons les définitions pour des fonctions d'une variable uniquement, excepté pour les espaces de Sobolev (requis entre autres pour estimer une densité conditionnelle au Chapitre 5). Dans cette section, la lettre $A$ désigne un intervalle de $\mathbb{R}$.

Pour plus de détails (propriétés et inclusions) sur tous les espaces cités ci-dessous, on pourra se référer (entre autres) à Adams (1975), Nikol′skiĭ (1975), DeVore & Lorentz (1993), et Triebel (2006).

#### Espaces de Hölder et de Nikol'skiĭ

Classiquement, la régularité d'une fonction est définie à travers ses dérivées (ou dérivées partielles). Les espaces les plus classiques sont les espaces $\mathcal{C}^k(A)$, pour $k \in \mathbb{N}$ des fonctions

$k$-fois continûment dérivables (ou continûment différentiables) sur $A$. On peut être plus précis quant aux propriétés des dérivées, et définir les espaces de Hölder, fondés sur un comportement local de type Lipschitz des dérivées (voir par exemple Tsybakov 2009).

**Définition 2.1.** *On dit qu'une fonction $t$ définie sur un intervalle $A$ appartient à la classe de Hölder $\mathcal{H}(\alpha, A, L)$ $(\alpha, L > 0)$, si $t$ est $\lfloor\alpha\rfloor-$fois dérivable sur $A$ et quels que soit $x, x' \in A$,*

$$\left|t^{(\lfloor\alpha\rfloor)}(x) - t^{(\lfloor\alpha\rfloor)}(x')\right| \leq L|x - x'|^{\alpha-\lfloor\alpha\rfloor}. \tag{2.12}$$

Si $\alpha \in (0; 1]$, l'espace $\mathcal{H}(\alpha, A, L)$ est l'ensemble des fonctions $t$ telles que $|t(x) - t(x')| \leq L|x - x'|^\alpha$, $x, x' \in A$. La même inégalité pour $\alpha > 1$ imposerait $t'(x) = 0$ pour tout $x$ et donc $t$ serait constante. La condition (2.12) représente donc une extension plus raisonnable, au cas $\alpha > 1$.

Les espaces de Nikol'skiĭ (Nikol′skiĭ, 1975) peuvent être compris comme une version intégrée des espaces de Hölder. Moins connus, ils sont pourtant relativement standard pour étudier le comportement global d'estimateurs à noyaux (voir par exemple Kerkyacharian *et al.* 2001, Goldenshluger & Lepski 2011a).

**Définition 2.2.** *On dit qu'une fonction $t$ définie sur $\mathbb{R}$ appartient à la classe de Nikol'skiĭ $\mathcal{N}_p(\alpha, L)$ $(p \geq 1,\ \alpha, L > 0)$, si $t$ est $\lfloor\alpha\rfloor$-fois dérivable sur $\mathbb{R}$ et quel que soit $x \in \mathbb{R}$,*

$$\left(\int_{\mathbb{R}} \left(t^{(\lfloor\alpha\rfloor)}(x' + x) - t^{(\lfloor\alpha\rfloor)}(x')\right)^p dx'\right)^{1/p} \leq L|x|^{\alpha-\lfloor\alpha\rfloor}.$$

Ces définitions s'étendent facilement à la dimension supérieure : voir par exemple Comte & Lacour (2013).

### Espaces de Sobolev périodisés (fonction d'une variable)

Commençons par définir les espaces de Sobolev pour un paramètre de régularité $\alpha$ entier naturel.

**Définition 2.3.** *Soient $\alpha \in \mathbb{N}\backslash\{0\}$, et $p \in [1; \infty]$. On dit qu'une fonction $t$ définie sur $A$ appartient à l'espace de Sobolev $W_p^\alpha(A)$ si $t$ est $\alpha$ fois faiblement dérivable, et si $t^{(l)}$ appartient à $L^p(A)$ pour tout $l \in \{0, \ldots, \alpha\}$. On définit également $W_p^\alpha(A, L)$ l'ensemble des fonctions de $W_p^\alpha(A)$ telles que $\|t^{(\alpha)}\|_{L^p(A)} \leq L$.*

**Remarque 2.3.**

1. La définition ci-dessus est celle donnée par Adams (1975) ou Härdle *et al.* (1998) entre autres. Elle est équivalente à la définition suivante, donnée par exemple par DeVore & Lorentz (1993) et Tsybakov (2009) : une fonction $t$ appartient à $W_p^\alpha(A)$ si $t$ est $\alpha-$fois faiblement dérivable, et si $t$ et $t^{(\alpha)}$ appartiennentà $L^p(A)$. On se réfèrera à DeVore & Lorentz (1993) p.37-39 pour la preuve de l'équivalence.
2. On a en particulier $W_p^0(A) = L^p(A)$.

Comme pour les espaces de Nikol'skiĭ, on considèrera ces classes pour $p = 2$. Dans ce cas, l'espace $W_2^\alpha(\mathbb{R})$ s'identifie à l'espace $H^\alpha$ des fonctions $t \in L^2(\mathbb{R})$ dont la transformée de Fourier $\mathcal{F}(t)$ vérifie $x \mapsto (1+x^2)^{\alpha/2}\mathcal{F}(t)(x) \in L^2(\mathbb{R})$ (ceci grâce à la formule de Fourier-Plancherel, reliant la norme $L^2$ d'une fonction à la norme $L^2$ de sa transformée de Fourier). C'est cette relation qui permet de démontrer les inclusions évoquées ci-dessous entre espaces de Besov et espaces de Sobolev (voir (2.14)).

La version suivante des espaces de Sobolev, dite périodisée, et définie par exemple par Tsybakov (2009) (p.49) est utilisée dans notre contexte.

**Définition 2.4.** *Soient $\alpha \in \mathbb{N}\backslash\{0\}$. On dit qu'une fonction $t$ définie sur $(0;1)$ appartient à l'espace de Sobolev $W_2^{\alpha,per}((0;1))$ (resp. à la boule $W_2^{\alpha,per}((0;1),L)$, $L>0$) si $t$ appartient à $W_2^\alpha((0;1))$ (resp. $W_2^\alpha((0;1),L)$) et si $t^{(l)}(0) = t^{(l)}(1)$ pour $l = 0,\dots,\alpha-1$.*

L'intérêt de ces espaces pour l'estimation non paramétrique par projection est leur caractérisation en terme de vitesse de convergence de la série de coefficients de Fourier. Suivant Tsybakov (2009), introduisons l'ellipsoïde suivant, pour $\alpha \in \mathbb{N}\backslash\{0\}$,et $Q > 0$,

$$\Theta(\alpha, Q) := \left\{ \theta \in \ell^2(\mathbb{N}),\ \sum_{j=1}^{\infty} \mu_{j,\alpha}^2 \theta_j^2 \le Q \right\},$$

où $(\mu_{j,\alpha})_{j\in\mathbb{N}\backslash\{0\}}$ est définie par

$$\mu_{j,\alpha} = \begin{cases} j^\alpha, \text{ si } j \text{ est pair,} \\ (j-1)^\alpha, \text{ sinon.} \end{cases}$$

On note $(\varphi_j)_{j\in\mathbb{N}\backslash\{0\}}$ la base de Fourier sur $(0;1)$ : $\varphi_1(x) = 1$, $\varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$, $\varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$, pour $x \in (0;1)$ et $k \in \mathbb{N}$. Le résultat suivant est énoncé dans Tsybakov (2009) (Lemme A.3). Sa démonstration repose principalement sur l'égalité de Parseval, reliant la norme $L^2$ d'une fonction à la série des carrés de ses coefficients de Fourier, ainsi que sur le lien entre les coefficients de Fourier d'une fonction et ceux de sa dérivée.

**Proposition 2.8.** *Soient $\alpha \in \mathbb{N}\backslash\{0\}$, et $L > 0$. Une fonction $t = \sum_{j=1}^\infty \theta_j\varphi_j$ est dans la boule $W_2^{\alpha,per}((0;1),L)$ si et seulement si la suite $(\theta_j)_j$ est dans l'ellipsoïde $\Theta(\alpha, L^2/\pi^{2\alpha})$.*

**Remarque 2.4.**

1. Pour $0 \le \alpha' \le \alpha$, l'inclusion suivante est vraie : $\Theta(\alpha, L^2/\pi^{2\alpha}) \subset \Theta(\alpha', L^2/\pi^{2\alpha})$. La Proposition 2.8 permet alors d'affirmer que l'inclusion analogue est valable pour les espaces de Sobolev périodisés : $W_2^{\alpha,per}((0;1),L) \subset W_2^{\alpha',per}((0;1),L)$.
2. La Proposition 2.8 permet d'étendre la définition des boules $W_2^{\alpha,per}((0;1),L)$ à des valeurs non entières de $\alpha$ : une fonction est dans un tel espace, si et seulement si la suite de ses coefficients de Fourier est dans l'ellipsoïde correspondant.

Enfin, une propriété importante de ces espaces est la suivante.

**Proposition 2.9.** *Soit $S_m = Vect\{\varphi_j,\ j = 1\dots,2m+1\}$, où $(\varphi_j)_j$ est la base de Fourier. Soit $s$ une fonction de classe $\mathcal{C}^1([0;1])$ appartenant à $W_2^{1,per}((0;1),L)$. En notant $\Pi_{S_m}$ l'opérateur de projection orthogonale sur $S_m$, on a*

$$\Pi_{S_m}(s') = (\Pi_{S_m}(s))'.$$

**Preuve de la Proposition 2.9.** Notons $D_m = 2m+1$. Puisque $(\varphi_j)_j$ forme une base orthonormée de $S_m$, $\Pi_{S_m}(s) = \sum_{j=1}^{D_m} a_j\varphi_j$, où $a_j = \langle h, \varphi_j\rangle$. Donc,

$$\begin{aligned}(\Pi_{S_m}(s))' &= \left(a_1 + \sum_{k=1}^{m} a_{2k}\varphi_{2k} + a_{2k+1}\varphi_{2k+1}\right)', \\ &= \sum_{k=1}^{m} a_{2k}\varphi'_{2k} + a_{2k+1}\varphi'_{2k+1} = \sum_{k=1}^{m} 2\pi k(-a_{2k}\varphi_{2k+1} + a_{2k+1}\varphi_{2k}).\end{aligned}$$

D'autre part, $\Pi_{S_m}(s') = \sum_{j=1}^{D_m} b_j\varphi_j$ où,

$$\begin{aligned}b_j &= \langle s', \varphi_j\rangle = \int_0^1 s'(x)\varphi_j(x)dx, \\ &= [s(x)\varphi_j(x)]_0^1 - \int_0^1 s(x)\varphi'_j(x)dx = -\int_0^1 s(x)\varphi'_j(x)dx,\end{aligned}$$

l'hypothèse $s \in \mathcal{C}^1$ justifiant l'intégration par partie, et l'hypothèse $s \in W_2^{1,per}(1,L)$ justifiant que le crochet est nul. Si $j = 2k$, on a donc

$$b_{2k} = 2\pi k \int_0^1 s(x)\varphi_{2k+1}(x)dx = 2\pi k a_{2k+1},$$

et de même, $b_{2k+1} = -2\pi k a_{2k}$, ce qui entraîne donc,

$$\Pi_{S_m}(s') = \sum_{k=1}^{m} 2\pi k(-a_{2k}\varphi_{2k+1} + a_{2k+1}\varphi_{2k}) = (\Pi_{S_m}(s))'.$$

□

**Espaces de Sobolev périodisés (fonction de deux variables)**

L'estimation de la densité conditionnelle en base trigonométrique déformée requière l'introduction d'espaces de Sobolev périodisés anisotropiques pour les fonctions de deux variables. On peut définir ceux-ci soit à l'aide de propriétés concernant les dérivées partielles analogues à celles de la Définition 2.4, soit à l'aide de propriétés des coefficients de Fourier. Pour des raisons de simplicité, nous choisissons la deuxième possibilité.

Soit $Q > 0$, et soit $\alpha = (\alpha_1, \alpha_2) \in (\mathbb{N}\backslash\{0\})^2$. On introduit le sous ensemble de $\ell^2((\mathbb{N}\backslash\{0\})^2)$ suivant :

$$\Theta(\alpha, Q) = \left\{ (\theta_{j,k})_{j,k\in\mathbb{N}\backslash\{0\}} \in \ell^2((\mathbb{N}\backslash\{0\})^2), \sum_{j,k\in\mathbb{N}\backslash\{0\}} \mu_{j,\alpha_1}^2\mu_{k,\alpha_2}^2|\theta_{j,k}|^2 \le Q^2 \right\}.$$

avec $\mu_{m,\alpha_l} = m^{\alpha_l}$ si $m$ est pair, $(m-1)^{\alpha_l}$ sinon $(l = 1, 2)$.

**Définition 2.5.** *Soit $\alpha = (\alpha_1, \alpha_2) \in (\mathbb{N}\backslash\{0\})^2$ et $Q > 0$. Soit $t \in L^2((0;1)\times(0;1))$. Soit $(\theta_{j,k})_{j,k}$ la suite de ses coefficients de Fourier, de telle sorte que dans $L^2$, $t = \sum_{j,k\in(\mathbb{N}\backslash\{0\})^2} \theta_{j,k}\varphi_j\otimes\varphi_k$. On dit que $t$ appartient à l'espace de Sobolev $W_2^{\alpha,per}((0;1)^2, L)$ $(L > 0)$, si la suite $(\theta_{j,k})_{j,k}$ appartient à l'ellipsoïde $\Theta(\alpha, L^2/\pi^{2(\alpha_1+\alpha_2)})$.*

Il est courant en statistique non paramétrique de définir des espaces de régularité à l'aide de propriétés de coefficients de développement d'une fonction dans une certaine base (on pourra trouver d'autres exemples dans Barron *et al.* 1999).

**Espaces de Besov**

La définition des espaces de Besov repose sur les notions de "module de régularité" ("moduli of smoothness"). Cette section est principalement fondée sur le Chapitre 2 de DeVore & Lorentz (1993). On note $\mathcal{F}_{p,A}$ l'ensemble $L^p(A)$, si $p < \infty$, et l'ensemble des fonctions uniformément continues sur $A$ sinon.

Pour $v > 0$, on définit $\Delta_v^1$ l'opérateur "différence du premier ordre", qui associe, à une fonction réelle $t$, la fonction $\Delta_v^1(t) : x \mapsto t(x+v) - t(x)$. On définit ensuite par récurrence, pour $r \in \mathbb{N}\backslash\{0\}$, l'opérateur "différence d'ordre $r$" : pour $r \geq 2$, $\Delta_v^r = \Delta_v^1[\Delta_v^{r-1}]$. Celui-ci s'exprime par la formule suivante :

$$\Delta_v^r t(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} t(x+kv). \tag{2.13}$$

**Remarque 2.5.** On peut montrer que pour $r \in \mathbb{N}\backslash\{0\}$, et $t \in \mathcal{C}^r(A)$, il y a convergence uniforme sur tout compact quand $v \to 0$ de $\Delta_v^r t$ vers $t$. L'opérateur différence d'ordre $r$ étend donc la notion de dérivabilité.

Le module de régularité d'ordre $r$ pour la norme $p$ est défini par la formule suivante :

$$\omega_r(t,x)_p := \sup_{0<v\leq x} \|\Delta_v^r(t)\|_{L^p(A)},\ x > 0,\ t \in L^p(A).$$

**Remarque 2.6.** Si $p = \infty$ et $r = 1$, on retrouve le module de continuité de la fonction $t$ :

$$\omega_1(t,x)_\infty := \sup_{0<v\leq x} \|\Delta_v^1(t)\|_{L^\infty(A)} = \sup_{0<v\leq x} \sup \operatorname{ess}_{x'\in A} |t(x+x') - t(x')|.$$

La notion de module de régularité étend donc cette notion, à la fois à tous les espaces $L^p$, et à des fonctions présentant un degré de régularité plus élevé que la simple continuité.

Le résultat suivant sera utile (voir ci-dessous la preuve de la Proposition 2.10, paragraphe (b)) :

**Lemme 2.2.** *Soient $t \in \mathcal{F}_{p,A}$, $x > 0$. Le module de régularité de $f$ vérifie les inégalités suivantes*

- *si $n \in \mathbb{N}$, $\omega_r(t,nx)_p \leq n^r \omega_r(t,x)_p$,*
- *si $\lambda > 0$, $\omega_r(t,\lambda x)_p \leq (\lambda+1)^r \omega_r(t,x)_p$.*

**Preuve du Lemme 2.2.** Pour la première inégalité, on va utiliser la généralisation suivante de l'expression (2.13) :

$$\Delta_{nh}^r t(x) = \sum_{k_1=0}^{n-1} \sum_{k_2=0}^{n-1} \cdots \sum_{k_r=0}^{n-1} \Delta_h^r(t)(x + k_1 h + k_2 h + \cdots + k_r h).$$

On majore ainsi la norme de l'opérateur différence d'ordre $r-k$ :

$$\begin{aligned}\omega_r(t,nx)_p &= \sup_{0<u\leq nx}\|\Delta_u^r t\|_p = \sup_{0<h\leq x}\|\Delta_{nh}^r t\|_p\,,\\ &= \sup_{0<h\leq x}\left\|\sum_{k_1=0}^{n-1}\sum_{k_2=0}^{n-1}\cdots\sum_{k_r=0}^{n-1}\Delta_h^r(t)(.+k_1h+k_2h+\cdots+k_rh)\right\|_p,\\ &\leq n^r\sup_{0<h\leq x}\|\Delta_h^r t\|_p\,.\end{aligned}$$

Pour la seconde, on désigne par $n$ la partie entière de $\lambda$, et on applique successivement la croissance de $\omega_r(t,.)_p$, l'inégalité que l'on vient de montrer, et la définition de la partie entière :

$$\omega_r(t,\lambda x)_p \leq \omega_r(t,(n+1)x)_p \leq (n+1)^r\omega_r(t,x)_p \leq (\lambda+1)^r\omega_r(t,x)_p.$$

□

**Définition 2.6.** *Soient $\alpha>0$ et $p,q\in[1,\infty]$. L'espace de Besov $\mathcal{B}_{p,q}^\alpha(A)$ est l'ensemble des fonctions $t\in\mathcal{F}_{p,A}$, telles que $|t|_{\mathcal{B}_{p,q}^\alpha}<\infty$, où*

$$|t|_{\mathcal{B}_{p,q}^\alpha} := \begin{cases}\displaystyle\left(\int_0^\infty\left[x^{-\alpha}\omega_r(t,x)_p\right]^q\frac{dx}{x}\right)^{1/q} & \text{si } q<\infty,\\ \displaystyle\sup_{x>0}x^{-\alpha}\omega_r(t,x)_p & \text{si } q=\infty,\end{cases}$$

*où $r=\lfloor\alpha\rfloor+1$.*

**Remarque 2.7.**

1. L'application $t\mapsto|t|_{\mathcal{B}_{p,q}^\alpha}$ est une semi-norme sur $\mathcal{B}_{p,q}^\alpha(A)$. Pour définir une norme, on pose
$$\|t\|_{\mathcal{B}_{p,q}^\alpha} = \|t\|_{L^p(A)} + |t|_{\mathcal{B}_{p,q}^\alpha}.$$
On notera, pour $L>0$, $\mathcal{B}_{p,q}^\alpha(A,L)$ la boule de rayon $L$ pour la semi-norme $|.|_{\mathcal{B}_{p,q}^\alpha}$.
2. Les espaces de Besov sont très utilisés dans l'approche par ondelettes : en effet, de la même façon que les espaces de Sobolev peuvent être caractérisés par l'appartenance des coefficients de Fourier à certains ellipsoïdes (Proposition 2.8), il est possible de définir les espaces de Besov à partir de coefficients d'une fonction dans une base d'ondelettes (voir Härdle *et al.* 1998).

On peut donner une interprétation "intuitive" (!) de ce qu'est une fonction dans un tel espace, lorsque $q=\infty$ : c'est une fonction de puissance $p-$ième intégrable et $\lfloor\alpha\rfloor-$fois dérivable, dont la dérivée d'ordre $\lfloor\alpha\rfloor$ vérifie la propriété de Lipschitz d'ordre $\alpha-\lfloor\alpha\rfloor$. Le paramètre $\alpha$ s'interprète donc bien ici aussi comme mesurant la régularité de la fonction que l'on considère.

DeVore & Lorentz (1993) (Chapitre 2, Section 7) prouvent que pour $p\geq 2$,

$$\mathcal{B}_{p,\infty}^\alpha(A)\subset\mathcal{B}_{2,\infty}^\alpha(A).$$

Ceci justifie que l'on considère uniquement la classe de régularité $\mathcal{B}_{2,\infty}^\alpha(A)$ pour le risque des estimateurs par projections considérés dans cette thèse.

**Liens entre ces espaces**

Les espaces de Besov sont les espaces de régularité les plus généraux, englobant les espaces de Hölder, Nikol'skiĭ et Sobolev. Citons, à titre d'exemple, les deux inclusions suivantes, prouvées dans DeVore & Lorentz (1993) :

$$L^\infty(\mathbb{R}) \cap \mathcal{H}(\alpha, L, \mathbb{R}) \subset \mathcal{B}^\alpha_{\infty,\infty}(\mathbb{R}, L) \text{ et } W_2^\alpha(\mathbb{R}, L) \subset \mathcal{B}^\alpha_{2,\infty}(\mathbb{R}, L). \tag{2.14}$$

### 2.2.2 Espaces de régularités et propriétés d'approximation : éléments pour le contrôle des termes de biais

La décomposition biais-variance du risque quadratique intégré des estimateurs bâtis dans cette thèse entraîne la nécessité de majorer le terme de biais, ou erreur d'approximation. Celui-ci se présente différemment selon que l'on considère des estimateurs par projection ou des estimateurs à noyaux. Nous détaillons de manière plus approfondie le cas de l'approximation par projection, puisque la majeure partie de nos estimateurs sont construits de cette façon.

**Approximation et projection orthogonale**

Considérons l'estimation d'une fonction $s$, par projection sur un sous-espace vectoriel $S_D$ de dimension $D$ de $L^2(A)$. Dans ce cadre, le terme de biais associé au risque quadratique est généralement égal à $E_{2,D}(s) = \|s - s_D\|_{L^2(A)}$ (ou tout au moins fait intervenir une quantité similaire), où $s_D$ désigne le projeté orthogonal de la fonction $s$ sur l'espace $S_D$.

Lorsque la dimension $D$ du modèle augmente, le terme de biais décroit car la fonction $s$ est de mieux en mieux approchée par sa projection orthogonale $s_D$. La vitesse de décroissance peut-être quantifiée, si l'on suppose que la fonction $s$ appartient à une boule d'un espace de Hölder, de Besov ou de Sobolev, et que l'on travaille avec des sous-espaces engendrés par des bases "classiques". Nous énonçons tout d'abord le résultat, puis en donnons une démontration dans quelques cas précis et simples, à titre d'exemples.

**(a) Enoncé du résultat** La proposition suivante est due à Barron *et al.* (1999) (Lemme 12 p.404).

**Proposition 2.10.** *Soit $s \in \mathcal{B}^\alpha_{p,\infty}(A)$, $1 \leq p \leq \infty$. On note $S_D$ l'un des trois espaces suivants :*

- *$S_D$ est l'espace des polynômes par morceaux de degré au plus $r$, avec $r > \alpha - 1$, basés sur une partition régulière à $D$ pièces, avec $A = (0;1)$*
- *$S_D$ est l'espace des polynômes trigonométriques de degré au plus $D$, et $A = \mathbb{T}$ (tore en dimension 1)*
- *$S_D$ est l'espace vectoriel engendré par les $(\varphi_\lambda)_{\lambda \in \bigcup_{j=0}^J \Lambda(j)}$, où $(\varphi_\lambda)_\lambda$ est une base d'ondelettes de régularité $r > \alpha - 1$, $\Lambda(j) = \{(j,k) | 1 \leq k \leq 2^j\}$, $D = 2^J$, $A = (0;1)$*

*Alors, il existe une constante $C(\alpha, p) > 0$, telle que l'erreur d'approximation $E_{p,D}(s)$ commise en approchant $s$ par $S_D$ est majorée par*

$$E_{p,D}(s) := \|s - s_D\|_{L^p(A)} \leq C(\alpha, p)|s|_{\mathcal{B}^\alpha_{p,\infty}} D^{-\alpha}.$$

**Remarque 2.8.** 1. Grâce aux inclusions entre les différents espaces, la propriété est en particulier vraie si $s$ appartient à la boule $W_2^{\alpha,per}((0;1),L)$ d'un espace de Sobolev périodisé.

2. Lacour (2007) (Lemme 9) énonce une version analogue de ce lemme pour les fonctions appartenant à des espaces de Besov anisotropiques multidimensionnels, fondé sur les travaux de Hochmuth (2002) et Nikol′skiĭ (1975). Nous l'utiliserons lors de l'estimation d'une fonction de deux variables, la densité conditionnelle. Pour une fonction appartenant à un espace de Besov anisotropique $\mathcal{B}_{2,\infty}^{(\alpha_1,\alpha_2)}((0;1)^2)$, l'inégalité obtenue est alors la suivante, pour des sous espaces de type $S_{D_1} \times S_{D_2}$, de dimensions $D_1 \times D_2$ (avec $S_{D_1}$ et $S_{D_2}$ comme dans la Proposition 2.10) :

$$\|s - s_{D_1,D_2}\|_{L^2((0;1)^2)} \leq C(\alpha, s)\left(D_1^{-\alpha_1} + D_2^{-\alpha_2}\right). \tag{2.15}$$

**(b) Preuve de la Proposition 2.10 dans un cas particulier** L'objectif de cette section est de démontrer la Proposition 2.10 dans le cas particulier où $S_D$ est l'espace des polynômes trigonométriques de degré au plus $D$. Remarquons que $E_{p,D}(s) = d_p(s, S_D) = \inf_{t \in S_D} \|s - t\|_{L^p(A)}$. Le problème se réduit à prouver

$$E_{p,D}(s) \leq C_{r,p}\omega_r(s, D^{-1})_p, \tag{2.16}$$

où $r$ désigne l'entier immédiatement supérieur à $\alpha$ et $C_{r,p}$ une constante. En effet, si (2.16) est montré, on a ensuite

$$\omega_r(s, D^{-1})_p = D^{-\alpha}D^{\alpha}\omega_r(s, D^{-1})_p \leq D^{-\alpha} \sup_{x \in (0;1]} t^{-\alpha}\omega_r(s,x)_p \leq C_{r,p}D^{-\alpha}|s|_{\mathcal{B}^{\alpha}_{p,\infty}}.$$

• **Introduction à la preuve de (2.16).** Le schéma de preuve est relativement simple : il suffit de trouver un élément $P \in S_D$ tel que $\|s - P\|_{L^p(\mathbb{T})} \leq C_r\omega_r(s, D^{-1})_p$. Pour cela, on va prendre $P$ sous la forme d'une convolée (ou d'une modification de celle-ci) d'un élément de $S_D$ avec $s$ ($T \in S_D$, $s \star T \in S_D$ aussi). Cet élément va être un noyau classique de l'analyse de Fourier, le noyau de Jackson, éventuellement modifié. On en rappelle donc sa définition et ses propriétés :

**Définition 2.7.** *Soient $r, D \in \mathbb{N}\backslash\{0\}$.*

- *Le noyau de Jackson d'ordre $D$ est défini par*

$$J_D : \ t \ \mapsto \ J_D(x) = \lambda_D \left(\frac{\sin(Dx/2)}{\sin(x/2)}\right)^4,$$

*où $\lambda_D$ est choisi de telle sorte que $\|J_D\|_{L^1((-\pi;\pi))} = \int_{-\pi}^{\pi} J_D(x)dx = 1$.*

- *Le noyau de Jackson modifié d'ordre $D$ est défini par*

$$J_{N,r} : \ x \ \mapsto \ J_{D,r}(x) = \lambda_{D,r} \left(\frac{\sin(Dx/2)}{\sin(x/2)}\right)^{2r},$$

*où $\lambda_{D,r}$ est choisi de telle sorte que $\int_{-\pi}^{\pi} J_{D,r}(x)dx = 1$.*

On a bien sûr $J_{D,2} = J_D$. Il est relativement simple de voir que $J_{D,r}$ est un polynôme trigonométrique de degré $rD$ pair et positif : en effet, $J_D = K_D^2/\|K_D^2\|_{L^1((-\pi;\pi))}$, où $K_D$ est le noyau de Féjer :

$$K_D(t) = \frac{1}{D}\left(\frac{\sin(Dt/2)}{\sin(t/2)}\right)^2 = \frac{1}{D}\sum_{n=0}^{D-1} Di_n,$$

avec $Di_n$ est le noyau de Dirichlet : $Di_n(t) = \sum_{k=-n}^{n} \exp(-ikt)$.

**Lemme 2.3.** *Avec les notations de la définition précédente, on a* $\lambda_{D,r} \underset{D\to\infty}{\sim} D^{-2r+1}$. *De plus,*

$$\int_0^\pi x^k J_{D,r}(x)dx = O(D^{-k}) \quad \textit{pour } k \in \{0,1,\ldots,2r-2\} \textit{ et } D \textit{ suffisamment grand.}$$

*En particulier, ceci est valable pour* $J_D$, *pour* $k = 0,1,2$.

On distingue dans la suite le cas $r = 2$ (rappel : $r$ est l'entier immédiatement supérieur à $\alpha$) du cas général.

• **Preuve de (2.16) dans le cas** $r = 2$**.** Supposons un instant qu'il existe une fonction $\Lambda_D \in S_D$, paire, positive, telle que

$$\int_{-\pi}^{\pi} \Lambda_D(x)dx = 1, \quad \text{et} \quad \int_0^\pi x^k \Lambda_D(x)dx \le C_1 D^{-k} \text{ pour k=0,1,2,}$$

où $C_1$ est une constante. On peut alors écrire, en notant $\check{\Lambda}(x) = \Lambda(-x)$, $x \in \mathbb{R}$ :

$$\begin{aligned} s \star \check{\Lambda}_D(x) - s(x) &= \int_{-\pi}^{\pi} (s(x+t) - s(x))\,\Lambda_D(t)dt, \\ &= \int_0^\pi (s(x+t) - s(x))\,\Lambda_D(t)dt + \int_0^\pi \big(s(x-t') - s(x)\big)\,\Lambda_D(-t')dt', \\ &= \int_0^\pi (s(x+t) + s(x-t) - 2s(x))\,\Lambda_D(t)dt. \end{aligned}$$

On passe à la norme $L^p$, puis on utilise l'Inégalité de Minkowski généralisée (voir Proposition 2.12) :

$$\begin{aligned} \left\|\check{\Lambda}_D \star s - s\right\|_{L^p(\mathbb{T})} &= \left[\int_{\mathbb{R}} \left(\int_{[0;\pi]} (s(x+t) + s(x-t) - 2s(x))\,\Lambda_D(t)dt\right)^p dx\right]^{1/p}, \\ &\le \int_{(0;\pi)} \Lambda_D(t) \left[\int_{\mathbb{R}} (s(x+t) + s(x-t) - 2s(x))^p\,dx\right]^{1/p} dt, \\ &= \int_{(0;\pi)} \Lambda_D(t) \|\Delta_t^1 \Delta_{-t}^1 s\|_{L^p((-\pi;\pi))} dt, \\ &\le \int_{(0;\pi)} \Lambda_D(t)\omega_2(s,t)_p dt, \\ &\le \omega_2(s, D^{-1})_p \int_{(0;\pi)} (Dt+1)^2 \Lambda_D(t)dt, \end{aligned}$$

en utilisant la seconde inégalité du Lemme 2.2. Finalement, on a donc

$$\|s - \check{\Lambda}_D \star s\|_{L^p((-\pi;\pi))} \leq C\omega_2(s, D^{-1})_p,$$

ce qui est l'assertion (2.16) cherchée. Il suffit maintenant de justifier l'existence d'une fonction $\Lambda_D$ ayant les propriétés requises. Le noyau de Jackson $J_{\lfloor D/2 \rfloor + 1}$ classique convient, comme le montre le Lemme 2.3, ce qui termine la preuve du cas $r = 2$.

• **Preuve de (2.16) dans le cas général.** La seule convolution d'un noyau avec la fonction $s$ ne fait apparaître que des opérateurs différences d'ordre 1 ou 2 et pas d'un ordre $r$ quelconque. On définit donc plutôt, pour $x \in \mathbb{R}$,

$$S_D(s)(x) = \int_{-\pi}^{\pi} \left[(-1)^r \Delta_t^r s(x) + s(x)\right] J_{M,r}(t)dt,$$

avec $M = \lfloor D/2 \rfloor + 1$. En utilisant la définition de $\Delta_t^r$ et le fait que $J_{M,r}$ est un polynôme trigonométrique, on obtient que $S_D(s)(x)$ est combinaison linéaire de termes de la forme $\int_0^{2\pi} s(x + kt)g(lt)dt$, où $g = \cos$ ou $\sin$. De plus, $s$ étant définie sur le tore $\mathbb{T}$, elle est $2\pi$ périodique, et donc $t \mapsto s(x + kt)$ est $2\pi/k$ périodique. On utilise donc le lemme suivant, obtenu par changement de variables :

**Lemme 2.4.** *Soient $k$ et $l$ deux entiers naturels non nuls. Soit $\xi \in L^1(\mathbb{T})$ une fonction de période $2\pi/k$. Si $k$ ne divise pas $l$*

$$\int_{\mathbb{T}} \xi(t)\cos(lt)dt = \int_{\mathbb{T}} \xi(t)\sin(lt)dt = 0.$$

Ceci permet d'affirmer, en faisant aussi le changement de variable $u = x + kt$ dans les termes dont $S_D(s)$ est combinaison linéaire, que $S_D(s)$ est un polynôme trigonométrique de degré inférieur ou égal à $N$. On majore ensuite $\|S_D(s) - s\|_{L^p((-\pi;\pi))}$, en suivant les mêmes étapes exactement que pour la majoration de $\left\|\check{\Lambda}_D \star s - s\right\|_{L^p((-\pi;\pi))}$ dans le cas $r = 2$, et cela suffit pour majorer la distance de $s$ à $S_D$.

□

• **Preuve du Lemme 2.3.** On utilise successivement d'abord la définition de $\lambda_{D,r}$ et la parité de la fonction intégrée pour écrire :

$$\begin{aligned}
\lambda_{D,r}^{-1} &= \int_{-\pi}^{\pi} \left(\frac{\sin(Dx/2)}{\sin(x/2)}\right)^{2r} dx = 2\int_0^{\pi} \left(\frac{\sin(Dx/2)}{\sin(x/2)}\right)^{2r} dx, \\
&\leq 2\pi^{2r} \int_0^{\pi} \left(\frac{\sin(Dx/2)}{x}\right)^{2r} dx \ \text{ car } \sin(x/2) \leq x/\pi, \\
&= 2^{2r-1}\pi^{2r} D^{2r-1} \int_0^{D\pi/2} \left(\frac{\sin u}{u}\right)^{2r} du, \\
&\leq 2^{2r-1}\pi^{2r} D^{2r-1} \int_0^{\infty} \left(\frac{\sin u}{u}\right)^{2r} du = CD^{2r-1},
\end{aligned}$$

où $C$ est une constante (indépendante de $D$).
Pour majorer $\lambda_{D,r}^{-1}$, on suit la même démarche mais en utilisant $\sin(x/2) \leq x/2$, et en minorant l'intégrale de 0 à $N\pi/2$ par l'intégrale de 0 à $\pi/2$ par exemple. Ceci donne l'équivalent pour $\lambda_{D,r}$. On écrit ensuite,

$$\begin{aligned}\int_0^\pi x^k J_{D,r}(x)dx &\leq \lambda_{D,r}\pi^{2r}\int_0^\pi x^k\left(\frac{\sin(Dx/2)}{x}\right)^{2r}dx,\\ &\leq \lambda_{D,r}\pi^{2r}2^{2r}D^{2r-k-1}\int_0^\infty u^k\left(\frac{\sin u}{u}\right)^{2r}du \leq CD^{-k},\end{aligned}$$

en utilisant l'équivalent de $\lambda_{D,r}$.

□

**(c) Preuve de l'Inégalité (2.15) dans un cas particulier** L'objectif de cette section est de démontrer l'Inégalité (2.15) dans le cas particulier où $s$ appartient à l'espace $W_2^{\alpha,per}((0;1)^2,L)$ ($\alpha=(\alpha_1,\alpha_2)$), et où $S_{D_1}$ (resp. $S_{D_2}$) est l'espace des polynômes trigonométriques de degré au plus $D_1$ (resp. $D_2$).

Dans ce cas, en notant $(\varphi_j)_{j\in\mathbb{N}\setminus\{0\}}$ la base de Fourier, on a $S_D = \text{Vect}\{\varphi_1,\ldots,\varphi_D\}$. On écrit $s=\sum_{j,k=1}^\infty \theta_{j,k}\varphi_j\otimes\varphi_k$, de telle sorte que la projection de $s$ sur $S_{D_1}\times S_{D_2}$ est $s_{D_1,D_2}=\sum_{j=1}^{D_1}\sum_{k=1}^{D_2}\theta_{j,k}\varphi_j\otimes\varphi_k$. Ainsi, en utilisant le caractère orthonormé des $\varphi_j\otimes\varphi_k$, et la définition de la boule $W_2^{\alpha,per}((0;1)^2,L)$ (voir Définition 2.5),

$$\begin{aligned}\|s-s_{D_1,D_2}\|_2^2 &= \left\|\sum_{\substack{j>D_1 \text{ ou}\\ k>D_2}}\theta_{j,k}\varphi_j\otimes\varphi_k\right\|_2^2 = \sum_{\substack{j>D_1 \text{ ou}\\ k>D_2}}|\theta_{j,k}|^2,\\ &= \sum_{\substack{j>D_1\\ k\geq 1}}|\theta_{j,k}|^2+\sum_{\substack{k>D_2\\ 1\leq j\leq D_1}}|\theta_{j,k}|^2,\\ &\leq \frac{1}{\mu_{D_1+1,\alpha_1}^2}\sum_{\substack{j>D_1\\ k\geq 1}}\mu_{j,\alpha_1}^2|\theta_{j,k}|^2+\frac{1}{\mu_{D_2+1,\alpha_2}^2}\sum_{\substack{k>D_2\\ 1\leq j\leq D_1}}\mu_{j,\alpha_2}^2|\theta_{j,k}|^2,\\ &\leq \left(\frac{1}{\mu_{D_1+1,\alpha_1}^2}+\frac{1}{\mu_{D_2+1,\alpha_2}^2}\right)\sum_{(j,k)\in(\mathbb{N}\setminus\{0\})^2}\mu_{j,\alpha_1}^2\mu_{k,\alpha_2}^2|\theta_{j,k}|^2,\\ &\leq \left(\frac{1}{\mu_{D_1+1,\alpha_1}^2}+\frac{1}{\mu_{D_2+1,\alpha_2}^2}\right)Q \leq \left(D_1^{-2\alpha_1}+D_2^{-2\alpha_2}\right)Q,\end{aligned}$$

en notant $Q=L^2/\pi^{2|\alpha|}$.

□

**Convolution et approximation de l'unité**

Considérons l'estimation d'une fonction $s$, à l'aide d'un noyau $K$, avec un paramètre de lissage (ou fenêtre) $h > 0$. En notant classiquement $K_h(x) = K(x/h)/h$, le terme de biais associé au risque quadratique est généralement égal à $\|s - K_h \star s\|_{L^2(\mathbb{R})}$. La famille de fonction $(K_h)_{h>0}$ étant une approximation de l'unité pour le produit de convolution, cette quantité tend vers 0 dans $L^2$ quand $h$ tend vers 0 (voir Briane & Pagès 2006, Théorème 13.5 p.270). On peut être plus précis, et quantifier la vitesse de décroissance en fonction de l'indice de régularité $\alpha$ de la fonction $s$, dans un espace de Nikols'kiĭ, à condition que le noyau soit suffisamment régulier. Rappelons la définition suivante :

**Definition 2.1.** *Un noyau $K$ est dit d'ordre $l$ si les fonctions $x \mapsto x^j K(x)$ sont intégrables, pour $j = 1, \dots, l$, et vérifient*

$$\int_{\mathbb{R}} x^j K(x) dx = 0, \quad j = 1, \dots, l.$$

Tsybakov (2009) donne des exemples de tels noyaux (voir Section 1.2.2 p.10). Une construction plus générale est également suggérée dans Kerkyacharian *et al.* (2001).

On peut alors énoncer le résultat suivant, démontré dans Tsybakov (2009) (Proposition 1.5 p.13) :

**Proposition 2.11.** *Soit $s \in \mathcal{N}(\alpha, L)$, pour $\alpha, L > 0$. Soit $K$ un noyau d'ordre $l = \lfloor \alpha \rfloor$, tel que $\int_{\mathbb{R}} |x|^\alpha |K(x)| dx < \infty$. Alors, l'erreur d'approximation commise en approchant $s$ par $s \star K_h$ est majorée par*

$$\|s - K_h \star s\|_{L^2(\mathbb{R})} \leq \frac{L}{l!} \int_{\mathbb{R}} |x|^\alpha |K(x)| dx h^\alpha.$$

Un autre cas intéressant et dont nous aurons besoin est celui où le biais s'écrit sous la forme $\|s\mathbf{1}_A - K_h \star s\mathbf{1}_A\|_{L^2(A)}$, avec $A$ un intervalle borné de $\mathbb{R}$. On peut alors prouver un résultat similaire, sous une hypothèse de régularité légèrement différente. C'est sous cette forme que ce résultat sera énoncé et démontré au Chapitre 4 (voir le Corollaire 4.1).

### 2.2.3 Autres résultats d'analyse utiles

On rappelle enfin dans cette section deux résultats techniques d'intégration, qui seront fréquemment utilisés au Chapitre 4.

**Proposition 2.12.** *(Inégalité de Minkowski généralisée) Pour toute fonction réelle $t$ intégrable sur $\mathbb{R}^2$,*

$$\int_{\mathbb{R}} \left( \int_{\mathbb{R}} t(x, y) dx \right)^2 dy \leq \left\{ \int_{\mathbb{R}} \left( \int_{\mathbb{R}} t^2(x, y) dy \right)^{1/2} dx \right\}^2.$$

On pourra trouver une démonstration de cette inégalité dans Tsybakov (2009) (Lemme A.1, p.191).

**Proposition 2.13.** *(Inégalité de Young) Soient $p, q \in [1; \infty)$ tels que $1/p + 1/q \geq 1$. Si $s \in L^p(\mathbb{R})$ et $t \in L^q(\mathbb{R})$, alors $s$ et $t$ sont convolables. De plus, si l'on pose $1/r = 1/p + 1/q - 1$ alors $f * g \in L^r(\mathbb{R})$ et*

$$\|s \star t\|_{L^r(\mathbb{R})} \leq \|s\|_{L^p(\mathbb{R})} \|t\|_{L^q(\mathbb{R})}. \tag{2.17}$$

Cette inégalité de convolution est démontrée dans Hirsch & Lacombe (1997) (Théorème 3.4 p.149).

# Première partie

# Estimation non paramétrique par déformation

# Chapitre 3

# Bases déformées pour l'estimation adaptative d'une fonction d'une variable réelle. Exemple détaillé de la régression additive.

**Sommaire**

Une partie de ce chapitre est une version modifiée de l'article *Penalization versus Goldenshluger-Lepski strategies in warped bases regression*, publié dans la revue *Esaim P&S.*, Vol. **17**, p.328-358 (2013).

**Résumé.** L'objectif principal de ce chapitre est de construire deux nouveaux estimateurs pour le problème classique de l'estimation d'une fonction de régression, dans le cas où le design est aléatoire. La procédure utilisée est la sélection de modèles, par pénalisation d'une part et par méthode de Goldenshluger-Lepski d'autre part : on considère des estimateurs par projection fondés sur le développement de la fonction à reconstruire dans des bases orthonormées de $L^2$. Il s'agit d'estimer un nombre fini des coefficients de ces développements, nombre déterminé ensuite de manière adaptative en utilisant un critère pénalisé ou un critère fondé sur les travaux de Goldenshluger & Lepski (2011a). La nouveauté des estimateurs construits réside dans l'utilisation de bases déformées par la fonction de répartition du design ou un estimateur de celle-ci, comme proposé par Kerkyacharian & Picard (2004), à la place des bases usuelles, pour éviter les problèmes liés au calcul des coefficients estimés (inversion de matrice dans le cas d'un estimateur des moindres carrés par exemple). Ceci permet d'obtenir des estimateurs dont les expressions sont explicites, qui sont facilement implémentables et pour lequels on démontre des inégalités-oracle prouvant que les estimateurs construits vérifient automatiquement le compromis biais-variance. Des vitesses de convergence sont également déduites. Des simulations permettent d'illustrer les résultats obtenus, de comparer les performances des deux estimateurs construits entre eux et avec des estimateurs des moindres carrés.

**Abstract.** This chapter mainly deals with the problem of estimating a regression function $s$, in a random design framework. We build two adaptive estimators based on model selection, and the use of warped bases. We start with a collection of finite dimensional linear spaces, spanned by orthonormal bases. Instead of expanding directly the target function $s$ on these bases, we rather consider the expansion of $g = s \circ F_X^{-1}$, where $F_X$ is the cumulative distribution function of the design, following Kerkyacharian & Picard (2004). The data-driven selection of the (best) space is done with two strategies: we use a penalization version of a "warped contrast", and a model selection device in the spirit of Goldenshluger & Lepski (2011a). Thus, we propose two functions, $\hat{g}_l$ $(l = 1, 2)$, easier to compute than least-squares estimators. We establish non-asymptotic mean-squared integrated risk bounds for the resulting estimators, $\hat{s}_l = \hat{g}_l \circ F_X$ if $F_X$ is known, or $\hat{s}_l = \hat{g}_l \circ \hat{F}_X$ otherwise, where $\hat{F}_X$ is the empirical distribution function. We also derive rates of convergence for the risk, and compare the theoretical and practical performances of the two selection rules.

## 3.1 Introduction

### 3.1.1 Statistical framework

Consider the observation sample $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ $(n \in \mathbb{N}\backslash\{0\})$ of couples of real random variables following the regression setting,

$$Y_i = s(X_i) + \varepsilon_i, 1 \leq i \leq n, \tag{3.1}$$

where $s : (a; b) \subset \mathbb{R} \to \mathbb{R}$ is the unknown function that we aim at recovering. The random variables $(\varepsilon_i)_{i\in\{1,\dots,n\}}$ are unobserved, centered, admitting a finite variance $\sigma^2$, and independent of the design $(X_i)_{i\in\{1,\dots,n\}}$. We assume that the latter are distributed with a density $f_X > 0$ with respect to the Lebesgue measure, supported on an interval $(a; b)$, $-\infty \leq a < b \leq +\infty$. We denote by $F_X$ the associated cumulative distribution function (c.d.f. in the sequel), and $F_X^{-1}$ its inverse, which exists thanks to the assumption $f_X > 0$.
The aim of this chapter is twofold: first, taking advantage of warped bases, we want to provide an adaptive non parametric strategy to recover the regression function $s$. Secondly, considering a new development of model selection theory, we are interested in the comparison of two selection strategies, from both theoretical and practical points of view: a classical penalization method and a recent selection device in the spirit of Goldenshluger & Lepski (2011a) (shortened by "GL method" in the sequel), applied in an original way to a projection estimator.

### 3.1.2 Motivation

Adaptive estimation of the regression function is a well-developed problem, and several procedures have been set. Historical methods are kernel strategies, initiated by Nadaraya (1964) and Watson (1964) who proposed kernel-type estimators, built as the ratio of an estimator of the product $sf_X$ divided by an estimator of the density $f_X$. The data-driven choice of the bandwidth, leading to adaptive estimators, was studied more accurately for example by Fan & Gijbels (1992) and Härdle & Tsybakov (1997), who provided asymptotic results (for methods also involving local polynomials). Nevertheless, estimators resulting of this strategy have the drawback of involving a ratio, with a denominator that can be small: this implies difficulties to study the risk and to implement the method.
In a different direction, estimators based on the expansion of the target function into bases, especially orthogonal-bases, have been proposed: spline bases (Golubev & Nussbaum 1992), wavelet bases (Donoho *et al.* 1995, Cai & Brown 1998 in the fixed design case, Antoniadis *et al.* 1997 in the random-design case), and also trigonometric bases (Efromovich 1999). Wavelet thresholding strategies offer a degree of localization leading to almost minimax but asymptotic rate of convergence (except in Donoho & Johnstone 1994). To obtain non-asymptotic risk bounds, all these estimators can be studied from the model selection point of view, initiated among others by Barron *et al.* (1999). The problem is to select a "best" estimator among a collection of projection estimators, for example least-squares estimators, to prove oracle inequalities for the risk. The selection is standardly done by the minimization of a penalized criterion (see for example Kohler & Krzyżak 2001, Wegkamp 2003, Birgé 2004, Baraud 2002).

But procedures based on the minimization of a least-squares contrast do not provide explicit estimators without matrix invertibility requirements (most of the time implicitly).

### 3.1.3 Estimation strategy

Adopting this model selection point of view, and using warped bases developed for building wavelet thresholding estimators by Kerkyacharian & Picard (2004), we provide adaptive estimators in this chapter. These estimates still satisfy non-asymptotic oracle-bounds and reach the exact optimal rate under mild assumptions while being easier to compute and more stable, even in case the amount of data can vary in the estimation domain. More precisely, denoting by $u \circ v$ the composition of functions $u$ and $v$, we define

$$g = s \circ F_X^{-1} = s\left(F_X^{-1}\right). \tag{3.2}$$

We assume that $g$ is squared integrable, we provide estimators for $g$ of the form

$$\hat{g}_D = \sum_{j=1}^{D} \hat{a}_j \varphi_j,$$

for a collection of possible $D$, with $(\varphi_j)_j$ a classical orthonormal family, and $\hat{a}_j$ an estimator of the scalar product $\langle g, \varphi_j \rangle$. Then we define

$$\hat{s}_D = \hat{g}_D \circ F_X \text{ or } \hat{s}_D = \hat{g}_D \circ \hat{F}_X,$$

as estimators of $s$, depending on whether we assume that $f_X$ is known or not (in this last case, $\hat{F}$ is the empirical distribution function). We get thus a development of the estimator in warped bases, that is,

$$\hat{s}_D = \sum_{j=1}^{D} \hat{a}_j \left(\varphi_j \circ F_X\right), \text{ or } \hat{s}_D = \sum_{j=1}^{D} \hat{a}_j \left(\varphi_j \circ \hat{F}_X\right).$$

The warping strategy brings a procedure computationally simple, without any matrix inversion (which are costly from the practical point of view), contrary to the projection estimators obtained with the classical least-squares criterion (see (1.14) in Chap. 1 and comments below). Developed first by Kerkyacharian & Picard (2004), the warped-bases have then been used in the regression setting by Pham Ngoc (2009) to build Bayesian estimators, and by Kulik & Raimondo (2009) to tackle the case of correlated errors. Our work focus on the adaptive selection of the "best" index $\hat{D}$ among all possible $D$. It is done in a second time with two strategies. First, we use a penalized version of a "warped contrast". Next, recent works of Goldenshluger & Lepski (2011a), in the case of density estimation can be explored to propose a new selection strategy. Thus we have at hand two data-driven estimators of the unknown function.

We prove that they both automatically realize the usual squared-bias/variance compromise, provide non-asymptotic oracle-inequalities for each estimator. We also give asymptotic rate of convergence on functional spaces, of Besov or Sobolev type. We find the classical non-parametric estimation rate, that is $n^{-2\alpha/(2\alpha+1)}$ where $\alpha$ is the regularity index. Thus, the

equivalence between the two adaptive estimators - one based on penalization, the other on the GL method - is obtained from the theoretical point of view. However, on our practical examples, the new GL strategy seems to outperform the penalization device.

### 3.1.4 Organisation of the chapter

We begin with the case of known design c.d.f in Section 3.2. In this simpler framework, we can easily explain how the estimators are built and state their adaptivity, while the general case of unknown design distribution is the subject of Section 3.3: it requires further technicalities, but similar results are proved. They are illustrated via simulations in Section 3.4. The proofs are gathered in Section 3.5. Finally, the work is completed by two appendices: the first one (Section 3.6) gives further information about the practical calibration of the constants that are involved in the selection devices. In the second one (Section 3.7), an oracle-type inequality is proved for the case of a random penalty.

## 3.2 Case of known design c.d.f.

To have a better understanding of the definition and properties of the estimators in the general case, we first focus on the simpler situation of known design distribution. This "toy-case", also used by other authors (see for example Pham Ngoc 2009) allows us to derive very simple results, with few assumptions and short proofs.
We first deal with the estimation of the function $g$ defined by (3.2). We consider a family of approximation spaces. In a first step, we estimate $g$ or more precisely its projection on these spaces. The second step is to ensure an automatic selection of the space, without any knowledge on $s$. Finally, we warp the function to estimate $s$ (and not $g$).

### 3.2.1 Assumptions on the models

The models are linear spaces of functions included in $L^2((0;1))$, the set of square-integrable real-valued functions on the interval $(0;1)$, equipped with $\|.\|$, the usual Hilbert norm, associated to the scalar-product $\langle .,. \rangle$. We denote the collection $\{S_m,\ m \in \mathcal{M}_n\}$, where $\mathcal{M}_n$ is a finite set of indices, with cardinality depending on the number of observations $n$. The assumptions and notations are the following:

- [$\mathcal{M}_1$ ] All the linear spaces $S_m$ are finite-dimensional. For all $m \in \mathcal{M}_n$, we denote by $D_m$ the dimension of the space $S_m$ and assume $1 \leq D_m \leq n$.
- [$\mathcal{M}_2$ ] The models are nested, that is, for all $(m_1, m_2) \in \mathcal{M}_n^2$, such that $D_{m_1} \leq D_{m_2}$, $S_{m_1} \subset S_{m_2}$. We denote by $(\varphi_j)_{j \in \{1,\dots,D_m\}}$ an orthonormal basis which spans $S_m$ ($m \in \mathcal{M}_n$), and by $m_{\max}$ the index of the largest model in the collection.
- [$\mathcal{M}_3$ ] There exists a positive constant $\phi_0$ such that for all indices $m \in \mathcal{M}_n$ and all functions $t \in S_m$, $\|t\|_{L^\infty((0;1))} \leq \phi_0 \sqrt{D_m} \|t\|$. This useful link between the $L^2$ norm and the infinite norm is equivalent to a property of the basis $(\varphi_j)_{j \in \{1,\dots,D_m\}}$: $\| \sum_{j=1}^{D_m} \varphi_j^2 \|_{L^\infty((0;1))} \leq \phi_0^2 D_m$. See Birgé & Massart (1998) for the proof of the equivalence.

The above assumptions are not rather weak. Indeed, they are verified by the spaces spanned by usual bases: trigonometric basis, regular compactly supported wavelet basis,

regular histogram basis and regular polynomial basis (with dyadic subdivisions in the last two examples). We refer to Section 3.3.2 for a description of trigonometric models, and to Barron *et al.* (1999), and Brunel & Comte (2005) for the other examples.

### 3.2.2 Estimation on a fixed model

#### Contrast and estimator on one model

We define the contrast function:

$$\gamma_n(t, F_X) := \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i \left(t \circ F_X(X_i)\right), \quad t \in L^2((0;1)). \tag{3.3}$$

Notice that $\gamma_n(., F_X)$ represents an empirical counterpart for the quadratic risk: for all $t \in L^2((0;1))$,

$$\begin{aligned}
\mathbb{E}\left[\gamma_n(t, F_X)\right] - \mathbb{E}\left[\gamma_n(g, F_X)\right] &= \|t\|^2 - \|g\|^2 - 2\mathbb{E}\left[s(X_1)\left\{(t-g)\circ F_X\right\}(X_1)\right], \\
&= \|t\|^2 - \|g\|^2 - 2\int_{(a;b)} s(x)\left\{(t-g)\circ F_X\right\}(x) f_X(x)dx, \\
&= \|t\|^2 - \|g\|^2 - 2\int_{(0;1)} g(u)(t-g)(u)du, \\
&= \|t\|^2 - \|g\|^2 - 2\langle g, t-g\rangle, \\
&= \|t-g\|^2,
\end{aligned}$$

so that $g$ minimizes $t \mapsto \mathbb{E}[\gamma_n(t, F_X)]$ over $L^2((0;1))$. This explains why a relevant strategy to estimate $g$ consists in minimizing $\gamma_n(., F_X)$ over each set $S_m$:

$$\hat{g}_m^{F_X} = \arg\min_{t \in S_m} \gamma_n(t, F_X). \tag{3.4}$$

The unique resulting estimator (for each index $m$) has a particularly simple expression,

$$\hat{g}_m^{F_X} = \sum_{j=1}^{D_m} \hat{a}_j^{F_X} \varphi_j, \text{ with } \forall j \in \{1, \ldots, D_m\}, \hat{a}_j^{F_X} = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j(F_X(X_i)). \tag{3.5}$$

Finally, we set

$$\hat{s}_m^{F_X, F_X} = \hat{g}_m^{F_X} \circ F_X$$

as an estimator of $s$. The explicit formula (3.5) is an unbiased estimator of the orthogonal projection of $g$ onto $S_m$. Compare for example to the classical least-squares estimator, which involves a matrix inversion (see Baraud 2002 and Section 3.4 for details). Notice that the notation for the estimator involves two super-indices $f_X$ to underline the dependence on the c.d.f. $f_X$ through both the coefficient $\hat{a}_j^{F_X}$ and the composition by $f_X$.

**Risk on one model**

In this section, we fix a model $S_m$ and briefly study the quadratic risk of the estimator $\hat{s}_m^{F_X,F_X}$. As for all the results stated in the sequel, we evaluate the risk with respect to the norm $\|.\|_{f_X}$ naturally associated to our estimation procedure:

$$\|v\|_{f_X}^2 = \int_{(a;b)} v^2(x) f_X(x) dx, \quad \langle v, w \rangle_{f_X} = \int_{(a;b)} v(x) w(x) f_X(x) dx,$$

for any functions $v, w \in L^2((a;b), f_X)$, the space of squared-integrable functions on $(a;b)$ with respect to the Lebesgue measure weighted by the density $f_X$. However, it is also possible to control the classical $L^2$ norm on $(a;b)$, under the assumption that $f_X$ is bounded from below by a strictly positive constant: if, for any $x \in (a;b)$, $f_X(x) > f_0 > 0$, then

$$\|v\|_{f_X}^2 \geq f_0 \int_{(a;b)} v^2(x) dx.$$

Notice besides that the following links hold between this weighted norm and the classical norm on $L^2((0;1))$ previously defined: for $t, s \in L^2((0;1))$, we compute, using $F_X' = f_X$,

$$\|t \circ F_X\|_{f_X} = \|t\|, \quad \langle t \circ F_X, s \circ F_X \rangle_{f_X} = \langle t, s \rangle.$$

Thus, the quadratic risk of $\hat{s}_m^{F_X,F_X}$ is given by

$$\begin{aligned} \mathbb{E}\left[\left\|\hat{s}_m^{F_X,F_X} - s\right\|_{f_X}^2\right] &= \left\|s - s_m^{F_X}\right\|_{f_X}^2 + \mathbb{E}\left[\left\|s_m^{F_X} - \hat{s}_m^{F_X,F_X}\right\|_{f_X}^2\right], \\ &= \|g - g_m\|^2 + \mathbb{E}\left[\left\|g_m - \hat{g}_m^{F_X}\right\|^2\right], \end{aligned} \tag{3.6}$$

where

$$s_m^{F_X} = g_m \circ F_X \text{ and } g_m \text{ is the orthogonal projection of } h \text{ onto } S_m. \tag{3.7}$$

Hence, we recover the usual decomposition into two terms: a squared bias term, which decreases when the dimension of the model $S_m$ grows (roughly, it is at most of order $D_m^{-2\alpha}$, where $\alpha$ is the index of smoothness of $g$), and a variance term, proportional to the dimension of the model $S_m$:

$$\begin{aligned} \mathbb{E}\left[\left\|s_m^{F_X} - \hat{s}_m^{F_X,F_X}\right\|_{f_X}^2\right] &= \sum_{j=1}^{D_m} \mathrm{Var}\left(\hat{a}_j^{F_X}\right), \\ &= \sum_{j=1}^{D_m} \frac{1}{n} \mathrm{Var}\left(Y_1 \left(\varphi_j \circ F_X\right)(X_1)\right) \leq \mathbb{E}\left[Y_1^2\right] \phi_0^2 \frac{D_m}{n}, \end{aligned} \tag{3.8}$$

where $\phi_0^2$ is defined in Assumption $[\mathcal{M}_3]$ (see Section 3.2.1).
Consequently, the best estimator among the family $(\hat{s}_m^{F_X,F_X})_{m \in \mathcal{M}_n}$ (in the sense that it achieves the smallest risk among the collection) is the one which realizes the trade-off between the two terms, without any knowledge of the index of smoothness $\alpha$.

### 3.2.3 Selection rules and main results

**Selection rules**

The aim is to realize a data-driven selection of the space $S_m$. For that purpose, we give a strategy to choose an estimator among the collection $(\hat{s}_m^{F_X,F_X})_{m\in\mathcal{M}_n}$. We propose two different strategies and build consequently two estimators.
First, the selection can be standardly done by

$$\hat{m}^{(1),F_X} = \arg\min_{m\in\mathcal{M}_n} \left[\gamma_n(\hat{g}_m^{F_X}, F_X) + \mathrm{pen}^{F_X}(m)\right], \tag{3.9}$$

with $\mathrm{pen}^{F_X}(.)$ a function to be properly chosen. As, $\gamma_n(\hat{g}_m^{F_X}, F_X) = -\|\hat{g}_m^{F_X}\|^2 = -\|\hat{s}_m^{F_X,F_X}\|_{f_X}^2$, and $\|g - g_m\|^2 = \|g\|^2 - \|g_m\|^2$, we can say that $\gamma_n(\hat{g}_m^{F_X}, F_X)$ estimates the bias term, up to an additive constant. This explains why the order of the penalty can be the upper bound on the variance term, that is

$$\mathrm{pen}^{F_X} : m \mapsto c_1\phi_0^2\mathbb{E}[Y_1^2]\frac{D_m}{n}, \tag{3.10}$$

with $c_1$ a purely numerical constant. In practice, we use a method inspired by the slope heuristic to find the value of this constant (see Section 3.4).
The second method follows the scheme developed by Goldenshluger & Lepski (2011a) for density estimation. The adaptive index is also chosen as the value which minimizes a sum of two terms:

$$\hat{m}^{(2),F_X} = \arg\min_{m\in\mathcal{M}_n} \left[A^{F_X}(m) + V^{F_X}(m)\right],$$

where $V^{F_X}$ is also the order of the variance term:

$$V^{F_X} : m \mapsto c_2\phi_0^2\mathbb{E}[Y_1^2]\frac{D_m}{n}, \tag{3.11}$$

with $c_2$ a purely numerical constant (adjusted in practice by simulations, see Appendix 1, Section 3.6). Here the function $A^{F_X}$ does not depend on the contrast: it is rather based on the comparison of the estimators built in the first stage:

$$A^{F_X}(m) = \max_{m'\in\mathcal{M}_n}\left(\left\|\hat{g}_{m'}^{F_X} - \hat{g}_{m\wedge m'}^{F_X}\right\|^2 - V^{F_X}(m')\right)_+, \tag{3.12}$$

where $x_+ = \max(x, 0)$, $x \in \mathbb{R}$. We will prove besides that $A^{F_X}(m)$ has the order of the bias term (see Lemma 3.2). Thus we get two estimators, explicitly expressed in a warped basis:

$$\tilde{s}_1^{F_X} = \hat{g}_{\hat{m}^{(1),G}}^{F_X} \circ F_X, \quad \tilde{s}_2^{F_X} = \hat{g}_{\hat{m}^{(2),F_X}}^{F_X} \circ F_X.$$

We stress out the fact that these estimators are simple to compute: their coefficients $\hat{a}_j^{F_X}$ are empirical means, and even if the "penalties" ($\mathrm{pen}^{F_X}$ and $V^{F_X}$) contain the unknown expectation $\mathbb{E}[Y_1^2]$, this term can easily be replaced in practice or theory by the empirical mean $(1/n)\sum_{i=1}^n Y_i^2$ (see Brunel & Comte 2005, proof of Theorem 3.4 p.465, and also Appendix 2, Section 3.7).
In addition to the advantage of the warped basis, the comparison of these two estimators, from both theoretical and practical point of view is new, and is of interest also for other statistical estimation problems.

### Oracle-type inequality

The first theorem provides non-asymptotic bounds for the risk of each estimator.

**Theorem 3.1.** *We assume that the regression function $s$ is bounded on the interval $(a;b)$. We consider models satisfying properties $[\mathcal{M}_1]$, $[\mathcal{M}_2]$ and $[\mathcal{M}_3]$, and finally suppose that there exists a real-number $p > 4$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right] < \infty$.*
*Then, the following inequality holds, for $i = 1, 2$,*

$$\mathbb{E}\left[\left\|\tilde{s}_i^{F_X} - s\right\|_{f_X}^2\right] \leq \min_{m\in\mathcal{M}_n}\left\{k_i\left\|s - s_m^{F_X}\right\|_{f_X}^2 + k_i'\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}\right\} + \frac{C_i}{n}, \qquad (3.13)$$

*where $s_m^{F_X}$ is defined by (3.7), $k_i$ and $k_i'$, ($i = 1, 2$) are numerical constants, and $C_i$ ($i = 1, 2$) are constants independent of $n$ and $m$, but depending on $\mathbb{E}[Y_1^2]$, $\phi_0^2$, $\sigma^2$, $\mathbb{E}[|\varepsilon_1|^{2+p}]$ and $\|s\|_{L^\infty((a;b))}$.*

Let us comment this result.

- These non-asymptotic risk bounds, also called oracle-type inequalities prove that both estimators automatically realize the squared bias/variance trade-off under few weak assumptions, up to some multiplicative constants (which are precised in the proof). This enhances the interest of warped bases: the risk of the estimators is smaller (up to the constant) than the risk of the best estimator in the family $(\hat{s}_m^{F_X,F_X})_m$. Moreover, the two estimators (the one selected by the GL method and the one selected by penalization) are theoretically equivalent in this context.
- Note that the assumptions for this result are particularly weak, compared to usual hypotheses in other statistical framework ($D_m$ in only supposed bounded by $n$). Moreover the proof is short, following the general setting of model selection methods (see for example Birgé & Massart 1998): it is mainly based on a concentration inequality due to Talagrand. The details can be found in Section 3.5. Remark also that the choice of $p = 4$ for the integrability of $\varepsilon_1$ (instead of $p > 4$) leads to the same inequality with a remainder of order $\ln^4(n)/n$ (instead of $1/n$). We can still relax this assumption: a moment of order $2 + p$, $p > 2$ for $\varepsilon_1$ is enough, if we suppose in compensation $D_m = O(\sqrt{n})$. These moment conditions may probably be improved, but we do not go further in this direction, to avoid additional technicalities. We also point out the fact that other results in regression model hold under weak conditions on the noise term (in the sense that no exponential moment for the $\varepsilon_i$ are required, contrary to the conditions in Barron *et al.* 1999): see for example recent works of Audibert & Catoni (2011a,b), in a prediction framework, and works of Wegkamp (2003) or Baraud (2002) for the model selection point of view.
- A result similar to Theorem 3.1 is obtained when the unknown expectation $\mathbb{E}[Y_1^2]$ involved in $\text{pen}^{F_X}$ and $V^{F_X}$ is replaced by the empirical mean $(1/n)\sum_{i=1}^n Y_i^2$. See Appendix 2, Section 3.7.

### Rate of convergence for the risk

Even if the novelty of our results is their non-asymptotic characters (compared to other warped-bases estimators in this framework, see for example Kerkyacharian & Picard 2004

and Pham Ngoc 2009), we can also deduce from Theorem 3.1 the rate of convergence of the risk. For that purpose, assume that $g = s \circ F_X^{-1}$ belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}$, for $\alpha$ a positive number.
Let us recall the definition of this space. First, for $r$ a positive integer and $v$ a positive number, the $r-$th order difference of a real-valued function $t$ on the interval $(0;1)$ is defined by

$$\Delta_v^r t(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} t(x+kv),$$

where $x$ is such that the $x+kv$ belongs to $(0;1)$, $k \in \{0, \dots r\}$. Next, for $u > 0$, the modulus of smoothness is given by $\omega_r(t,u)_2 = \sup_{0<v\leq u} \|\Delta_v^r t\|$. We say that the function $t$ belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}$ if $t$ belongs to the space $L^2((0;1))$ and if, for $r = \lfloor \alpha \rfloor + 1$ ($\lfloor . \rfloor$ is the integer part function), $|t|_{\mathcal{B}^{\alpha}_{2,\infty}} = \sup_{u>0} u^{-\alpha} \omega_r(t,u)_2 < \infty$. We refer to DeVore & Lorentz (1993) for general definitions and properties of this space. Finally, for all $L > 0$, we denote by $\mathcal{B}^{\alpha}_{2,\infty}(L)$ the space of functions $t$ which satisfies: $|t|_{\mathcal{B}^{\alpha}_{2,\infty}} \leq L$. Details are also given in Chapter 2.
It is well known that for all the collections of models described in Section 3.2.1 (trigonometric models, regular polynomial bases, regular and compactly supported wavelet bases), the projection $g_m$ of $g$ on the subspace $S_m$ achieves the rate of approximation for the Besov class of functions $\mathcal{B}^{\alpha}_{2,\infty}(L)$ (see Lemma 12, Barron *et al.* 1999 and Proposition 2.10, Chapter 2):

$$\|g - g_m\|^2 \leq C(\alpha) L^2 D_m^{-2\alpha}, \tag{3.14}$$

where $C(\alpha)$ is a constant depending on $\alpha$ and also on the basis. Therefore, the minimization of the left side of Inequality (3.13) leads to the following corollary:

**Corollary 3.1.** *We suppose that the function $g = s \circ F_X^{-1}$ belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}(L)$, for some fixed $\alpha > 0$ and $L > 0$. We also assume that $g$ is bounded over the interval $(0;1)$. We consider one of the models defined in Section 3.2.1: trigonometric model, dyadic piecewise polynomials (with a regularity $r$ such that $r \geq \alpha - 1$) or compactly supported regular wavelets. Then, under the assumptions of Theorem 3.1,*

$$\mathbb{E}\left[ \left\| \tilde{s}_i^{F_X} - s \right\|_{f_X}^2 \right] \leq C(L,\alpha) n^{\frac{-2\alpha}{2\alpha+1}}, \ i = 1, 2,$$

*with $C(L,\alpha)$ a numerical constant which only depends on $L$ and $\alpha$.*

Thus, the model selection procedure not only leads to a non-asymptotic squared bias/variance trade-off but also to an adaptive estimator: indeed, it automatically reaches the asymptotic rate of order $n^{-2\alpha/(2\alpha+1)}$, the minimax rate, in regression setting.
Theorem 2 in Kerkyacharian & Picard (2004) states a rate $(n/\ln(n))^{-2\alpha/(2\alpha+1)}$ for an estimator obtained in the same framework ($f_X$ known, warped basis) by a thresholding algorithm on wavelet coefficients: thus, the rate we get does not suffer from a loss of a $\ln(n)$ factor. Therefore, our method provides an improvement. Moreover, Theorem 3.1 and Corollary 3.1 are valid for several models (wavelets models, but also trigonometric models...) and, contrary to Kerkyacharian & Picard (2004), for a noise $\varepsilon_1$ not necessarily Gaussian (only weak integrability assumptions are required).

Notice also that the assumptions in Corollary 3.1 are set on the function $g = s \circ F_X^{-1}$, like Proposition 2 of Kerkyacharian & Picard (2004). Proper regularity conditions on function $s$ can also be used to get the same asymptotic result, by defining "weighted" Besov spaces. We refer to Section 4.3 in Kerkyacharian & Picard (2004) in which such spaces are precisely described and their properties stated.

## 3.3 Case of unknown design c.d.f.

### 3.3.1 Estimator on a fixed model

The obvious question resulting of Section 3.2 is: what is to be done if the c.d.f. is not known? To adapt the previous estimation procedure, we replace $F_X$ by its empirical counterpart. But instead of estimating $F_X$ over the whole sample, we assume that we observe $(X_{-i})_{i\in\{1,\dots,n\}}$, a sample of random variables distributed as the $(X_i)_i$, and independent of them, and we define,

$$\hat{F}_n : \ x \ \mapsto \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_{-i}\leq x}.$$

This sample splitting simplifies the proofs. We propose a simple "plug-in" strategy to define the estimators. First, for each index $m \in \mathcal{M}_n$, we set

$$\hat{g}_m^{\hat{F}} = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{F}} \varphi_j, \text{ with } \forall j \in \{1,\dots,D_m\}, \ \hat{a}_j^{\hat{F}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j\left(\hat{F}_n(X_i)\right), \tag{3.15}$$

which is the minimizer of the contrast function $t \mapsto \gamma_n(t, \hat{F}_n)$ on $S_m$ (see (3.3)). Note that the $\hat{g}_m^{\hat{F}}$, $m \in \mathcal{M}_n$, are still easily available for the statistician, like the estimators of $s$: $\hat{s}_m^{\hat{F},\hat{F}} = \hat{g}_m^{\hat{F}} \circ \hat{F}_n$.

### 3.3.2 Risk on one model

We propose to state an upper-bound for the risk of these estimators, which is similar to the bias-variance decomposition and upper-bound obtained for $\hat{s}_m^{F_X,F_X}$, $m \in \mathcal{M}_n$, in Inequalities (3.6) and (3.8). The challenge comes from the plug-in of the empirical $\hat{F}_n$: it seems natural but involves non straightforward computations. For instance, it requires control of terms of the form $\varphi_j(\hat{F}_n) - \varphi_j(F_X)$. That is why we select one of the bases only, and not any of the ones used in Section 3.2. Following the example of Efromovich (1999), we use models based on the trigonometric basis, that is $S_m = \text{Span}\{\varphi_1,\dots,\varphi_{D_m}\}$, with $D_m = 2m+1$, $m \in \mathcal{M}_n = \{1,\dots,\lfloor n/2\rfloor - 1\}$, and for all $j \in \{1,\dots,m\}$ and all $x \in (0;1)$,

$$\varphi_1(x) = 1, \ \ \varphi_{2j}(x) = \sqrt{2}\cos(2\pi j x), \ \ \varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi j x).$$

Notice that the assumption $[\mathcal{M}_3]$ is satisfied with $\phi_0 = 1$. This choice is guided among other things by the following property: let $\xi$ be a function continuously derivable on the interval $(0;1)$, such that $\xi(0) = \xi(1)$. The orthogonal projection of the derivative $\xi'$ of $\xi$ onto $S_m$ coincides with the derivative of the projection of $\xi$ onto $S_m$. Formally, if we denote by

$\Pi_{S_m}$ the operator of orthogonal projection onto $S_m$, $\Pi_{S_m}(\xi') = (\Pi_{S_m}(\xi))'$ (see Proposition 2.9, Chapter 2).

The proof of the following result is thus postponed to Section 3.5.

**Proposition 3.1.** *We assume that the regression function $s$ and the density $f_X$ admit both continuous derivative on $[a;b]$ (respectively $[0;1]$). We also assume that $\|s\|_{f_X} \leq L$ ($L > 0$) and that $s(a) = s(b)$. We consider the trigonometric models, and suppose that for any $m \in \mathcal{M}_n$, $D_m = O(n^{1/3}/\ln^{1/3}(n))$.*
*Then, the following inequality holds,*

$$\mathbb{E}\left[\left\|\hat{s}_m^{\hat{F},\hat{F}} - s\right\|_{f_X}^2\right] \leq 6\left\|s - s_m^{F_X}\right\|_{f_X}^2 + C\frac{D_m}{n},$$

*where $s_m^{F_X}$ is defined by (3.7), and $C$ is a constant independent on $n$ and $m$, but depending on $\|\varphi_2^{(l)}\|$ ($l = 1, 3$), $\|g\|$, $\|g'\|$, and $\mathbb{E}[Y_1^2]$.*

The assumption on the model dimension, that is $D_m = O(n^{1/3}/\ln^{1/3}(n))$, is required to handle the replacement of the c.d.f. $F_X$ by its empirical counterpart (it is not necessary when $F_X$ is known, see e.g. (3.8)). The result of Proposition 3.1 is actually obtained by computing upper-bound for terms of the form $\sum_{j=1}^{D_m}(\varphi_j(\hat{F}_n) - \varphi_j(F_X))$: this can be done thanks to Taylor developments on the function $\varphi_j$. This leads to quantities which look like $\|\sum_{j=1}^{D_m}\varphi_j^{(k)}\|_{L^\infty((0;1))}\|\hat{F}_n - F_X\|_{L^\infty((a;b))}$ ($k$ is a nonnegative integer), which are bounded by $D_m^k/\sqrt{n}$. To control this by $D_m/n$, we thus have to limit the size of $D_m$. We do not actually know if our assumption $D_m \leq Cn^{1/3}$ (up to logarithmic factor) are optimal: deeper developments in the proofs might improved it, but will required cumbersome computations. We do not go further in this direction since the proofs are already long and technical.

### 3.3.3 Adaptive estimation

#### Selection of the estimators

The selection rules follow exactly the same scheme as previously, and allow us to build two estimators. Define, for each $m \in \mathcal{M}_n$,

$$\begin{aligned}&\text{pen}: \ m \mapsto c_1'\phi_0^2\mathbb{E}[Y_1^2]D_m/n,\\ &V: \ m \mapsto c_2'\phi_0^2\mathbb{E}[Y_1^2]D_m/n, A(m) = \max_{m' \in \mathcal{M}_n}\left(\left\|\hat{g}_{m'}^{\hat{F}} - \hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2 - V(m')\right)_+,\end{aligned} \tag{3.16}$$

with $c_1'$ and $c_2'$ purely numerical constants (adjusted in practice, see Section 3.4), and set

$$\hat{m}^{(1)} = \arg\min_{m \in \mathcal{M}_n}\left[\gamma_n(\hat{g}_m^{\hat{F}}, \hat{F}_n) + \text{pen}(m)\right], \ \hat{m}^{(2)} = \arg\min_{m \in \mathcal{M}_n}\left[A(m) + 2V(m)\right]. \tag{3.17}$$

Finally, the selected estimators are

$$\tilde{s}_1^{\hat{F}} = \hat{g}_{\hat{m}^{(1)}}^{\hat{F}} \circ \hat{F}_n, \ \ \tilde{s}_2^{\hat{F}} = \hat{g}_{\hat{m}^{(2)}}^{\hat{F}} \circ \hat{F}_n. \tag{3.18}$$

### Main result

The goal of this section is to establish adaptive properties for both estimators $\tilde{s}_i^{\hat{F}}$, $i = 1, 2$. As already said, they depend on the empirical c.d.f. $\hat{F}_n$ at two stages, which leads to complexity in the proof. In this framework, we get a similar result to the one obtained when $f_X$ was supposed to be known.

**Theorem 3.2.** *We assume that the regression function $s$ and the density $f_X$ admit both continuous derivative on $[a; b]$. We also assume that $\|s\|_{f_X} \leq L$ ($L > 0$) and that $s(a) = s(b)$. We consider the trigonometric models, and suppose that there exists a real-number $p > 8/3$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right] < \infty$, and that for any $m \in \mathcal{M}_n$, $D_m = O(n^{1/3}/\ln(n))$.*
*Then, the following inequality holds: for all $n \geq n_0 = \exp(\|g'\|^2)$,*

$$\mathbb{E}\left[\left\|\tilde{s}_i^{\hat{F}} - s\right\|_{f_X}^2\right] \leq \min_{m \in \mathcal{M}_n} \left\{ k_i \left\|s - s_m^{F_X}\right\|_{f_X}^2 + k_i' \phi_0^2 \mathbb{E}\left[Y_1^2\right] \frac{D_m}{n} \right\} + \frac{C_i \ln(n)}{n}, \ i = 1, 2, \tag{3.19}$$

*where $s_m^{F_X}$ is defined by (3.7), $k_i$ and $k_i'$, ($i = 1, 2$) are numerical constants, and $C_i$ ($i = 1, 2$) are constants independent on $n$ and $m$, but depending on $\|\varphi_2^{(l)}\|$ ($l = 1, 3$), $\|g\|$, $\|g'\|$, and $\mathbb{E}[Y_1^2]$.*

First, notice that the assumption on the model dimension, $D_m = O(n^{1/3}/\ln(n))$ is similar to the one required in Proposition 3.1: it is thus not a consequence of adaptation (see details above).

Theorem 3.2 proves that the warped-bases selected estimators have exactly the same behaviour as the least-squares estimator (see for instance Inequality (15), in Baraud 2002): both estimators realize the squared bias/variance compromise. Consequently, a model selection strategy with warped-bases has the advantage of providing estimators easier to compute than least-squares estimators and with analogous theoretical properties.
Notice that the upper bound we provide for the risk holds for any $n \geq n_0$ so it can still be considered as a non-asymptotic result. This is an advantage compared to other procedures based on the thresholding of the estimated coefficients in wavelet bases, even if the bases are also warped (see for example Kerkyacharian & Picard 2004).

### Rate of convergence for the risk

As a consequence of the choice of trigonometric models, it is natural to consider spaces of periodic functions, that is Sobolev spaces. Following Tsybakov (2009), we first define, for $\alpha$ a positive integer and $L$ a positive number, the space $W_2^\alpha(L)$ of real-valued functions $g$ on the interval $(0; 1)$ such that $g^{(\alpha-1)}$ is absolutely continuous and

$$\left\|g^{(\alpha)}\right\|^2 = \int_0^1 \left(g^{(\alpha)}(x)\right)^2 dx \leq L^2.$$

Then, we say that a function $g$ belongs to the space $W_2^{\alpha,per}(L)$ if it belongs to $W_2^\alpha(L)$ and

$$\forall j = 0, 1, \ldots, \alpha - 1, \ g^{(j)}(0) = g^{(j)}(1).$$

This definition can be extended to positive real-number $\alpha$ (see Tsybakov 2009, and Section 2.2.1 of Chapter 2 for details).
The standard rate of convergence is then achieved if smoothness properties are supposed for $g$. In fact, the approximation error orders can also be bounded in the case of Sobolev spaces. If $g$ belongs to the space $W_2^{\alpha,per}(L)$ for $\alpha \geq 1$ and $L > 0$, and if we denote by $g_m$ its orthogonal projection (for the usual scalar product of $L^2((0;1))$) on the trigonometric model $S_m$, then Tsybakov (2009) (see Lemma A.3) proves the following inequality:

$$\|g - g_m\|^2 \leq \frac{L^2}{\pi^{2\alpha}} D_m^{-2\alpha}.$$

Consequently, we state the following result, which is similar to Corollary 3.1:

**Corollary 3.2.** *We suppose that the function $g = s \circ F_X^{-1}$ belongs to the Sobolev space $W_2^{\alpha,per}(L)$, for some fixed $\alpha \geq 1$ and $L > 0$. Then, under the assumptions of Theorem 3.2,*

$$\mathbb{E}\left[\left\|\tilde{s}_i^{\hat{F}} - s\right\|_{f_X}^2\right] \leq C(L,\alpha) n^{\frac{-2\alpha}{2\alpha+1}},\ i = 1, 2,$$

*where $C(L,\alpha)$ is a constant which depends on $L$ and $\alpha$.*

Most of the comments following Corollary 3.1 also apply to this result. The order of the rate, $n^{(-2\alpha)/(2\alpha+1)}$ in place of the rate $(n/\ln(n))^{(-2\alpha)/(2\alpha+1)}$ achieved by the estimator $\hat{s}^{@}$ in Kerkyacharian & Picard (2004) is a consequence of the model selection strategy, by penalization or by the GL method. But the assumptions for their result are different of ours. We decide to concentrate on the trigonometric models (instead of the wavelet setting of Kerkyacharian & Picard (2004)). Consequently, the estimators are adaptive for Sobolev regularities. This, and the fact that the index $\alpha$ of regularity has to be larger than 1 can seem to be a little more restrictive than the assumptions of Theorem 2 in Kerkyacharian & Picard (2004): $g$ is there assumed to belong to a Besov space with index $\alpha \geq 1/2$, and to a Hölder space (with regularity 1/2), and these spaces are larger than the one we use. But contrary to them, and in addition to the convergence rate improvement (no additional $\ln(n)$), our methods allow general noise and not only Gaussian noise. Moreover, the trigonometric basis enables to consider other regularities, and to get faster rates. For example, if $g$ belongs to an analytic space, its Fourier's coefficients decrease with exponential rate: $\|g - g_m\| \leq C\exp(-\epsilon D_m)$, for some $\epsilon > 0$ and $C$ a positive constant, leading to the rate $\ln(n)/n$.
Finally, let us notice that assumptions can probably be stated with regularity conditions directly on $s$ instead of $g$, by defining "weighted" spaces. But, as our main contribution is to provide non-asymptotic results which do not require the control of the bias term (and thus, the regularity assumption), this construction is beyond the scope of the chapter.

## 3.4 Simulations

### 3.4.1 Implementation

The simulation study is mainly conducted in order to compare from the practical point of view the penalized estimator $\tilde{s}_1^{\hat{F}}$ and the one defined with the GL method $\tilde{s}_2^{\hat{F}}$, when using

the trigonometric basis $(\varphi_j)_j$. This comparison is new and beyond the classical regression setting: the study would be of interest in many other contexts.
We also compute the adaptive least-squares estimator, denoted by $\tilde{s}^{LS}$, to investigate the difference between the classical orthonormal bases and the warped bases. Let us recall briefly its definition. First, we set, for $t \in L^2((0;1))$, and $m \in \mathcal{M}_n$:

$$\gamma_n^{LS}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 \text{ and } \mathrm{pen}^{LS}(m) = C\sigma^2\frac{D_m}{n}, \tag{3.20}$$

with $C$ a numerical constant. We define for each $m$, $\hat{s}_m^{LS} = \arg\min_{t\in S_m}\gamma_n^{LS}(t)$, and select $\hat{m}^{LS} = \arg\min_{t\in S_m}\gamma_n^{LS}(t) + \mathrm{pen}^{LS}(m)$. Then we have $\hat{s}_{\hat{m}^{LS}}^{LS} = \sum_{j=1}^{D_{\hat{m}^{LS}}}\hat{a}_j^{LS}\varphi_j$, where $\hat{a}^{LS} = (\hat{a}_j^{LS})_j$ is computed by inverting the matrix $M_{\hat{m}} = (M_{\hat{m},j,k})_{j,k\in\{1,\dots,D_{\hat{m}}\}}$, that is $\hat{a}^{LS} = M_{\hat{m}}^{-1}b$, with

$$M_{m,j,k} = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)\varphi_k(X_i), \text{ and } b = (b_j)_{j\in\{1,\dots,D_m\}}, \ b_j = \frac{1}{n}\sum_{i=1}^{n}Y_i\varphi_j(X_i). \tag{3.21}$$

We refer to Baraud (2002) for the theory and to Comte & Rozenholc (2004) for the practical computation. We have thus three estimators to compute, from data $(X_i, Y_i)_{i\in\{1,\dots,n\}}$. We first notice that their common expression is:

$$\hat{s}_{\hat{m}} = \sum_{j=1}^{D_{\hat{m}}}\hat{a}_j\psi_j,$$

with, for $\tilde{s}_1^{\hat{F}}$ and $\tilde{s}_2^{\hat{F}}$, $\hat{a}_j = \hat{a}_j^{\hat{F}}$ defined by equation (3.15) and $\psi_j = \varphi_j \circ \hat{F}_n$, and for $\tilde{s}^{LS}$, $\hat{a}_j = \hat{a}_j^{LS}$ and $\psi_j = \varphi_j$. In the first case, we generate another sample $(X_{-i})_{i\in\{1,\dots,n\}}$, to find the empirical c.d.f $\hat{F}_n$, and to compute the coefficients $\hat{a}_j^{\hat{F}}$. Concretely, choosing $m_{\max} = 8$, we use the following steps:

- For each $m \in \{1,\dots,m_{\max}\}$, compute $\mathrm{crit}(m)$, for the three following definitions:
  - $\mathrm{crit}(m) = \gamma_n(\hat{g}_m^{\hat{F}}, \hat{F}_n) + \mathrm{pen}(m)$ in the warped-bases case, with penalization. Notice that $\gamma_n(\hat{g}_m^{\hat{F}}) = -\sum_{j=1}^{\hat{D}_m}(\hat{a}_j^{\hat{F}})^2$.
  - $\mathrm{crit}(m) = A(m) + 2V(m)$ in the warped-bases case, with the GL method. Notice that $A(m) = \max_{m'>m}\{\sum_{j=D_m+1}^{D_{m'}}(a_j^{\hat{F}})^2 - V(m')\}_+$.
  - $\mathrm{crit}(m) = \gamma_n^{LS}(\hat{s}_m^{LS}) + \mathrm{pen}^{LS}(m)$ in the least-squares case. The least-squares contrast is computed like the warped-bases criterion. The penalty defined by (3.20) is implemented, with $\sigma^2$ replaced by the unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n-(2mm+1)}\sum_{i=1}^{n}(Y_i - \hat{s}_{mm}^{LS}(X_i))^2, \text{ with } mm = [\sqrt{n}].$$

- In the three cases, select $\hat{m}$ (that is $\hat{m} = \hat{m}^{(1)}, \hat{m}^{(2)}, \hat{m}^{LS}$) such that $\mathrm{crit}(m)$ is minimum.
- Compute then the three estimators $\tilde{s}_l = \sum_{j=1}^{D_{\hat{m}^{(l)}}}\hat{a}_j^{\hat{F}}(\varphi_j\circ\hat{F}_n)$, $l = 1, 2$ and $\tilde{s}^{LS} = \sum_{j=1}^{D_{\hat{m}^{LS}}}\hat{a}_j^{LS}\varphi_j$, at a sequence of equispaced points in $(a;b)$.

**Remark 3.1.** To implement crit($m$), the numerical constants $c_1'$ (of pen), $C$ (of pen$^{LS}$), and $c_2'$ (of $V$) have to be calibrated. The constant $C$ is chosen equal to 2.5, which is a value often found in the literature (constants of the $C_p$ criterion of Mallows, for example). We decide to concentrate on the data-driven calibration of the constants involved in the definition of the new estimators, that is $c_1'$ and $c_2'$. The constant $c_1'$ is useful for the penalized warped bases estimator $\tilde{s}_1^{\hat{F}}$: it can thus be carried out for each simulated sample using a method inspired by the slope heuristic (developed first by Birgé & Massart 2007). But this data-driven solution can not be used for the recent method of GL, leading to the estimator $\tilde{s}_2^{\hat{F}}$. So, to compare in the same way the two estimators, we choose to experiment it with fixed constants, previously stated. The constant $c_1'$ is adjusted prior to the comparison, using however the slope heuristic: we use the graphical interface CAPUSHE developed by Baudry *et al.* (2012), to conduct an experimentation over 100 samples (see our examples, Section 3.4.2), with the so-called "dimension-jump" method. We choose then the largest constant over all attempts proposed by the software, that is $c_1' = 4$ (recall that in penality calibration, it is more secure to overpenalize). For the constant of the GL method, we looked at the quadratic risk with respect to its value $c_2'$, and chose one of the first values leading to reasonable risk and complexity of the selected model, $c_2' = 0.5$ (for the computation of the risk, see Section 3.4.2 below). Notice finally that the specific factor 2 involved in the definition of $\hat{m}^{(2)}$ (see definition (3.17)) could be also adjusted: it plays a technical role in the proof but might have been replaced by any other constant larger than 1. Additional explanations and graphs related to the choice of the constants can be found in Appendix 1, Section 3.6.

### 3.4.2 Examples

The procedure is applied for different regression functions, design and noise. To concentrate on the comparison of the three methods, we decide to present the estimation results for two very smooth functions, on the interval $(0;1)$: a polynomial function, $s_1 : x \mapsto x(x-1)(x-0.6)$, and an exponential function, $s_2 : x \mapsto -\exp(-200(x-0.1)^2) - \exp(-200(x-0.9)^2)$. The sensibility of the method to the underlying design is tested with the following densities, all supported by $(0;1)$. In the definitions, $c$ is a constant adjusted to obtain density function in each case:

– $\mathcal{U}_{(0;1)}$, the classical uniform distribution,
– $\mathcal{DU}_{(0;1)}$, a probability distribution with density $x \mapsto cx\mathbf{1}_{(0;1)}$,
– $\mathcal{E}_c(1)$, a truncated exponential distribution with mean 1 that is with density $x \mapsto ce^{-x}\mathbf{1}_{(0;1)}$,
– $\mathcal{N}_c(0.5, 0.01)$, a truncated Gaussian distribution with density $x \mapsto c\exp(-(x-0.5)^2/0.02)\mathbf{1}_{(0;1)}(x)$,
– $\mathcal{NBM}_t$, a truncated bimodal Gaussian distribution, with density $x \mapsto c(\exp(-200(x-0.05)^2) + \exp(-200(x-0.95)^2))\mathbf{1}_{(0;1)}(x)$,
– $\mathcal{CM}$, a distribution with piecewise constant density $2.485\mathbf{1}_{[0;0.2]} + 0.01\mathbf{1}_{]0.1;0.8]} + 2.485\mathbf{1}_{]0.8;1]}$,

Finally, the variables $\varepsilon_i$ are generated following either a Gaussian distribution, or a Laplace distribution, with mean 0. They are denoted respectively by $\mathcal{N}(0, v)$ ($v$ the variance) and by $\mathcal{L}(0, \ell)$ ($\ell$ a positive real such that the Laplace density is $x \mapsto 1/(2\ell)\exp(-|x|/\ell)$). The parameters $\ell$ and $v$ are adjusted for each of the functions $s_1$ and $s_2$: it is natural to choose cases

in which there is a little more signal than noise. Precisely, the values are chosen such that the ratio of the variance of the signal ($\mathrm{Var}(s(X_1))$) over the variance of the noise ($\mathrm{Var}(\varepsilon_1)$) belongs to $[1.6; 2.4]$, whatever the design distribution. This ratio, denoted by "s2n", will be precised in Tables 3.1 and 3.2.

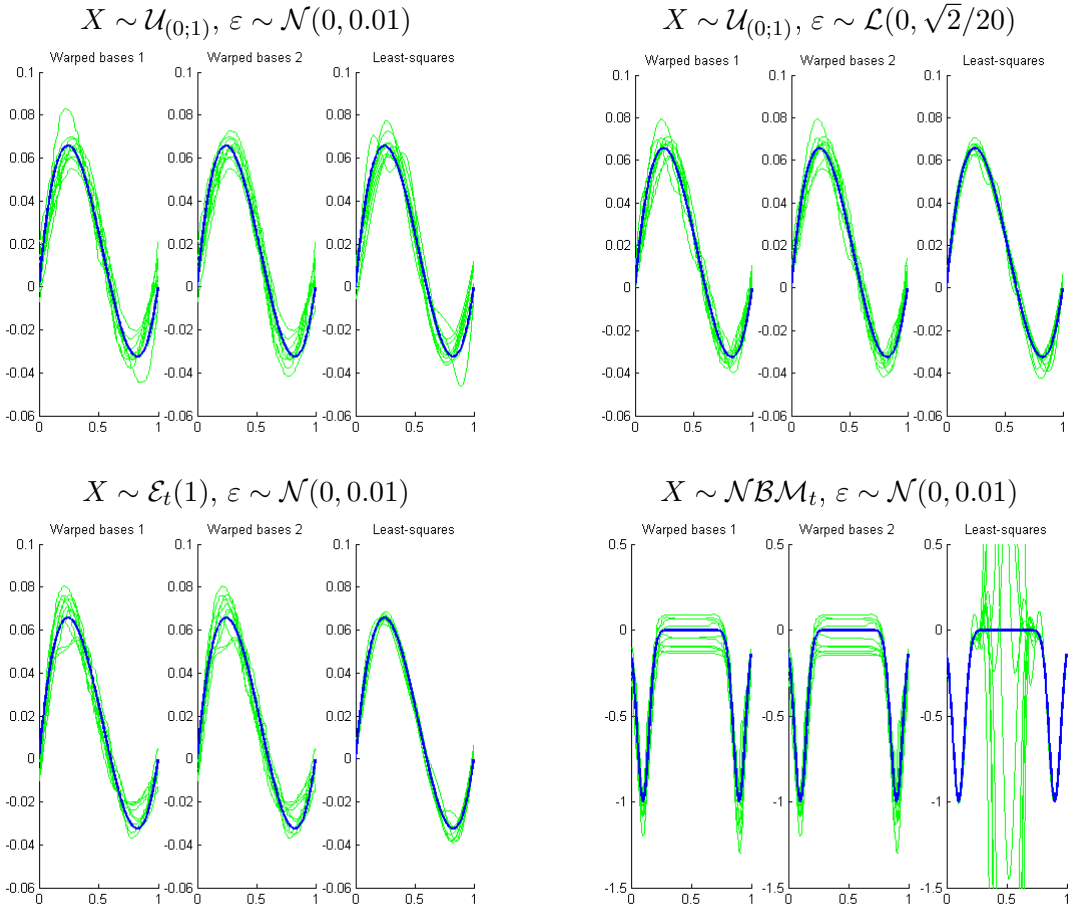



Figure 3.1: Plots of 20 estimators $\tilde{s}_1^{\hat{F}}$ (Warped bases 1), $\tilde{s}_2^{\hat{F}}$ (Warped bases 2) and $\tilde{s}^{LS}$ (Least-squares) of function $s_1$ or $s_2$, built from i.i.d. sample in trigonometric bases. Bold line: True function, Thin lines: Estimators.

We first compare the visual quality of the reconstruction, for the three estimators. Figure Figure 3.1 shows beams of estimated functions versus true functions in four cases. Precisely, for each figure, we plot 20 estimators of each kind, built from i.i.d samples of data of size $n = 500$. The three first plots show that the results are quite good for all the estimators. The noise distribution does not seem to affect significantly the results. Notice that the computation of the estimators $\tilde{s}^{LS}$ requires much more time than the others. It is due to the computation of the inverse of the matrix $M_{\hat{m}}$, while the warped-bases methods are simpler.

So one can easily use warped bases for estimation problems with large data samples sizes (see for example domains as fluorescence, physics, neuronal models...). The last plot of Figure Figure 3.1 shows that the warped-bases estimators behave still correctly if the design density is very inhomogeneous (we obtain the same type of plots when the $X_i$ is distributed with $\mathcal{CM}$). In fact, if we implement the least-squares method without taking additionnal precautions and without numerical approximation for the computation of $M_{\hat{m}}^{-1}$, the estimator can not adapt to a design density which nearly vanishes on a long interval. This highlights the interest for warping the bases: this method seems to be very stable, whatever the design distribution, and even if it is very inhomogeneous: it tends to detect better the hole which can occur in the design density. Let us notice that specific methods exist, taking into account the inhomogeneity of the data to obtain upper bounds for the quadratic pointwise risk, see for example Gaïffas (2007).
The beams of estimators seem to enhance the equivalence we found in the theory between the GL method and the penalization method. For more precise results concerning these selection rules, we compare $L^2$ risk, in the different models (the two functions estimated, the possible design and noise). The ISE (Integrated Squared Error) for one estimator $\tilde{s}$ is $\text{ISE} = \int_a^b (s(x) - \tilde{s}(x))^2 dx$. It is computed as follows:

$$ISE \approx \frac{b-a}{K} \sum_{k=0}^{K} \left[ \tilde{s}\left(a + k\frac{b-a}{K}\right) - s\left(a + k\frac{b-a}{K}\right)\right]^2,$$

where $K$ is an integer (we choose $K = 1000$). The mean ISE (MISE) is the mean of those values over $N = 100$ independent simulated samples.

The risks (multiplied by 1000) are displayed in Table 3.1 (estimation of $s_1$) and 3.2 (estimation of $s_2$) for the estimators $\tilde{s}_1^{\hat{F}}$ (WB1) and $\tilde{s}_2^{\hat{F}}$ (WB2) are computed for different sample sizes going from $n = 100$ to 2000. Notice first that the difference of order of size between the values of the two tabulars is explained by the difference of amplitude between the two functions ($s_1$ takes its values in the interval $[-0.04; 0.07]$ and $s_2$ in $[-1; 0]$). As expected, the values of MISE get smaller when the sample size increases, and they are similar for both estimators, in most cases. The GL method gives slightly smaller risks in 59% of the cases (in bold-blue in the tables). But it seems that the values are better than those of the penalized estimator in 76% of the cases for the large sample sizes ($n = 500$ to 2000). We have to put this result into perspective: larger classes of functions and models would have to be studied to confirm this and we keep in mind that the methods are equivalent from the theoretical point of view.

## 3.5 Proofs of the main results

### 3.5.1 A key result

One of the main argument of the proof of Theorem 3.1 and Theorem 3.2 is the control of the centered empirical process defined by

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(t \circ F_X\right)(X_i) - \langle (t \circ F_X), s\rangle_{f_X}, \quad t \in L^2((0;1)), \tag{3.22}$$

| $\varepsilon$ | $X$ | n=100 | 200 | 500 | 1000 | 1500 | 2000 | Estimator |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{N}(0, 0.0006)$ | $\mathcal{U}_{(0;1)}$ | 0.238 | 0.116 | 0.058 | 0.029 | 0.017 | 0.017 | WB1 |
| | s2n=2.07 | 0.462 | 0.227 | 0.087 | 0.045 | 0.028 | 0.024 | WB2 |
| | $\mathcal{DU}_{(0;1)}$ | 0.407 | 0.254 | 0.144 | 0.09 | 0.069 | 0.058 | WB1 |
| | s2n=1.74 | 0.55 | 0.276 | **0.141** | **0.084** | **0.064** | **0.054** | WB2 |
| | $\mathcal{E}_t(1)$ | 0.231 | 0.152 | 0.052 | 0.032 | 0.021 | 0.018 | WB1 |
| | s2n=1.9 | 0.501 | 0.248 | 0.09 | 0.042 | 0.027 | 0.024 | WB2 |
| | $\mathcal{N}_t(0.5, 0.1)$ | 0.473 | 0.181 | 0.089 | 0.052 | 0.036 | 0.028 | WB1 |
| | s2n=1.98 | 0.68 | 0.243 | 0.097 | 0.053 | **0.036** | **0.027** | WB2 |
| | $\mathcal{NBM}_t$ | 0.957 | 0.788 | 0.561 | 0.448 | 0.436 | 0.395 | WB1 |
| | s2n=1.94 | 1.037 | **0.785** | **0.537** | **0.436** | **0.433** | **0.393** | WB2 |
| | $\mathcal{CM}$ | 1.012 | 0.943 | 0.775 | 0.718 | 0.692 | 0.68 | WB1 |
| | s2n=2.07 | 1.267 | 0.968 | **0.773** | **0.711** | **0.688** | **0.679** | WB2 |
| $\mathcal{L}(0, 0.0173)$ | $\mathcal{U}_{(0;1)}$ | 0.235 | 0.102 | 0.051 | 0.026 | 0.02 | 0.016 | WB1 |
| | | 0.44 | 0.215 | 0.085 | 0.04 | 0.031 | 0.023 | WB2 |
| | $\mathcal{DU}_{(0;1)}$ | 0.352 | 0.268 | 0.13 | 0.084 | 0.069 | 0.059 | WB1 |
| | | 0.494 | 0.28 | **0.122** | **0.074** | **0.062** | **0.054** | WB2 |
| | $\mathcal{E}_t(1)$ | 0.278 | 0.133 | 0.065 | 0.031 | 0.024 | 0.018 | WB1 |
| | | 0.576 | 0.244 | 0.099 | 0.043 | 0.033 | 0.023 | WB2 |
| | $\mathcal{N}_t(0.5, 0.1)$ | 0.338 | 0.2 | 0.092 | 0.05 | 0.036 | 0.03 | WB1 |
| | | 0.539 | 0.254 | 0.101 | 0.052 | **0.036** | **0.028** | WB2 |
| | $\mathcal{NBM}_t$ | 1.104 | 0.699 | 0.562 | 0.453 | 0.425 | 0.412 | WB1 |
| | | 1.221 | **0.662** | **0.532** | **0.442** | **0.418** | **0.406** | WB2 |
| | $\mathcal{CM}$ | 1.078 | 0.889 | 0.801 | 0.716 | 0.688 | 0.683 | WB1 |
| | | 1.207 | 0.919 | **0.797** | **0.707** | **0.686** | **0.682** | WB2 |

Table 3.1: Values of MISE $\times 1000$ averaged over 100 samples, for the estimation of $s_1$

| $\varepsilon$ | $X$ | n=100 | 200 | 500 | 1000 | 1500 | 2000 | Estimator |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{N}(0, 0.05)$ | $\mathcal{U}_{(0;1)}$ | 73.979 | 37.574 | 13.557 | 6.606 | 4.088 | 3.126 | WB1 |
| | s2n=2.33 | **72.02** | **34.761** | **13.32** | **6.506** | **3.975** | **3.109** | WB2 |
| | $\mathcal{DU}_{(0;1)}$ | 65.367 | 54.668 | 43.972 | 36.923 | 32.499 | 29.707 | WB1 |
| | s2n=2.33 | 73.101 | **53.149** | **39.232** | **32.683** | **29.873** | **28.252** | WB2 |
| | $\mathcal{E}_t(1)$ | 74.224 | 41.907 | 17.365 | 9.384 | 6.925 | 5.187 | WB1 |
| | s2n=2.37 | 76.55 | **37.431** | **16.401** | **9.074** | **6.842** | 5.307 | WB2 |
| | $\mathcal{N}_t(0.5, 0.1)$ | 75.906 | 53.158 | 30.022 | 16.046 | 13.3 | 12.119 | WB1 |
| | s2n=1.76 | 88.46 | 54.046 | **27.695** | **15.861** | **13.21** | **12.119** | WB2 |
| | $\mathcal{NBM}_t$ | 86.712 | 29.374 | 14.892 | 6.949 | 4.368 | 3.502 | WB1 |
| | s2n=2.06 | 73.514 | 32.237 | **12.609** | **6.529** | **4.054** | **2.935** | WB2 |
| | $\mathcal{CM}$ | 125.098 | 47.224 | 29.851 | 20.533 | 20.016 | 17.296 | WB1 |
| | s2n=1.69 | 111.872 | 53.719 | 31.766 | 20.593 | **18.595** | **16.1** | WB2 |
| $\mathcal{L}(0, 0.1581)$ | $\mathcal{U}_{(0;1)}$ | 77.489 | 35.98 | 13.657 | 6.47 | 3.808 | 3.032 | WB1 |
| | | **73.596** | **32.823** | 13.667 | **6.392** | **3.772** | **3.026** | WB2 |
| | $\mathcal{DU}_{(0;1)}$ | 70.605 | 55.9 | 43.65 | 37.967 | 33.642 | 30.021 | WB1 |
| | | 80.886 | **54.544** | **38.695** | **32.008** | **29.473** | **27.925** | WB2 |
| | $\mathcal{E}_t(1)$ | 64.881 | 44.879 | 17.774 | 10.31 | 6.987 | 5.856 | WB1 |
| | | 71.622 | **38.003** | **16.928** | **9.897** | **6.761** | **5.689** | WB1 |
| | $\mathcal{N}_t(0.5, 0.1)$ | 82.315 | 50.384 | 27.537 | 15.931 | 13.474 | 12.523 | WB1 |
| | | 90.932 | **48.743** | **25.119** | 16.24 | **13.403** | **12.523** | WB2 |
| | $\mathcal{NBM}_t$ | 98.027 | 33.034 | 13.593 | 7.472 | 4.697 | 3.604 | WB1 |
| | | **83.533** | **32.761** | **12.162** | **6.437** | **4.484** | **3.119** | WB2 |
| | $\mathcal{CM}$ | 113.635 | 48.175 | 25.483 | 21.765 | 18.833 | 18.229 | WB1 |
| | | **95.868** | 49.138 | **24.812** | **18.662** | **16.717** | **16.011** | WB2 |

Table 3.2: Values of MISE $\times 1000$ averaged over 100 samples, for the estimation of $s_2$

on the unit sphere

$$\mathcal{S}(m) = \{t \in S_m,\ \|t\| = 1\}$$

of a fixed model $S_m$. Let us first state the following result, which we use for both theorems.

**Proposition 3.2.** *Under the assumptions of Theorem 3.1, with* $p(m') = 6(1+2\delta)\phi_0^2\mathbb{E}[Y_1^2]\frac{D_{m'}}{n}$, *(*$\delta > 0$*) for any* $m' \in \mathcal{M}_n$, *there exists a constant* $C$ *depending on* $\phi_0^2$, $\|s\|_\infty$, $\mathbb{E}[s^2(X_1)]$, $\sigma^2$, $\mathbb{E}[|\varepsilon_1|^p]$ *and* $\delta$ *such that,*

$$\mathbb{E}\left[\sum_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t) - p(m')\right)_+\right] \leq \frac{C}{n}.$$

***Proof of Proposition 3.2***

We split the process $\nu_n$ into three parts, writing $\nu_n = \nu_n^{(1)} + \nu_n^{(2,1)} + \nu_n^{(2,2)}$, with

$$\begin{aligned}
\nu_n^{(1)}(t) &= \frac{1}{n}\sum_{i=1}^{n} s(X_i)\,(t\circ F_X)\,(X_i) - \langle (t\circ F_X), s\rangle_{f_X},\\
\nu_n^{(2,1)}(t) &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{|\varepsilon_i|\leq\kappa_n}\,(t\circ F_X)\,(X_i) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{|\varepsilon_i|\leq\kappa_n}\,(t\circ F_X)\,(X_i)\right],\\
\nu_n^{(2,2)}(t) &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{|\varepsilon_i|>\kappa_n}\,(t\circ F_X)\,(X_i) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{|\varepsilon_i|>\kappa_n}\,(t\circ F_X)\,(X_i)\right],
\end{aligned}$$

with $c$ a constant depending on the collection of models and where we define

$$\kappa_n = c\frac{\sqrt{n}}{\ln(n)}. \tag{3.23}$$

We obtain,

$$\begin{aligned}
\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t) - p(m')\right)_+ \leq\ & 3\left\{\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2 - \frac{p_1(m')}{3}\right)_+\right.\\
&+\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,1)}(t)\right)^2 - \frac{p_2(m')}{3}\right)_+\\
&\left.+\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,2)}(t)\right)^2\right\}
\end{aligned} \tag{3.24}$$

with $p_1(.) + p_2(.) = p(.)$.

We upper bound the first two terms by applying the Talagrand Inequality, stated in Proposition 2.2 of Chapter 2. We apply Inequality (2.1) to the first term of equation (3.24), with function $r$ replaced by $r_t : x \mapsto s(x)(t\circ F_X)(x)$, $t \in \mathcal{R} = \mathcal{S}_{m'}$, and $\xi_i = X_i$. Let us first compute the constants $M_1^{(1)}$, $H^{(1)}$, and $v^{(1)}$. We observe first that $\|r_t\|_\infty \leq \|s\|_\infty\|t\|_\infty$ and we use assumption $[\mathcal{M}_3]$ to get $\|r_t\|_\infty \leq \phi_0\sqrt{D_{m'}}\|t\|\|s\|_\infty = \phi_0\sqrt{D_{m'}}\|s\|_\infty := M_1^{(1)}$.

Then, noting that $t \in \mathcal{S}(m')$ can be written $t = \sum_{j=1}^{D_{m'}} b_j\varphi_j$ with $\sum_j b_j^2 = 1$, we apply

Cauchy-Schwarz's inequality to get $\sup_{t\in\mathcal{S}(m')}\nu_n^{(1)}(t)^2 \leq \sum_{j=1}^{D_{m'}}\nu_n^{(1)}(\varphi_j)^2$. Since assumptions $[\mathcal{M}_2]$ and $[\mathcal{M}_3]$ hold, we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\nu_n^{(1)}(t)^2\right] \leq \sum_{j=1}^{D_{m'}}\frac{1}{n}\mathrm{Var}(s(X_1)\left(\varphi_j\circ F_X\right)(X_1)) \leq \phi_0^2\mathbb{E}\left[s^2(X_1)\right]\frac{D_{m'}}{n} := \left(H^{(1)}\right)^2.$$

Finally, $\mathrm{Var}(r_t(X_1)) \leq \mathbb{E}[r_t^2(X_1)] \leq \|s\|_\infty^2 := v^{(1)}$. Replacing the quantities $M_1^{(1)}$, $H^{(1)}$ and $v^{(1)}$ by the values derived above, Inequality (2.1) becomes, for a $c>0$

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2 - \frac{p_1(m')}{3}\right)_+\right]$$

$$\leq c\|s\|_{L^\infty((a;b))}\left\{\frac{1}{n}\sum_{m'\in\mathcal{M}_n}\exp\left(-\bar{k}D_{m'}\right) + \phi_0^2\frac{1}{n^2}\sum_{m'\in\mathcal{M}_n}D_{m'}\exp\left(-\bar{\bar{k}}\sqrt{n}\right)\right\},$$

with $\bar{k}$ and $\bar{\bar{k}}$ two constants (independent of $m'$ and $n$) and $p_1(m') = 3\times 2(1+2\delta)\left(H^{(1)}\right)^2$. Therefore, using that the cardinality of $\mathcal{M}_n$ is bounded by $n$ and also that $D_{m'}\leq n$ , the following upper bound holds, for $c_1$ a constant,

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2 - \frac{p_1(m')}{3}\right)_+\right] \leq \frac{c_1}{n}. \tag{3.25}$$

Similarly, we apply Inequality (2.1) to the second process $\nu_n^{(2,1)}$. We replace $r$ by $r_t : (\varepsilon, x)\mapsto \varepsilon\mathbf{1}_{\varepsilon\leq\kappa_n}t\circ F_X(x)$, and $\xi_i = (\varepsilon_i, X_i)$. Thus we compute

$$M_1^{(2)} = \kappa_n\phi_0\sqrt{D_{m'}},\quad H^{(2)} = \phi_0\sigma\sqrt{\frac{D_{m'}}{n}},\quad v^{(2)} = \sigma^2.$$

With $p_2(m') = 3\times 2(1+2\delta)\left(H^{(2)}\right)^2$, we get

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,1)}(t)\right)^2 - \frac{p_2(m')}{3}\right)_+\right] \leq \frac{c_2}{n}, \tag{3.26}$$

for $c_2$ a constant.
Finally, we look for an upper bound for the process $\nu_n^{(2,2)}$. We can not apply the concentration inequality, because it is not bounded. However, following the same line as in computations above, we write

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,2)}(t)\right)^2\right] \leq \sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\nu_n^{(2,2)}(\varphi_j)^2\right)\right] \leq \frac{1}{n}\mathbb{E}\left[|\varepsilon_1|^{2+p}\mathbf{1}_{|\varepsilon_1|>\kappa_n}\right]\phi_0^2\frac{\kappa_n^{-p}D_{m'}}{n} \leq \frac{c_3}{n}, \tag{3.27}$$

with $c_3>0$, since $\kappa_n$ is defined by (3.23) and $p>4$.
We conclude the proof of Proposition 3.2 by gathering in the equation (3.24) the three inequalities (3.25), (3.26), and (3.27).

□

We also set the following technical lemma, which will be useful several times, with $\nu$ an empirical process.

**Lemma 3.1.** *Let $\nu : L^2((0;1)) \mapsto \mathbb{R}$ be a linear functional. Let also $m$ be an index of the collection $\mathcal{M}_n$. Then,*

$$\sup_{t \in \mathcal{S}(m)} \nu^2(t) = \sum_{j=1}^{D_m} \nu^2(\varphi_j).$$

***Proof of Lemma 3.1.***
If $t$ belongs to $\mathcal{S}(m)$, it can be written $t = \sum_{j=1}^{D_m} b_j \varphi_j$, with $\sum_{j=1}^{D_m} b_j^2 = 1$. Thus, by the linearity of $\nu$ and the Cauchy-Schwarz Inequality,

$$\nu^2(t) = \left( \sum_{j=1}^{D_m} b_j \nu(\varphi_j) \right)^2 \leq \sum_{j=1}^{D_m} \nu^2(\varphi_j).$$

This leads to $\sup_{t \in \mathcal{S}(m)} \nu^2(t) \leq \sum_{j=1}^{D_m} \nu^2(\varphi_j)$. The equality is obtained by choosing $t = \sum_{j=1}^{D_m} b_j \varphi_j \in L^2((0;1))$, with $b_j = \nu(\varphi_j)/(\sum_{k=1}^{D_m} \nu^2(\varphi_k))$.

□

### 3.5.2 Proof of Theorem 3.1

In all the proofs, the letter $C$ denotes a nonnegative real that may change from line to line.

#### Main part of the proof, case of the estimator $\tilde{s}_1^{F_X}$

The proof for the estimator $\tilde{s}_1^{F_X}$ is classical of the model selection device (see Massart 2007). The typical sketch we describe is the same as the proof of Theorem 3.1 page 185 in Brunel & Comte (2005) for example. For the sake of simplicity, we denote in this section $\gamma_n(.) = \gamma_n(., F_X)$, $\text{pen} = \text{pen}^{F_X}$, $\hat{m} = \hat{m}^{(1),F_X}$. Let $m \in \mathcal{M}_n$ be fixed. By the definition of $\hat{m}$ (see (3.9)),

$$\gamma_n(\hat{g}_{\hat{m}}^{F_X}) + \text{pen}(\hat{m}) \leq \gamma_n(\hat{g}_m^{F_X}) + \text{pen}(m).$$

Then, since $g_m^{F_X}$ minimizes the contrast on the space $S_m$, and since the orthogonal projection $g_m$ belongs to $S_m$, we also have $\gamma_n(\hat{g}_m^{F_X}) \leq \gamma_n(g_m)$. This leads to

$$\gamma_n(\hat{g}_m^{F_X}) - \gamma_n(g_m) \leq \text{pen}(m) - \text{pen}(\hat{m}). \tag{3.28}$$

We now replace the contrast by its definition: notice that, for $t \in L^2((0;1))$,

$$\|t\|^2 = \|t \circ F_X\|_{f_X}^2 = \|t \circ F_X - s\|_{f_X}^2 - \|s\|_{f_X}^2 + 2\langle t \circ F_X, s \rangle_{f_X},$$

we obtain, thanks to (3.28),

$$\|\tilde{s}_1^{F_X} - s\|_{f_X}^2 \leq \|s_m^{F_X} - s\|_{f_X}^2 + \text{pen}(m) - \text{pen}(\hat{m}) + 2\nu_n(\hat{g}_{\hat{m}}^{F_X} - g_m), \tag{3.29}$$

where $s_m^{F_X} = g_m \circ F_X$, and with $\nu_n$ defined by (3.22). Now, define the ball $\mathcal{S}(p) = \{t \in S_p,\ \|t\| = 1\}$, for $p \in \mathcal{M}_n$. We write

$$\begin{aligned} 2\nu_n(\hat{g}_{\hat{m}}^{F_X} - g_m) &= 2\|\hat{g}_{\hat{m}}^{F_X} - g_m\| \nu_n\left(\frac{\hat{g}_{\hat{m}}^{F_X} - g_m}{\|\hat{g}_{\hat{m}}^{F_X} - g_m\|}\right), \\ &\leq 2\|\hat{g}_{\hat{m}}^{F_X} - g_m\| \sup_{\substack{t \in S_m + S_{\hat{m}} \\ \|t\|=1}} |\nu_n(t)|, \\ &\leq 2\|\hat{g}_{\hat{m}}^{F_X} - g_m\| \sup_{t \in \mathcal{S}(m \vee \hat{m})} |\nu_n(t)|, \\ &= 2\|\tilde{s}_1^{F_X} - s_m^{F_X}\|_{f_X} \sup_{t \in \mathcal{S}(m \vee \hat{m})} |\nu_n(t)|, \end{aligned}$$

because the models are nested: $S_m + S_{\hat{m}} \subset S_{m \vee \hat{m}}$. Introduce a real number $\theta > 0$, and recall that $2xy \leq x^2/\theta + \theta y^2$, for $x, y \in \mathbb{R}$, we deduce

$$\begin{aligned} 2\nu_n(\hat{g}_{\hat{m}}^{F_X} - g_m) &\leq \frac{\|\tilde{s}_1^{F_X} - s_m^{F_X}\|_{f_X}^2}{\theta} + \theta \sup_{t \in \mathcal{S}_{(m \vee \hat{m})}} (\nu_n(t))^2, \\ &\leq \frac{2}{\theta}\left(\|\tilde{s}_1^{F_X} - s\|_{f_X}^2 + \|s_m^{F_X} - s\|_{f_X}^2\right) + \theta \sup_{t \in \mathcal{S}_{(m \vee \hat{m})}} (\nu_n(t))^2. \end{aligned}$$

Gathering this with (3.29), we get

$$\left(1 - \frac{2}{\theta}\right)\|\tilde{s}_1^{F_X} - s\|_{f_X}^2 \leq \left(1 + \frac{2}{\theta}\right)\|s - s_m^{F_X}\|_{f_X}^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \theta \sup_{t \in \mathcal{S}(m \vee \hat{m})} (\nu_n(t))^2.$$

We denote by $p(m \vee \hat{m}) = 6(1 + 2\delta)\phi_0^2 \mathbb{E}[Y_1^2] D_{m \vee m'}/n$, for $\theta > 2$. We thus have

$$\begin{aligned} \|\tilde{s}_1^{F_X} - s\|_{f_X}^2 &\leq \frac{\theta}{\theta - 2}\left\{\frac{\theta + 2}{\theta}\|s - s_m^{F_X}\|_{f_X}^2 + \mathrm{pen}(m) + \right. \\ &\qquad \left. \theta\left(\sup_{t \in \mathcal{S}_{(m \vee \hat{m})}} (\nu_n(t))^2 - p(m, \hat{m})\right)_+ + \theta p(m, \hat{m}) - \mathrm{pen}(\hat{m})\right\}. \end{aligned}$$

We bound the supremum on the random model $\mathcal{S}_{(m \vee \hat{m})}$ roughly:

$$\begin{aligned} \left\|\tilde{s}_1^{F_X} - s\right\|_{f_X}^2 &\leq \frac{\theta + 2}{\theta - 2}\|s - s_m^{F_X}\|_{f_X}^2 + \frac{\theta}{\theta - 2}\mathrm{pen}(m) \\ &\quad + \frac{\theta^2}{\theta - 2} \sum_{m' \in \mathcal{M}_n} \left(\sup_{t \in \mathcal{S}(m \vee m')} (\nu_n(t))^2 - p(m')\right)_+ \\ &\quad + \frac{\theta}{\theta - 2}\left(\theta p(m \vee \hat{m}) - \mathrm{pen}(\hat{m})\right). \end{aligned}$$

It remains to apply Proposition 3.2:

$$\mathbb{E}\left[\left\|\tilde{s}_1^{F_X} - s\right\|_{f_X}^2\right] \leq \frac{\theta + 2}{\theta - 2}\|s - s_m^{F_X}\|_{f_X}^2 + \frac{\theta}{\theta - 2}\mathrm{pen}(m) + \frac{\theta^2}{\theta - 2}\frac{C}{n} + \frac{\theta}{\theta - 2}\mathbb{E}\left[\theta p(m \vee \hat{m}) - \mathrm{pen}(\hat{m})\right].$$

Choosing $c_1$ equal to $6\theta(1 + 2\delta)$ in the definition of $\mathrm{pen}(m)$ (see (3.10)) permits to obtain

$$\mathbb{E}\left[\theta p(m \vee \hat{m}) - \mathrm{pen}(\hat{m})\right] \leq \mathrm{pen}(m).$$

since $p(m \vee \hat{m}) \leq 6(1+2\delta)\phi_0^2 \mathbb{E}[Y_1^2](D_m + D_{\hat{m}})/n$. We have thus proved the result:

$$\mathbb{E}\left[\left\|\tilde{s}_1^{F_X} - s\right\|_{f_X}^2\right] \leq \frac{\theta+2}{\theta-2}\|s - s_m^{F_X}\|_{f_X}^2 + \frac{2\theta}{\theta-2}\mathrm{pen}(m) + \frac{\theta^2}{\theta-2}\frac{C}{n},$$

which is Theorem 3.1 with $k_1 = (\theta+2)(\theta-2)$, $k_1' = 2\theta/(\theta-2) \times c_1$, and $C_1 = C\theta^2/\theta - 2$.

□

**Main part of the proof, case of the estimator $\tilde{s}_2^{F_X}$**

The proof for the second estimator is less classical, since it is inspired both by the model selection strategy, and by the definition of the GL method. For the sake of simplicity, we denote in this section by $V = V^{F_X}$, $A = A^{F_X}$, $\hat{m} = \hat{m}^{(2),F_X}$. Let $S_m$ be a fixed model in the collection indexed by $\mathcal{M}_n$. We decompose the loss of the estimator as follows:

$$\begin{aligned}\left\|\tilde{s}_2^{F_X} - s\right\|_{f_X}^2 &= \left\|\hat{g}_{\hat{m}}^{F_X} - g\right\|^2, \\ &\leq 3\left\|\hat{g}_{\hat{m}}^{F_X} - \hat{g}_{m\wedge\hat{m}}^{F_X}\right\|^2 + 3\left\|\hat{g}_{m\wedge\hat{m}}^{F_X} - \hat{g}_m^{F_X}\right\|^2 + 3\left\|\hat{g}_m^{F_X} - g\right\|^2.\end{aligned}$$

By definition of $A$ and $\hat{m}$,

$$\begin{aligned}\left\|\tilde{s}_2^{F_X} - s\right\|_{f_X}^2 &\leq 3\left(A(m) + V(\hat{m})\right) + 3\left(A(\hat{m}) + V(m)\right) + 3\left\|\hat{g}_m^{F_X} - g\right\|^2, \\ &\leq 6\left(A(m) + V(m)\right) + 3\left\|\hat{g}_m^{F_X} - g\right\|^2.\end{aligned}$$

We have already bounded the risk of the estimator on a fixed model (see Section 3.2.2, Inequalities (3.6) and (3.8)): $\mathbb{E}[\|\hat{g}_m^{F_X} - g\|^2] \leq \phi_0^2 \mathbb{E}[Y_1^2] D_m/n + \|g_m - g\|^2$. Therefore we get

$$\mathbb{E}\left[\left\|\tilde{s}_2^{F_X} - s\right\|_{f_X}^2\right] \leq 6\mathbb{E}\left[A(m)\right] + 6V(m) + 3\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n} + 3\left\|g_m - g\right\|^2.$$

Next, we have to control the term $A(m)$: we use the following lemma, proved just below, to conclude.

**Lemma 3.2.** *Under the assumptions of Theorem 3.1, there exists a constant $C > 0$ depending on $\phi_0^2$, $\|s\|_\infty$, $\mathbb{E}[s^2(X_1)]$, $\sigma^2$, $\mathbb{E}[|\varepsilon_1|^p]$ such that, for each index $m \in \mathcal{M}_n$,*

$$\mathbb{E}\left[A(m)\right] \leq \frac{C}{n} + 12\left\|g_m - g\right\|^2.$$

□

**Proof of Lemma 3.2**

For each index $m \in \mathcal{M}_n$, we decompose,

$$\left\|\hat{g}_{m'}^{F_X} - \hat{g}_{m\wedge m'}^{F_X}\right\|^2 \leq 3\left\|\hat{g}_{m'}^{F_X} - g_{m'}\right\|^2 + 3\left\|g_{m'} - g_{m\wedge m'}\right\|^2 + 3\left\|g_{m\wedge m'} - \hat{g}_{m\wedge m'}^{F_X}\right\|^2.$$

Thus we have

$$\begin{aligned} A(m) &\leq 3 \max_{m' \in \mathcal{M}_n} \left[ \left\| \hat{g}_{m'}^{F_X} - g_{m'} \right\|^2 - \frac{V(m')}{6} \right]_+ + 3 \max_{m' \in \mathcal{M}_n} \left[ \left\| g_{m \wedge m'} - \hat{g}_{m \wedge m'}^{F_X} \right\|^2 - \frac{V(m')}{6} \right]_+ \\ &\quad + 3 \max_{m' \in \mathcal{M}_n} \left\| g_{m'} - g_{m \wedge m'} \right\|^2, \\ &:= 3 \left( T_a + T_b^m + T_c^m \right), \end{aligned} \tag{3.30}$$

and study the terms of the above decomposition.

***Upper-bound for*** $T_a$**.** We roughly simplify the problem by writing first

$$\mathbb{E}\left[T_a\right] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[ \left\{ \left\| \hat{g}_{m'}^{F_X} - g_{m'} \right\|^2 - \frac{V(m')}{6} \right\}_+ \right].$$

Let us notice that

$$\|\hat{g}_{m'}^{F_X} - g_{m'}\|^2 = \sum_{j=1}^{D_{m'}} (\hat{a}_j^{F_X} - a_j)^2 = \sum_{j=1}^{D_{m'}} \nu_n^2(\varphi_j), \tag{3.31}$$

with $\nu_n$ the empirical process defined by (3.22). By Lemma 3.1, this last quantity is equal to $\sup_{t \in \mathcal{S}(m')} \nu_n^2(t)$. Consequently, $\mathbb{E}[T_a] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E}[\{\sup_{t \in \mathcal{S}(m')} \nu_n^2(t) - \frac{V(m')}{6}\}_+]$. We apply then Proposition 3.2: the latter is bounded by $C/n$, for the choice $V(m') = 6 \times p(m')$, which means the choice of $c_2 = 36(1+2\delta)$ in the definition (3.11).

***Upper-bound for*** $T_b^m$**.** To study this term, we write, distinguish whether $m' \leq m$ or $m' > m$,

$$\begin{aligned} T_b^m &= \max \left( \max_{\substack{m' \in \mathcal{M}_n \\ m' \leq m}} \left\{ \left\| g_{m'} - \hat{g}_{m'}^{F_X} \right\|^2 - \frac{V(m')}{6} \right\}_+, \max_{\substack{m' \in \mathcal{M}_n \\ m' > m}} \left\{ \left\| g_m - \hat{g}_m^{F_X} \right\|^2 - \frac{V(m')}{6} \right\}_+ \right), \\ &\leq \max \left( T_a, \left\{ \left\| g_m - \hat{g}_m^{F_X} \right\|^2 - \frac{V(m)}{6} \right\}_+ \right) \leq T_a + \left\{ \left\| g_m - \hat{g}_m^{F_X} \right\|^2 - \frac{V(m)}{6} \right\}_+, \end{aligned}$$

using $-V(m') \leq -V(m)$ for $m' > m$. The last computation proves that $\mathbb{E}[T_a] \leq C/n$ and the same bound holds for the second term, as a consequence of Proposition 3.2. Finally, $\mathbb{E}[T_b^m] \leq C/n$.

***Upper-bound for*** $T_c^m$**.** This term is a bias term. We notice that

$$T_c^m = \max_{\substack{m' \in \mathcal{M}_n \\ m \leq m'}} \|g_{m'} - g_m\|^2 \leq 2 \max_{\substack{m' \in \mathcal{M}_n \\ m \leq m'}} \|g_{m'} - g\|^2 + 2 \|g - g_m\|^2.$$

But assuming $m \leq m'$, we have $S_m \subset S_{m'}$, thus, the orthogonal projections $g_m$ and $g_{m'}$ of $g$ onto $S_m$ and $S_{m'}$ satisfy $\|g_{m'} - g\|^2 \leq \|g_m - g\|^2$. So we have $T_c^m \leq 4\|g_m - g\|^2$, which conclude the proof.

□

### 3.5.3 Proofs of Proposition 3.1 and of Theorem 3.2

In this section, we denote the estimator by $\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} = \hat{g}_{\hat{m}}^{\hat{F}} \circ \hat{F}_n$ (with shortened $\hat{m}^{(2)}$ in $\hat{m}$), and coherently:

$$\hat{s}_{\hat{m}}^{\hat{F},F_X} = \hat{g}_{\hat{m}}^{\hat{F}} \circ F_X,$$

which is an intermediate between the two estimators $\hat{s}_{\hat{m}}^{\hat{F},\hat{F}}$ and $\hat{s}_{\hat{m}}^{F_X,F_X}$. We will also use this notation for fixed index $m \in \mathcal{M}_n$. The letter $C$ is a numerical constant, which may vary from line to line.

#### Notations, and properties of the empirical distribution function

Let us recall some useful tools for the sequel, which are introduced in Section 2.1.2 of Chapter 2. Denoting by $U_{-i} = F_X(X_{-i})$ the uniform variable associated to $X_{-i}$, for any $i \in \{1, \dots, n\}$, we define the empirical distribution function

$$\hat{U}_n : u \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_{-i} \leq u}. \tag{3.32}$$

The following equality holds for any coefficient $\hat{a}_j^{\hat{F}}$ of the estimator (see equation (3.5)):

$$\mathbb{E}\left[\hat{a}_j^{\hat{F}} \,|(X_{-l})_l\right] = \int_0^1 \left(s \circ F_X^{-1}\right)(u) \left(\varphi_j \circ \hat{U}_n\right)(u) du. \tag{3.33}$$

Moreover, we will use several inequalities to control the deviations of the empirical c.d.f. $\hat{U}_n$ or $\hat{F}_n$, which are all introduced in Section 2.1.2 of Chapter 2.

#### Proof of Proposition 3.1

We split the risk of the estimator into five terms: for any index $m$,

$$\mathbb{E}\left[\left\|\hat{s}_m^{\hat{F},\hat{F}} - s\right\|_{f_X}^2\right] \leq \sum_{l=0}^{4} T_l^m,$$

whith

$$\begin{aligned}
T_0^m &= \|s - s_m^{F_X}\|_{f_X}^2 + \|s_m^{F_X} - \hat{s}_m^{F_X,F_X}\|_{f_X}^2, \\
T_1^m &= \left\|\hat{s}_m^{F_X,F_X} - \hat{s}_m^{\hat{F},F_X} - \mathbb{E}\left[\hat{s}_m^{F_X,F_X} - \hat{s}_m^{\hat{F},F_X} \,|(X_{-l})_l\right]\right\|_{f_X}^2, \\
T_2^m &= \left\|\hat{s}_m^{\hat{F},F_X} - \hat{s}_m^{\hat{F},\hat{F}} - \mathbb{E}\left[\hat{s}_m^{\hat{F},F_X} - \hat{s}_m^{\hat{F},\hat{F}} \,|(X_{-l})_l\right]\right\|_{f_X}^2, \\
T_3^m &= \left\|\mathbb{E}\left[\hat{s}_m^{F_X,F_X} - \hat{s}_m^{\hat{F},F_X} \,|(X_{-l})_l\right]\right\|_{f_X}^2, \quad T_4^m = \left\|\mathbb{E}\left[\hat{s}_m^{\hat{F},F_X} - \hat{s}_m^{\hat{F},\hat{F}} \,|(X_{-l})_l\right]\right\|_{f_X}^2.
\end{aligned} \tag{3.34}$$

Let us remark that $T_0^m$ is the bias-variance decomposition for the risk of an estimator $\hat{s}_m^{F_X,F_X}$ (on the fixed model $S_m$). The bound for its expectation is given by Inequalities (3.6) and (3.8). For the other terms, upper-bounds are stated by the following Lemmas, which are proved in the following paragraphs.

**Lemma 3.3.** *Under the assumptions of Proposition 3.2,*

$$\mathbb{E}\left[T_1^m\right] \leq C_2 \|\varphi_2'\|^2_{L^\infty((0;1))} \mathbb{E}\left[Y_1^2\right] \frac{D_m^3}{n^2}$$

*If* $D_m \leq n^{1/2}$,

$$\mathbb{E}\left[T_1^m\right] \leq C_2 \|\varphi_2'\|^2_{L^\infty((0;1))} \mathbb{E}\left[Y_1^2\right] \frac{D_m}{n}$$

**Lemma 3.4.** *Under the assumptions of Proposition 3.2,*

$$\mathbb{E}\left[T_2^m\right] \leq C_2 \mathbb{E}\left[Y_1^2\right] \phi_0^2 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \frac{D_m^4}{n^2}$$

*If* $D_m \leq n^{1/3}$ *in particular,*

$$\mathbb{E}\left[T_2^m\right] \leq C_2 \mathbb{E}\left[Y_1^2\right] \phi_0^2 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \frac{D_m}{n}$$

**Lemma 3.5.** *Under the assumptions of Proposition 3.2,*

$$\begin{aligned} \mathbb{E}[T_3^m] &\leq 6\left(C_2\|g'\|^2\frac{1}{n} + \phi_0^2\mathbb{E}[Y_1^2]\frac{D_m}{n}\right) + 3C_4(\pi^4/4)\|g\|^2\frac{D_m^4}{n^2} \\ &\quad + \frac{C_6}{2}\left\|\varphi_2^{(3)}\right\|_\infty^2 \|g\|^2\frac{D_m^7}{n^3}. \end{aligned} \tag{3.35}$$

*Particularly, if* $D_m \leq n^{1/3}$,

$$\mathbb{E}[T_3^m] \leq \left(6\left(C_2\|g'\|^2 + \phi_0^2\mathbb{E}[Y_1^2]\right) + 3C_4(\pi^4/4)\|g\|^2 + \frac{C_6}{2}\left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;1))}\|g\|^2\right)\frac{D_m}{n}.$$

**Lemma 3.6.** *Under the assumptions of Proposition 3.2,*

$$\mathbb{E}\left[T_4^m\right] \leq C\left(\frac{D_m^4\ln(n)}{n^2} + \frac{D_m^7\ln(n)}{n^3} + \frac{D_m^{10}\ln(n)}{n^4} + \frac{D_m^6}{n^4} + \frac{1}{n} + \frac{D_m^3}{n^2} + \frac{D_m^2}{n^{3/2}}\right)$$

*where* $C$ *depend on* $\|g\|, \|g'\|, \|\varphi_2'\|_{L^\infty((0;1))}, \|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\mathbb{E}[Y_1^2]$ *and,* $\phi_0^2$. *Particularly, if* $D_m \leq n^{1/3}/\ln^{1/3}(n)$,

$$\mathbb{E}\left[T_4^m\right] \leq C\frac{D_m}{n}.$$

**Proof of Lemma 3.3.** We can write

$$T_1^m = \left\|\sum_{j=1}^{D_m}\left(\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}} - \mathbb{E}[\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}}|(X_{-l})_l]\right)(\varphi_j \circ F_X)\right\|^2_{f_X}.$$

As the functions $\varphi_j$ are orthonormal, it becomes

$$T_1^m = \sum_{j=1}^{D_m}\left(\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2. \tag{3.36}$$

Now, $\mathbb{E}[T_1^m|(X_{-l})_l] = \sum_{j=1}^{D_m} \mathrm{Var}(\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}}|(X_{-l})_l)$, where $\mathrm{Var}(.|(X_{-l})_l)$ is the conditional variance with respect to the sample $(X_{-l})_{l\in\{1,\dots,n\}}$ (we denote by a similar notation the conditional expectation in the sequel). We work out it, for any index $j \in \{1,\dots,D_m\}$,

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_j^{F_X} - \hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right) &= \frac{1}{n}\mathrm{Var}\left(Y_1\left(\varphi_j\left(F_X(X_1)\right) - \varphi_j\left(\hat{F}_n(X_1)\right)\right)|(X_{-l})_l\right),\\ &\le \frac{1}{n}\mathbb{E}\left[s(X_1)^2\left(\varphi_j\left(F_X(X_1)\right) - \varphi_j\left(\hat{F}_n(X_1)\right)\right)^2|(X_{-l})_l\right]\\ &\quad + \frac{\sigma^2}{n}\mathbb{E}\left[\left(\varphi_j\left(F_X(X_1)\right) - \varphi_j\left(\hat{F}_n(X_1)\right)\right)^2|(X_{-l})_l\right].\end{aligned}$$

We use the mean value theorem: $(\varphi_j(F_X(X_1))-\varphi_j(\hat{F}_n(X_1)))^2 \le \|\varphi_j'\|^2_{L^\infty((0;1))}\|F_X-\hat{F}_n\|^2_{L^\infty((0;1))}$. This leads to

$$\begin{aligned}\mathbb{E}\left[T_1^m\,|(X_{-l})_l\right] &\le \frac{1}{n}\left(\mathbb{E}\left[s^2(X_1)\right]+\sigma^2\right)\sum_{j=1}^{D_m}\left\|\varphi_j'\right\|^2_{L^\infty((0;1))}\left\|F_X-\hat{F}_n\right\|^2_{L^\infty((a;b))},\\ &= \frac{1}{n}\mathbb{E}\left[Y_1^2\right]\sum_{j=1}^{D_m}\left\|\varphi_j'\right\|^2_{L^\infty((0;1))}\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))}.\end{aligned}$$

The sum is bounded by $D_m \times (D_m\|\varphi_2'\|_{L^\infty((0;1))})^2$, and we apply Proposition 2.4 (Chapter 2) with $p=2$, to conclude $\mathbb{E}[T_1^m] \le C_2\|\varphi_2'\|^2_{L^\infty((0;1))}\mathbb{E}[Y_1^2]D_m^3/n^2$.

□

**Proof of Lemma 3.4.** We write

$$\begin{aligned}&T_2^m\\ &= \int_{(a;b)}\left(\hat{g}_m^{\hat{F}}\left(F_X(x)\right) - \hat{g}_m^{\hat{F}}\left(\hat{F}_n(x)\right) - \mathbb{E}\left[\hat{g}_m^{\hat{F}}\left(F_X(x)\right) - \hat{g}_m^{\hat{F}}\left(\hat{F}_n(x)\right)|(X_{-l})_l\right]\right)^2 f_X(x)dx,\\ &= \int_{(0;1)}\left\{\sum_{j=1}^{D_m}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)\left(\varphi_j(u) - \varphi_j\left(\hat{U}_n(u)\right)\right)\right\}^2 du,\end{aligned}$$

We use the Cauchy-Schwarz Inequality, and by computations analogous of those of Lemma 3.3, we get

$$T_2^m \le \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} D_m^3\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))}\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2. \quad (3.37)$$

Moreover, $\mathbb{E}\left[\sum_{j=1}^{D_m}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2|(X_{-l})_l\right]$

$$\begin{aligned}
&= \sum_{j=1}^{D_m}\mathrm{Var}\left(\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right) = \frac{1}{n}\sum_{j=1}^{D_m}\mathrm{Var}\left(\varphi_j\left(\hat{F}_n(X_1)\right)Y_1\,|(X_{-l})_l\right),\\
&\leq \frac{1}{n}\sum_{j=1}^{D_m}\mathbb{E}\left[\varphi_j^2\left(\hat{F}_n(X_1)\right)Y_1^2\,|(X_{-l})_l\right] \leq \frac{1}{n}\left\|\sum_{j=1}^{D_m}\varphi_j\right\|^2_{L^\infty((0;1))}\mathbb{E}\left[Y_1^2\,|(X_{-l})_l\right],\\
&\leq \phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}.
\end{aligned}$$

We end the proof by applying Proposition 2.4 (Chapter 2) with $p=2$:

$$\begin{aligned}
\mathbb{E}\left[T_2^m\right] &\leq \mathbb{E}\left[Y_1^2\right]\phi_0^2\left\|\varphi_2'\right\|^2_{L^\infty((0;1))}\frac{D_m^4}{n}\mathbb{E}\left[\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))}\right],\\
&= C_2\mathbb{E}\left[Y_1^2\right]\phi_0^2\left\|\varphi_2'\right\|^2_{L^\infty((0;1))}\frac{D_m^4}{n^2}.
\end{aligned}$$

□

**Proof of Lemma 3.5.** The term $\mathbb{E}[T_3^m]$ requires more computations than the two terms above. Let us first notice that $T_3^m = \sum_{j=1}^{D_m}\{\int_0^1 s(F_X^{-1}(u))(\varphi_j(u) - \varphi_j(\hat{U}_n(u)))du\}^2$. We apply Taylor formula with Lagrange form for the remainder rest: there exists a random number depending on $j$, $\hat{\alpha}_{j,n,u}$, such that the following splitting holds:

$$T_3^m \leq 3T_{3,1}^m + 3T_{3,2}^m + 3T_{3,3}^m, \tag{3.38}$$

with notations

$$\begin{aligned}
T_{3,1}^m &= \sum_{j=1}^{D_m}\left\{\int_0^1 g(u)\left(\hat{U}_n(u) - u\right)\varphi_j'(u)du\right\}^2,\\
T_{3,2}^m &= (1/4)\sum_{j=1}^{D_m}\left\{\int_0^1 g(u)\left(\hat{U}_n(u) - u\right)^2\varphi_j''(u)du\right\}^2,\\
T_{3,3}^m &= (1/6)\sum_{j=1}^{D_m}\left\{\int_0^1 g(u)\left(\hat{U}_n(u) - u\right)^3\varphi_j^{(3)}(\hat{\alpha}_{j,n,u})du\right\}^2.
\end{aligned}$$

Writing the definition of $\hat{U}_n(u)$, and noting that $u = \mathbb{E}[\mathbf{1}_{U_i\leq u}]$ $(i=1,\ldots,n)$, we get for the first term

$$T_{3,1}^m = \sum_{j=1}^{D_m}\left(\frac{1}{n}\sum_{i=1}^{n}A_{i,j} - \mathbb{E}[A_{i,j}]\right)^2, \quad \text{with } A_{i,j} = \int_{U_i}^1 g(u)\varphi_j'(u)du.$$

An integration by parts so as to compute $A_{i,j}$ leads to

$$T_{3,1}^m \leq 2T_{3,1,1}^m + 2T_{3,1,2}^m, \tag{3.39}$$

with notations

$$\begin{aligned} T_{3,1,1}^m &= \sum_{j=1}^{D_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(U_i)\varphi_j(U_i) - \mathbb{E}\left[g(U_i)\varphi_j(U_i)\right] \right\}^2, \\ T_{3,1,2}^m &= \sum_{j=1}^{D_m} \left\{ \int_0^1 g'(u) \left(\hat{U}_n(u) - u\right) \varphi_j(u) du \right\}^2. \end{aligned} \tag{3.40}$$

The same study as the one done for $T_1^m$ gives

$$\begin{aligned} \mathbb{E}\left[T_{3,1,1}^m\right] &\leq \frac{1}{n} \sum_{j=1}^{D_m} \mathbb{E}\left[\left(g(U_1)\varphi_j(U_1)\right)^2\right] \leq \frac{1}{n} \left\| \sum_{j=1}^{D_m} \varphi_j^2 \right\|_{L^\infty((0;1))} \int_0^1 g^2(u) du, \\ &= \int_0^1 g(u)^2 du \phi_0^2 \frac{D_m}{n} = \phi_0^2 \mathbb{E}[s(X_1)^2] \frac{D_m}{n} \leq \phi_0^2 \mathbb{E}[Y_1^2] \frac{D_m}{n}. \end{aligned}$$

Besides, using definition and properties of the orthogonal projection on $S_m$,

$$T_{3,1,2}^m = \sum_{j=1}^{D_m} \left( \langle g'(\hat{U}_n - id), \varphi_j \rangle \right)^2 = \left\| \Pi_{S_m}(g'(\hat{U}_n - id)) \right\|^2 \leq \|g'\|^2 \|\hat{U}_n - id\|_{L^\infty((0;1))}^2.$$

Concluding with Proposition 2.4 (Chapter 2), $p = 2$, we obtain $\mathbb{E}[T_{3,1,2}^m] \leq C_2 \|g'\|^2 / n$. Hence,

$$\mathbb{E}\left[T_{3,1}^m\right] \leq 2 \left( C_2 \|g'\|^2 \frac{1}{n} + \phi_0^2 \mathbb{E}[Y_1^2] \frac{D_m}{n} \right) \leq C \frac{D_m}{n}.$$

Let us deal with $T_{3,2}^m$. We notice that for any $j \geq 2$, $\varphi_j'' = -(\pi\mu_j)^2 \varphi_j$, with $\mu_j = j$ for even $j$, and $\mu_j = j - 1$ for odd $j$. Consequently,

$$\begin{aligned} \mathbb{E}\left[T_{3,2}^m\right] &= (\pi^4/4)\mathbb{E}\left[ \sum_{j=1}^{D_m} \left\{ \int_0^1 g(u) \left(\hat{U}_n(u) - u\right)^2 \mu_j^2 \varphi_j(u) du \right\}^2 \right], \\ &\leq (\pi^4/4) D_m^4 \mathbb{E}\left[ \sum_{j=1}^{D_m} \left\{ \int_0^1 g(u) \left(\hat{U}_n(u) - u\right)^2 \varphi_j(u) du \right\}^2 \right], \\ &= (\pi^4/4) D_m^4 \mathbb{E}\left[ \sum_{j=1}^{D_m} \left\{ \langle g\left(\hat{U}_n - id\right)^2, \varphi_j \rangle \right\}^2 \right]. \end{aligned}$$

Proceeding as in the term $T_{3,1,2}$, we get $\mathbb{E}[T_{3,2}^m] \leq C_4(\pi^4/4)\|g\|^2 D_m^4 / n^2$. Last, we bound roughly

$$\mathbb{E}\left[T_{3,3}^m\right] \leq (1/6) \sum_{j=1}^{D_m} \left\| \varphi_j^{(3)} \right\|_{L^\infty((0;1))}^2 \|g\|^2 \mathbb{E}\left[ \left\| \hat{U}_n - id \right\|_{L^\infty((0;1))}^6 \right] \leq \frac{C_6}{6} \left\| \varphi_2^{(3)} \right\|_{L^\infty((0;1))}^2 \|g\|^2 \frac{D_m^7}{n^3}.$$

Finally, we gather the three bounds for $\mathbb{E}[T_{3,l}]$, $l = 1, 2, 3$, to end the proof of the inequality.

□

**Proof of Lemma 3.6.** First,

$$\begin{aligned} T_4^m &= \left\| \mathbb{E}\left[ \sum_{j=1}^{D_m} \hat{a}_j^{\hat{F}} \left( (\varphi_j \circ F_X) - (\varphi_j \circ \hat{F}_n) \right) | (X_{-l})_l \right] \right\|_{f_X}^2, \\ &\leq 2 \left\| \mathbb{E}\left[ \sum_{j=1}^{D_m} (\hat{a}_j^{\hat{F}} - a_j) \left( (\varphi_j \circ F_X) - (\varphi_j \circ \hat{F}_n) \right) | (X_{-l})_l \right] \right\|_{f_X}^2 \\ &\quad + 2 \left\| \mathbb{E}\left[ \sum_{j=1}^{D_m} a_j \left( (\varphi_j \circ F_X) - (\varphi_j \circ \hat{F}_n) \right) | (X_{-l})_l | \right] \right\|_{f_X}^2 := 2T_{4,1}^m + 2T_{4,2}^m. \end{aligned}$$

Then,

$$\begin{aligned} T_{4,1}^m &\leq \int_{(a;b)} \mathbb{E}\left[ \sum_{j=1}^{D_m} \left( \hat{a}_j^{\hat{F}} - a_j \right)^2 \sum_{j=1}^{D_m} \left( \varphi_j(F_X(x)) - \varphi_j(\hat{F}_n(x)) \right)^2 | (X_{-l})_l \right] f_X(x) dx, \\ &= \mathbb{E}\left[ \sum_{j=1}^{D_m} \left( \hat{a}_j^{\hat{F}} - a_j \right)^2 \sum_{j=1}^{D_m} \int_{(0;1)} \left( \varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 du \, | (X_{-l})_l \right], \\ &\leq 2T_{4,1,1}^m + 2T_{4,1,2}^m, \end{aligned}$$

with

$$\begin{aligned} T_{4,1,1}^m &= \int_{(0;1)} \mathbb{E}\left[ \left\{ \sum_{j=1}^{D_m} \left( \hat{a}_j^{\hat{F}} - \mathbb{E}\left[ \hat{a}_j^{\hat{F}} | (X_{-l})_l \right] \right)^2 \right\} \left\{ \sum_{j=1}^{D_m} \left( \varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 \right\} | (X_{-l})_l \right] du, \\ T_{4,1,2}^m &= \int_{(0;1)} \mathbb{E}\left[ \left\{ \sum_{j=1}^{D_m} \left( \mathbb{E}\left[ \hat{a}_j^{\hat{F}} | (X_{-l})_l \right] - a_j \right)^2 \right\} \left\{ \sum_{j=1}^{D_m} \left( \varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 \right\} | (X_{-l})_l \right] du. \end{aligned}$$

Let us now bound each term. The first one is

$$T_{4,1,1}^m = \sum_{j=1}^{D_m} \mathrm{Var}\left( \hat{a}_j^{\hat{F}} | (X_{-l})_l \right) \int_{(0;1)} \sum_{j=1}^{D_m} \left( \varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 du,$$

which is bounded using the mean value theorem:

$$T_{4,1,1}^{m'} \leq \sum_{j=1}^{D_m} \mathrm{Var}\left( \hat{a}_j^{\hat{F}} | (X_{-l})_l \right) D_m^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \left\| \hat{U}_n - id \right\|_{L^\infty((0;1))}^2.$$

As

$$\begin{aligned} \mathrm{Var}\left( \hat{a}_j^{\hat{F}} | (X_{-l})_l \right) &= \frac{1}{n} \mathrm{Var}\left\{ Y_1 \varphi_j\left( \hat{F}_n(X_1) \right) | (X_{-l})_l \right\}, \\ &\leq \frac{1}{n} \|\varphi_j\|_{L^\infty((0;1))}^2 \left( \mathbb{E}\left[ s^2(X_1) \right] + \sigma^2 \right) = \frac{1}{n} \|\varphi_j\|_{L^\infty((0;1))}^2 \mathbb{E}\left[ Y_1^2 \right], \end{aligned}$$

we obtain

$$T_{4,1,1}^m \leq \phi_0^2 \mathbb{E}\left[ Y_1^2 \right] \frac{D_m}{n} \times D_m^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \left\| \hat{U}_n - id \right\|_{L^\infty((0;1))}^2, \tag{3.41}$$

and finally, with Proposition 2.4 (Chapter 2), $p = 2$, we obtain

$$\mathbb{E}\left[T_{4,1,1}^m\right] \leq \phi_0^2 \mathbb{E}\left[Y_1^2\right] C_2 \|\varphi_2'\|_{L^\infty((0;1))}^2 \frac{D_m^4}{n^2}.$$

The second term can be written

$$T_{4,1,2}^m = T_3^m \int_{(0;1)} \sum_{j=1}^{D_m} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u))\right)^2 du,$$

and again by the mean value theorem

$$T_{4,1,2}^m \leq T_3^m D_m^3 \left\|\varphi_2'\right\|_{L^\infty((0;1))}^2 \left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2. \tag{3.42}$$

We have now to be more precise to bound $\|\hat{U}_n - id\|_{L^\infty((0;1))}$: we introduce the set $\{\|\hat{U}_n - id\|_{L^\infty((0;1))} \leq \alpha_n\}$, and $\{\|\hat{U}_n - id\|_{L^\infty((0;1))} > \alpha_n\}$, with $\alpha_n = \sqrt{2\ln(n)/n}$, which leads to $T_{4,1,2}^m \leq T_{4,1,2,1}^m + T_{4,1,2,2}^m$, with

$$\begin{aligned}
T_{4,1,2,1}^m &= D_m^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \alpha_n^2 T_3^m, \\
T_{4,1,2,2}^m &= D_m^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2 \mathbf{1}_{\|\hat{U}_n - id\|_{L^\infty((0;1))} > \alpha_n} T_3^m.
\end{aligned}$$

For $T_{4,1,2,1}^m$, we use Lemma 3.5 to bound $T_3^m$:

$$\begin{aligned}
\mathbb{E}\left[T_{4,1,2,1}^m\right] &\leq D_m^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \alpha_n^2 \times C\left(\frac{D_m}{n} + \frac{D_m^7}{n^3} + \frac{D_m^4}{n^2}\right), \\
&= C\left(\frac{D_m^4 \ln(n)}{n^2} + \frac{D_m^{10} \ln(n)}{n^4} + \frac{D_m^7 \ln(n)}{n^3}\right),
\end{aligned}$$

where $C$ depends on $\|g\|, \|g'\|, \|\varphi_2'\|_{L^\infty((0;1))}, \|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\mathbb{E}[Y_1^2]$, and $\phi_0^2$.

For $T_{4,1,2,2}^m$, we roughly bound $T_3^m$:

$$\begin{aligned}
\left(\mathbb{E}\left[\hat{a}_j^{\hat{F}} | \left(X_{-l}\right)_l\right] - a_j\right)^2 &= \left(\int_0^1 g(u)\left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u)\right) du\right)^2, \\
&\leq \|g\|^2 \|\varphi_j'\|_{L^\infty((0;1))}^2 \left\|id - \hat{U}_n\right\|_{L^\infty((0;1))}^2,
\end{aligned}$$

and thus $T_3^m \leq \|g\|^2 \|\varphi_2'\|_{L^\infty((0;1))}^2 \|id - \hat{U}_n\|_{L^\infty((0;1))}^2 D_m^3$. Applying the Cauchy-Schwarz Inequality and Propositions 2.3 and 2.4 (Chapter 2), we obtain

$$\begin{aligned}
\mathbb{E}\left[T_{4,1,2,2}^m\right] &\leq D_m^6 \|g\|^2 \|\varphi_2'\|_{L^\infty((0;1))}^4 \mathbb{E}\left[\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^4 \mathbf{1}_{\|\hat{U}_n - id\|_{L^\infty((0;1))} > \alpha_n}\right], \\
&\leq D_m^6 \|g\|^2 \|\varphi_2'\|_{L^\infty((0;1))}^4 \frac{\sqrt{C_8}}{n^2} \exp\left(-n\alpha_n^2\right), \\
&= \|g\|^2 \sqrt{C_8} \|\varphi_2'\|_{L^\infty((0;1))}^4 \frac{D_m^6}{n^4}.
\end{aligned}$$

Moreover, $T_{4,2}^m = \| \sum_{j=1}^{D_m} a_j((\varphi_j \circ F_X) - (\varphi_j \circ \hat{F}_n))\|_{f_X}^2$, so

$$\begin{aligned} \mathbb{E}\left[T_{4,2}^m\right] &\leq \mathbb{E}\left[\left\| \sum_{j=1}^{D_m} a_j \left((\varphi_j \circ F_X) - (\varphi_j \circ \hat{F}_n)\right)\right\|_{f_X}^2\right], \\ &= \mathbb{E}\left[\sum_{j,k=1}^{D_m} a_j a_k \int_0^1 (\varphi_j(u) - \varphi_j \circ \hat{U}_n(u))(\varphi_k(u) - \varphi_k \circ \hat{U}_n(u))du\right]. \end{aligned} \tag{3.43}$$

This yields, with Taylor formula, $\mathbb{E}[T_{4,2}^m] \leq \mathbb{E}[T_{4,2,1}^m + T_{4,2,2}^m + T_{4,2,3}^m]$, with

$$\begin{aligned} T_{4,2,1}^m &= \sum_{j,k=1}^{D_m} a_j a_k \int_0^1 (u - \hat{U}_n(u))^2 \varphi_j'(u)\varphi_k'(u)du, \\ T_{4,2,2}^m &= (1/4)\sum_{j,k=1}^{D_m} a_j a_k \int_0^1 (u - \hat{U}_n(u))^4 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_k''(\hat{\alpha}_{k,n,u})du, \\ T_{4,2,3}^m &= \sum_{j,k=1}^{D_m} a_j a_k \int_0^1 (u - \hat{U}_n(u))^3 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_k'(u)du, \end{aligned}$$

recalling that $a_l = \langle g, \varphi_l \rangle$. First,

$$\begin{aligned} T_{4,2,1}^m &= \int_0^1 (u - \hat{U}_n(u))^2 \left\{(\Pi_{S_m}(g))'(u)\right\}^2 du, \\ &= \int_0^1 (u - \hat{U}_n(u))^2 \left\{\Pi_{S_m}(g')(u)\right\}^2 du, \\ &\leq \left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2 \left\|\Pi_{S_m}(g')\right\|^2 \leq \left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2 \left\|g'\right\|^2. \end{aligned}$$

Therefore, $\mathbb{E}[T_{4,2,1}^m] \leq C_2\|g'\|^2/n$. Then, notice that

$$T_{4,2,2}^m = (1/4)\mathbb{E}\left[\int_0^1 (u - \hat{U}_n(u))^4 \left(\sum_{j=1}^{D_m} a_j \varphi_j''(\hat{\alpha}_{j,n,u})\right)^2 du\right],$$

we bound the Fourier's coefficients of the function $g$. To that end, we introduce the real numbers $\mu_j$, for $j \in \{1, \ldots, D_m\}$, defined by $\mu_j = j$ if $j$ is even, $\mu_j = j - 1$ otherwise. We obtain:

$$\left(\sum_{j=1}^{D_m} a_j \varphi_j''(\hat{\alpha}_{j,n,u})\right)^2 = \|\varphi_2''\|_{L^\infty((0;1))}^2 \left(\sum_{j=1}^{D_m} a_j \mu_j^2\right)^2 \leq \|\varphi_2''\|_{L^\infty((0;1))}^2 \left(\sum_{j=1}^{D_m} a_j^2 \mu_j^2\right) \sum_{j=1}^{D_m} \mu_j^2.$$

The function $g$ belongs to the Sobolev space $W_2^{1,per}(L)$, because $g(0) = g(1)$, $g$ belongs to $\mathcal{C}^1((0;1))$, and $\|g\|^2 = \|s\|_{f_X}^2 \leq L^2$. Thus we use Lemma A.3 (p. 162) from Tsybakov (2009) (see also Proposition 2.8 of Chapter 2): the sequence $(a_j)_j$ belongs to the ellipsoïd $\Theta(1, L^2/\pi^2)$, so

$$T_{4,2,2}^m \leq C\mathbb{E}\left[\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^4 D_m^3\right] \leq C\frac{D_m^3}{n^2}.$$

Following the same line of computations, we write,

$$T^m_{4,2,3} = \mathbb{E}\left[\int_0^1 (u-\hat{U}_n(u))^3 \left(\sum_{j=1}^{D_m} a_j \varphi_j''(\hat{\alpha}_{j,n,u})\right)\left(\sum_{k=1}^{D_m} a_k \varphi_k'(u)\right) du\right],$$

and bound as follows, for $u \in (0;1)$

$$\left|\sum_{j=1}^{D_m} a_j \varphi_j''(\hat{\alpha}_{j,n,u})\right| \leq \|\varphi_2''\|_{L^\infty((0;1))} \frac{L}{\pi} D_m^{3/2}, \quad \left|\sum_{k=1}^{D_m} a_k \varphi_k'(u)\right| \leq \|\varphi_2'\|_{L^\infty((0;1))} \frac{L}{\pi} D_m^{1/2}.$$

Consequently, $\mathbb{E}[T^m_{4,2,3}] \leq C\mathbb{E}[\|\hat{U}_n - id\|^3_{L^\infty((0;1))} D_m^2] \leq C D_m^2/n^{3/2}$. The proof of Lemma 3.6 is thus completed.

□

**Main part of the proof of Theorem 3.2**

We only study the most original estimator, $\tilde{s}_2^{\hat{F}}$, selected with the GL method (the proof for the other estimator can be found in Chagny 2011). The proof follows almost the same line as the one of Theorem 3.1, with further technicalities. We use the notations introduced in Section 3.5.3 (proof of Proposition 3.1). Let $S_m$ be a fixed model in the collection indexed by $\mathcal{M}_n$. To recover the framework of the proof of Theorem 3.1, we begin with the decomposition

$$\begin{aligned}
\left\|\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} - s\right\|_g^2 &\leq 3\left\|\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{f}_{\hat{m}}^{\hat{F},F_X} - \mathbb{E}\left[\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{s}_{\hat{m}}^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_g^2 \\
&\quad +3\left\|\mathbb{E}\left[\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{s}_{\hat{m}}^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_g^2 + 3\left\|\hat{s}_{\hat{m}}^{\hat{F},F_X} - s\right\|_g^2, \\
&= 3T_2^{\hat{m}} + 3T_4^{\hat{m}} + 3\left\|\hat{g}_{\hat{m}}^{\hat{F}} - g\right\|^2.
\end{aligned}$$

Thus, we can introduce $A$ and $V$, in the last term, in a similar way as previously:
$\left\|\hat{g}_{\hat{m}}^{\hat{F}} - g\right\|^2$

$$\begin{aligned}
&\leq 3\left\|\hat{g}_{\hat{m}}^{\hat{F}} - \hat{g}_{m\wedge\hat{m}}^{\hat{F}}\right\|^2 + 3\left\|\hat{g}_{m\wedge\hat{m}}^{\hat{F}} - \hat{g}_m^{\hat{F}}\right\|^2 + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2, \\
&\leq 3\left(A(m) + V\left(\hat{m}\right)\right) + 3\left(A\left(\hat{m}\right) + V\left(\hat{m}\right)\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2, \\
&= 3\left(A(m) + 2V\left(m\right)\right) + 3\left(A\left(\hat{m}\right) + 2V\left(\hat{m}\right)\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2 - 3V\left(\hat{m}\right) - 3V\left(m\right), \\
&\leq 6\left(A(m) + 2V(m)\right) - 2V\left(\hat{m}\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2,
\end{aligned}$$

using the definition of $\hat{m}$. The last term of this decomposition is bounded by:

$$\left\|\hat{g}_m^{\hat{F}} - g\right\|^2 = \left\|\hat{s}_m^{\hat{F},F_X} - s\right\|_g^2 \leq 3T_1^m + 3T_3^m + 3T_0^m,$$

where $T_l^m$ $(l = 0, 1, 3)$ are defined by (3.34). As a result, we get

$$\begin{aligned}
\left\|\hat{s}_{\hat{m}}^{\hat{F},\hat{F}} - s\right\|_g^2 &\leq 3T_2^{\hat{m}} + 3T_4^{\hat{m}} - 3 \times 2V(\hat{m}) + 3 \times 6\left(A(m) + V(m)\right) \\
&\quad +3 \times 3 \times \left(3T_1^m + 3T_3^m + 3T_0^m\right).
\end{aligned}$$

Therefore, it follows from Inequalities (3.6) and (3.8) that

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{s}_{\hat{m}}^{\hat{F},\hat{F}}-s\right\|_{f_X}^2\right] &\leq 18\left(\mathbb{E}\left[A(m)\right]+V(m)\right)+3\mathbb{E}\left[\left(T_2^{\hat{m}}-V(\hat{m})\right)_+\right]+3\mathbb{E}\left[\left(T_4^{\hat{m}}-V(\hat{m})\right)_+\right]\\ &\quad+\mathbb{E}\left[T_1^m\right]+\mathbb{E}\left[T_3^m\right]+27\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}+27\left\|s-s_m^{F_X}\right\|_{f_X}^2.\end{aligned}$$

A bound for $A(m)$ is given by the following lemma, whose proof is deferred to Section 3.5.3.

**Lemma 3.7.** *Under the assumptions of Theorem 3.1, there exists a constant $C>0$ depending on $\|\varphi_2^{(l)}\|$ $(l=1,3)$, $\|g\|$, $\|g'\|$, and $\mathbb{E}[Y_1^2]$, such that, for each index $m\in\mathcal{M}_n$,*

$$\begin{aligned}\mathbb{E}\left[A(m)\right] &\leq 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{48}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{48}\right)_+\right]\\ &\quad+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m'}\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m\wedge m'}\right]+12\left\|s_m^{F_X}-s\right\|_{f_X}^2+\frac{C}{n}.\end{aligned}$$

Applying this lemma, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{s}_{\hat{m}}^{\hat{F},\hat{F}}-s\right\|_{f_X}^2\right] &\leq C\left(\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m'}\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m\wedge m'}\right]+\mathbb{E}\left[T_1^m\right]\right.\\ &\quad+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{48}\right)_+\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{48}\right)_+\right]\\ &\quad\left.+\mathbb{E}\left[T_3^m\right]+\mathbb{E}\left[\left(T_2^{\hat{m}}-V(\hat{m})\right)_+\right]+\mathbb{E}\left[\left(T_4^{\hat{m}}-V(\hat{m})\right)_+\right]\right)\\ &\quad+C\left(\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}+\left\|s-s_m^{F_X}\right\|_{f_X}^2+\frac{1}{n}\right).\end{aligned}$$

It remains to study the terms $T_l^m$, $l=1,\dots,4$. Bounding $(T_l^{\hat{m}}-V(\hat{m}))_+\leq\max_{m'}(T_l^{m'}-V(m'))_+$ $(l=2,4)$, it is enough to apply Lemmas 3.8 to 3.11 below to conclude: we have just to choose the constant in the definition of $V$ larger than the ones of $V_l$ $(l=2,3,4)$.

**Lemma 3.8.** *Assuming that the models are trigonometric, and that $D_{m_{\max}}=O(n^{1/2})$, there exists a constant $C>0$ (depending on $\|\varphi_2'\|_{L^\infty((0;1))}$ and $\mathbb{E}[Y_1^2]$) such that*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m'}\right]\leq C\frac{D_{m_{\max}}}{n}.$$

**Lemma 3.9.** *Assuming that the models are trigonometric, that $D_{m_{\max}}=O(n^{1/3})$ and that there exists a real-number $p>8/3$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right]<\infty$, there exists a constant $C>0$ (depending on $\|\varphi_2'\|_{L^\infty((0;1))}$ and $\mathbb{E}[Y_1^2]$) such that*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_2^{m'}-V_2(m')\right)_+\right]\leq C\frac{\ln(n)}{n},$$

*with $V_2(m')=\kappa\kappa' D_{m'}^4\ln^2(n)/n^2$, and $\kappa'=7/3$, $\kappa=96\phi_0^2\mathbb{E}[Y_1^2]\|\varphi_2'\|_{L^\infty((0;1))}^2$.*

Assuming that $D_{m'}=O((n/\ln(n)^2)^{1/3})$, we get

$$V_2(m')\leq\kappa\kappa'\frac{D_{m'}}{n}:=V_2^{bis}(m').$$

The result of Lemma 3.9 holds with $V_2^{bis}$ in place of $V_2$.

**Lemma 3.10.** *Assume that the models are trigonometric, that* $D_{m_{\max}} = O(n^{1/3}/\ln(n))$, *and that* $g \in \mathcal{C}^1((0;1))$. *For all* $m \in \mathcal{M}_n$, *the following inequality holds, for* $p_{m'} = m'$ *or* $p_{m'} = m \wedge m'$:

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(T_3^{p_{m'},b} - V_3(m')\right)_+\right] \leq \frac{C}{n}.$$

*with* $V_3(m') = k_3 D_{m'}/n$, *and* $k_3$ *a numerical constant depending only (and linearly) on* $\mathbb{E}[Y_1^2]$, *and where the constant* $C > 0$ *depends on* $\|\varphi_2'\|_{L^\infty((0;1))}$, $\|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\|g\|$, $\|g'\|$, $\mathbb{E}[Y_1^2]$.

**Lemma 3.11.** *Assuming that the models are trigonometric, that* $D_{m_{\max}} = O(n^{1/3}/\ln(n))$, *and that* $g \in W_2^{1,per}(L)$ *(*$L > 0$*), there exists a constant* $C > 0$ *(depending on* $\|\varphi_2'\|_{L^\infty((0;1))}$, $\|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\|g\|$, $\|g'\|$, $\mathbb{E}[Y_1^2]$*) such that, for all* $m \in \mathcal{M}_n$, $n \geq n_0 = \exp\left(\|g'\|^2\right)$,

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(T_4^{m'} - V_4(m')\right)_+\right] \leq C\frac{\ln(n)}{n},$$

*with* $V_4(m') = k_4 D_{m'}/n$, *and* $k_4$ *a numerical constant depending only (and linearly) on* $\mathbb{E}[Y_1^2]$.

Notice that it is also possible to obtain the result for any $n \in \mathbb{N}$. But the price to pay is a penalty $V_4$ depending on $\|g'\|^2$.

□

**Proof of Lemma 3.7**

The following proof is close to the proof of Lemma 3.2. Fix an index $m' \in \mathcal{M}_n$. We split

$$\left\|\hat{g}_{m'}^{\hat{F}} - \hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2 \leq 3\left\|\hat{g}_{m'}^{\hat{F}} - g_{m'}\right\|^2 + 3\left\|g_{m'} - g_{m\wedge m'}\right\|^2 + 3\left\|g_{m\wedge m'} - \hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2.$$

Relation (3.31) still holds for an other empirical process, and by applying Lemma 3.1, we have, for $p = m'$ or $p = m \wedge m'$ $\|g_p - \hat{g}_p^{\hat{F}}\|^2 = \sup_{t \in \mathcal{S}(p)} \tilde{\nu}_n(t)^2$, with, for $t \in L^2((0;1))$,

$$\tilde{\nu}_n(t) = \frac{1}{n}\sum_{i=1}^n Y_i\left(t \circ \hat{F}_n\right)(X_i) - \mathbb{E}\left[Y_i\left(t \circ F_X\right)(X_i)\right].$$

We split $\tilde{\nu}_n$ into $\tilde{\nu}_n = \nu_n + R_n$, with

$$R_n(t) = \frac{1}{n}\sum_{i=1}^n Y_i t(\hat{F}_n(X_i) - F_X(X_i)).$$

This yields to $\tilde{\nu}_n^2 \leq 2\nu_n^2 + 2R_n^2$. If $t$ belongs to $\mathcal{S}(p)$, $t = \sum_{j=1}^{D_p} \theta_j \varphi_j$ with $\sum_{j=1}^{D_p} \theta_j^2 = 1$, so that

$$\begin{aligned}\sup_{t \in \mathcal{S}(p)} R_n^2(t) &= \sup_{\substack{\theta \in \mathbb{R}^p \\ \sum_j \theta_j^2 = 1}} \left(\sum_{j=1}^{D_p} \theta_j \frac{1}{n}\sum_{i=1}^n Y_i \varphi_j(\hat{F}_n(X_i) - F_X(X_i))\right)^2, \\ &= \sup_{\substack{\theta \in \mathbb{R}^p \\ \sum_j \theta_j^2 = 1}} \left(\sum_{j=1}^{D_p} \theta_j\left(\hat{a}_j^{\hat{F}} - \hat{a}_j^{F_X}\right)\right)^2 = \sum_{j=1}^{D_p}\left(\hat{a}_j^{\hat{F}} - \hat{a}_j^{F_X}\right)^2,\end{aligned}$$

by using the same arguments as in the proof of Lemma 3.1. Introducing the conditional expectation of $\hat{a}_j^{\hat{F}} - \hat{a}_j^{F_X}$, we note that $\sup_{t\in\mathcal{S}(p)} R_n^2(t) \leq 2T_1^p + 2T_3^p$. We obtain,

$$\left\| g_p - \hat{g}_p^{\hat{F}} \right\|^2 \leq 2 \sup_{t\in\mathcal{S}(p)} (\nu_n(t))^2 + 4T_1^p + 4T_3^p.$$

and thus, substracting $V(m')$ and taking expectation, this yields
$\mathbb{E}[A(m)]$

$$\begin{aligned}
\leq\ & 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2 - \frac{V(m')}{24}\right)_+\right] + 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2 - \frac{V(m')}{24}\right)_+\right] \\
& +12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'} - \frac{V(m')}{48}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'} - \frac{V(m')}{48}\right)_+\right] \\
& +12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m'}\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m\wedge m'}\right] + 3\max_{m'\in\mathcal{M}_n}\|g_{m'} - g_{m\wedge m'}\|^2.
\end{aligned}$$

The last term is denoted by $T_c^m$ in (3.30) and proved to be bounded by $4\|g_m - g\|^2$ (see the proof of Lemma 3.2). Moreover, applying Proposition 3.2 yields to

$$\begin{aligned}
\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2 - p(m')\right)_+\right] &\leq \frac{C}{n}, \\
\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2 - p(m')\right)_+\right] &\leq \frac{C}{n},
\end{aligned}$$

using $-p(m') \leq -p(m\wedge m')$ (remember that $p(m') = C\phi_0^2\mathbb{E}\left[Y_1^2\right] D_{m'}/n$). By gathering the last bounds, and noting that the constant $c_v'$ (in the definition of $V(m')$) can be chosen larger than the one of $p(m')$, we obtain the result of Lemma 3.7.

□

**Proof of Lemmas 3.8 to 3.11**

In this section we state upper bounds for $T_l^m$, $l = 1,\ldots,4$ (see (3.34)). Recall that $m_{\max}$ is the index of the largest model in the collection. Notice that $D_{m_{\max}} \geq m_{\max}$, since we work with the trigonometric model. Recall that we denote by $a_j$ the Fourier coefficients of the function $g$, that is, $g_m = \sum_{j=1}^{D_m} a_j\varphi_j$, where $g_m$ is the orthogonal projection on the space $S_m$, $m\in\mathcal{M}_n$.
The sketch of all the proofs can be described by the following cases:

($\mathcal{A}$) Some of the terms are less than $CD_m/n$, under the constraint $D_m \leq Cn^{1/3}/\ln(n)$, and so we do not need to center them. For example, they involve expectations of form $\mathbb{E}[\sum_{j=1}^{D_m}(\varphi_j(F_X(X_1)) - \varphi_j(\hat{F}_n(X_1)))^2]$. By using a Taylor formula, we come down to terms of form $\sum_{j=1}^{D_m}(\varphi_j^{(k)})^2\mathbb{E}[\|\hat{U}_n - id\|^{2k}_{L^\infty((0;1))}]$ ($k$ an integer), and bound them with Proposition 2.4 (Chapter 2). This is the case for $T_1^m$ (Lemma 3.8), $T_3^m$, first inequality (first part of Lemma 3.10), and for some terms of the decomposition of $T_4^m$ (see proof of Lemma 3.11).

($\mathcal{B}$) The other terms have to be centered to be negligible. Then, there are two possibilities:
- ($\mathcal{B}_1$) The first one is to make emerge the supremum of an empirical process (with Lemma 3.1) and the to use the Talagrand Inequality (2.1). This is the case for a part of $T_2^m$ and $T_3^m$ (Lemmas 3.9 and 3.10, second inequality).
- ($\mathcal{B}_2$) The second is to bound these terms by quantity of form $C(D_m)\|\hat{U}_n - id\|^k_{L^\infty((0;1))}$ ($k$ an integer, $C(D_m)$ a constant depending on $D_m$), and to use Inequality (2.5) or (2.6) of Corollary 2.1 (Chapter 2).This is the case for the other parts of $T_2^m$ and $T_3^m$ (Lemmas 3.9 and 3.10, second inequality).

For sake of conciseness, we do not detail all of the proofs, especially the ones which follow a line already described. However, the lector can find all the details in Chagny (2011).

**Proof of Lemma 3.8.** Inequality (3.36) shows that $T_1^{m'} \leq T_1^{m_{\max}}$ and $\mathbb{E}[\max_{m'\in\mathcal{M}_n} T_1^{m'}] \leq \mathbb{E}[T_1^{m_{\max}}]$. Thus it is sufficient to bound $\mathbb{E}[T_1^{m_{\max}}]$. This has already be done in Lemma 3.3. We apply it, this completes the proof.

□

**Proof of Lemma 3.9.** Begining with $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_2^{m'} - V_2(m'))_+] \leq \sum_{m'\in\mathcal{M}_n}\mathbb{E}[(T_2^{m'} - V_2(m'))_+]$, we have just to study this quantity for each index $m'$. Thanks to Inequality (3.37) of Lemma 3.4,

$$\begin{aligned}\mathbb{E}\left[\left(T_2^{m'} - V_2(m')\right)_+\right] &\leq D_{m'}^3 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \mathbb{E}\left[\left(\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2 \left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))}\right.\right.\\ &\qquad \left.\left. - \frac{\kappa\kappa'}{\|\varphi_2'\|^2_{L^\infty((0;1))}}\frac{D_{m'}}{n^2}\ln^2(n)\right)_+\right],\\ &\leq T_{2,a}^{m'} + T_{2,b}^{m'},\end{aligned}$$

denoting by

$$T_{2,a}^{m'} = D_{m'}^3 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \mathbb{E}\left[\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2 \left(\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))} - \kappa'\frac{\ln(n)}{n}\right)_+\right],$$

$$T_{2,b}^{m'} = D_{m'}^3 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \kappa'\frac{\ln(n)}{n}\mathbb{E}\left[\left(\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^2 - \frac{\kappa}{\|\varphi_2'\|^2_{L^\infty((0;1))}}\frac{D_{m'}}{n}\ln(n)\right)_+\right].$$

For the term $T_{2,a}^{m'}$, we first obtain

$$T_{2,a}^{m'} = D_{m'}^3 \left\|\varphi_2'\right\|^2_{L^\infty((0;1))} \sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^4\right]^{1/2} \mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^2_{L^\infty((0;1))} - \kappa'\frac{\ln(n)}{n}\right)^2\right]^{1/2},$$

and roughly bound

$$\sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\hat{a}_j^{\hat{F}} - \mathbb{E}\left[\hat{a}_j^{\hat{F}}\,|(X_{-l})_l\right]\right)^4\right] \leq 16\phi_0^4\mathbb{E}\left[Y_1^4\right] D_{m'}.$$

Gathering this bound with Inequality (2.5) of Corollary 2.1 (Chapter 2) leads to,

$$\sum_{m' \in \mathcal{M}_n} T_{2,a}^{m'} \leq C \sum_{m' \in \mathcal{M}_n} D_{m'}^4 n^{-1-\kappa'} \leq C n^{4/3-\kappa'} \leq C n^{-1}$$

as soon as $D_{m'} \leq C n^{1/3}$ and for $\kappa' = 7/3$. For the second term $T_{2,b}^{m'}$, thanks to Lemma 3.1, we notice first that $\sum_{j=1}^{D_{m'}} \left( \hat{a}_j^{\hat{F}} - \mathbb{E}\left[ \hat{a}_j^{\hat{F}} \, | (X_{-l})_l \right] \right)^2 = \sup_{t \in \mathcal{S}(m')} \bar{\nu}_n^2(t)$, with, for $t \in L^2((0;1))$,

$$\bar{\nu}_n(t) = \frac{1}{n} \sum_{i=1}^{n} Y_i t\left( \hat{F}_n(X_i) \right) - \mathbb{E}\left[ Y_i t\left( \hat{F}_n(X_i) \right) | (X_{-l})_l \right],$$

a process which is centered conditionally to the sample $(X_{-l})_l$. We must now bound its deviations, exactly as we bound the one of the process $\nu_n$, in the proof of Proposition 3.2, but conditionally to the variables $X_{-l}$. Let us just recall the sketch of the proof: we split $\bar{\nu}_n$ into three parts, taking into account that $Y_i = s(X_i) + \varepsilon_i(\mathbf{1}_{|\varepsilon| \leq \kappa_n} + \mathbf{1}_{|\varepsilon| > \kappa_n})$. We get thus three terms: the two main terms are bounded, and are hence controlled with the Talagrand Inequality (2.1). We finally obtain,

$$\sum_{m' \in \mathcal{M}_n} T_{2,b}^{m'} \leq C \frac{\ln(n)}{n},$$

which completes the proof.

□

**Proof of Lemma 3.10** Let us begin with $V_3(p_{m'}) \leq V_3(m')$. Therefore $\mathbb{E}[\max_{m' \in \mathcal{M}_n} (T_3^{p_{m'},b} - V_3(m'))_+] \leq \mathbb{E}[\max_{m' \in \mathcal{M}_n} (T_3^{p_{m'},b} - V_3(p_{m'}))_+]$. In the sequel, we simplify the notations by setting $p = p_{m'}$. As previously, we get $T_3^{p,b} \leq 6T_{3,1,1}^p + 6T_{3,1,2}^p + 3T_{3,2}^p + 3T_{3,3}^p$. Thus

$$\begin{aligned} \mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( T_3^{p,b} - V_3(p) \right)_+ \right] \leq & \; \mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( 6T_{3,1,1}^p - V_3(p)/3 \right)_+ \right] + \mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} 6T_{3,1,2}^p \right] \\ & + \mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( 3T_{3,2}^p - V_3(p)/3 \right)_+ \right] \\ & + \mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( 3T_{3,3}^p - V_3(p)/3 \right)_+ \right]. \end{aligned} \tag{3.44}$$

The term that we have not centered is directly negligible: its definition (see (3.40)) proves that $T_{3,1,2}^p \leq T_{3,1,2}^{m_{\max}}$, thus we obtain

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} 6T_{3,1,2}^p \right] \leq \frac{C}{n}. \tag{3.45}$$

It remains to bound the three other terms. Let us distinguish $T_{3,1,1}^p$ of the two others: Equality (3.31) and Lemma 3.1 lead to $T_{3,1,1}^p = \sup_{t \in \mathcal{S}(p)} (\nu_n^{(1)}(t))^2$, for the process defined by

$$\nu_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^{n} s(X_i) \left( t \circ F_X \right)(X_i) - \mathbb{E}\left[ f(X_i) \left( t \circ F_X \right)(X_i) \right].$$

Thus we apply Talagrand Inequality (2.1), as in the proof of Proposition 3.2. The useful quantities are the following:

$$M_1^{(1)} = \phi_0 \|s\|_{L^\infty((a;b))} \sqrt{D_p},\ \left(H^{(1)}\right)^2 = \frac{D_p}{n} \mathbb{E}\left[s^2(X_1)\right] \phi_0^2,\ v^{(1)} = \|s\|^2_{L^\infty((a;b))}.$$

We have again

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(6T^p_{3,1,1} - V_{3,1,1}(p)\right)_+\right] \leq \frac{C}{n}, \tag{3.46}$$

with $V_{3,1,1}(p) = 6 \times 2(1+2\delta)\mathbb{E}\left[s^2(X_1)\right] \phi_0^2 D_p/n$. But as

$$V_{3,1,1}(p) \leq 12(1+2\delta)\mathbb{E}\left[Y_1^2\right] \phi_0^2 \frac{D_p}{n} := V^{bis}_{3,1,1}(p),$$

the result holds with $V^{bis}_{3,1,1}$.
For the two other terms, the strategy is the one described in $(\mathcal{B}_2)$ (beginning of this section). For example, using $T^p_{3,2} \leq (\pi^4/4)\|g\|^2 D_p^4 \left\|\hat{U}_n - id\right\|^4_{L^\infty((0;1))}$ implies, for $V_{3,2}(p) = \kappa D_p^4 \ln^2(n)/n^2$,
$\mathbb{E}\left[\left(3T^p_{3,2} - V_{3,2}(p)\right)_+\right]$

$$\begin{aligned} &\leq (3\pi^4/4)\|g\|^2 D_p^4 \mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^4_{L^\infty((0;1))} - \frac{\kappa}{(3\pi^4/4)\|g\|^2}\frac{\ln^2(n)}{n^2}\right)_+\right], \\ &\leq C D_p^4 n^{-\kappa_b^{1/2} 2^{-1/2}}, \end{aligned} \tag{3.47}$$

for $\kappa_b = \kappa/(3\pi^4/4)\|g\|^2$. Thus, if $D_p \leq Cn^{1/3}$,

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(3T^p_{3,2} - V_{3,2}(p)\right)_+\right] \leq Cn \times n^{4/3} \times n^{-\kappa_b^{1/2} 2^{-1/2}}.$$

The choice of $\kappa = 50\pi^4/3\|g\|^2$ leads successively to $\kappa_b \geq 200/9$, and to $7/3 - \sqrt{\kappa_b/2} \leq -1$, so that the last upper-bound is $O(1/n)$. If $D_p \leq Cn^{1/3}/\ln(n)$, we have

$$V_{3,2}(p) \leq 50\pi^4/3\mathbb{E}\left[Y_1^2\right] \frac{D_p}{n} := V^{bis}_{3,2}(p),$$

which can also be used. We do not detail the control for the term $T^m_{3,3}$. Similarly, we get

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(3T^p_{3,3} - V^{bis}_{3,3}(p)\right)_+\right] \leq C/n, \tag{3.48}$$

with $V^{bis}_{3,3}(p) = (13^3 \times 2/27)\|\varphi_2^{(3)}\|^2_{L^\infty((0;1))}\mathbb{E}\left[Y_1^2\right] D_p/n$. We conclude the proof of Lemma 3.10 by gathering Inequalities (3.45), (3.46), (3.47), and (3.48), in the bound (3.44), and choosing the constant $k_3$ such that $V_3 \geq 3V^{bis}_{3,1,1}$, $V_3 \geq 3V^{bis}_{3,2}$, and $V_3 \geq 3V^{bis}_{3,3}$.

□

**Proof of Lemma 3.11.** The sketch of the proof is the same as the proof of Lemma 3.10, that is why we do not detail all the computations. We split

$$T_4^{m'} \leq 4T_{4,1,1}^{m'} + 4T_{4,1,2}^{m'} + 2T_{4,2,1}^{m'} + 2T_{4,2,2}^{m'} + 2T_{4,2,3}^{m'},$$

where the different terms are defined in the proof of Lemma 3.6 (see Paragraph 3.5.3), and thus, $\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_4^{m'} - V_4(m')\right)_+\right]$

$$\begin{aligned}\leq\ & \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T_{4,1,1}^{m'} - V_4(m')/3\right)_+\right] + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T_{4,1,2}^{m'} - V_4(m')/3\right)_+\right] \\ & + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(2T_{4,2,3}^{m'} - V_4(m')/3\right)_+\right] + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} 2T_{4,2,1}^{m'}\right] + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} 2T_{4,2,2}^{m'}\right].\end{aligned}$$

We show that the two terms which we have not centered are negligible (less than $C\ln(n)/n$ if $D_{m_{\max}} = O(n^{1/3})$. For the three others we apply the strategy $(\mathcal{B}_2)$.

First, Inequality (3.41) allows us to conclude that as announced,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T_{4,1,1}^{m'} - V_4(m')/3\right)_+\right] \leq \frac{C}{n};$$

by Inequality (2.3) (Corollary 2.1, Chapter 2). Then, for the term involving $T_{4,1,2}^m$, we use (3.42), and replace in this inequality $T_3^{m'}$ by its detailed bound. We obtained by gathering Inequalities (3.38) and (3.39):

$$T_3^{m'} \leq 6T_{3,1,1}^{m'} + 6T_{3,1,2}^{m'} + 3T_{3,2}^{m'} + 3T_{3,3}^{m'}.$$

This leads to $T_{4,1,2}^{m'} \leq \sum_{l=1}^4 T_{4,1,2,l}^{m'}$, and then $T_{4,1,2,l}^{m'} \leq C\|\hat{U}_n - id\|_{L^\infty((0;1))}^{p_l}$ ($p_l$ an integer), so that we can use the method $(\mathcal{B}_2)$, for each of this four terms. We prove easily $\mathbb{E}[\max_{m'\in\mathcal{M}_n} T_{4,2,1}^{m'}] \leq C_2\|g'\|^2/n$. The same type of inequality remains true for $T_{4,2,2}^{m'}$. And finally, $T_{4,2,3}^{m'}$ can be handled with the methode $(\mathcal{B}_2)$.

□

## 3.6 Appendix 1: More about the practical calibration of the penalty constants

We detail Remark 3.1 in this section, and go deeper into the choice of the penalty constants.

Both of the estimation methods we implement in the simulation study of Section 3.4 heavily depend on the choice of the numerical constants involved in the definition of the penalties ($c_1$ for $\text{pen}^{F_X}$, and $c_1'$ for pen, in the case of unknown c.d.f. $F_X$), and of the terms $V$ of the GL method ($c_1'$ and $c_2'$). From the theoretical point of view, the constants are purely numerical. From the practical point of view, values for these constants are needed for the

algorithm to run. Since we present simulation results in the general case of unknown design distribution, we focus on the choice of the value of $c_1'$ (penalization method), and of $c_2'$ (GL method). Recall Remark 3.1: the slope heuristic (Birgé & Massart 2007) can be used only to choose $c_1'$, and not to calibrate the constant $c_2'$ of $V$ in the recent GL method. It would have been quite unfair to experiment one method with a constant adjusted for each sample and the other with a fixed constant. That is why we fix constants for both strategies, but try to do a reasonable choice:

- For the constant $c_1'$ of the penalty, we use a preliminary study, based on slope heuristic, with the so-called "dimension jump method". We choose two models $Y_i = s(X_i) + \varepsilon_i$, which are also experimented in the chapter: in the first one, $X_i$ has a uniform distribution on $(0;1)$, $s = s_1$ is the polynomial function of Section 3.4.2, $\varepsilon_i$ has the standard Gaussian distribution; in the second one, $X_i$ has a truncated bimodal Gaussian distribution, $s = s_2$ is the second function of Section 3.4.2, and $\varepsilon_i$ has still Gaussian distribution. For each model, we conduct the experimentation with 50 samples with the graphical interface CAPUSHE (CAlibrating Penalty Using Slope HEuristics) developed by Baudry *et al.* (2012): for each sample, the software proposes a value of the constant (with dimension jump method). We plot in Figure Figure 3.2 the histograms of all the proposed constant values, for each model.

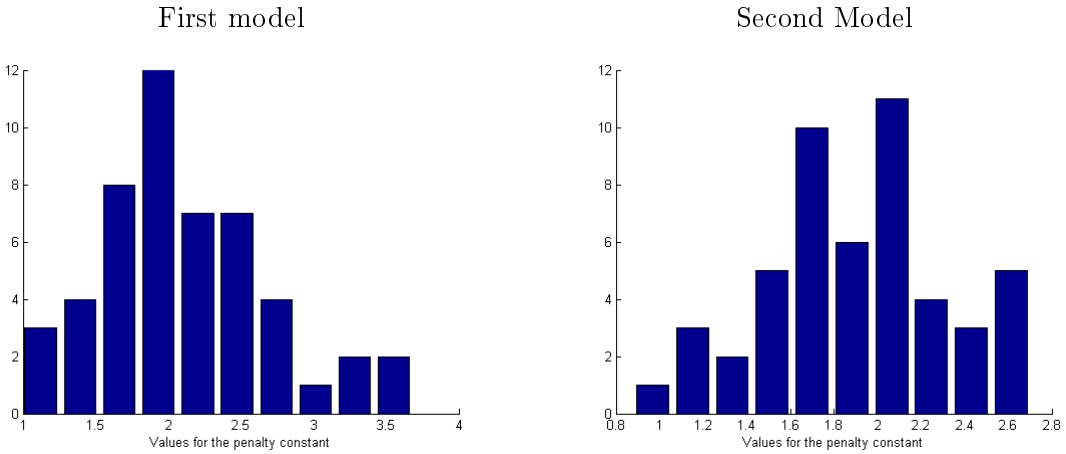



Figure 3.2: Histogram of the constant values proposed by CAPUSHE with the 'dimension jump' method in the two studied models

  As usual in penalty calibration, it is more secure to choose the constant too large than too small, since too small penalties may lead to explosive MISEs. Therefore, we choose the largest constant $c_1' = 4$.
- For the constant $c_2'$ of the term $V$ of the GL method, we experiment several values for the same two models: the MISE was plotted with respect to possible values of $c_2'$ from 0.0001 to 30, see Figure Figure 3.3.
  We decide to choose a value leading to reasonable risk and complexity of the selected space $S_m$. Therefore, all values between 0.01 and 7 (approximately) were admissible, and we choose $c_2' = 0.5$. Another value in the range just improves the results for some

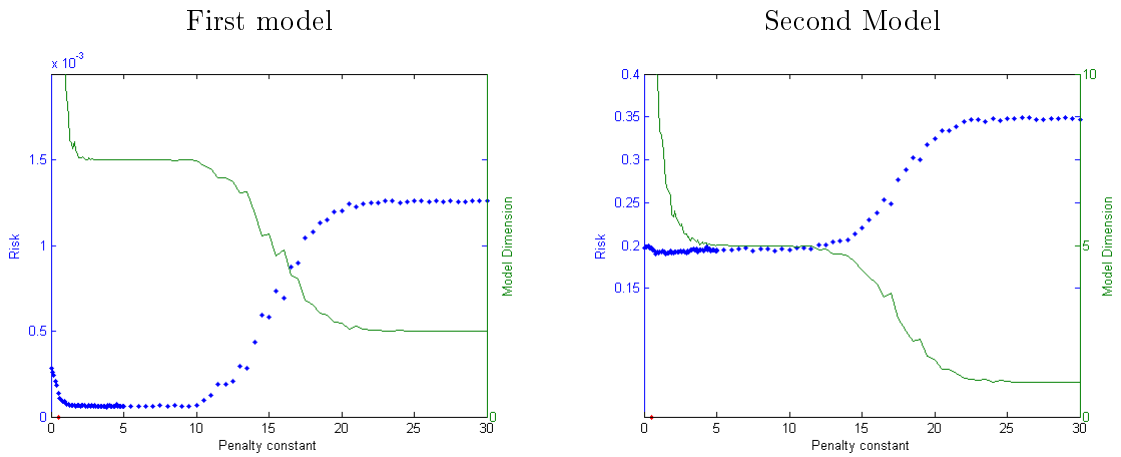



Figure 3.3: Plots of the risk (labeling on the left of each graph) and of selected model dimension versus the value of the constant $c_2'$ (axis label) for the estimation in the two models. Bold dotted line: risk. Solid line: selected model dimension.

other models (other function $f$, design density, or noise distribution), and deteriorate them for some others. We can not reasonably put all the trials we ran.

## 3.7 Appendix 2: An example of random penalty

All the penalty terms defined to prove the main results ($\text{pen}^{F_X}$ or pen for the model selection device, and $V^{F_X}$ or $V$ for the GL method, see (3.10), (3.11) and (3.16)) depend on the unknown expectation $\mathbb{E}[Y_1^2]$. We have already remarked that it can be replaced by its empirical version. It enables to prove analogous upper-bounds as the bounds of Theorem 3.1 and Theorem 3.2 for the risk of all the estimators we built with these new random penalty. As an example, to show how such bounds are derived, we state one oracle-type inequality, in the case of a random penalty in the model selection device with known c.d.f. $F_X$. Notice that another replacement of an unknown quantity in a penalty term is proved in the following chapter, for the GL method (see Chapter 6, Theorem 6.2).

Recall that $\text{pen}^{F_X} = c_1\phi_0^2\mathbb{E}[Y_1^2]D_m/n$ (definition (3.10)). Denoting by $\overline{Y_n^2} = (1/n)\sum_{i=1}^n Y_i^2$, we also define

$$\widehat{\text{pen}}^{F_X}(m) := \frac{c_1}{1-\kappa}\phi_0^2\overline{Y_n^2}\frac{D_m}{n}, \tag{3.49}$$

with $c_1$ the constant of $\text{pen}^{F_X}$, and $\kappa$ a fixed real number in the interval $(0;1)$. Choose

$$\hat{m}^a \in \text{argmin}_{m\in\mathcal{M}_n}\gamma_n(\hat{g}_m^{F_X}, F_X) + \widehat{\text{pen}}^{F_X}(m),$$

similarly to (3.9). The associated estimator of $s$ is $\tilde{\tilde{s}}_1^{F_X} = \hat{g}_{\hat{m}^a}^{F_X} \circ F_X$. We are now able to prove:

**Theorem 3.3.** *Under the assumptions of Theorem 3.1, there exists two numerical constants $\bar{k}_1$ and $\bar{k}_1'$, and a constant $\bar{C}_1$ which depends on the same quantities as the constant $C_1$ of Theorem 3.1, such that*

$$\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2\right] \leq \min_{m\in\mathcal{M}_n}\left\{\bar{k}_1\left\|s - s_m^{f_X}\right\|_{f_X}^2 + \bar{k}_1' pen^{F_X}(m)\right\} + \frac{\bar{C}_1}{n}. \tag{3.50}$$

**Proof of Theorem 3.3.** For the sake of clarity, we denote by $\widehat{\text{pen}}$ and pen the two penalties. We introduce the set

$$\Omega_\kappa = \left\{\left|\frac{\overline{Y_n^2}}{\mathbb{E}[Y_1^2]} - 1\right| < \kappa\right\},$$

and split the risk:

$$\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2\right] = \mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}\right] + \mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa^c}\right].$$

The sketch of the proof is the following: first, on $\Omega_\kappa$, the proof is quite similar to the proof of Theorem 3.1. Second the probability of the set $\Omega_\kappa^c$ is negligible compared to $1/n$.

***Upper-bound for the term*** $\mathbb{E}[\|\tilde{\tilde{s}}_1^{F_X} - s\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}]$**.** First, notice that on $\Omega_\kappa$, the two following inequalities hold:

$$\mathbb{E}\left[Y_1^2\right] \leq \frac{1}{1-b}\overline{Y_n^2}, \quad \overline{Y_n^2} \leq (1+b)\mathbb{E}\left[Y_1^2\right].$$

We begin like in the proof of Theorem 3.2 (see Section 3.5.2):

$$\begin{aligned}\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \leq\ & \frac{\theta}{\theta-2}\left\{\frac{\theta+2}{\theta}\left\|s - s_m^{F_X}\right\|_{f_X}^2 + \widehat{\text{pen}}(m) + \right.\\ & \left.\frac{\theta^2}{\theta-2}\sum_{m'\in\mathcal{M}_n}\left(\sup_{\substack{t\in S_{m\vee m'}\\ \|t\|=1}}(\nu_n(t))^2 - p(m\vee m')\right)_+ + \theta p(m\vee\hat{m}^a) - \widehat{\text{pen}}(\hat{m}^a)\right\}.\end{aligned}$$

Thanks to Proposition 3.2, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}\right] \leq\ & \frac{\theta+2}{\theta-2}\left\|s - s_m^{F_X}\right\|_{f_X}^2 + \mathbb{E}\left[\mathbf{1}_{\Omega_\kappa}\frac{\theta}{\theta-2}\widehat{\text{pen}}(m)\right] + \frac{\theta^2}{\theta-2}\frac{C}{n}\\ & + \frac{\theta}{\theta-2}\mathbb{E}\left[\mathbf{1}_{\Omega_\kappa}\left(\theta p(m\vee\hat{m}^a) - \widehat{\text{pen}}\left(\hat{m}^a\right)\right)\right].\end{aligned}$$

Moreover, on the set $\Omega_\kappa$

$$\begin{aligned}\theta p(m\vee\hat{m}^b) &= 6\theta(1+2\delta)\mathbb{E}\left[Y_1^2\right]\frac{D_m + D_{\hat{m}^a}}{n},\\ &\leq 6\theta(1+2\delta)\frac{\overline{Y_n^2}}{1-\kappa}\frac{D_m + D_{\hat{m}^a}}{n},\\ &= \widehat{\text{pen}}(m) + \widehat{\text{pen}}(\hat{m}^a),\end{aligned}$$

since $c_1$ has been defined equal to $6\theta(1+2\delta)$ in Section 3.5.2. Therefore,

$$\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}\right] \leq \frac{\theta+2}{\theta-2}\left\|s - s_m^{F_X}\right\|_{f_X}^2 + \frac{2\theta}{\theta-2}\mathbb{E}\left[\mathbf{1}_{\Omega_\kappa}\widehat{\text{pen}}(m)\right] + \frac{\theta^2}{\theta-2}\frac{C}{n}.$$

Finally, on $\Omega_\kappa$, $\widehat{\text{pen}}(m) \leq ((1+\kappa)/(1-\kappa))\text{pen}(m)$, and we conclude

$$\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}\right] \leq \frac{\theta+2}{\theta-2}\left\|s - s_m^{F_X}\right\|_{f_X}^2 + \frac{2(1+\kappa)\theta}{(\theta-2)(1-\kappa)}\text{pen}(m) + \frac{\theta^2}{\theta-2}\frac{C}{n}. \tag{3.51}$$

***Upper-bound for the term*** $\mathbb{E}[\|\tilde{\tilde{s}}_1^{F_X} - s\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa}]$**.** Let us begin with a bound for $\|\tilde{\tilde{s}}_1^{F_X} - s\|_{f_X}^2$. We start with $\gamma_n(g_{\hat{m}^a}^{F_X}, F_X) \leq \gamma_n(g_{\hat{m}^a}, F_X)$, and obtain, like in Section 3.5.2,

$$\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \leq \left\|s_{\hat{m}^a}^{F_X} - s\right\|_{f_X}^2 + 2\nu_n\left(\hat{g}_{\hat{m}^a}^{F_X} - g_{\hat{m}^a}\right),$$

Thus, for any $\theta' > 0$,

$$\begin{aligned} 2\nu_n\left(\hat{g}_{\hat{m}^a}^{F_X} - g_{\hat{m}^a}\right) &\leq \frac{1}{\theta'}\left\|\tilde{\tilde{s}}_1^{F_X} - s_{\hat{m}^a}^{F_X}\right\|_{f_X}^2 + \theta' \sup_{t\in S_{\hat{m}^a},\, \|t\|=1} \nu_n^2(t), \\ &\leq \frac{2}{\theta'}\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 + \frac{2}{\theta'}\left\|s - s_{\hat{m}^a}^{F_X}\right\|_{f_X}^2 + \theta' \sup_{t\in S_{\hat{m}^a},\, \|t\|=1} \nu_n^2(t). \end{aligned}$$

If $\theta' > 2$, this leads to

$$\begin{aligned} \left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 &\leq \frac{\theta'+2}{\theta'-2}\left\|s_{\hat{m}^a}^{F_X} - s\right\|_{f_X}^2 + \frac{\theta'^2}{\theta'-2} \sup_{t\in S_{\hat{m}^a},\, \|t\|=1} \nu_n^2(t), \\ &\leq \frac{\theta'+2}{\theta'-2}\|s\|_{f_X}^2 + \frac{\theta'^2}{\theta'-2} \sup_{t\in S_{\hat{m}^a},\, \|t\|=1} \nu_n^2(t), \end{aligned}$$

and by taking expectation,

$$\begin{aligned} &\mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa^c}\right] \\ &\leq \frac{\theta'+2}{\theta'-2}\|s\|_{f_X}^2 \mathbb{P}(\Omega_\kappa^c) + \frac{\theta'^2}{\theta'-2}\left(\mathbb{E}\left[\sum_{m\in\mathcal{M}_n}\left(\sup_{t\in S_m,\, \|t\|=1} \nu_n^2(t) - \text{pen}(m)\right)_+\right] + \mathbb{E}\left[\text{pen}(\hat{m}^a)\mathbf{1}_{\Omega_\kappa^c}\right]\right). \end{aligned}$$

Thanks to $D_{\hat{m}^a} \leq n$, we also have $\text{pen}(\hat{m}^a) \leq 6\theta(1+2\delta)\phi_0^2\mathbb{E}[Y_1^2]$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\left\|\tilde{\tilde{s}}_1^{F_X} - s\right\|_{f_X}^2 \mathbf{1}_{\Omega_\kappa^c}\right] &\leq \left(\frac{\theta'+2}{\theta'-2}\|s\|_{f_X}^2 + \frac{\theta'^2}{\theta'-2}6\theta(1+2\delta)\phi_0^2\mathbb{E}[Y_1^2]\right)\mathbb{P}(\Omega_\kappa^c) \\ &\quad + \frac{\theta'^2}{\theta'-2}\mathbb{E}\left[\sum_{m\in\mathcal{M}_n}\left(\sup_{t\in S_m,\, \|t\|=1} \nu_n^2(t) - \text{pen}(m)\right)_+\right]. \end{aligned}$$

By applying Proposition 3.2, the second term of this bound is shown to be less than $C/n$. For the first one, the main tool is the Bienaymé-Tchebitcheff Inequality:

$$\begin{aligned} \mathbb{P}(\Omega_\kappa^c) &= \mathbb{P}\left(\left|\overline{Y_n^2} - \mathbb{E}\left[Y_1^2\right]\right| < \kappa\mathbb{E}\left[Y_1^2\right]\right), \\ &\leq \frac{1}{\kappa^2\mathbb{E}\left[Y_1^2\right]^2}\text{Var}\left(\overline{Y_n^2}\right) \leq \frac{\mathbb{E}\left[Y_1^4\right]}{\kappa^2\mathbb{E}\left[Y_1^2\right]^2}\frac{1}{n}. \end{aligned}$$

Since the last bound is $O(1/n)$, gathering this with (3.51) ends the proof.

□

# Chapitre 4

# Noyaux déformés pour l'estimation adaptative d'une fonction d'une variable réelle.

**Sommaire**

Ce chapitre est une version modifiée de l'article *Adaptive warped kernel estimators*, soumis pour publication.

**Résumé.** Nous proposons dans ce chapitre d'étendre la méthode de déformation introduite au Chapitre 3 pour l'estimation adaptative par projection de la fonction de régression à l'estimation adaptative par noyaux dans des contextes plus variés. Le problème peut toujours être formulé de la façon suivante : le but est d'estimer une fonction d'une variable réelle, sur la base d'un échantillon de couples de variables aléatoires $(X, Y)$. On considère d'abord une collection d'estimateurs à noyaux construits à partir d'observations transformées de type $(\Phi(X), Y)$, pour $\Phi$ une transformation bijective liée à la loi de $X$. Puis on met en œuvre une stratégie adaptative de sélection de la fenêtre inspirée de la méthode de Goldenshluger & Lepski (2011a). L'intérêt de la déformation est double. D'un point de vue pratique, les estimateurs obtenus sont, comme ceux du Chapitre 3 numériquement stables. D'un point de vue théorique, le compromis biais-variance est automatiquement réalisé : des inégalités de type oracle sont établies. La pertinence de la méthode est illustrée pour l'estimation des fonctions suivantes : régression additive et multiplicative, fonction de répartition dans un modèle de censure par intervalle, risque instantané pour des données censurées à droite. Enfin, nous montrons comment une stratégie fondée sur la minimisation d'un contraste aurait également pu être mise en œuvre.

**Abstract.** In this chapter, we propose to extend the "warping" device introduced in Chapter 3 to recover a regression function with adaptive projection estimators. The aim is to estimate different real-valued functions from a sample of random couples $(X, Y)$, with "warped"-kernel estimates. We thus use transformed data $(\Phi(X), Y)$, with $\Phi$ a one-to-one function, to build a collection of kernel estimators. The data-driven selection of the best bandwidth is done with a method inspired by Goldenshluger and Lepski (2011). The method permits to handle various problems such as additive and multiplicative regression, conditional density estimation, hazard rate estimation based on randomly right censored data, and cumulative distribution function estimation from current-status data. The interest is threefold. First, the squared-bias/variance trade-off is automatically realized. Next, non-asymptotic risk bounds are derived. Last, the estimator is easily computed thanks to its simple expression: a short simulation study is presented. Finally, a section illustrates how a warped-bases strategy could have also been developed in each of the studied examples.

## 4.1 Introduction

Let $(X, Y)$ be a couple of real random variables, and $(X_i, Y_i)_{i=1,\dots,n}$ an *i.i.d.* sample drawn as $(X, Y)$. The main goal of nonparametric estimation is to recover an unknown function $s$, linked with $(X, Y)$, such as the regression function, from the data. Among the huge variety of methods that have been investigated, the use of transformed data $(F_X(X_i), Y_i)$, with $F_X$ the cumulative distribution function (c.d.f.) of $X$, has received attention in the past decades. In this context, both kernel and projection estimators have been studied in random design regression estimation (Yang 1981, Stute 1984, Kerkyacharian & Picard 2004, Kulik & Raimondo 2009, Pham Ngoc 2009), conditional density or c.d.f estimation (Stute 1986a, Mehra *et al.* 2000) or for the white noise model (Chesneau 2007). However, to our knowledge, few papers focus on the problem of adaptivity of such "warped estimators". The aim of the present work is twofold: first, we want to show that a warping kernel device can be applied to various estimation problems, including survival analysis models (see examples below). Secondly, we address the problem of bandwidth selection, with the intention of providing an adaptive "warped" estimator, which satisfies non-asymptotic risk bounds.

The basic idea, which motivates the study of warped kernel estimators introduced by Yang (1981), can be first explained in the classical regression framework. Here, the target function is the conditional expectation, $s : x \mapsto \mathbb{E}[Y|X = x]$ *i.e.*

$$s(x) = \frac{1}{f_X(x)} \int_{\mathbb{R}} y f_{(X,Y)}(x, y) dy, \tag{4.1}$$

when a density $f_{(X,Y)}$ for the couple $(X, Y)$ exists, and where $f_X$ is the marginal density of the design $X$. Historical kernel methods were initiated by Nadaraya (1964) and Watson (1964). The famous estimator named after them is built as the ratio of a kernel estimator of the product $sf_X$ divided by a kernel estimator of the density $f_X$:

$$\tilde{s}^{NW} : x \mapsto \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)},$$

where $K_h : x \mapsto K(x/h)/h$, for $h > 0$, and $K : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K(x)dx = 1$. Adaptive estimation then requires the automatic selection of the bandwidth $h$, and the ratio form of the estimate suggests that two such parameters should be selected: one for the numerator, and one for the denominator. From the theoretical point of view, there is no reason to choose the same. Nevertheless, non-asymptotic results such as oracle-inequality are difficult to derive for an estimator defined with two different data-driven smoothing parameters. See Penskaya (1995) for a thorough study of the ratio-form estimators. Moreover, when the design $X$ is very irregular (for example when a "hole" occurs in the data), a ratio may lead to instability (see Pham Ngoc 2009). The warped kernel estimators introduced by Yang (1981) and Stute (1984) avoid the ratio-form. Indeed denote by $\hat{F}_n$ the empirical c.d.f. of the $X_i$'s and let

$$\hat{s}_h = \frac{1}{n}\sum_{i=1}^{n} Y_i K_h \left(F_X(x) - F_X(X_i)\right), \text{ or } \hat{s}_h = \frac{1}{n}\sum_{i=1}^{n} Y_i K_h \left(\hat{F}_n(x) - \hat{F}_n(X_i)\right), \tag{4.2}$$

depending on whether the c.d.f. $F_X$ is known or not. The following equality (see Proposition 4.1) holds:

$$\mathbb{E}\left[YK_h\left(u - F_X(X)\right)\right] = K_h \star (s \circ F_X^{-1})(u),$$

where $\star$ is the convolution product and $\circ$ is the composition symbol. Thus, the first estimator of (4.2) can be viewed as $\hat{s}_h = \widehat{s \circ F_X^{-1}} \circ F_X$. The main advantage is that its expression involves one bandwidth $h$ only.

In this chapter, we generalize the warping strategy to various functional estimation problems: as a first extension of (4.1), we propose to recover functions $s$ of the form

$$s(x) = \frac{1}{\phi(x)} \int \theta(y) f_{(X,Y)}(x,y)dy, \tag{4.3}$$

for $\theta : \mathbb{R} \to \mathbb{R}$, and $\phi : \mathbb{R} \to \mathbb{R}_+\backslash\{0\}$. In this case, the warping device brings into play the transformation $(\Phi(X), Y)$ of the data, with $\Phi' = \phi$. The form (4.3) covers the additive regression model described above, by setting $\Phi = F_X$, and $\theta(y) = y$. But it also permits to deal with the simplified heteroskedastic model $Y = \sqrt{s(X)}\varepsilon$, where $\varepsilon$ is an unobserved noise, centered, with variance equals to 1. In this case, $\Phi = F_X$, and $\theta(y) = y^2$.

In several examples however, the couple $(X, Y)$ does not admit a density, but $X$ admits a marginal density. Then (4.3) can be extended and the target function $s$ takes the form:

$$s(x) = \frac{f_X(x)}{\phi(x)} \mathbb{E}[\theta(Y)|X = x]. \tag{4.4}$$

This allows to handle two classical settings in survival analysis: the interval censoring case 1, and right censored data. In the interval censoring model, case 1, the target function is $s(x) = \mathbb{P}(Z \leq x)$, where $Z$ is a survival time, which is not observed, and we only know a current status at the observed time $X$ of examination. We also know $Y = \mathbf{1}_{Z \leq X}$, which indicates whether $Z$ occurs before $X$ or not. We refer to Jewell & van der Laan (2004) for a review of the estimation methods in this setting, and more recently to Ma & Kosorok (2006), Brunel & Comte (2009) or Plancade (2013) for investigations including adaptivity. In right-censored data, the function of interest at time $x$ is the hazard rate function, that is the risk of death at time $x$, given that the patient is alive until $x$. This model has been studied by Tanner & Wong (1983), Müller & Wang (1994) and Patil (1993), among all. Adaptive results are available for projection-type estimators (see Brunel & Comte 2005, 2008, Reynaud-Bouret 2006 or Akakpo & Durot 2010), but to our knowledge not for kernel estimators.

The chapter is organized as follows. We present in Section 4.2 the estimation method, detail the examples illustrating the relevance of the introduction of a general target function $s$ defined by (4.4). We also study the global risk of the warped kernel estimators with fixed bandwidth. Section 4.3 is devoted to adaptive estimation: we define a data-driven choice of the bandwidth, inspired by Goldenshluger & Lepski (2011a) which allows to derive non-asymptotic results for the adaptive estimators. Oracle-type inequalities are provided for the M.I.S.E., and convergence rates are deduced under regularity assumptions. Sections 4.2 and 4.3 deal with the case of known deformation $\Phi$. Section 4.4 discusses briefly the case of unknown $\Phi$, for which details are given in Appendix 1 (Section 4.7). In Section 4.5, the

method is illustrated through numerical simulations. Proofs are gathered in Section 4.6. A more general proof of the main result is given in Appendix 2 (Section 4.8), under slightly different assumptions. The bandwidth selection rule is discussed in Appendix 3 (Section 4.9). Finally, we extend the method to warped-bases estimation in Appendix 4 (Section 4.10).

## 4.2 Estimation method

### 4.2.1 Warped kernel strategy

Consider a sample $(X_i, Y_i)_{i=1,\dots,n}$ of *i.i.d.* random couples with values in $A \times B$, where $A$ is an open interval of $\mathbb{R}$ and $B$ a Borel subset of $\mathbb{R}$. We assume that $X_i$ has a marginal density $f_X$ and we aim at recovering a function $s : A \to \mathbb{R}$ linked with the distribution of $(X_i, Y_i)$. To estimate $s$, we replace the explanatory variable $X_i$ by $\Phi(X_i)$, where $\Phi : A \to \Phi(A) \subset \mathbb{R}$ is one-to-one and absolutely continuous. The data $(\Phi(X_i), Y_i)_{i=1,\dots,n}$ are called the warped sample with deformation function $\Phi$. The sets $A$ and $B$ are supposed to be given. The target function can be written as:

$$s(x) = g \circ \Phi(x) = g(\Phi(x)), \text{ with } g : \Phi(A) \to \mathbb{R}. \tag{4.5}$$

We first estimate the auxiliary function $g = s \circ \Phi^{-1}$ with $\Phi^{-1}$ the inverse function of $\Phi$. In the general case, $\Phi$ is unknown and we must estimate it also. Let $K$ be a function such that $\int_{\mathbb{R}} K(u)du = 1$ and set $K_h : u \mapsto K(u/h)/h$, for $h > 0$. We define, for $u \in \Phi(A)$,

$$\hat{g}_h(u) = \frac{1}{n} \sum_{i=1}^{n} \theta(Y_i) K_h \left( u - \hat{\Phi}(X_i) \right), \tag{4.6}$$

where $\theta : \mathbb{R} \to \mathbb{R}$ is a given function, $\hat{\Phi}$ is an empirical counterpart for $\Phi$, and for $x \in A$

$$\hat{s}_h(x) = \hat{g}_h \circ \hat{\Phi}(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(Y_i) K_h \left( \hat{\Phi}(x) - \hat{\Phi}(X_i) \right). \tag{4.7}$$

Let us give examples covered by the above framework. They are sum up in Table 4.1, with the corresponding estimators.

**Example 1 (Additive random design regression):** we observe $(X_i, Y_i)$ with $Y_i = s(X_i) + \varepsilon_i$, $(\varepsilon_i)_{i=1,\dots,n}$ is independent of $(X_i)_{i=1,\dots,n}$, $\mathbb{E}[\varepsilon_i^2] < \infty$ and $\mathbb{E}[\varepsilon_i] = 0$. We choose $\Phi(x) = F_X(x)$, the cumulative distribution function (c.d.f. in the sequel) of $X$ and assume that $\Phi : A \to \Phi(A)$ is invertible.

**Example 2 (Heteroskedastic model):** $Y_i = \sigma(X_i)\varepsilon_i$, $(\varepsilon_i)_{i=1,\dots,n}$ independent of $(X_i)_{i=1,\dots,n}$, $\mathbb{E}[\varepsilon_i^2] = 1$, $\mathbb{E}[\varepsilon_i] = 0$, $\Phi(x) = F_X(x)$, with $\Phi : A \to \Phi(A)$ invertible. Here $s(x) = \sigma^2(x) = \mathbb{E}[Y_i^2 | X_i = x]$.

**Example 3 (Interval censoring, Case 1):** the observation is $(X_i, Y_i)$ where $Y_i = \mathbf{1}_{Z_i \leq X_i}$, $Z_i, X_i \geq 0$ are independent event occurence times, $Y_i$ indicates whether $Z_i$ (the time of

| Example | $s$ | $\Phi$ | $\hat{s}_h$ |
|---|---|---|---|
| 1. $Y = s(X) + \varepsilon$ | $s$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} Y_i K_h(F_X(x) - F_X(X_i))$ |
| 2. $Y = \sigma(X)\varepsilon$ | $\sigma^2$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 K_h(F_X(x) - F_X(X_i))$ |
| 3. $(X, \mathbf{1}_{Z\leq X})$ | $F_Z$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq X_i} K_h(F_X(x) - F_X(X_i))$ |
| 4. $(X = Z \wedge C, \mathbf{1}_{Z\leq C})$ | $\frac{f_Z}{1-F_Z}$ | $\Phi(x) = \int_0^x (1 - F_X(t))dt$ | $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq C_i} K_h(\Phi(x) - \Phi(X_i))$ |

Table 4.1: Summary of the studied examples and of the "warping" function used in each case.

interest) occurs before $X_i$ (the so-called "examination time") or not and $Z_i$ is not observed. The target function is $s(x) = \mathbb{P}(Z_i \leq x) = \mathbb{E}[Y_i | X_i = x]$. We choose $\Phi = F_X$.

**Example 4 (Hazard rate estimation from right censored-data):** the observation is $X_i = Z_i \wedge C_i$, $Y_i = \mathbf{1}_{Z_i\leq C_i}$, where $Z_i$ and $C_i$ are not observed and independent, $Z_i \geq 0$ is a lifetime and $C_i \geq 0$ is a censoring time. The function $s$ of interest is the hazard rate function $s(x) = f_Z(x)/(1 - F_Z(x))$, where $f_Z$ (resp. $F_Z$) is the density (resp. the c.d.f.) of $Z$. This function satisfies

$$s(x) = \frac{f_X(x)}{1 - F_X(x)} \mathbb{E}[Y | X = x], \tag{4.8}$$

a relation which is proved in Section 4.6.1. In this case, we assume $F_X(t) < 1$ for all $t \geq 0$, and take $\Phi(x) = \int_0^x (1 - F_X(t))dt$.

The following equality is the cornerstone of the method and justifies the introduction of (4.6):

**Proposition 4.1.** *Let $(X, Y)$ be a random couple with values in $A \times B \subset \mathbb{R}^2$. Assume that $X$ has a density $f_X$ and let $\Phi : A \to \Phi(A)$ be a one-to-one absolutely continuous function. Let $s : A \to \mathbb{R}$ defined by (4.4) and $g = s \circ \Phi^{-1}$. Then, if $\theta$ satisfies $\mathbb{E}|\theta(Y)K_h(u - \Phi(X))| < \infty$, for all $u \in \Phi(A)$,*

$$\mathbb{E}\left[\theta(Y)K_h(u - \Phi(X))\right] = K_h \star \left(g\mathbf{1}_{\Phi(A)}\right)(u) := g_h(u), \tag{4.9}$$

*where $\star$ is the convolution product.*

Equality (4.9) shows that $\hat{g}_h$, defined by (4.6) is an empirical version of $g_h$ and thus $\hat{s}_h$ in (4.7) suits well to estimate $s$.

Hereafter, for the sake of clarity, we assume that $\Phi$ is known: thus we choose $\hat{\Phi} = \Phi$ in (4.6) and (4.7). In Section 4.4, we discuss the case of an unknown deformation $\Phi$. The theoretical results are the same, up to further technicalities due to the plug-in of an empirical version for $\Phi$.

### 4.2.2 Risk of the fixed bandwidth estimator

In this section, we study the global properties of $\hat{s}_h$ as an estimate of $s$ on $A$, with a fixed bandwidth $h$. The quadratic risk weighted by the derivative $\phi$ of the warping function $\Phi$ is the natural criterion in our setting. Let us introduce, for a measurable function $t$ on $A$,

$$\|t\|_\phi^2 = \int_A t^2(x)\phi(x)dx, \tag{4.10}$$

and denote by $L^2(A,\phi)$ the space of functions $t$ for which the quantity (4.10) exists and is finite. We also use the corresponding scalar product $\langle .,. \rangle_\phi$. For $t_1, t_2$ belonging to $L^2(A,\phi)$, we have

$$\|t_1 \circ \Phi\|_\phi = \|t_1\|_{L^2(\Phi(A))}, \quad \langle t_1 \circ \Phi, t_2 \circ \Phi \rangle_\phi = \langle t_1, t_2 \rangle_{\Phi(A)},$$

where $\|t_1\|^2_{L^2(\Phi(A))} = \int_{\Phi(A)} t_1^2(x)dx$ and $\langle .,. \rangle_{\Phi(A)}$ denotes the usual scalar product on $L^2(\Phi(A))$. Therefore,

$$\|\hat{s}_h - s\|_\phi^2 = \|\hat{g}_h - g\|^2_{L^2(\Phi(A))}.$$

The following bias-variance decomposition of the risk holds:

**Proposition 4.2.** *Let $K$ belong to $L^2(\mathbb{R})$. Assume that $s$ belongs to $L^2(A,\phi)$, and that $\mathbb{E}[\theta^2(Y_1)] < \infty$. Then (recall that $g_h$ is defined in (4.9)),*

$$\mathbb{E}\left[\|\hat{s}_h - s\|_\phi^2\right] \leq \|g - g_h\|^2_{L^2(\Phi(A))} + \frac{1}{nh}\mathbb{E}\left[\theta^2(Y_1)\right]\|K\|^2_{L^2(\mathbb{R})}. \tag{4.11}$$

If $s$ is bounded on $A$, $s \in L^2(A,\phi)$. This is the case for Examples 1-3, as $\phi = f_X$. In Example 4, we can check that $s \in L^2(A,\phi)$ for all classical distributions for $C$ and $Z$ used in survival analysis (such as exponential, Weibull, Gamma...). The general condition to be checked in this example is $\int_A f_X^2(x)/(1 - F_X(x))dx < \infty$.

## 4.3 Adaptive estimation

### 4.3.1 Data-driven choice of the bandwidth

As usual, we must choose a bandwidth $h$ which realizes the best compromise between the squared-bias and the variance terms (see Proposition 4.2). Moreover, we need define a data-driven choice of the bandwidth. For this, we use a method described in Goldenshluger & Lepski (2011a). Let $\mathcal{H}_n$ be a finite collection of bandwidths, with cardinality depending

on $n$ and properties precised below (Assumptions (H2)-(H3)). We introduce the auxiliary estimators, involving two kernels,

$$\hat{s}_{h,h'}(x) = \hat{g}_{h,h'}(\Phi(x)) \quad \text{with} \quad \hat{g}_{h,h'} = K_{h'} \star \left(\hat{g}_h \mathbf{1}_{\Phi(A)}\right).$$

For a constant $\kappa > 0$ to be precised later on, we define, for $h \in \mathcal{H}_n$,

$$V(h) = \kappa \left(1 + \|K\|^2_{L^1(\mathbb{R})}\right) \|K\|^2_{L^2(\mathbb{R})} \mathbb{E}\left[\theta^2(Y_1)\right] \frac{1}{nh}. \tag{4.12}$$

Next, we set

$$A(h) = \max_{h' \in \mathcal{H}_n} \left\{ \|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_\phi - V(h') \right\}_+, \tag{4.13}$$

which is an estimation of the squared-bias term (see Lemma 4.1). Note that $\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_\phi = \|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2$. Lastly, the adaptive estimator is defined in the following way:

$$\hat{s} = \hat{s}_{\hat{h}} \quad \text{with} \quad \hat{h} = \arg\min_{h \in \mathcal{H}_n} \{A(h) + V(h)\}. \tag{4.14}$$

The selected bandwidth $\hat{h}$ is data-driven. In $V(h)$, the expectation $\mathbb{E}[\theta^2(Y_1)]$ can be replaced by the corresponding empirical mean (see Brunel & Comte 2005, proof of Theorem 3.4 p.465, Appendix 2 of Chapter 3, or Theorem 6.2 of Chapter 6). In Examples 3-4, it can be replaced by 1, its upper-bound. Notice that the bandwidth selection rule is discussed below, see Section 4.9 (Appendix 3).

### 4.3.2 Results

We consider the following assumptions:

(H1) The function $s$ is bounded. Denote by $\|s\|_{L^\infty(A)}$ its sup-norm.

(H2) There exist $\alpha_0 > 0$ and a constant $k_0 \geq 0$ such that $\displaystyle\sum_{h \in \mathcal{H}_n} \frac{1}{h} \leq k_0 n^{\alpha_0}$.

(H3) For all $\kappa_0 > 0$, there exists $C_0 > 0$, such that $\displaystyle\sum_{h \in \mathcal{H}_n} \exp\left(-\frac{\kappa_0}{h}\right) \leq C_0$.

(H4) The kernel $K$ is of order $l$, *i.e.* for all $j \in \{1, \ldots, l+1\}$, the function $x \mapsto x^j K(x)$ is integrable, and for $1 \leq j \leq l$, $\int_{\mathbb{R}} x^j K(x) dx = 0$.

Assumption (H1) is required to obtain Theorem 4.1 below. Nevertheless the value $\|s\|_{L^\infty(A)}$ is not needed to compute the estimator (see (4.14)). This assumption holds in Example 3 ($s \leq 1$ in this case), and in Example 4, for instance when $Z$ has exponential or Gamma distribution. Assumptions (H2)-(H3) mean that the bandwidth collection should not be too large. For instance, the following classical collections satisfy these assumptions:

1. $\mathcal{H}_{n,1} = \left\{k^{-1},\ k = 1, \ldots, \chi(n)\right\}$ with $\alpha_0 = 2$, $\chi(n) = n$ or $\alpha_0 = 1$, $\chi(n) = \sqrt{n}$.
2. $\mathcal{H}_{n,2} = \left\{2^{-k},\ k = 1, \ldots, [\ln(n)/\ln(2)]\right\}$, with $\alpha_0 = 1$.

Assumption (H4) is required only to deduce convergence rate from the main non-asymptotic result. We need a moment assumption linked with (H2):

(H5) With $\alpha_0$ given by (H2), there exists $p > 2\alpha_0$, such that $\mathbb{E}[|\theta(Y) - \mathbb{E}[\theta(Y)|X]|^{2+p}] < \infty$.

If $\theta$ is bounded, (H5) evidently holds. In Examples 1 and 2, (H5) is a moment assumption on the noise which is usual in regression settings. Notice also that the smaller $\alpha_0$, the less restrictive the integrability constraint $p$ on the noise moments.
We prove the following oracle-type inequality:

**Theorem 4.1.** *We assume that* (H1)-(H3) *hold in Examples 1-4, and additionally that* (H5) *is fulfilled for Examples 1-2. Then there exist two constants $c_1 > 0$ and $c_2 > 0$, such that:*

$$\mathbb{E}\left[\|\hat{s} - s\|_\phi^2\right] \leq c_1 \min_{h \in \mathcal{H}_n} \left\{ \|s - s_h\|_\phi^2 + \frac{\mathbb{E}\left[\theta^2(Y_1)\right] \|K\|_{L^2(\mathbb{R})}^2}{nh} \right\} + \frac{c_2}{n}, \tag{4.15}$$

*with $s_h = g_h \circ \Phi$ and $\hat{s}$ defined by (4.14). The constant $c_1$ depends only on $\|K\|_{L^1(\mathbb{R})}$.*

The constant $c_2$ depends on $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$ in Examples 3-4, and also on the moment of $\varepsilon_1$ and $\mathbb{E}[s^2(X_1)]$ for Examples 1-2. The adaptive estimator $\hat{s}$ automatically makes the squared-bias/variance compromise. The selected bandwidth $\hat{h}$ is performing as well as the unknown oracle:

$$\tilde{h} := \arg\min_{h \in \mathcal{H}_n} \mathbb{E}[\|s - \hat{s}_h\|_\phi^2].$$

up to the multiplicative constant $c_1$ and up to a remainding term of order $1/n$, which is negligible.
The interest of Inequality (4.15) is that it is non-asymptotic. Moreover, contrary to usual kernel estimation results, Assumption (H4) is not needed. This is one of the advantages of the bandwidth selection method.

To deduce convergence rates, smoothness classes must be defined to quantify the bias term. Recall the definitions given in Section 2.2.1 of Chapter 2. Define the Hölder class with order $\beta > 0$ and constant $L > 0$ by

$$\mathcal{H}(\beta, L) = \left\{ t : \mathbb{R} \to \mathbb{R},\ t^{(\lfloor\beta\rfloor)} \text{ exists}, \forall x, x' \in \mathbb{R},\ \left| t^{(\lfloor\beta\rfloor)}(x) - t^{(\lfloor\beta\rfloor)}(x') \right| \leq L|x - x'|^{\beta - \lfloor\beta\rfloor} \right\},$$

where $\lfloor\beta\rfloor$ is the largest integer less than $\beta$. We also need the Nikol'skiĭ class of functions:

$$\mathcal{N}_2(\beta, L) = \left\{ t : \mathbb{R} \to \mathbb{R},\ t^{(\lfloor\beta\rfloor)} \text{ exists}, \forall x \in \mathbb{R},\ \int_{\mathbb{R}} \left( t^{(\lfloor\beta\rfloor)}(x' + x) - t^{(\lfloor\beta\rfloor)}(x') \right)^2 dx' \leq L^2 |x|^{2\beta - 2\lfloor\beta\rfloor} \right\}$$

We can now deduce from Theorem 4.1 the convergence rate of the risk, under regularity assumptions for the auxiliary function $g$.

**Corollary 4.1.** *Let $\tilde{g} = g\mathbf{1}_{\phi(A)}$ on $\mathbb{R}$. Assume that*
- *$\tilde{g}$ belongs to the Hölder class $\mathcal{H}(\beta, L)$, with $\tilde{g}(0) = \tilde{g}(1)$ in Examples 1-3,*
- *$\tilde{g}$ belongs to the Nikol'skiĭ class $\mathcal{N}_2(\beta, L)$ in Example 4.*

*Assume* (H4) *with $l = \lfloor\beta\rfloor$. Then, under the assumptions of Theorem 4.1,*

$$\mathbb{E}\left[\|\hat{s} - s\|_\phi^2\right] \leq C n^{-\frac{2\beta}{2\beta+1}}, \tag{4.16}$$

*where $C$ is a constant which does not depend on $n$ and $\beta$.*

In Examples 1-3, $\Phi(A) = (0;1)$ and the Hölder condition is enough. In Example 4, $\Phi(A) = \mathbb{R}_+$ and we need the Nikol'skiĭ condition. Both spaces are standard in kernel estimation, see *e.g.* Tsybakov (2009) and Goldenshluger & Lepski (2011a).
We recover the classical optimal rates in nonparametric estimation. Note however that our regularity assumptions are set on $g$ and not $s$, as long as we do not consider specific warped spaces defined in Kerkyacharian & Picard (2004).

**Remark 4.1.** We have strong conditions on $g$ at the boundary of the support $[0;1]$, in Examples 1-3. This is nevertheless well-known in kernel estimation, which are rarely "free of boundary effects". This also explains why we restrict the estimation interval for the simulation study, by using the quantiles of the observations $X_i$ (see Section 4.5). Notice that recent methods provide boundary corrections in kernel estimation, see Karunamuni & Alberts (2005) and Bertin & Klutchnikoff (2011) for example.

## 4.4 The general case of unknown $\Phi$

Up to now we have considered the case of a known "warping" function $\Phi$. This is also the framework of *e.g.* Pham Ngoc (2009) or Chesneau (2007). It allows to derive the main result with few assumptions and short proofs. To deal with the general case, we use a plug-in device. Let $\hat{F}_n$ be the empirical c.d.f. of $X$. We estimate $\Phi$ by $\hat{\Phi}(x) = \hat{F}_n(x)$ for Examples 1-3, and by $\hat{\Phi}(x) = \int_0^x \left(1 - \hat{F}_n(t)\right) dt$ for Example 4. Now $\hat{g}_h$ is given by (4.6). To define $\hat{h}$, we replace $\|\hat{s}_{h,h'} - \hat{s}_{h'}\|_\phi^2$ by $\|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2$ in $A(h)$ (see (4.13)). Theorem 4.1 holds under stronger assumptions on the bandwidth collection $\mathcal{H}_n$. However the proof requires lengthy and cumbersome technicalities. To deal with the difference $\hat{\Phi} - \Phi$, we use the deviation inequality of Dvoretzky *et al.* (1956): for any $\lambda > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F_X(x)\right| \geq \lambda\right) \leq K \exp\left(-2n\lambda^2\right),$$

with $K$ an universal constant. Details are given in Chapter 3 and Chaper 5 for regression and conditional density estimation using warped bases, which require analogous arguments. Moreover, a non adaptive bound for the risk is proved in the appendix, Section 4.7, as an example of the required tools.

## 4.5 Illustration

To illustrate the procedure, we focus only on two of the four examples: the additive regression (Example 1), and the estimation of c.d.f. under interval censoring case I (Example 3). In each case, we compare the warped kernel strategy, denoted by WK in this section, with another adaptive method: a regression-type one, based on the minimization of a penalized least-squares contrast. We denote it by LS.

### 4.5.1 Implementation of the warped-kernel estimators

The theoretical study allows the choice of several kernels and bandwidth collections. For practical purpose, we consider the Gaussian kernel, $K : x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, which satisfies Assumption $(K_1)$. It has the advantage of having simple convolution-products:

$$\forall h, h' > 0, \quad K_h \star K_{h'} = K_{\sqrt{h^2+h'^2}}. \tag{4.17}$$

The experiment is conducted with the dyadic collection $\mathcal{H}_{n,2}$ defined above. The larger collection $\mathcal{H}_{n,1}$ has also been tested: since it does not really improve the results but increases the computation time, we only keep the other collection. Besides, the simulations are performed in the case of unknown $\Phi$. Therefore in Examples 1 and 3, the estimator is

$$\hat{s} \; : \; x \; \mapsto \; \frac{1}{n}\sum_{i=1}^{n} \theta(Y_i) K_{\hat{h}}(\hat{F}_n(x) - \hat{F}_n(X_i)),$$

with $\hat{F}_n$ the empirical c.d.f. of the $X_i$'s. Then, the estimation procedure can be decomposed in some steps:

• Simulate a data sample $(X_i, Y_i)$, $i = 1, \ldots, n$, fitting Example 1 or 3.

• Compute $V(h)$ and $A(h)$ for each $h \in \mathcal{H}_{n,1}$.

- For $V(h)$: its computation require a value for $\kappa$ (see (4.12)). A lower bound for its theoretical value is provided by the proof: it is very pessimistic due to rough upper-bounds (for the sake of clarity). A practical calibration is required, like in most model selection devices. Since classical techniques such as the slope heuristic are not currently well developed for the Goldenshluger-Lepki method, we adjust $\kappa$ on simulations, prior to the comparison with the other estimates. We set $\kappa = 0.05$ in Example 1, and $\kappa = 0.3$ in Example 3.

- For $A(h)$: thanks to (4.17), the auxiliary estimates are easily computed: $\hat{s}_{h,h'} = \hat{s}_{\sqrt{h^2+h'^2}}$. The $L^2-$norm is then approximated by a Riemann sum:

$$\|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\Phi(A))} \approx \frac{1}{N}\sum_{k=1}^{N} \left(\hat{g}_{h,h'}(u_k) - \hat{g}_{h'}(u_k)\right)^2,$$

where $N = 50$, and $(u_k)_k$ are grid points evenly distributed across $(0; 1)$.

• Select $\hat{h}$ such that $A(h) + V(h)$ is minimum.

• Compute $\hat{s}_{\hat{h}}$.

### 4.5.2 Example 1: additive regression

We compare the warped kernel method (WK) with the adaptive estimator studied in Baraud (2002). It is a projection estimator, developed in an orthogonal basis of $L^2(A)$, and built with a penalized least-squares contrast. The experiment is carried out with the Matlab toolbox FY3P, written by Yves Rozenholc, and available on his web page `http://www.math-info.univ-paris5.fr/~rozen/YR/Softwares/Softwares.html`. A regular piecewise polynomial basis is used, with degrees chosen in an adaptive way. Since the kernel we choose has only one vanishing moment, the comparison is fair if we consider polynomials with degrees equal to or less than 1. We denote by LS1 the resulting estimator.

However, as shown below, we will see that the warped-kernel generally outperforms the least-square, even if we use polynomials with degree at most 2 (LS2). We also experiment the Fourier basis, but the results are not as good as the polynomial basis for the least-squares estimator. Thus, we do not mention the values.

The procedure is applied for different regression functions, design and noise. We focus on the three following regression functions

$$\begin{aligned} s_1 &: x \mapsto x(x-1)(x-0.6) \\ s_2 &: x \mapsto -\exp(-200(x-0.1)^2) - \exp(-200(x-0.9)^2) + 1 \\ s_3 &: x \mapsto \cos(4\pi x) + \exp(-x^2) \end{aligned}$$

The influence of the design is explored through four distributions:

- $\mathcal{U}_{(0;1)}$, the uniform distribution on the interval $(0;1)$,
- $\gamma(4, 0.08)$, the Gamma distribution, with parameters 4 and 0.08 (0.08 is the scale parameter),
- $\mathcal{N}(0.5, 0.01)$, the Gaussian distribution with mean 0.5 and variance 0.01,
- $\mathcal{BN}$ a bimodal Gaussian distribution, with density $x \mapsto c(\exp(-200(x-0.05)^2) + \exp(-200(x-0.95)^2))$ ($c$ is a constant adjusted to obtain a density function).

We also test the sensibility of the method to the noise distribution: contrary to the underlying design distribution, it does not seem to affect the results. Thus, we present the simulation for a Gaussian centered noise, with variance $\sigma^2$. The value of $\sigma$ is chosen in such a way that the signal-to-noise ratio (the ratio of the variance of the signal $\mathrm{Var}(s(X_1))$ over the variance of the noise $\mathrm{Var}(\varepsilon_1)$) approximately equals 2.

Beams of estimators (WK, LS1, and LS2) are presented in Figures Figure 4.1, Figure 4.2, and Figure 4.3, with the generated data-sets and the function to estimate. Precisely, Figure Figure 4.1 and Figure 4.2 show some regular cases, while Figure Figure 4.3 depicts the case where a hole occurs in the design density: the estimator built with warped kernel behaves still correctly, even if the data are very inhomogeneous.

A study of the risk is reported in Table 4.2, for the sample sizes $n = 60, 200, 500$ and 1000. The MISE is obtained by averaging the following approximations of the ISE values, for $j \in \{1, \ldots, J = 200\}$, computed with $J$ sample replications:

$$ISE_j = \frac{b-a}{N} \sum_{k=1}^{N} \left(\tilde{s}(x_k) - s(x_k)\right)^2,$$

where $\tilde{s}$ stands for one of the estimators, $b$ is the quantile of order 95% of the $X_i$ and $a$ is the quantile of order 5%. The $(x_k)_{k=1,\ldots,N}$ are the sample points falling in $[a;b]$. In 56% of the examples, the risks of the warped-kernel estimator are smaller than the ones of the least-squares estimator, in piecewise polynomials basis with degrees at most 2 (LS2). Besides, if we consider the comparison with LS1, which is more fair as explained above, the WK estimators give better results in 77% of the cases.

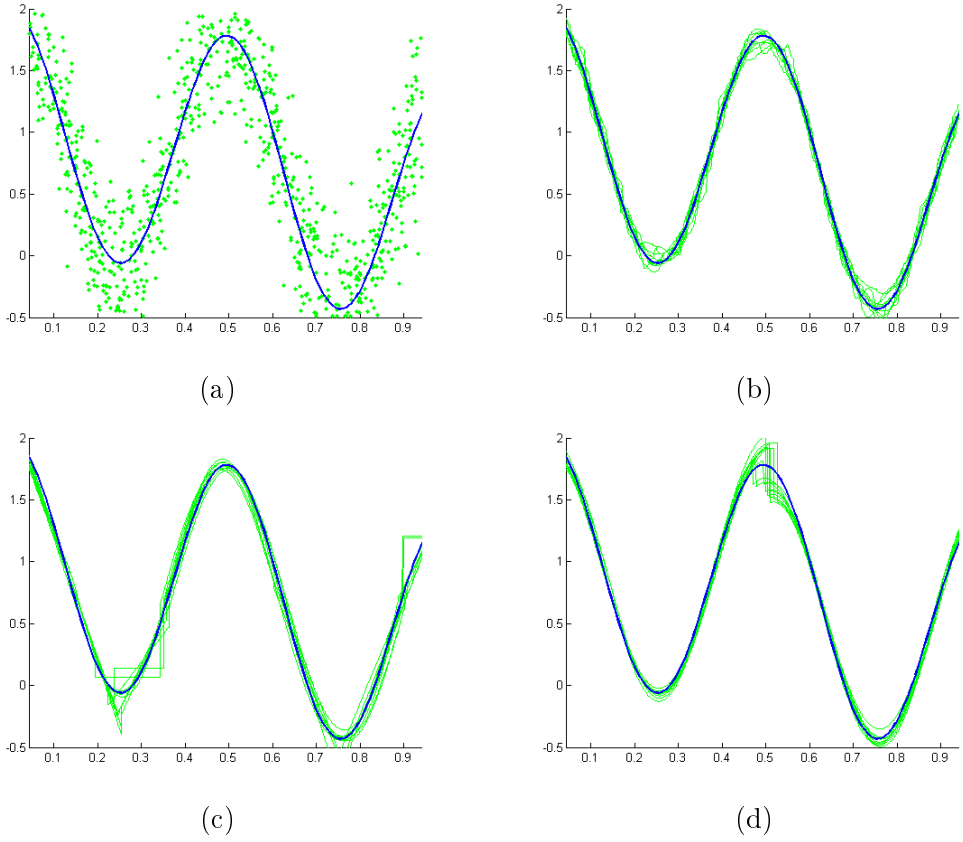



Figure 4.1: Estimation in Example 1, with true regression function $s_3$, design distribution $\mathcal{U}_{(0;1)}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function $s_3$. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

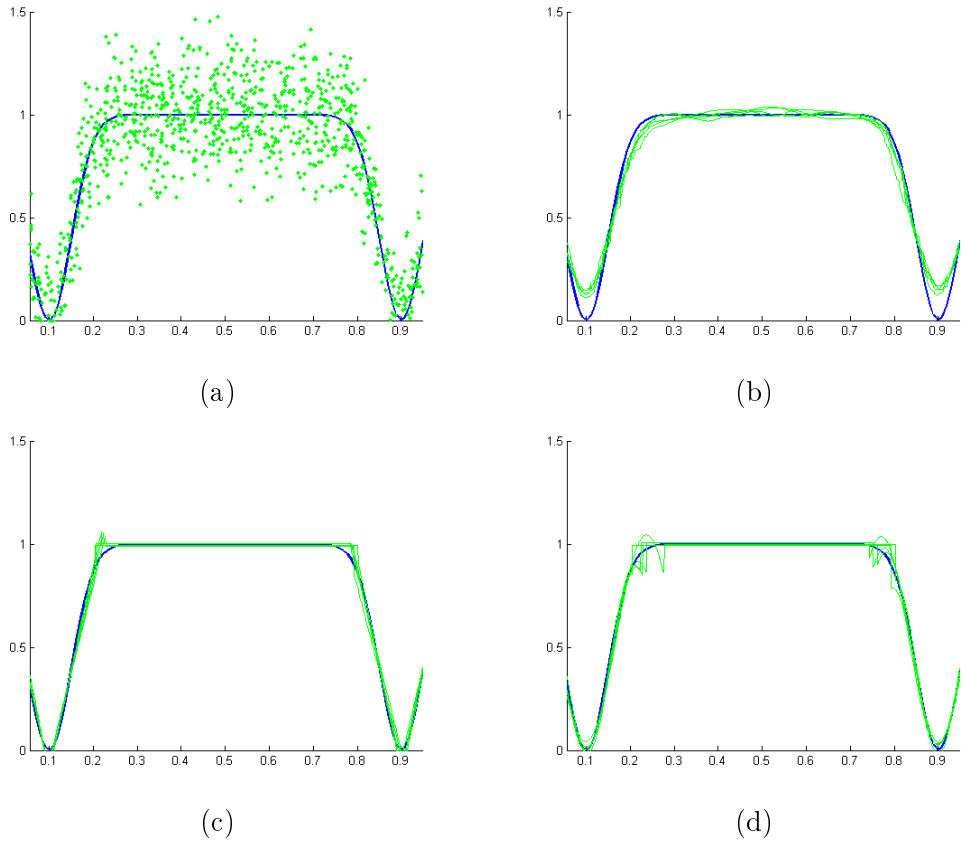



Figure 4.2: Estimation in Example 1, with true regression function $s_2$, design distribution $\mathcal{U}_{(0;1)}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function $s_2$. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

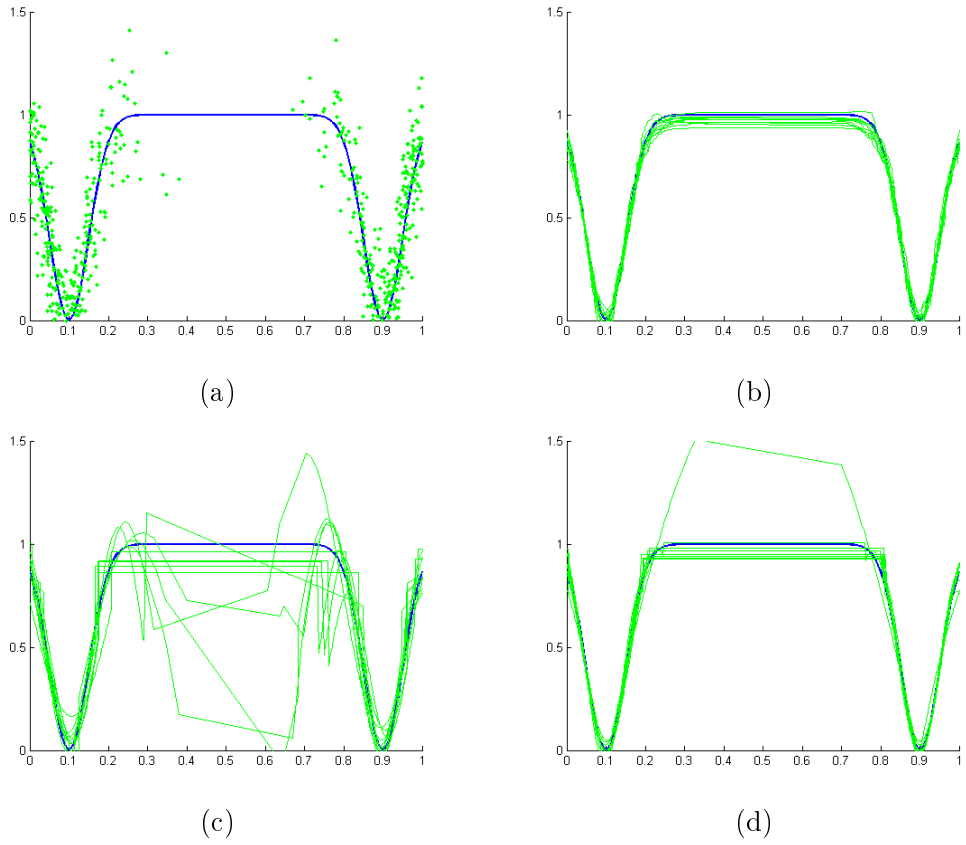
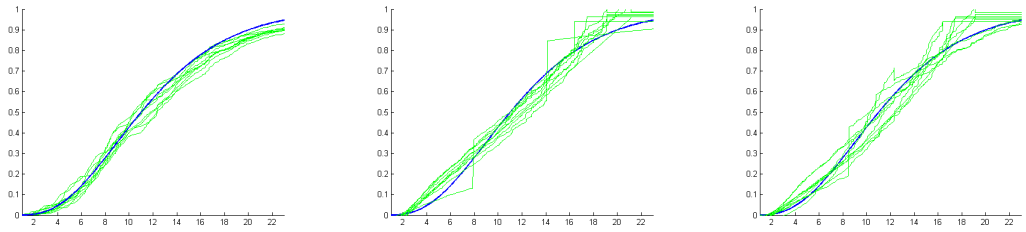



Figure 4.3: Estimation in Example 1, with true regression function $s_2$, design distribution $\mathcal{BN}$, and $n = 1000$. (a) points: data $(X_i, Y_i)_i$, thick line: true function $s_2$. (b)-(c)-(d) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (b)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (c)) or 2 (subplot (d)).

| $s$ | $X$ | $\sigma$ | $n = 60$ | 200 | 500 | 1000 | Method |
|---|---|---|---|---|---|---|---|
| $s_1$ | $\mathcal{U}_{(0;1)}$ | $\sqrt{.0006}$ | 0.3719 | 0.1341 | 0.1957 | 0.2454 | WK |
| | | | 0.3892 | 0.1293 | 0.0681 | 0.0446 | LS2 |
| | $\gamma(4, 0.08)$ | $5.10^{-5}$ | 0.0052 | 0.0033 | 0.0004 | 0.0003 | WK |
| | | | 0.0097 | 0.004 | 0.0017 | 0.0012 | LS2 |
| | $\mathcal{N}(0.5, 0.01)$ | 0.011 | 0.0049 | 0.0020 | 0.0008 | 0.0005 | WK |
| | | | 0.0020 | 0.0012 | 0.0010 | 0.0008 | LS2 |
| | $B\mathcal{N}$ | 0.022 | 0.524 | 0.422 | 0.267 | 0.205 | WK |
| | | | 0.166 | 0.054 | 0.038 | 0.029 | LS2 |
| $s_2$ | $\mathcal{U}_{(0;1)}$ | 0.17 | 16.35 | 6.791 | 3.51 | 0.837 | WK |
| | | | 33.212 | 2.058 | 0.691 | 0.407 | LS2 |
| | $\gamma(4, 0.08)$ | 0.08 | 1.885 | 0.354 | 0.204 | 0.147 | WK |
| | | | 4.047 | 0.801 | 0.552 | 0.429 | LS2 |
| | $\mathcal{N}(0.5, 0.01)$ | 0.01 | 0.0619 | 0.0186 | 0.0079 | 0.0006 | WK |
| | | | 0.0078 | 0.0014 | 0.0001 | 0.0001 | LS2 |
| | $B\mathcal{N}$ | 0.18 | 12.052 | 5.279 | 1.698 | 1.041 | WK |
| | | | 52.668 | 11.009 | 5.817 | 1.215 | LS2 |
| $s_3$ | $\mathcal{U}_{(0;1)}$ | 0.35 | 28.03 | 10.55 | 4.63 | 2.747 | WK |
| | | | 125.055 | 45.298 | 12.607 | 5.713 | LS1 |
| | | | 31.073 | 7.477 | 4.199 | 3.319 | LS2 |
| | $\gamma(4, 0.08)$ | 0.44 | 19.615 | 6.283 | 3.869 | 3.309 | WK |
| | | | 41.261 | 13.34 | 4.808 | 3.727 | LS1 |
| | | | 23.213 | 5.549 | 2.059 | 0.86 | LS2 |
| | $\mathcal{N}(0.5, 0.01)$ | 0.44 | 6.341 | 2.452 | 1.28 | 0.861 | WK |
| | | | 10.453 | 3.961 | 2.098 | 1.078 | LS1 |
| | | | 3.753 | 1.386 | 1.028 | 0.644 | LS2 |
| | $B\mathcal{N}$ | 0.32 | 44.381 | 13.618 | 9.637 | 7.928 | WK |
| | | | 182.525 | 58.787 | 24.229 | 12.317 | LS1 |
| | | | 66.663 | 30.377 | 8.521 | 4.574 | LS2 |

Table 4.2: Values of MISE $\times 1000$ averaged over 200 samples, for the estimators of the regression function (Example 1), built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).



(a) (b) (c)

Figure 4.4: Estimation in Example 3, in model M7, and $n = 1000$. (a)-(b)-(c) beams of 20 estimators built from i.i.d. sample (thin lines) versus true function (thick line): warped kernel estimators (subplot (a)), least-squares estimator in piecewise polynomial bases with degree at most 1 (subplot (b)) or 2 (subplot (c)).

### 4.5.3 Example 3: Interval censoring, case 1

The same comparison is carried out for the estimation of the c.d.f. under interval censoring. The adaptive least-squares estimate is provided by Brunel & Comte (2009), and the same Matlab toolbox is used for its implementation: recall that the target function can be seen as a regression function: $s(x) = \mathbb{P}(Z \leq x) = \mathbb{E}[\mathbf{1}_{Z \leq x} | X = x]$. Different models are considered for generating the data. The estimation set $A$ is calibrated, such that most of the data belong to this interval, as it is done in Brunel & Comte (2009). We shorten "follows the distribution" by the symbol "$\sim$".

- M1: $X \sim \mathcal{U}_{(0;1)}$, and $Z \sim \mathcal{U}_{(0;1)}$, $A = (0;1)$ (for instance, the target function is $F_Z : x \mapsto x$),
- M2: $X \sim \mathcal{U}_{(0;1)}$, and $Z \sim \chi_2(1)$ (Chi-squared distribution with 1 degree of freedom), $A = (0;1)$,
- M3: $X \sim \mathcal{E}(1)$ (exponential distribution with mean 1), and $Z \sim \chi_2(1)$, $A = (0;1.2)$,
- M4: $X \sim \beta(4,6)$ (Beta distribution of parameter (4,6)), $Z \sim \beta(4,8)$, $A = (0;0.5)$,
- M5: $X \sim \beta(4,6)$, $Z \sim \mathcal{E}(10)$ (exponential distribution with mean 0.1), $A = (0;0.5)$,
- M6: $X \sim \gamma(4, 0.08)$, $Z \sim \mathcal{E}(10)$, $A = (0, 0.5)$,
- M7: $X \sim \mathcal{E}(0.1)$, $Z \sim \gamma(4,3)$, $A = (1;23)$.

The first two models, and the fourth, were also used by Brunel & Comte (2009). All these models allow to investigate thoroughly the sensibility of the method to the distribution of the examination time $X$, and to the range of the estimation interval.

Figure Figure 4.4 shows the smoothness of warped-kernel estimates. We also explore the difference between the estimators by computing the MISE for the different models. Table 4.3 reveals that the warped-kernel estimates can advantageously be used as soon as the design $X_i$ has not a uniform distribution: it always outperforms the least-squares estimators in these cases.

| Model | $X$ | $Z$ | (a;b) | $n = 60$ | 200 | 500 | 1000 | Method |
|---|---|---|---|---|---|---|---|---|
| 1 | $\mathcal{U}_{(0;1)}$ | $\mathcal{U}_{(0;1)}$ | $(0;1)$ | 2.41 | 1.125 | 0.975 | 0.533 | WK |
| | | | | 0.63 | 0.111 | 0.056 | 0.024 | LS2 |
| 2 | $\mathcal{U}_{(0;1)}$ | $\chi_2(1)$ | $(0;1)$ | 1.558 | 0.804 | 0.57 | 0.415 | WK |
| | | | | 1.602 | 0.44 | 0.244 | 0.13 | LS2 |
| 3 | $\mathcal{E}(1)$ | $\chi_2(1)$ | $(0;1.2)$ | 1.285 | 0.614 | 0.243 | 0.247 | WK |
| | | | | 2.385 | 0.893 | 0.651 | 0.365 | LS2 |
| 4 | $\mathcal{B}(4,6)$ | $\mathcal{B}(4,8)$ | $(0;0.5)$ | 0.423 | 0.236 | 0.09 | 0.094 | WK |
| | | | | 0.449 | 0.271 | 0.117 | 0.105 | LS2 |
| 5 | $\mathcal{B}(4,6)$ | $\mathcal{E}(10)$ | $(0;0.5)$ | 0.388 | 0.229 | 0.119 | 0.103 | WK |
| | | | | 0.467 | 0.261 | 0.13 | 0.095 | LS2 |
| 6 | $\gamma(4,0.08)$ | $\mathcal{E}(10)$ | $(0;0.5)$ | 0.424 | 0.166 | 0.102 | 0.069 | WK |
| | | | | 0.698 | 0.286 | 0.162 | 0.095 | LS2 |
| 7 | $\mathcal{E}(0.1)$ | $\gamma(4,3)$ | $(1;23)$ | 14.955 | 5.145 | 3.973 | 2.113 | WK |
| | | | | 19.825 | 11.797 | 9.738 | 5.898 | LS2 |

Table 4.3: Values of MISE $\times 100$ averaged over 100 samples, for the estimators of the c.d.f. from current status data (Example 3) built with the warped kernel method (WK) or the least-squares methods, with piecewise polynomials of degree at most 1 or 2 (LS1 or LS2).

## 4.6 Proofs

### 4.6.1 Proof of Equality (4.8)

Equality (4.8) comes down to compute a conditional expectation: precisely, we prove that

$$\mathbb{E}\left[Y|X=x\right] = \frac{f_Z(x)}{f_X(x)}\frac{1-F_X(x)}{1-F_Z(x)}.$$

To do so, let $H$ be a test function. Recall that $X = Z \wedge C$, $Y = \mathbf{1}_{Z\leq C}$, and denote by $f_Z$ (resp. $f_C$) the density of $Z$ (resp. $C$). We compute

$$\begin{aligned}\mathbb{E}\left[YH(X)\right] &= \mathbb{E}\left[\mathbf{1}_{Z\leq C}H(Z)\right] = \int_{\mathbb{R}_+^2} \mathbf{1}_{z\leq c}H(z)f_Z(z)f_C(c)dzdc,\\ &= \int_{\mathbb{R}_+} \left(1-F_C(z)\right)H(z)f_Z(z)dz = \int_{\mathbb{R}_+} \frac{f_Z(z)}{f_X(z)}\frac{1-F_X(z)}{1-F_Z(z)}H(z)f_X(z)dz,\end{aligned}$$

taking into account the equality $1-F_X = (1-F_Z)(1-F_C)$. Thus, we identify $\mathbb{E}[Y|X]$.

□

### 4.6.2 Proof of Proposition 4.1

We have:

$$\begin{aligned}\mathbb{E}\left[\theta(Y)K_h(u-\Phi(X))\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta(Y)|X\right]K_h(u-\Phi(X))\right],\\ &= \int_A K_h(u-\Phi(x))\mathbb{E}\left[\theta(Y)|X=x\right]f_X(x)dx.\end{aligned}$$

We set $u' = \Phi(x)$, thus $du' = \phi(x)dx$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\theta(Y)K_h(u-\Phi(X))\right] &= \int_{\Phi(A)} K_h(u-u')\mathbb{E}\left[\theta(Y)|X=\Phi^{-1}(u)\right]f_X\left(\Phi^{-1}(u)\right)\frac{du}{\phi\circ\Phi^{-1}(u)},\\ &= \int_{\Phi(A)} K_h(u-u')s\circ\Phi^{-1}(u)du.\end{aligned}$$

□

### 4.6.3 Proof of Proposition 4.2

The following classical bias-variance decomposition holds:

$$\mathbb{E}\left[\|\hat{s}_h - s\|_\phi^2\right] = \|g-g_h\|^2_{L^2(\Phi(A))} + \mathbb{E}\left[\|g_h-\hat{g}_h\|^2_{L^2(\Phi(A))}\right],$$

since, thanks to (4.9), $\mathbb{E}[\hat{g}_h(u)] = g_h(u)$. We bound the variance term as follows:

$$\mathbb{E}\left[\|g_h-\hat{g}_h\|^2_{L^2(\Phi(A))}\right] = \mathbb{E}\left[\int_{\Phi(A)}\left(\hat{g}_h(u)-\mathbb{E}\left[\hat{g}_h(u)\right]\right)^2 du\right] = \int_{\Phi(A)}\mathrm{Var}\left(\hat{g}_h(u)\right)du,$$

and for each $u \in \Phi(A)$,

$$\mathrm{Var}\left(\hat{g}_h(u)\right) = \frac{1}{n}\mathrm{Var}\left(\theta(Y_1)K_h\left(u-\Phi(X_1)\right)\right) \leq \frac{1}{n}\mathbb{E}\left[\theta^2(Y_1)K_h^2\left(u-\Phi(X_1)\right)\right].$$

Therefore, by integrating with respect to $u$, we get

$$\mathbb{E}\left[\|g_h - \hat{g}_h\|^2_{L^2(\Phi(A))}\right] \leq \mathbb{E}\left[\theta^2(Y_1)\right]\|K\|^2_{L^2(\mathbb{R})}\frac{1}{nh}.$$

### 4.6.4 Proof of Theorem 4.1

Let $h \in \mathcal{H}_n$ be fixed. We start with the following decomposition for the loss of the estimator $\tilde{s} = \hat{s}_{\hat{h}}$:

$$\begin{aligned}\left\|\hat{s}_{\hat{h}} - s\right\|^2_\phi &= \left\|\hat{g}_{\hat{h}} - g\right\|^2_{L^2(\Phi(A))}, \\ &\leq 3\left\|\hat{g}_{\hat{h}} - \hat{g}_{h,\hat{h}}\right\|^2_{L^2(\Phi(A))} + 3\left\|\hat{g}_{h,\hat{h}} - \hat{g}_h\right\|^2_{L^2(\Phi(A))} + 3\left\|\hat{g}_h - g\right\|^2_{L^2(\Phi(A))}.\end{aligned}$$

The definitions of $A(h)$ and $A(\hat{h})$ enable us to write, using the definition of $\hat{h}$,

$$\begin{aligned}3\left\|\hat{g}_{\hat{h}} - \hat{g}_{h,\hat{h}}\right\|^2_{L^2(\Phi(A))} + 3\left\|\hat{g}_{h,\hat{h}} - \hat{g}_h\right\|^2_{L^2(\Phi(A))} &\leq 3\left(A(h) + V\left(\hat{h}\right)\right) + 3\left(A\left(\hat{h}\right) + V(h)\right), \\ &\leq 6\left(A(h) + V(h)\right),\end{aligned}$$

Besides, applying also Proposition 4.2, we obtain

$$\mathbb{E}\left[\left\|\hat{s}_{\hat{h}} - s\right\|^2_\phi\right] \leq 6\mathbb{E}\left[A(h)\right] + 6V(h) + \frac{\mathbb{E}[\theta^2(Y_1)]\|K\|^2_{L^2(\mathbb{R})}}{nh} + 3\|g_h - g\|^2_{L^2(\Phi(A))}. \tag{4.18}$$

Therefore, the remainding part of the proof follows from the lemma hereafter.

**Lemma 4.1.** *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 4.1, there exist constants $C_1, C_2$ such that,*

$$\mathbb{E}\left[A(h)\right] \leq C_1\|g_h - g\|^2_{L^2(\Phi(A))} + \frac{C_2}{n}, \tag{4.19}$$

*where the constant $C_1$ only depends on $\|K\|_{L^1(\mathbb{R})}$.*

Applying Inequality (4.19) in (4.18) implies (4.15) by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 4.1.

□

### 4.6.5 Proof of Lemma 4.1

To study $A(h)$, we introduce the auxiliary quantities $g_{h,h'} := K_{h'} \star (g_h \mathbf{1}_{\Phi(A)}) = K_{h'} \star ((K_h \star g\mathbf{1}_{\Phi(A)})\mathbf{1}_{\Phi(A)})$, for any $h' \in \mathcal{H}_n$, and we first split

$$\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_\phi = \|\hat{g}_{h,h'} - \hat{g}_{h'}\|^2_{L^2(\Phi(A))} \leq 3\left(T_a + T_b + \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}\right), \tag{4.20}$$

where

$$T_a = \|\hat{g}_{h,h'} - g_{h,h'}\|^2_{L^2(\Phi(A))}, \;\; T_b = \|g_{h,h'} - g_{h'}\|^2_{L^2(\Phi(A))}.$$

The first term can be bounded as follows, using the Young Inequality (2.17) (Proposition 2.13 of Chapter 2) with $p = 2$, $q = 1$, and $r = 2$:

$$\begin{aligned} T_a &\leq \left\| K_h \star \left(\hat{g}_{h'}\mathbf{1}_{\Phi(A)} - g_{h'}\mathbf{1}_{\Phi(A)}\right)\right\|^2_{L^2(\mathbb{R})}, \\ &\leq \|K\|^2_{L^1(\mathbb{R})} \left\|\hat{g}_{h'}\mathbf{1}_{\Phi(A)} - g_{h'}\mathbf{1}_{\Phi(A)}\right\|^2_{L^2(\mathbb{R})} = \|K\|^2_{L^1(\mathbb{R})} \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}. \end{aligned}$$

In the same way, $T_b \leq \|K_{h'}\|^2_{L^1(\mathbb{R})}\|g_h - g\|^2_{L^2(\Phi(A))}$. Therefore, the decomposition (4.20) becomes:

$$\|\hat{s}_{h,h'} - \hat{s}_{h'}\|^2_{f_X} \leq 3\|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\Phi(A))} + 3(1 + \|K\|^2_{L^1(\mathbb{R})}) \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))}.$$

Now, we get back to the definition of $A(h)$ given by (4.13):

$$\begin{aligned} A(h) \leq 3 \;& \|K\|^2_{L^1(\mathbb{R})}\|g - g_h\|^2_{L^2(\Phi(A))} \\ & +3(1 + \|K\|^2_{L^1(\mathbb{R})}) \max_{h' \in \mathcal{H}_n} \left( \|\hat{g}_{h'} - g_{h'}\|^2_{L^2(\Phi(A))} - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ . \end{aligned} \tag{4.21}$$

We apply Lemma 4.3: $\|\hat{g}_{h'} - g_{h'}\|_{L^2(\Phi(A))} = \sup_{t \in \bar{S}(0,1)} \langle \hat{g}_{h'} - g_{h'}, t\rangle_{\Phi(A)}$, with $\bar{S}(0,1)$ a dense countable subset of $\tilde{S}(0,1) = \{t \in L^1(\Phi(A)) \cap L^2(\Phi(A)),\ \|t\|_{L^2(\Phi(A))} = 1\}$. Now,

$$\begin{aligned} \langle \hat{g}_{h'} - g_{h'}, t\rangle_{\Phi(A)} &= \frac{1}{n}\sum_{i=1}^{n} \int_{\Phi(A)} \{\theta(Y_i)K_{h'}(u - \Phi(X_i)) - \mathbb{E}\left[\theta(Y_i)K_{h'}(u - \Phi(X_i))\right]\} t(u)du \\ &= \nu_{n,h'}(t), \end{aligned}$$

where $\nu_{n,h'}$ is an empirical process. Thus, thanks to (4.21), it remains to bound the deviations of $\sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t)$. First, we have

$$\begin{aligned} &\mathbb{E}\left[ \max_{h' \in \mathcal{H}_n} \left( \sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t) - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ \right] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}\left[ \left( \sup_{t \in \bar{S}(0,1)} \nu^2_{n,h'}(t) - \frac{V(h')}{3(1 + \|K\|^2_{L^1(\mathbb{R})})} \right)_+ \right]. \end{aligned}$$

Then, the conclusion results from the following lemma:

**Lemma 4.2.** *Under the assumptions of Theorem 4.1, there exists a constant $C$ such that,*

$$\sum_{h \in \mathcal{H}_n} \mathbb{E}\left[ \left( \sup_{t \in \bar{S}(0,1)} \nu^2_{n,h}(t) - \tilde{V}(h) \right)_+ \right] \leq \frac{C}{n},$$

*with $\tilde{V}(h) = \delta'\|K\|_{L^2(\mathbb{R})}\mathbb{E}[\theta(Y_1)^2]/(nh)$ for a numerical $\delta' > 0$.*

We choose the constant $\kappa$ involved in the definition of $V$ such that $\tilde{V}(h) \leq V(h)(1 + \|K\|^2_{L^1(\mathbb{R})})/3$. Thus, the proof is complete.

□

### 4.6.6 Proof of Lemma 4.2

We write the empirical process

$$\nu_{n,h}(t) = \frac{1}{n}\sum_{i=1}^{n}\psi_{t,h}(X_i,Y_i) - \mathbb{E}\left[\psi_{t,h}(X_i,Y_i)\right], \tag{4.22}$$
$$\text{with } \psi_{t,h}(X_i,Y_i) = \theta(Y_i)\int_{\Phi(A)} K_h\left(u - F_X(X_i)\right)t(u)du.$$

The guiding idea is to apply Talagrand's Inequality (Proposition 2.2 of Chapter 2). If $\theta$ is bounded, this inequality can be applied. Otherwise, we have to introduce a truncation.

**Example 1**

Recall that $\Phi = F_X$ and $\Phi(A) = (0;1)$. We split the process $\nu_{n,h}$ into three parts, writing $\nu_{n,h} = \nu_{n,h}^{(1)} + \nu_{n,h}^{(2,1)} + \nu_{n,h}^{(2,2)}$, with, for $l = 1, (2,1), (2,2)$,

$$\nu_{n,h}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{t,h}^{(l)}(Z_i) - \mathbb{E}\left[\varphi_{t,h}^{(l)}(Z_i)\right],$$

$Z_i = X_i$ or $(X_i,\varepsilon_i)$, and

$$\begin{array}{l}\varphi_{t,h}^{(1)} : x \mapsto s(x)\int_0^1 K_h(u - F_X(x))t(u)du,\\ \varphi_{t,h}^{(2,1)} : (x,\varepsilon) \mapsto \varepsilon\mathbf{1}_{|\varepsilon|\le\kappa_n}\int_0^1 K_h(u - F_X(x))t(u)du,\\ \varphi_{t,h}^{(2,2)} : (x,\varepsilon) \mapsto \varepsilon\mathbf{1}_{|\varepsilon|>\kappa_n}\int_0^1 K_h(u - F_X(x))t(u)du,\end{array}$$

where we define, for a constant $c$ which will be specified below,

$$\kappa_n = c\frac{\sqrt{n}}{\ln(n)}. \tag{4.23}$$

We apply Talagrand's Inequality to the first two bounded empirical processes, and bound roughly the last one. Thus, we split:

$$\begin{aligned}\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h}^2(t) - \tilde{V}(h)\right)_+\right] \le\ & 3\sum_{h\in\mathcal{H}_n}\left\{\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2 - \frac{\tilde{V}_1(h)}{3}\right)_+\right]\right.\\ &+\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1)}(t)\right)^2 - \frac{\tilde{V}_2(h)}{3}\right)_+\right]\\ &\left.+\ \mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,2)}(t)\right)^2\right]\right\},\end{aligned} \tag{4.24}$$

with the decomposition $\tilde{V}(h) = \tilde{V}_1(h) + \tilde{V}_2(h)$, and, denoting by $\delta'' = \delta'/2$,

$$\tilde{V}_1(h) = 3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[s^2(X_1)\right]}{nh},\ \text{and}\ \tilde{V}_2(h) = 3\delta''\frac{\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}\left[\varepsilon_1^2\right]}{nh}.$$

Actually, recall that we have $\mathbb{E}[\theta^2(Y_1)] = \mathbb{E}[Y_1^2] = \mathbb{E}[s^2(X_1)] + \mathbb{E}[\varepsilon_1^2]$ here.

We now show that each of the three terms of the right hand-side of (4.24) is upper-bounded by a quantity of order $1/n$. This will end the proof.

• ***First term of (4.24).***

Let us begin with $\nu_{n,h}^{(1)}$. To do so, we compute $H^{(1)}$, $M^{(1)}$ and $v^{(1)}$, involved in Inequality (2.1) of Proposition 2.2 in Chapter 2.

– For $M^{(1)}$, let $t \in \bar{S}(0,1)$ and $x \in A$ be fixed:

$$\begin{aligned}\left|\varphi_{t,h}^{(1)}(x)\right| &\leq |s(x)| \int_0^1 |K_h(u - F_X(x))t(u)|\, du \leq |s(x)| \|K_h\|_{L^2(\mathbb{R})} \|t\|_{L^2(\Phi(A))}, \\ &= |s(x)| \frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}} \leq \|s\|_{L^\infty(A)} \frac{\|K\|_{L^2(\mathbb{R})}}{\sqrt{h}} := M^{(1)}.\end{aligned}$$

– For $H^{(1)}$, notice that

$$\nu_{n,h}^{(1)}(t) = \langle \hat{d}_h - g_h, t \rangle_{\Phi(A)}, \text{ with } \hat{d}_h = \frac{1}{n} \sum_{i=1}^n s(X_i) K_h\left(. - F_X(X_i)\right).$$

Thus, thanks to Lemma 4.3, we obtain,

$$\begin{aligned}\mathbb{E}\left[\sup_{t \in \bar{S}(0,1)} \left(\nu_{n,h}^{(1)}(t)\right)^2\right] &= \mathbb{E}\left[\left\|\hat{d}_h - g_h\right\|_{L^2((0;1))}^2\right], \\ &= \int_0^1 \mathrm{Var}\left(\hat{d}_h(u)\right) du, \text{ since } g_h(u) = \mathbb{E}\left[\hat{d}_h(u)\right], \\ &\leq \int_0^1 \frac{1}{n} \mathbb{E}\left[s^2(X_1) K_h^2\left(u - F_X(X_1)\right)\right] du.\end{aligned}$$

Then, we use the same computation as the one done to bound the variance term in the proof of Proposition 4.2, and set $(H^{(1)})^2 = \|K\|_{L^2(\mathbb{R})}^2 \mathbb{E}[s^2(X_1)]/(nh)$.

– For $v^{(1)}$, we also fix $t \in \bar{S}(0,1)$. Hereafter, we set $\check{K}_h(u) = K_h(-u)$. First,

$$\mathrm{Var}\left(\varphi_{t,h}^{(1)}(X_1)\right) \leq \mathbb{E}\left[\left(\varphi_{t,h}^{(1)}(X_1)\right)^2\right] \leq \|s\|_{L^\infty(A)}^2 \mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right],$$

and the expectation can be written

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^1 K_h(u - F_X(X_1))t(u)du\right)^2\right] &= \mathbb{E}\left[\left(\check{K}_h * \left(t\mathbf{1}_{(0;1)}\right)\right)^2 \left(F_X(X_1)\right)\right], \\ &= \int_0^1 \left(\check{K}_h * \left(t\mathbf{1}_{(0;1)}\right)\right)^2 (u)du \leq \left\|\check{K}_h * \left(t\mathbf{1}_{(0;1)}\right)\right\|_{L^2(\mathbb{R})}^2, \\ &\leq \left\|\check{K}_h\right\|_{L^1(\mathbb{R})}^2 \left\|t\mathbf{1}_{(0;1)}\right\|_{L^2(\mathbb{R})}^2 = \left\|\check{K}_h\right\|_{L^1(\mathbb{R})}^2 \|t\|_{L^2((0;1))}^2,\end{aligned}$$

thanks to the Young Inequality (2.17) (Proposition 2.13 of Chapter 2). Therefore,

$$\mathrm{Var}\left(\varphi_{t,h}^{(1)}(X_1)\right) \leq \|s\|_{L^\infty(A)} \|K\|_{L^1(\mathbb{R})}^2 := v^{(1)}.$$

Then, the Talagrand Inequality gives, for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2 - 2(1+2\delta)\left(H^{(1)}\right)^2\right)_+\right] \leq k_1\left\{\frac{1}{n}\exp\left(-k_2\frac{1}{h}\right) + \frac{1}{n^2h}\exp\left(-k_3\sqrt{n}\right)\right\},$$

where $k_1, k_2, k_3$ are three constants which depend on $\mathbb{E}[s^2(X_1)]$, $\|s\|_{L^\infty(A)}$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. Assumptions (H2)-(H3) lead to

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2 - 2(1+2\delta)\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}[s^2(X_1)]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n},$$

with $C$ a constant (which also depends on the previous quantities).

- ***Second term of (4.24).***

For the second empirical process $\nu_{n,h}^{(2,1)}$, the sketch of the proof is the same: similarly, we compute the quantities involved in the Talagrand Inequality (Proposition 2.2 of Chapter 2),

$$M^{(2)} = \kappa_n\|K\|_{L^2(\mathbb{R})}\frac{1}{\sqrt{h}},\quad H^{(2)} = \|K\|_{L^2(\mathbb{R})}\left(\mathbb{E}[\varepsilon_1^2]\right)^{1/2}\frac{1}{\sqrt{nh}},\quad v^{(2)} = \|K\|_{L^1(\mathbb{R})}^2\mathbb{E}[\varepsilon_1^2],$$

and we obtain, by the concentration inequality, for $\delta > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1)}(t)\right)^2 - 2(1+2\delta)\left(H^{(2)}\right)^2\right)_+\right] \leq k_1\left\{\frac{1}{n}\exp\left(-k_2\frac{1}{h}\right) + \frac{\kappa_n^2}{n^2h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right)\right\},$$

where $k_1, k_2, k_3$ are three constants which depend on $\mathbb{E}[\varepsilon_1^2]$, $\|K\|_{L^1(\mathbb{R})}$ and $\|K\|_{L^2(\mathbb{R})}$. The first term of the right hand-side is like above. With the definition (4.23) of $\kappa_n$, the sum over $h \in \mathcal{H}_n$ of the second term of the upper bound can be written

$$\sum_{h\in\mathcal{H}_n}\frac{\kappa_n^2}{n^2h}\exp\left(-k_3\frac{\sqrt{n}}{\kappa_n}\right) = \frac{c^2}{n^{1+k_3/c}\ln^2(n)}\sum_{h\in\mathcal{H}_n}\frac{1}{h}.$$

Consequently, using Assumptions (H2)-(H3) and choosing $c$ in the definition of $\kappa_n$ such that $c \leq k_3/\alpha_0$, we also obtain for a constant $C$,

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1)}(t)\right)^2 - 2(1+2\delta)\|K\|_{L^2(\mathbb{R})}^2\mathbb{E}[\varepsilon_1^2]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n}.$$

- ***Third term of (4.24).***

The last empirical process is $\nu_{n,h}^{(2,2)}(t) = \int_0^1 t(u)\psi(u)du$, with

$$\psi(u) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}K_h\left(u - F_X(X_i)\right) - \mathbb{E}\left[\varepsilon_i\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}K_h\left(u - F_X(X_i)\right)\right].$$

It is not bounded. Nevertheless, we use the Cauchy-Schwarz Inequality, and the equality $\|t\|_{L^2(\Phi(A))} = 1$, for $t \in \bar{S}(0,1)$

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\tilde{S}(0,1)}\left(\nu_{n,h}^{(2,2)}(t)\right)^2\right] &\leq \mathbb{E}\left[\int_0^1\psi^2(u)du\right],\\
&\leq \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2\mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right]\mathbb{E}\left[\int_0^1 K_h^2\left(u - F_X(X_1)\right)du\right],\\
&\leq \frac{\|K\|_{L^2(\mathbb{R})}^2}{nh}\mathbb{E}\left[\varepsilon_1^2\mathbf{1}_{\{|\varepsilon_1|>\kappa_n\}}\right] \leq \frac{\|K\|_{L^2(\mathbb{R})}^2\kappa_n^{-p}}{nh}\mathbb{E}\left[\varepsilon_1^{2+p}\right].
\end{aligned}$$

Thus, there exists a constant $k_1$ which depends on $\|K\|_{L^2(\mathbb{R})}$ and $\mathbb{E}[\varepsilon_1^{2+p}]$,

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,2)}(t)\right)^2\right]\leq k_1\frac{\kappa_n^{-p}}{n}\sum_{h\in\mathcal{H}_n}\frac{1}{h}=c_1\kappa^{-p}\frac{\ln^p(n)}{n^{1+p/2}}\sum_{h\in\mathcal{H}_n}\frac{1}{h}.$$

The conclusion comes from Assumptions (H2)-(H3), and the choice of $p\geq 2\alpha_0$.

□

**Examples 2-4**

For the multiplicative regression model (Example 2), we split the process into two terms: $\nu_{n,h}=\nu_{n,h}^{(1)}+\nu_{n,h}^{(2)}$, with

$$\begin{aligned}\nu_{n,h}^{(1)}(t)=\frac{1}{n}\sum_{i=1}^n\Big\{&\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\\&-\mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|\leq\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right]\Big\},\\ \nu_{n,h}^{(2)}(t)=\frac{1}{n}\sum_{i=1}^n\Big\{&\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\\&-\mathbb{E}\left[\sigma^2(X_i)\varepsilon_i^2\mathbf{1}_{\{|\varepsilon_i|>\kappa_n\}}\int_0^1 K_h\left(u-F_X(X_i)\right)t(u)du\right]\Big\},\end{aligned}$$

where $\kappa_n$ is still a constant for the proof, which equals $\sqrt{c\frac{\sqrt{n}}{\ln(n)}}$ and $c>0$ is obtained by the computations, like in Example 1. We exactly recover the framework of this previous example: the deviations of the process $\nu_{n,h}^{(1)}$ are bounded thanks to Talagrand's Inequality of Proposition 2.2 (Chapter 2), and the second one is bounded in the same way as the process $\nu_{n,h}^{(2,2)}$ of the additive regression setting.

For Examples 3-4, there is no point in splitting the process (4.22), since it is already bounded (recall that $\theta(Y_1)$ is bounded by 1). Thus, we apply the concentration inequality.

Recall that $\Phi(A)=\mathbb{R}_+$. In both of these cases, the quantity $M_1$ involved in the assumptions of the Talagrand Inequality (2.1) equals $M_1=\|K\|_{L^2(\mathbb{R})}/\sqrt{h}$. Moreover, $H^2$ can be chosen as the upper-bound of the variance term of the estimator $\hat{g}_h$, that is $H^2=\|K\|_{L^2(\mathbb{R})}/nh$. Finally, $v$ equals $\|K\|_{L^1(\mathbb{R})}$ for Example 3, and $\|g\|_{L^\infty(\mathbb{R}_+)}\|K\|_{L^1(\mathbb{R})}$ for Example 4.

As an example, let us detail the computation of $v$ in Example 4. Recall that $X=C\wedge Z$, $Y=\mathbf{1}_{Z\leq C}$, $s$ is the hazard rate, and the warping $\Phi$ is the function $x\mapsto\int_0^x(1-F_X(t))dt$. Thus, denoting by $f_C$ (respectively $f_Z$) a density of the variable $C$ (respectively $Z$), and $F_C$ (respectively $F_Z$) its c.d.f.,

$$\begin{aligned}\mathrm{Var}\left(\psi_{t,h}(X_1,Y_1)\right)&\leq\mathbb{E}\left[\left(\psi_{t,h}(X_1,Y_1)\right)^2\right]=\mathbb{E}\left[Y_1\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(X_1))t(u')du'\right)^2\right],\\&=\int_{\mathbb{R}_+\times\mathbb{R}}\mathbf{1}_{z\leq c}\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(z))t(u')du'\right)^2f_C(c)f_Z(z)dzdc,\\&=\int_{\mathbb{R}_+}\left(\int_{\mathbb{R}_+}K_h(u'-\Phi(z))t(u')du'\right)^2f_Z(z)\left(1-F_C\right)(z)dz.\end{aligned}$$

We set $z = \Phi^{-1}(u)$. The integral becomes

$$\int_{\mathbb{R}_+} \left( \int_{\mathbb{R}_+} K_h(u' - \Phi(z)) t(u') du' \right)^2 f_Z(z) \left(1 - F_C\right)(z) dz$$
$$= \int_{\mathbb{R}_+} \left( \int_{\mathbb{R}_+} K_h(u' - u) t(u') du' \right)^2 f_Z \circ \Phi^{-1}(u) \left(1 - F_C\right) \circ \Phi^{-1}(u) \frac{du}{((1 - F_X) \circ \Phi^{-1}(u))}.$$

Thanks to the same arguments as the ones used to prove Proposition 4.1 in Section 4.6.2, we obtain:

$$\begin{aligned}
\mathrm{Var}\left(\varphi_{t,h}(X_1, Y_1)\right) &\leq \int_{\mathbb{R}_+} g(u) \left( \int_{\mathbb{R}_+} K_h(u' - u) t(u') du' \right)^2 du, \\
&= \int_{\mathbb{R}_+} g(u) \left( K_h * \left( t \mathbf{1}_{\mathbb{R}_+} \right)(u) \right)^2 du \leq \|g\|_{L^\infty(\mathbb{R}_+)} \left\| \check{K}_h * \left( t \mathbf{1}_{\mathbb{R}_+} \right) \right\|_{L^2(\mathbb{R})}, \\
&\leq \|g\|_{L^\infty(\mathbb{R}_+)} \left\| \check{K}_h \right\|_{L^1(\mathbb{R})} \left\| \left( t \mathbf{1}_{\mathbb{R}_+} \right) \right\|_{L^2(\mathbb{R})} = \|g\|_{L^\infty(\mathbb{R}_+)} \|K\|_{L^1(\mathbb{R})} := v.
\end{aligned}$$

Once we have the three quantities, we easily apply The Talagrand Inequality (2.1) and the proof is complete by using Assumptions (H2)-(H3), like above (see the computations in Example 1).

□

### 4.6.7 Proof of Corollary 4.1

We must bound the bias term of the right hand-side of Inequality (4.15) (Theorem 4.1). Actually, if we prove that

$$\|s - s_h\|_\phi^2 \leq C h^{2\beta},$$

where $C$ is a constant, then the proof of the Corollary will be completed by computing the minimum which is involved in (4.15). By definition,

$$\|s - s_h\|_\phi^2 = \|g - g_h\|_{L^2(\Phi(A))}^2 = \int_{\Phi(A)} \left(g_h(u) - g(u)\right)^2 du.$$

We distinguish two cases in the sequel, depending on the considered examples.

**Examples 1-3**

Here, $\Phi(A) = (0; 1)$. We start with the definition of $g_h$: for $u \in \Phi(A)$,

$$\begin{aligned}
g_h(u) &= \frac{1}{h} \int_0^1 g(u') K\left( \frac{u - u'}{h} \right) du' = \int_{\frac{u-1}{h}}^{\frac{u}{h}} g(u - hz) K(z) dz, \\
&= \int_{\frac{u-1}{h}}^{\frac{u}{h}} \tilde{g}(u - hz) K(z) dz = \int_{\mathbb{R}} \tilde{g}(u - hz) K(z) dz.
\end{aligned}$$

Thus, since $\int_{\mathbb{R}} K(u) du = 1$,

$$\tilde{g}_h(u) - g(u) = \int_{\mathbb{R}} K(z) \tilde{g}(u - hz) dz - \tilde{g}(u) = \int_{\mathbb{R}} K(z) \left[ \tilde{g}(u - hz) - \tilde{g}(u) \right] dz. \quad (4.25)$$

We use a Taylor-Lagrange formula for $\tilde{g}$: for $u \in (0;1)$, and $z \in \mathbb{R}$, there exists $\theta \in (0;1)$ such that

$$\tilde{g}(u-hz) - g(u) = -hz\tilde{g}'(u) + \frac{(-hz)^2}{2!}\tilde{g}''(u) + \cdots + \frac{(-hz)^{l-1}}{(l-1)!}\tilde{g}^{(l-1)}(u) + \frac{(-hz)^l}{l!}\tilde{g}^{(l)}(u-\theta hz),$$

with $l = \lfloor \beta \rfloor$. With Assumption $(K_l)$, we obtain

$$\|s - s_h\|_\phi^2 \leq \left( \int_{z\in\mathbb{R}} |K(z)| \frac{|hz|^l}{l!} \left\{ \int_{u=0}^1 \left\{ \tilde{g}^{(l)}(u-\theta hz) - \tilde{g}^{(l)}(u) \right\}^2 du \right\}^{1/2} dz \right)^2.$$

Since $\tilde{g}$ belongs to the Hölder space $\mathcal{H}(\beta, L)$,

$$\begin{aligned} \left[ \int_{u=0}^1 \left\{ \tilde{g}^{(l)}(u-\theta hz) - \tilde{g}^{(l)}(u) \right\}^2 du \right]^{1/2} &\leq \left[ \int_{u=0}^1 L^2 (\theta hu)^{2(\beta-l)} du \right]^{1/2}, \\ &= L|hz|^{\beta-l}, \end{aligned}$$

which enables us to conclude.

□

**Example 4**

Here, $\Phi(A) = \mathbb{R}_+$. Similarly, we first obtain Equality (4.25). Then, the idea is the same as in Examples 1-3, but since we integrate over an unbounded subset, we choose an integrated remainding term in the Taylor formula:

$$\begin{aligned} \tilde{g}(u-hz) - \tilde{g}(u) &= -hz\tilde{g}'(u) + \frac{(-hz)^2}{2!}\tilde{g}''(u) + \cdots + \frac{(-hz)^{l-1}}{(l-1)!}\tilde{g}^{(l-1)}(u) \\ &\quad + \frac{(-hz)^l}{(l-1)!} \int_0^1 (1-\theta)^{l-1} \tilde{g}^{(l)}(u-\theta hz) d\theta. \end{aligned}$$

The reasoning is then the same as in density estimation (see Tsybakov 2009 for details).

□

### 4.6.8 A technical tool

We state a lemma which allow us to replace a $L^2-$norm by the supremum of an empirical process.

**Lemma 4.3.** *Let $B$ be a borelian subset of $\mathbb{R}$. Denote by $\tilde{S}_B(0,1)$ the set of functions $t \in L^2(B)$ such that $\|t\|_{L^2(B)} = 1$. Then, for any function $v \in L^2(B)$,*

$$\|v\|_{L^2(B)} = \sup_{t \in \tilde{S}_B(0,1)} \langle v, t \rangle_B.$$

*Moreover, the supremum over $\tilde{S}_B(0,1)$ equals the supremum over a countable subset $\bar{S}_B(0,1)$ of $\tilde{S}_B(0,1)$.*

*Proof of Lemma 4.3.* The Cauchy-Schwarz Inequality leads to

$$\sup_{t\in\tilde{S}_B(0,1)} \langle v,t\rangle_B \leq \sup_{t\in\tilde{S}_B(0,1)} \|v\|_{L^2(B)}\|t\|_{L^2(B)} = \|v\|_{L^2(B)}.$$

Besides, if we set $t = v/\|v\|_{L^2(B)}$, then $t$ belongs to $\tilde{S}_B(0,1)$, and $\langle t,v\rangle_B = \|v\|_{L^2(B)}$. This ends the proof of the equality. Finally, we can replace $\tilde{S}_B(0,1)$ by one of its dense countable subset: such a set exists thanks to the separability of $L^2(\mathbb{R})$.

□

## 4.7 Appendix 1: Additional materials and results for the general case of unknown Φ

In this section, we give some details about the general case of an unknown transformation Φ.

### 4.7.1 Notations

For the sake of clarity, we begin with an overview of the notations. The warping functions Φ and their estimator $\hat{\Phi}_n$ are defined in Table 4.4, for $x \in A$, with $\hat{F}_n$ is the empirical counterpart for $F_X$. But instead of estimating $F_X$ with the whole sample $(X_i)_{i=1,\dots,n}$, we assume that another sample $(X_{-i})_{i=1,\dots,n}$, independent of the $X_i$'s, but distributed like them, is available. Thus, we set

$$\hat{F}_n : x \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_{-i}\leq x}.$$

The introduction of the second sample of variable $X$ is an artefact of the theory: it only allows to avoid dependency problems in the proof of the results, which are technical and cumbersome enough (see below). Using a single sample would have required totally different statistic and probabilistic tools. However, we have obviously used only one sample to compute the estimator in the simulation study, see Section 4.5 (otherwise the comparison with other methods would not have been fair).
Let us now recall the following definitions of the estimators we will studied in this section: for $h > 0$,

$$\hat{s}_h = \hat{g}_h \circ \hat{\Phi}_n, \quad \text{with } \hat{g}_h : u \mapsto \frac{1}{n}\sum_{i=1}^{n} \theta(Y_i) K_h\left(u - \hat{\Phi}_n(X_i)\right). \tag{4.26}$$

We aim at providing an upper bound for the risk of this estimator. The main challenge of the plug-in device is to bound the difference $\hat{\Phi}_n - \Phi$, in order to come down to the risk of the estimator with known Φ, bounded in Proposition 4.2.

### 4.7.2 Properties of Φ and $\hat{\Phi}_n$

We introduce and recall in this section properties and deviation inequalities which will be repeatedly used in the proof of Proposition 4.3.

| | $\Phi(x)$ | $\hat{\Phi}_n(x)$ |
|---|---|---|
| Examples 1-3. | $F_X(x)$ | $\hat{F}_n(x)$ |
| Example 4. | $\int_0^x (1 - F_X(t))\, dt$ | $\int_0^x \left(1 - \hat{F}_n(t)\right) dt$ |

Table 4.4: The warping functions and their estimators

First, recall from Chapter 2, Section 2.1.2 that we set $U_i = F_X(X_{-i})$, for $i \in \{1, \ldots, n\}$ and denote by $\hat{U}_n$ the empirical c.d.f related to the sample $(U_i)_{1 \leq i \leq n}$.

The Dvoretzky-Kiefer-Wolfowitz Inequalities (Propositions 2.3 and 2.4) of Section 2.1.2 (Chapter 2) are sufficient to handle the case $\Phi = F_X$, $\hat{\Phi}_n = \hat{F}_n$ (Examples 1-3). In Example 4, recall that the deformation is $\Phi(x) = \int_0^x (1 - F_X(t))dt$. We assume that the support $A$ can be written $A = (0; \tau)$ with finite $\tau$ (still in Example 4). Before studying the deviations of its empirical counterpart, we first state the following equalities, which will be useful, even though simple:

**Lemma 4.4.** *Denote by* $\Phi(x) = \int_0^x (1 - F_X(t))dt$, $x \in (0; \tau)$, *and* $\Phi' = \phi$. *The function* $\Phi$ *satisfies*

1. $\Phi(x) = \displaystyle\int_0^\tau (y \wedge x) f_X(y) dy$, $x \in (0; \tau)$,
2. $\mathbb{E}[X] = \int_0^\tau \phi(x) dx$.

***Proof of Lemma 4.4***

1. Fix $x \in (0; \tau)$, and compute

$$\begin{aligned} \Phi(x) &= \int_0^x \mathbb{P}(X > t) dt = \int_0^x \left( \int_0^\tau \mathbf{1}_{y>t} f_X(y) dy \right) dt, \\ &= \int_0^\tau f_X(y) \left( \int_0^x \mathbf{1}_{y>t} dt \right) dy = \int_0^\tau f_X(y)\, (y \wedge x)\, dy. \end{aligned}$$

2. To recover the expectation of $X$, we compute

$$\begin{aligned} \int_0^\tau \phi(x) dx &= \int_0^\tau (1 - F_X(x))\, dx = \int_0^\tau \left( \int_x^\infty f_X(t) dt \right) dx, \\ &= \int_0^\infty f_X(t) \int_0^\tau \mathbf{1}_{t>x} dx dt = \int_0^\infty f_X(t) t dt = \mathbb{E}[X]. \end{aligned}$$

□

Now, consider the estimator $\hat{\Phi}_n = \int_0^x (1 - \hat{F}_n(t))dt$. It can also be written $\hat{\Phi}_n(x) = \frac{1}{n}\sum_{i=1}^n X_{-i} \wedge x$. Thus it has no bias for the estimation of $\Phi$ (see the first point of Lemma 4.4). Moreover, recalling that $A = (0;\tau)$ with finite $\tau > 0$ (still in Example 4) permits to write,

$$\sup_{x\in(0;\tau)} \left|\hat{\Phi}_n(x) - \Phi(x)\right| \leq \tau \sup_{x\in\mathbb{R}_+} \left|\hat{F}_n(x) - F_X(x)\right|. \tag{4.27}$$

Thus, Proposition 2.3 and Corollary 2.4 are also useful to bound $\hat{\Phi}_n - \Phi$ in Example 4.
To standardize the notations in the proofs below, we finally set $\tau = 1$ when $\Phi = F_X$ and $\hat{\Phi}_n = \hat{F}_n$, that is in Examples 1-3. Therefore, Inequality (4.27) holds whatever $\Phi$ is.

### 4.7.3 Result

Before stating the result, we introduce new assumptions:

(H1') the function $s$ is continuously derivable on $A$.

(H2') the kernel $K$ is twice continuously derivable, with bounded derivatives $K'$ and $K''$ on $\mathbb{R}$.

Assumption (H1') is somehow restrictive but required for integration by parts (see Section 4.7.4). Assumption (H2') permits to use Taylor formulas to deal with the difference $K_h(u - \hat{\Phi}_n(X_i)) - K_h(u - \Phi(X_i))$. This is not a problem as we choose the kernel in practice.

We now illustrate how the plug-in device suits well to recover the function $s$ by providing an upper bound for the risk of $\hat{s}_h$.

**Proposition 4.3.** *Assume* (H1') *and* (H2'). *Assume also that* $A = (0;\tau)$ *with* $\tau < \infty$ *for Example 4. Then, there exist three constants* $c_1$, $c_2$ *and* $c_3$ *such that*

$$\mathbb{E}\left[\|\hat{s}_h - s\|_\phi^2\right] \leq 5\,\|g - g_h\|^2_{L^2(\Phi(A))} + c_1\frac{1}{nh} + c_2\frac{1}{n^2h^4} + c_3\frac{1}{n^2h^6}. \tag{4.28}$$

*If moreover* $h > n^{-1/5}$, *there exists* $c > 0$ *such that*

$$\mathbb{E}\left[\|\hat{s}_h - s\|_\phi^2\right] \leq 5\,\|g - g_h\|^2_{L^2(\Phi(A))} + c\frac{1}{nh}. \tag{4.29}$$

Notice that the additional assumption $A = (0;\tau)$ with $\tau < \infty$ is needed to control the difference $\hat{\Phi}_n - \hat{\Phi}$ in Example 4 (see Section above). Inequality (4.29) is immediatly deduced from (4.28) with the additional assumption $h > n^{-1/5}$. It shows that the same result as Proposition 4.2 holds when the warping function $\Phi$ is unknown, under mild assumptions. The main adaptive result (Theorem 4.1) in this general framework can then be deduced from this bound.

We can also compare the assumption on the bandwidth $h$ (required to handle the substitution of $\hat{\Phi}_n$ to $\Phi$) to the assumption on the model dimension $D_m$ in the previous chapter: to prove the analogous result in Chapter 3 (Proposition 3.1), $D_m$ is supposed to be bounded by $n^{1/3}$ (up to a logarithmic factor). Keeping in mind that $h$ plays the role of the inverse $1/D_m$, one could expect the assumption $h > n^{-1/3}$. However, in Proposition 4.3, we require

the stronger inequality $h > n^{-1/5}$. The control of the risk of the warped-kernel estimators is actually more technical than for the warped-bases estimators: we cannot use here the main tricks of projection estimation ($\|\Pi_{S_m} s\| \leq \|s\|$, or the other advantages of the trigonometric basis, see Section 3.3.2 for details). We thus use rough bounds which leads to the stronger constraint.

### 4.7.4 Proof of Proposition 4.3

**Main part of the proof of Proposition 4.3**

Let us specify the notations of this section. Our goal is to study the risk of $\hat{s}_h$ defined by (4.26): we denote it by $\hat{s}^{\hat{\Phi}_n,\hat{\Phi}_n}$. We have

$$\hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} = \hat{g}_h^{\hat{\Phi}_n} \circ \hat{\Phi}_n, \text{ with } \hat{g}_h^{\hat{\Phi}_n}(u) = \frac{1}{n}\sum_{i=1}^{n} \theta(Y_i) K_h\left(u - \hat{\Phi}_n(X_i)\right)$$

Moreover, $\hat{s}_h^{\Phi,\Phi} = \hat{g}_h^{\Phi} \circ \Phi$ with $\hat{g}_h^{\Phi}(u) = (1/n)\sum_{i=1}^{n} \theta(Y_i) K_h(u - \Phi(X_i))$ is the estimator studied in the main part of the chapter. Coherently, we also introduce $\hat{s}^{\hat{\Phi}_n,\Phi} = \hat{g}_h^{\hat{\Phi}_n} \circ \Phi$. The following decomposition is the key of the proof:

$$\left\| \hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - s \right\|_\phi^2 \leq 5 \sum_{l=0}^{3} T_l^h,$$

with

$$\begin{aligned}
T_0^h &= \left\| \hat{s}_h^{\Phi,\Phi} - s_h \right\|_\phi^2 + \left\| s_h^{\Phi} - s \right\|_\phi^2, \\
T_1^h &= \left\| \hat{s}_h^{\hat{\Phi}_n,\Phi} - \hat{s}_h^{\Phi,\Phi} - \mathbb{E}\left[ \hat{s}_h^{\hat{\Phi}_n,\Phi} - \hat{s}_h^{\Phi,\Phi} \,|(X_{-i})\right] \right\|_\phi^2, \\
T_2^h &= \left\| \hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\hat{\Phi}_n,\Phi} - \mathbb{E}\left[ \hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\hat{\Phi}_n,\Phi} \,|(X_{-i})\right] \right\|_\phi^2, \\
T_3^h &= \left\| \mathbb{E}\left[ \hat{s}_h^{\hat{\Phi}_n,\hat{\Phi}_n} - \hat{s}_h^{\Phi,\Phi} \,|(X_{-i})\right] \right\|_\phi^2,
\end{aligned}$$

where $\mathbb{E}[Z|(X_{-i})]$ is the conditional expectation of a variable $Z$ given the sample $(X_{-i})_{i=1,\dots,n}$. The term $T_0^h$ has been bounded in Proposition 4.2. For the three others, we set the following lemmas, which end the proof.

**Lemma 4.5.** *Under the assumptions of Proposition 4.3,*

$$\mathbb{E}\left[T_l^h\right] \leq \kappa_1 \mathbb{E}\left[\theta^2(Y_1)\right] \|K'\|_{L^\infty(\mathbb{R})} C_2 \frac{1}{n^2 h^4},$$

*with $\kappa_1 = 1$ in Examples 1-3, $\kappa_1 = \tau^2 \mathbb{E}[X]$ in Example 4, and $C_2$ defined by Proposition 2.4.*

**Lemma 4.6.** *Under the assumptions of Proposition 4.3,*

$$\mathbb{E}\left[T_3^h\right] \quad \leq \quad \frac{\kappa}{nh} + 16\kappa_1 \mathbb{E}\left[\theta^2(Y_1)\right] \left\|K''\right\|_{L^\infty(\mathbb{R})}^2 C_4 \frac{1}{n^2 h^6}.$$

*with* $\kappa_1 = \tau^2\mathbb{E}[X]$ *and*

$$\kappa = \begin{cases} 12\,\|K\|^2_{L^2(\mathbb{R})}\left(2C_2\|s\|^2_{L^\infty(A)} + 2\,\|s\|^2_\phi + \left(\frac{1}{4}+3C_2\right)\|s'\|^2_{L^2(A)}\right) & \textit{(Examples 1-3),} \\ \|K\|^2_{L^2(\mathbb{R})}\Big(\|s\|^2_{L^\infty(A)}(\mathbb{E}[X_{-1}^2]+2C_2\tau^2) & \\ \qquad +\|s'\|^2_{L^1(A)}(\mathbb{E}[X_{-1}^2]+C_2\tau^2)+\mathbb{E}[X_{-1}]\|s\|^2_{L^2(A)}\Big) & \textit{(Example 4).} \end{cases}$$

**Remark 4.2.** To prove Lemma 4.5, Assumption (H1') is not required and Assumption (H2') can be weakened: it is sufficient to assume that the kernel $K$ is continuously derivable, with bounded derivative $K'$.

□

**Proof of Lemma 4.5**

The first term to bound is

$$T_1^h = \int_A \left(\hat{g}_h^{\hat{\Phi}_n}\circ\Phi(x) - \hat{g}_h^{\Phi}\circ\Phi(x) - \mathbb{E}\left[\hat{g}_h^{\hat{\Phi}_n}\circ\Phi(x) - \hat{g}_h^{\Phi}\circ\Phi(x)\,|(X_{-i})\right]\right)^2\phi(x)dx,$$

and its conditional expectation is

$$\mathbb{E}\left[T_1^h\,|(X_{-i})\right] = \int_A \mathrm{Var}\left(\hat{g}_h^{\hat{\Phi}_n}(\Phi(x)) - \hat{g}_h^{\Phi}(\Phi(x))\,|(X_{-i})\right)\phi(x)dx.$$

For any $x\in A$, we compute

$$\begin{aligned}
&\mathrm{Var}\left(\hat{g}_h^{\hat{\Phi}_n}(\Phi(x)) - \hat{g}_h^{\Phi}(\Phi(x))\,|(X_{-i})\right) \\
&\quad= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \theta(Y_i)\left\{K_h\left(\Phi(x)-\hat{\Phi}_n(X_i)\right) - K_h\left(\Phi(x)-\Phi(X_i)\right)\right\}|(X_{-i})\right), \\
&\quad= \frac{1}{n}\mathrm{Var}\left(\theta(Y_1)\left\{K_h\left(\Phi(x)-\hat{\Phi}_n(X_1)\right) - K_h\left(\Phi(x)-\Phi(X_1)\right)\right\}|(X_{-i})\right), \\
&\quad\le \frac{1}{nh^2}\mathbb{E}\left[\theta^2(Y_1)\left\{K\left(\frac{\Phi(x)-\hat{\Phi}_n(X_1)}{h}\right) - K\left(\frac{\Phi(x)-\Phi(X_1)}{h}\right)\right\}^2|(X_{-i})\right].
\end{aligned}$$

The mean value theorem for the kernel $K$ between to real numbers $a$ and $b$ leads to:

$$|K(b)-K(a)| \le \|K'\|_{L^\infty(\mathbb{R})}|b-a|.$$

By choosing $a = (\Phi(x)-\hat{\Phi}_n(X_1))/h$ and $b = (\Phi(x)-\Phi(X_1))/h$, we obtain

$$\begin{aligned}
&\mathrm{Var}\left(\hat{g}_h^{\hat{\Phi}_n}(\Phi(x)) - \hat{g}_h^{\Phi}(\Phi(x))\,|(X_{-i})\right) \\
&\qquad\le \frac{1}{nh^2}\mathbb{E}\left[\theta^2(Y_1)\|K'\|^2_{L^\infty(\mathbb{R})}\frac{\left\{\Phi(X_i)-\hat{\Phi}_n(X_i)\right\}^2}{h^2}|(X_{-i})\right], \\
&\qquad\le \frac{1}{nh^4}\|K'\|^2_{L^\infty(\mathbb{R})}\left\|\hat{\Phi}_n-\Phi\right\|^2_{L^\infty(A)}\mathbb{E}\left[\theta^2(Y_1)\,|(X_{-i})\right], \\
&\qquad= \frac{1}{nh^4}\|K'\|^2_{L^\infty(\mathbb{R})}\left\|\hat{\Phi}_n-\Phi\right\|^2_{L^\infty(A)}\mathbb{E}\left[\theta^2(Y_1)\right].
\end{aligned}$$

Thus,

$$\mathbb{E}\left[T_1^h \,|(X_{-i})\right] \leq \frac{1}{nh^4}\int_A \phi(x)dx\|K'\|^2_{L^\infty(\mathbb{R})}\left\|\hat{\Phi}_n - \Phi\right\|^2_{L^\infty(A)} \mathbb{E}\left[\theta^2(Y_1)\right],$$

and consequently

$$\mathbb{E}\left[T_1^h\right] \leq \frac{1}{nh^4}\int_A \phi(x)dx\|K'\|^2_{L^\infty(\mathbb{R})}\mathbb{E}\left[\left\|\hat{\Phi}_n - \Phi\right\|^2_{L^\infty(A)}\right] \mathbb{E}\left[\theta^2(Y_1)\right].$$

With Inequality (4.27) and Proposition 2.4 successively, we obtain

$$\mathbb{E}\left[T_1^h\right] \leq \frac{\tau^2 C_2}{n^2h^4}\mathbb{E}\left[\theta^2(Y_1)\right]\int_A \phi(x)dx\|K'\|^2_{L^\infty(\mathbb{R})}.$$

To conclude, it remains to compute $\int_A \phi(x)dx$: it equals 1 in Examples 1-3 and $\mathbb{E}[X]$ in Example 4 (see Lemma 4.4).

To deal with the second term, we first write

$$T_2^h = \int_A \left(\hat{g}_h^{\hat{\Phi}_n}\circ\hat{\Phi}_n(x) - \hat{g}_h^{\hat{\Phi}_n}\circ\Phi(x) - \mathbb{E}\left[\hat{g}_h^{\hat{\Phi}_n}\circ\hat{\Phi}_n(x) - \hat{g}_h^{\hat{\Phi}_n}\circ\Phi(x)\,|(X_{-i})\right]\right)^2\phi(x)dx.$$

We now argue as for $T_1^h$:

$$\begin{aligned}
\mathbb{E}\left[T_2^h\,|(X_{-i})\right] &= \int_A \mathrm{Var}\left(\hat{g}_h^{\hat{\Phi}_n}\circ\hat{\Phi}_n(x) - \hat{g}_h^{\hat{\Phi}_n}(\Phi(x))\,|(X_{-i})\right)\phi(x)dx,\\
&\leq \int_A \frac{1}{nh^2}\mathbb{E}\left[\theta^2(Y_1)\left\{K\left(\frac{\hat{\Phi}_n(x)-\hat{\Phi}_n(X_1)}{h}\right)\right.\right.\\
&\quad \left.\left. -K\left(\frac{\Phi(x)-\hat{\Phi}_n(X_1)}{h}\right)\right\}^2 |(X_{-i})\right]\phi(x)dx,\\
&\leq \int_A \frac{1}{nh^2}\mathbb{E}\left[\theta^2(Y_1)\|K'\|^2_{L^\infty(\mathbb{R})}\frac{\left\{\hat{\Phi}_n(x)-\Phi(x)\right\}^2}{h^2}|(X_{-i})\right]\phi(x)dx,\\
&\leq \int_A \phi(x)dx\frac{1}{nh^4}\mathbb{E}\left[\theta^2(Y_1)\right]\|K'\|_{L^\infty(\mathbb{R})}\left\|\hat{\Phi}_n-\Phi\right\|^2_{L^\infty(A)}.
\end{aligned}$$

The conclusion is the same as for the first term.

□

**Proof of Lemma 4.6**

We have

$$\begin{aligned}
T_3^h &= \mathbb{E}\left[\int_A\left(\hat{s}^{\hat{\Phi}_n,\hat{\Phi}_n}(x) - \hat{s}^{\Phi,\Phi}(x)\right)^2\phi(x)dx\,|(X_{-i})\right],\\
&= \int_A\left(\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^n\theta(Y_i)\left\{K\left(\frac{\hat{\Phi}_n(x)-\hat{\Phi}_n(X_i)}{h}\right)-K\left(\frac{\Phi(x)-\Phi(X_i)}{h}\right)\right\}|(X_{-i})\right]\right)^2\phi(x)dx,\\
&= \int_A\left(\mathbb{E}\left[\frac{1}{h}\theta(Y_1)\left\{K\left(\frac{\hat{\Phi}_n(x)-\hat{\Phi}_n(X_1)}{h}\right)-K\left(\frac{\Phi(x)-\Phi(X_1)}{h}\right)\right\}|(X_{-i})\right]\right)^2\phi(x)dx,
\end{aligned}$$

since $(X_i, Y_i)_i$ are *i.i.d.* given the sample $(X_{-i})_i$. We apply the Taylor formula with Lagrange form for the remainder term $K$: for $b = (\hat{\Phi}_n(x) - \hat{\Phi}_n(X_1))/h$ and $a = (\Phi(x) - \Phi(X_1))/h$, there exists $\hat{\alpha}_{n,x,h}$ between $a$ and $b$ such that

$$K(b) - K(a) = (b-a)K'(a) + (b-a)^2 \frac{K''(\hat{\alpha}_{n,x,h})}{2}.$$

This leads to the decomposition $T_3^h \leq 2T_{3,1}^h + 2T_{3,2}^h$, with

$$T_{3,1}^h = \int_A \left( \mathbb{E}\left[ \frac{1}{h}\theta(Y_1) \left\{ \frac{\hat{\Phi}_n(x) - \hat{\Phi}_n(X_1) - \Phi(x) + \Phi(X_1)}{h} \right.\right.\right.$$
$$\left.\left.\left. \times\ K'\left( \frac{\Phi(x) - \Phi(X_1)}{h} \right) \right\} |(X_{-i}) \right] \right)^2 \phi(x)dx,$$
$$T_{3,2}^h = \int_A \left( \mathbb{E}\left[ \frac{1}{h}\theta(Y_1) \left\{ \frac{\left(\hat{\Phi}_n(x) - \hat{\Phi}_n(X_1) - \Phi(x) + \Phi(X_1)\right)^2}{h^2} K''(\hat{\alpha}_{n,x,h}) \right\} |(X_{-i}) \right] \right)^2 \phi(x)dx.$$

We now bound each of the terms.

• **Upper-bound for $\mathbf{T_{3,1}^h}$.**
Let us begin with the splitting $T_{3,1}^h \leq 2T_{3,1,1}^h + 2T_{3,1,2}^h$, where

$$T_{3,1,1}^h = \int_A \left( \mathbb{E}\left[ \frac{1}{h}\theta(Y_1) \left\{ \frac{\left(\hat{\Phi}_n(x) - \Phi(x)\right)}{h} K'\left( \frac{\Phi(x) - \Phi(X_1)}{h} \right) \right\} |(X_{-i}) \right] \right)^2 \phi(x)dx,$$
$$T_{3,1,2}^h = \int_A \left( \mathbb{E}\left[ \frac{1}{h}\theta(Y_1) \left\{ \frac{\left(\hat{\Phi}_n(X_1) - \Phi(X_1)\right)}{h} K'\left( \frac{\Phi(x) - \Phi(X_1)}{h} \right) \right\} |(X_{-i}) \right] \right)^2 \phi(x)dx.$$

⋆ **Upper-bound for $\mathbf{T_{3,1,1}^h}$.**
We have

$$\begin{aligned} T_{3,1,1}^h &= \frac{1}{h^2} \int_A \left( \frac{\left(\hat{\Phi}_n(x) - \Phi(x)\right)}{h^2} \mathbb{E}\left[ \theta(Y_1) K'\left( \frac{\Phi(x) - \Phi(X_1)}{h} \right) |(X_{-i}) \right] \right)^2 \phi(x)dx, \\ &= \frac{1}{h^2} \int_A \frac{\left(\hat{\Phi}_n(x) - \Phi(x)\right)^2}{h^2} \left( \mathbb{E}\left[ \theta(Y_1) K'\left( \frac{\Phi(x) - \Phi(X_1)}{h} \right) \right] \right)^2 \phi(x)dx. \end{aligned}$$

We now need a slightly different version of Proposition 4.1. We replace $K_h(u - \Phi(X))$ by any function $t \circ \Phi(X)$ such that the expectation exists, in the proof of Section 4.6.2, to obtain:

$$\mathbb{E}\left[\theta(Y_1) t \circ \Phi(X_1)\right] = \int_{\Phi(A)} t(u')g(u')du' = \int_A t(\Phi(x'))s(x')\phi(x')dx', \tag{4.30}$$

where the last inequality is the consequence of the change of variable $x = \Phi^{-1}(x')$. Here, with $t = K'((\Phi(x) - \Phi(X_1))/h)$, (4.30) leads to

$$\begin{aligned} T_{3,1,1}^h &= \frac{1}{h^2}\int_A \left(\hat{\Phi}_n(x) - \Phi(x)\right)^2 \left(\int_A \frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right) s(x')\phi(x')dx'\right)^2 \phi(x)dx, \\ &\leq \frac{1}{h^2}\int_A \left\|\hat{\Phi}_n - \Phi\right\|^2_{L^\infty(A)} \left(\int_A \frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right) s(x')\phi(x')dx'\right)^2 \phi(x)dx, \end{aligned}$$

We anew apply Inequalities (4.27) and Proposition 2.4, which give

$$\mathbb{E}\left[T_{3,1,1}^h\right] \leq \frac{\tau^2 C_2}{nh^2}\int_A (J_x)^2\,\phi(x)dx, \text{ with } J_x = \int_A \frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right) s(x')\phi(x')dx'.$$

We now discuss two cases, depending on the value of $\Phi(A)$ (see Table 4.4).

- **Example 1-3** ($\Phi(A) = (0;1)$). By the changes of variables $u = \Phi(x)$ and $u' = \Phi(x')$, the last inequality can be written

$$\mathbb{E}\left[T_{3,1,1}^h\right] \leq \frac{C_2}{nh^2}\int_{(0;1)} (I_u)^2\,du, \text{ avec } I_u = \int_{(0;1)} \frac{1}{h}K'\left(\frac{u - u'}{h}\right) g(u')du'.$$

An integration by parts permits to compute $I_u$ ($s$, and thus $g$ are assumed to be continuously derivable):

$$\begin{aligned} I_u &= \left[-g(u')K\left(\frac{u - u'}{h}\right)\right]_0^1 + \int_{(0;1)} K\left(\frac{u - u'}{h}\right) g'(u')du', \\ &= -g(1)K\left(\frac{u-1}{h}\right) + g(0)K\left(\frac{u}{h}\right) + \int_{(0;1)} K\left(\frac{u - u'}{h}\right) g'(u')du'. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[T_{3,1,1}^h\right] &\leq \frac{C_2}{nh^2}\int_{(0;1)} \left(-g(1)K\left(\frac{u-1}{h}\right) + g(0)K\left(\frac{u}{h}\right) + \int_{(0;1)} K\left(\frac{u - u'}{h}\right) g'(u')du'\right)^2 du, \\ &= \frac{C_2}{nh}\int_{(0;h)} \left(-g(1)K\left(v - \frac{1}{h}\right) + g(0)K(v) + \int_{(0;1)} K\left(v - \frac{u'}{h}\right) g'(u')du'\right)^2 du, \\ &\leq \frac{3C_2}{nh}\left\{g^2(1)\int_{(0;1)} K^2\left(v - \frac{1}{h}\right)dv + g^2(0)\int_{(0;1)} K^2(v)dv \right. \\ &\quad \left. + \int_{(0;1)} \left(g'(u')\right)^2 du' \int_{(0;1)} \left(\int_{(0;1)} K^2\left(v - \frac{u'}{h}\right)du'\right)dv\right\}, \\ &\leq \frac{3}{C_2 nh}\left\{g^2(1)\|K\|^2_{L^2(\mathbb{R})} + g^2(0)\|K\|^2_{L^2(\mathbb{R})} + \int_{(0;1)} \left(g'(u')\right)^2 du' \|K\|^2_{L^2(\mathbb{R})}\right\}, \\ &\leq \frac{3C_2}{4nh}\left(2\|g\|^2_{L^\infty((0;1))} + \|g'\|^2_{L^2((0;1))}\right)\|K\|^2_{L^2(\mathbb{R})}, \\ &\leq \frac{3C_2}{4nh}\left(2\|s\|^2_{L^\infty(A)} + \|s'\|^2_{L^2(A)}\right)\|K\|^2_{L^2(\mathbb{R})} \end{aligned}$$

- **Example 4** ($A = (0;\tau)$). We also integrate by parts:

$$\begin{aligned} J_x &= \left[-s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\right]_0^\tau + \int_0^\tau s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx', \\ &= -s(\tau)K\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right) + s(0)K\left(\frac{\Phi(x)}{h}\right) + \int_0^\tau s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'. \end{aligned}$$

With similar computations as in the previous examples, but with variables $x$ and $x'$ in the integrals,

$$\begin{aligned} \mathbb{E}\left[T_{3,1,1}^h\right] \leq \frac{3C_2\tau^2}{nh^2}\Bigg\{ & s^2(\tau)\int_0^\tau K^2\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right)\phi(x)dx \\ & + s^2(0)\int_0^\tau K^2\left(\frac{\Phi(x)}{h}\right)\phi(x)dx \\ & + \int_0^\tau\left(\int_0^\tau s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx\Bigg\}. \end{aligned}$$

There are three terms to bound. First,

$$\begin{aligned} s^2(\tau)\int_0^\tau K^2\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right)\phi(x)dx &\leq \|s\|^2_{L^\infty(A)}\|K\|^2_{L^2(\mathbb{R})}, \\ s^2(0)\int_0^\tau K^2\left(\frac{\Phi(x)}{h}\right)\phi(x)dx &\leq \|s\|^2_{L^\infty(A)}\|K\|^2_{L^2(\mathbb{R})}. \end{aligned}$$

For the most complicated term, we apply the generalized Minkowski Inequality (see Proposition 2.12, Chapter 2):

$$\begin{aligned} &\int_0^\tau\left(\int_0^\tau s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx \\ &\qquad \leq \left[\int_0^\tau |s'(x')|\left(\int_0^\tau K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x)dx\right)^{1/2}dx'\right]^2. \end{aligned}$$

Moreover,

$$\begin{aligned} \left(\int_0^\tau K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x)dx\right)^{1/2} &= \left(\int_0^{\Phi(\tau)} K^2\left(\frac{u-\Phi(x')}{h}\right)du\right)^{1/2}, \\ &= \left(\int_0^{\Phi(\tau)/h} K^2\left(v-\frac{\Phi(x')}{h}\right)hdv\right)^{1/2}, \\ &\leq \sqrt{h}\|K\|_{L^2(\mathbb{R})}, \end{aligned}$$

and therefore

$$\begin{aligned} \int_0^\tau\left(\int_0^\tau s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx &\leq h\|K\|^2_{L^2(\mathbb{R})}\left[\int_0^\tau s'(x')dx'\right]^2, \\ &\leq h\|K\|^2_{L^2(\mathbb{R})}\|s'\|^2_{L^1(A)}. \end{aligned}$$

At the end,

$$\mathbb{E}\left[T_{3,1,1}^h\right] \leq \frac{3C_2\tau^2}{nh}\|K\|^2_{L^2(\mathbb{R})}\left\{2\|s\|^2_{L^\infty(A)} + \|s'\|^2_{L^1(A)}\right\}.$$

Finally, whatever the example we consider, we have shown that

$$\mathbb{E}\left[T^h_{3,1,1}\right] \leq \frac{3C_2\tau^2}{nh}\|K\|^2_{L^2(\mathbb{R})}\left\{2\|s\|^2_{L^\infty(A)} + C_s\right\}, \tag{4.31}$$

with $C_s = \|s'\|^2_{L^2(A)}$ (Examples 1-3), and $C_s = \|s'\|^2_{L^1(A)}$ (Example 4).

⋆ **Upper-bound for $\mathbf{T^h_{3,1,2}}$.**

Recall first the definition of this term:

$$T^h_{3,1,2} = \int_A \left(\mathbb{E}\left[\frac{1}{h}\theta(Y_1)\left\{\frac{\left(\hat{\Phi}_n(X_1) - \Phi(X_1)\right)}{h}K'\left(\frac{\Phi(x) - \Phi(X_1)}{h}\right)\right\}|(X_{-i})\right]\right)^2 \phi(x)dx.$$

By using that $\hat{\Phi}_n$ is measurable with respect to the sample $(X_{-i})_i$, and by noticing that the $X_i$'s and $Y_i$'s are independent of $(X_{-i})_i$, we can derive a conditional version of (4.30): for any function $t$ for which the expectation exists,

$$\mathbb{E}\left[\theta(Y_1)t\left(\hat{\Phi}_n(X_1), \Phi(X_1)\right)|(X_{-i})\right] = \int_A t\left(\hat{\Phi}_n(x'), \Phi(x')\right)s(x')\phi(x')dx'.$$

This leads to

$$T^h_{3,1,2} = \frac{1}{h^2}\int_A \left(\int_A \frac{\left(\hat{\Phi}_n(x') - \Phi(x')\right)}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right)s(x')\phi(x')dx'\right)^2 \phi(x)dx.$$

Now, $\hat{\Phi}_n(x')$ is an empirical mean of variables with expectation $\Phi(x')$:

- If $\hat{\Phi}_n$ is the empirical c.d.f, it is the mean of $\mathbf{1}_{X_{-i}\leq x'}$, with expectation $F_X(x')$,
- If $\hat{\Phi}_n(x) = \int_0^x(1 - \hat{F}_n(t))dt$, it is the mean of $X_{-i} \wedge x'$, with expectation $\Phi(x')$ (see Lemma 4.4).

Thus, we have

$$\mathbb{E}\left[T^h_{3,1,2}\right] = \frac{1}{h^2}\int_A \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \int_A T_{i,x'}\frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right)s(x')\phi(x')dx'\right)\phi(x)dx,$$

where $T_{i,x'} = \mathbf{1}_{X_{-i}\leq x'}$ for Examples 1-3, and $T_{i,x'} = X_{-i} \wedge x'$ for Example 4 (see Table 4.4 and Lemma 4.4). Then,

$$\begin{aligned}\mathbb{E}\left[T^h_{3,1,2}\right] &= \frac{1}{nh^2}\int_A \mathrm{Var}\left(\int_A T_{1,x'}\frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right)s(x')\phi(x')dx'\right)\phi(x)dx, \\ &= \frac{1}{nh^2}\int_A \mathbb{E}\left[\left(\int_A \left(T_{1,x'} - \mathbb{E}\left[T_{1,x}\right]\right)\frac{1}{h}K'\left(\frac{\Phi(x) - \Phi(x')}{h}\right)s(x')\phi(x')dx'\right)^2\right]\phi(x)dx.\end{aligned}$$

We must now separate the two cases, depending on the definition of $T_{1,x}$.

- **Examples 1-3**.

We set $u = \Phi(x)$ and $u' = \Phi(x')$ in the integrals. With $F_X(X_{-1}) = U_1$, the last inequality becomes

$$\mathbb{E}\left[T^h_{3,1,2}\right] \leq \frac{1}{nh^2}\int_{(0;1)} \mathbb{E}\left[(I_{U_1,u})^2\right] du,$$

$$\text{with } I_{U_1,u} = \int_0^1 \left(\mathbf{1}_{U_1\leq u'} - u\right) g(u')\frac{1}{h}K'\left(\frac{u-u'}{h}\right) du'.$$

An integration by parts leads to:

$$\begin{aligned}
I_{U_1,u} &= \int_{U_1}^1 g(u')\frac{1}{h}K'\left(\frac{u-u'}{h}\right) du' - \int_0^1 u'g(u')\frac{1}{h}K'\left(\frac{u-u'}{h}\right) du', \\
&= \left[-g(u')K\left(\frac{u-u'}{h}\right)\right]_{U_1}^1 + \int_{U_1}^1 g'(u')K\left(\frac{u-u'}{h}\right) du' + \left[u'g(u')K\left(\frac{u-u'}{h}\right)\right]_0^1 \\
&\quad - \int_0^1 \left[g(u') + u'g'(u')\right] K\left(\frac{u-u'}{h}\right) du', \\
&= g(U_1)K\left(\frac{u-U_1}{h}\right) + \int_0^1 g'(u')K\left(\frac{u-u'}{h}\right)\mathbf{1}_{U_1\leq u'}du' \\
&\quad - \int_0^1 \left[g(u') + u'g'(u')\right] K\left(\frac{u-u'}{h}\right) du', \\
&= g(U_1)K\left(\frac{u-U_1}{h}\right) + \int_0^1 g'(u')K\left(\frac{u-u'}{h}\right)\left(\mathbf{1}_{U_1\leq u'} - u'\right) du' \\
&\quad - \int_0^1 g(u')K\left(\frac{u-u'}{h}\right) du'.
\end{aligned}$$

It follows that $\mathbb{E}[T^h_{3,1,2}] \leq 3/(nh^2)\{\mathbb{E}[T^h_{3,1,2,1}] + \mathbb{E}[T^h_{3,1,2,2}] + \mathbb{E}[T^h_{3,1,2,3}]\}$, with

$$\begin{aligned}
T^h_{3,1,2,1} &= \int_0^1 \left(g(U_1)K\left(\frac{u-U_1}{h}\right)\right)^2 du, \\
T^h_{3,1,2,2} &= \int_0^1 \left(\int_0^1 g'(u')K\left(\frac{u-u'}{h}\right)\left(\mathbf{1}_{U_1\leq u'} - u'\right) du'\right)^2 du, \\
T^h_{3,1,2,3} &= \int_0^1 \left(\int_0^1 g(u')K\left(\frac{u-u'}{h}\right) du'\right)^2 du.
\end{aligned}$$

We now show that the expectations of these three terms are roughly of size $h$. The main argument is that $h\|K_h\|_{L^2(\mathbb{R})} = \|K\|_{L^2(\mathbb{R})}$. Precisely, for the first term,

$$\begin{aligned}
\mathbb{E}\left[T^h_{3,1,2,1}\right] &= \int_0^1\int_0^1 g^2(u')K^2\left(\frac{u-u'}{h}\right) du'du, \\
&\leq \int_0^1 g^2(u')h\,\|K\|^2_{L^2(\mathbb{R})}\,du' = h\,\|K\|^2_{L^2(\mathbb{R})}\,\|g\|^2_{L^2((0;1))}\,.
\end{aligned}$$

The second term is bounded by,

$$\begin{aligned}\mathbb{E}\left[T^h_{3,1,2,2}\right] &\leq \int_0^1 \mathbb{E}\left[\int_0^1 K^2\left(\frac{u-u'}{h}\right)\left(\mathbf{1}_{U_1\leq u'}-u'\right)^2\left(g'(u')\right)^2 du'\right] du,\\ &= \int_0^1\int_0^1 K^2\left(\frac{u-u'}{h}\right)u'(1-u')\left(g'(u')\right)^2 du'du,\\ &\leq \frac{1}{4}\int_0^1\left(\int_0^1 K^2\left(\frac{u-u'}{h}\right)du\right)\left(g'(u')\right)^2 du',\\ &\leq h\frac{1}{4}\left\|K\right\|^2_{L^2(\mathbb{R})}\left\|g'\right\|^2_{L^2((0;1))}.\end{aligned}$$

Finally, the same method leads to

$$\mathbb{E}\left[T^h_{3,1,2,2}\right] \leq \int_0^1\int_0^1 g^2(u')K^2\left(\frac{u-u'}{h}\right)du'du \leq h\left\|K\right\|^2_{L^2(\mathbb{R})}\left\|g\right\|^2_{L^2((0;1))}.$$

We have proved that

$$\begin{aligned}\mathbb{E}\left[T^h_{3,1,2}\right] &\leq \frac{3}{nh}\left\|K\right\|^2_{L^2(\mathbb{R})}\left(2\left\|g\right\|^2_{L^2((0;1))}+\frac{1}{4}\left\|g'\right\|^2_{L^2((0;1))}\right),\\ &= \frac{3}{nh}\left\|K\right\|^2_{L^2(\mathbb{R})}\left(2\left\|s\right\|^2_{\phi}+\frac{1}{4}\left\|s'\right\|^2_{L^2(A)}\right).\end{aligned}$$

We gather this inequality with (4.31):

$$\mathbb{E}\left[T^h_{3,1}\right] \leq \frac{6}{nh}\left\|K\right\|^2_{L^2(\mathbb{R})}\left(2C_2\|s\|^2_{L^\infty(A)}+2\left\|s\right\|^2_{\phi}+\left(\frac{1}{4}+3C_2\right)\left\|s'\right\|^2_{L^2(A)}\right).$$

- **Example 4**.

The term to bound is

$$\begin{aligned}\mathbb{E}\left[T^h_{3,1,2}\right] &= \frac{1}{nh^2}\int_0^\tau \mathrm{Var}\left(\int_0^\tau X_{-1}\wedge x'\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)s(x')\phi(x')dx'\right)\phi(x)dx,\\ &\leq \frac{1}{nh^2}\int_0^\tau \mathbb{E}\left[\left(\underbrace{\int_0^\tau X_{-1}\wedge x'\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)s(x')\phi(x')dx'}_{J_x}\right)^2\right]\phi(x)dx.\end{aligned}$$

We can split $J_x$ into two terms: $J_x = J_{x,1}+J_{x,2}$, with

$$\begin{aligned}J_{x,1} &= \int_0^\tau x'\mathbf{1}_{x'\leq X_{-1}}\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)s(x')\phi(x')dx',\\ J_{x,2} &= X_{-1}\int_0^\tau \mathbf{1}_{x'>X_{-1}}\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)s(x')\phi(x')dx'.\end{aligned}$$

We integrate $J_{x,1}$ by parts:

$$\begin{aligned}J_{x,1} &= \int_0^{X_{-1}} x's(x')\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x')dx',\\ &= \left[-x's(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\right]_0^{X_{-1}}+\int_0^{X_{-1}}\left[x's(x')\right]'K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx',\\ &= -X_{-1}s(X_{-1})K\left(\frac{\Phi(x)-\Phi(X_{-1})}{h}\right)+\int_0^{X_{-1}}\left[s(x')+x's'(x')\right]K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx',\end{aligned}$$

and compute similarly $J_{x,2}$,

$$\begin{aligned}
J_{x,2} &= X_{-1}\int_{X_{-1}}^{\tau} s(x')\frac{1}{h}K'\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x')dx', \\
&= X_{-1}\left[-s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\right]_{X_{-1}}^{\tau} + X_{-1}\int_{X_{-1}}^{\tau} s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx', \\
&= X_{-1}s(X_{-1})K\left(\frac{\Phi(x)-\Phi(X_{-1})}{h}\right) - X_{-1}s(\tau)K\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right) \\
&\quad + X_{-1}\int_{X_{-1}}^{\tau} s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'.
\end{aligned}$$

We add the two results:

$$\begin{aligned}
J_x &= -X_{-1}s(\tau)K\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right) \\
&\quad + \int_0^{\tau}\left(\mathbf{1}_{x'>X_{-1}}X_{-1} + x'\mathbf{1}_{x'\leq X_{-1}}\right)s'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx' \\
&\quad + \int_0^{X_{-1}} s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx', \\
&= -X_{-1}s(\tau)K\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right) + \int_0^{\tau} X_{-1}\wedge x's'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx' \\
&\quad + \int_0^{X_{-1}} s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'.
\end{aligned}$$

It follows that

$$\mathbb{E}\left[T_{3,1,2}^h\right] = \frac{1}{nh^2}\int_0^{\tau}\mathbb{E}\left[J_x^2\right]\phi(x)dx \leq \frac{3}{nh^2}\left(\mathbb{E}\left[T_{3,1,2,1}^h\right] + \mathbb{E}\left[T_{3,1,2,2}^h\right] + \mathbb{E}\left[T_{3,1,2,3}^h\right]\right)$$

with

$$\begin{aligned}
T_{3,1,2,1}^h &= \int_0^{\tau} X_{-1}^2 s^2(\tau)K^2\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right)\phi(x)dx, \\
T_{3,1,2,2}^h &= \int_0^{\tau}\left(\int_0^{\tau} X_{-1}\wedge x's'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx, \\
T_{3,1,2,3}^h &= \int_0^{\tau}\left(\int_0^{X_{-1}} s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx.
\end{aligned}$$

For the first term

$$\begin{aligned}
\mathbb{E}\left[T_{3,1,2,1}^h\right] &= \mathbb{E}\left[X_{-1}^2\right]s^2(\tau)\int_0^{\tau} K^2\left(\frac{\Phi(x)-\Phi(\tau)}{h}\right)\phi(x)dx, \\
&= \mathbb{E}\left[X_{-1}^2\right]s^2(\tau)\int_0^{\Phi(\tau)} K^2\left(\frac{u-\Phi(\tau)}{h}\right)du, \\
&= h\mathbb{E}\left[X_{-1}^2\right]s^2(\tau)\int_0^{\Phi(\tau)/h} K^2\left(v-\frac{\Phi(\tau)}{h}\right)dv, \\
&\leq h\mathbb{E}\left[X_{-1}^2\right]\|s\|_{L^\infty(A)}^2\|K\|_{L^2(\mathbb{R})}^2.
\end{aligned}$$

For the second, we apply the generalized Minkowski Inequality (Proposition 2.12 of Chapter 2),

$$\begin{aligned}
\mathbb{E}\left[T^h_{3,1,2,2}\right] &= \mathbb{E}\left[\int_0^\tau \left(\int_0^\tau X_{-1}\wedge x's'(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\phi(x)dx\right],\\
&\leq \mathbb{E}\left[\left\{\int_0^\tau\left(\int_0^\tau \left(X_{-1}\wedge x's'(x')\right)^2K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x)dx\right)^{1/2}dx\right\}^2\right],\\
&= \mathbb{E}\left[\left\{\int_0^\tau X_{-1}\wedge x'|s'(x')|\left(\int_0^\tau K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)\phi(x)dx\right)^{1/2}dx\right\}^2\right],\\
&\leq \mathbb{E}\left[X_{-1}^2\right]\left\{\int_0^\tau |s'(x')|\sqrt{h}\|K\|_{L^2(\mathbb{R})}dx'\right\}^2,\\
&= h\mathbb{E}\left[X_{-1}^2\right]\|K\|_{L^2(\mathbb{R})}\|s'\|^2_{L^1(A)}.
\end{aligned}$$

And finally,

$$\begin{aligned}
\mathbb{E}\left[T^h_{3,1,2,3}\right] &= \int_0^\tau \mathbb{E}\left[\left(\int_0^{X_{-1}} s(x')K\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\right)^2\right]\phi(x)dx,\\
&\leq \mathbb{E}\left[X_{-1}\int_0^\tau\int_0^{X_{-1}} s^2(x')K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\phi(x)dx\right],\\
&\leq \mathbb{E}\left[X_{-1}\right]\int_0^\tau\int_0^\tau s^2(x')K^2\left(\frac{\Phi(x)-\Phi(x')}{h}\right)dx'\phi(x)dx,\\
&\leq h\mathbb{E}\left[X_{-1}\right]\|K\|^2_{L^2(\mathbb{R})}\|s\|^2_{L^2(A)}.
\end{aligned}$$

At the end,

$$\mathbb{E}\left[T^h_{3,1,2}\right] \leq \frac{3}{nh}\|K\|^2_{L^2(\mathbb{R})}\left(\mathbb{E}\left[X_{-1}^2\right]\|s\|^2_{L^\infty(A)}+\mathbb{E}\left[X_{-1}^2\right]\|s'\|^2_{L^1(A)}+\mathbb{E}\left[X_{-1}\right]\|s\|^2_{L^2(A)}\right).$$

We gather this inequality with (4.31), like for the first three examples:

$$\begin{aligned}
\mathbb{E}\left[T^h_{3,1}\right] \leq \frac{6}{nh}\|K\|^2_{L^2(\mathbb{R})}&\left(\|s\|^2_{L^\infty(A)}(\mathbb{E}[X_{-1}^2]+2C_2\tau^2)\right.\\
&\left.+\|s'\|^2_{L^1(A)}(\mathbb{E}[X_{-1}^2]+C_2\tau^2)+\mathbb{E}[X_{-1}]\|s\|^2_{L^2(A)}\right).
\end{aligned}$$

The upper-bound of $T^h_{3,1}$ is thus completed in each of the examples.

- **Upper-bound for $\mathbf{T^h_{3,2}}$.**

We again split the term: $T^h_{3,2}\leq 4T^h_{3,2,1}+4T^h_{3,2,2}$, with

$$T^h_{3,2,1}=\int_A\left(\mathbb{E}\left[\frac{1}{h}\theta(Y_1)\left\{\frac{\left(\hat\Phi_n(x)-\Phi(x)\right)^2}{h^2}K''\left(\hat\alpha_{n,x,h}\right)\right\}|(X_{-i})\right]\right)^2\phi(x)dx,$$

$$T^h_{3,2,2}=\int_A\left(\mathbb{E}\left[\frac{1}{h}\theta(Y_1)\left\{\frac{\left(\hat\Phi_n(X_1)-\Phi(X_1)\right)^2}{h^2}K''\left(\hat\alpha_{n,x,h}\right)\right\}|(X_{-i})\right]\right)^2\phi(x)dx.$$

The two terms can be bounded in the same way. We detail the computation for the first one:

$$\begin{aligned} T_{3,2,1}^h &\leq \frac{1}{h^6}\int_A \mathbb{E}\left[\theta^2(Y_1)\left\{\left(\hat{\Phi}_n(x)-\Phi(x)\right)^4\left(K''\left(\hat{\alpha}_{n,x,h}\right)\right)^2\right\}|(X_{-i})\right]\phi(x)dx, \\ &\leq \frac{1}{h^6}\int_A \mathbb{E}\left[\theta^2(Y_1)\,|(X_{-i})\right]\left\|\hat{\Phi}_n-\Phi\right\|_{L^\infty(A)}^4\left\|K''\right\|_{L^\infty(\mathbb{R})}^2\phi(x)dx, \\ &= \frac{1}{h^6}\int_A \phi(x)dx\mathbb{E}\left[\theta^2(Y_1)\right]\left\|\hat{\Phi}_n-\Phi\right\|_{L^\infty(A)}^4\left\|K''\right\|_{L^\infty(\mathbb{R})}^2. \end{aligned}$$

The conclusion is analogous as the one of the proof of Lemma 4.5 (see Section 4.7.4). We obtain

$$\mathbb{E}\left[T_{3,2,1}^h\right] \leq \kappa_1\mathbb{E}\left[\theta^2(Y_1)\right]\left\|K''\right\|_{L^\infty(\mathbb{R})}^2 C_4\frac{1}{n^2h^6},$$

where $\kappa_1 = 1$ for Examples 1-3, and $\kappa_1 = \tau^4\mathbb{E}[X]$ in Example 4. The same bound holds for $T_{3,2,2}^h$. At the end,

$$\mathbb{E}\left[T_{3,2}^h\right] \leq 8\kappa_1\mathbb{E}\left[\theta^2(Y_1)\right]\left\|K''\right\|_{L^\infty(\mathbb{R})}^2 C_4\frac{1}{n^2h^6}.$$

□

## 4.8 Appendix 2: Generalization of the proof of Lemma 4.2

### 4.8.1 Objective

Recall that Lemma 4.2 gives a bound of size $1/n$ for

$$\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\nu_{n,h}^2(t)-\tilde{V}(h)\right)_+\right],$$

with, for any $t \in \bar{S}(0,1)$ (a subset of $\tilde{S}(0,1) = \{t \in L^1(\Phi(A)) \cap L^2(\Phi(A)),\ \|t\|_{L^2(\Phi(A))} = 1\}$)

$$\nu_{n,h}(t) = \frac{1}{n}\sum_{i=1}^n\int_{\Phi(A)}\left\{\theta(Y_i)K_{h'}\left(u-\Phi(X_i)\right)-\mathbb{E}\left[\theta(Y_i)K_{h'}\left(u-\Phi(X_i)\right)\right]\right\}t(u)du,$$

and where $\tilde{V}(h) = \delta'\|K\|_{L^2(\mathbb{R})}\mathbb{E}[\theta(Y_1)^2]/(nh)$ for a numerical $\delta' > 0$. We have provided in Section 4.6.6 a proof which depends on the different considered examples, mainly under Assumption (H1) ($s$ is bounded). The aim of this section is to show that a general proof is possible with (H1) replaced by the following slightly stronger assumption:

(H1$_{bis}$) (i) The conditional expectation $\mathbb{E}[\theta(Y)|X]$ is bounded,

(ii) The ratio $f_X/\phi$ is bounded.

To see that (H1$_{bis}$) is more restrictive than (H1), one must recall that the target function $s$ can be defined by (4.4): $s = (f_X/\phi)\mathbb{E}[\theta(Y)|X]$.

Notice that Assumption (i) is automatically fulfilled in Example 3-4, since $\theta(Y) = Y$, with $Y$ bounded by 1. Moreover, if $\Phi = F_X$, the ratio $f_X/\phi$ equals 1 and (ii) is satisfied.

Under (H1$_{bis}$), we give a proof which embraces Examples 1-4 and moreover, permits to handle all the statistical settings covered by the formula (4.4). In the sequel, $s_0(X) := \mathbb{E}[\theta(Y)|X]$.

### 4.8.2 Generalized proof

We begin with the following splitting of the empirical process: $\nu_{n,h} = \nu_{n,h}^{(1b)} + \nu_{n,h}^{(2,1b)} + \nu_{n,h}^{(2,2b)}$, where $\nu_{n,h}^{(l)}$ has the form

$$\nu_{n,h}^{(l)}(t) = \frac{1}{n}\sum_{i=1}^{n} \xi_{t,h}^{(l)} - \mathbb{E}\left[\xi_{t,h}^{(l)}\right], \quad l \in \{(1b),(2,1b),(2,2b)\},$$

with, for $\kappa_n^b = c^b\sqrt{n}/\ln(n)$ $(c^b > 0)$,

$$\begin{aligned}
\xi_{t,h}^{(1b)}(X_i) &= s_0(X_i)\int_{\Phi(A)} K_{h'}\left(u - \Phi(X_i)\right) t(u)du, \\
\xi_{t,h}^{(2,1b)}(X_i,Y_i) &= \left(\theta(Y_i) - s_0(X_i)\right)\mathbf{1}_{|\theta(Y_i)-s_0(X_i)|\leq\kappa_n^b}\int_{\Phi(A)} K_{h'}\left(u - \Phi(X_i)\right) t(u)du, \\
\xi_{t,h}^{(2,2b)}(X_i,Y_i) &= \left(\theta(Y_i) - s_0(X_i)\right)\mathbf{1}_{|\theta(Y_i)-s_0(X_i)|>\kappa_n^b}\int_{\Phi(A)} K_{h'}\left(u - \Phi(X_i)\right) t(u)du.
\end{aligned}$$

The guideline is to apply the Talagrand Inequality (Proposition 2.2) to bound the first two empirical processes $\nu_{n,h}^{(1b)}$ and $\nu_{n,h}^{(2,1b)}$, and bound the third one roughly. Details can be found in the three following sections, in which we often refer to Section 4.6.6, since the computations are similar. Finally, gathering the three upper bounds of size $1/n$ (up to constant) permits to complete the proof of Lemma 4.2.

**Remark 4.3.** Clearly, if $\theta$ is bounded, the proof can be greatly simplified, since there is no use in splitting the empirical process in this case. However, we choose to give the most general formulation.

#### Upper-bound for $\nu_{n,h}^{(1b)}$

To apply the concentration results of Proposition 2.2, we first need to compute appropriate values for the bounds $M_1^{(1b)}$, $H^{(1b)}$ and $v^{(1b)}$. The computation for the first quantity is totally similar to the one done for $M_1^{(1)}$ in Section 4.6.6, and leads to $M_1^{(1b)} = \|s_0\|_{L^\infty(A)}\|K\|_{L^2(\mathbb{R})}\sqrt{h}$. In an analogous way, we first obtain

$$\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1b)}(t)\right)^2\right] \quad\leq\quad \|K\|_{L^2(\mathbb{R})}^2\frac{\mathbb{E}[s_0^2(X_1)]}{nh}.$$

But, $\mathbb{E}[s_0^2(X_1)] = \mathbb{E}[\mathbb{E}[\theta(Y_1)|X_1]^2] \leq \mathbb{E}[\theta^2(Y_1)]$. Thus, we set $(H_1^{(1b)})^2 = \mathbb{E}[\theta^2(Y_1)]\|K\|_{L^2(\mathbb{R})}^2/nh$. Finally, $v^{(1b)}$ is a bound for the following variance:

$$\begin{aligned}
\mathrm{Var}\left(\xi_{t,h}^{(1)}(X_1)\right) &\leq \mathbb{E}\left[s_0^2(X_1)\left(\int_{\Phi(A)} K_h(u-\Phi(X_1))t(u)du\right)^2\right], \\
&\leq \|s_0\|_{L^\infty(A)}^2\mathbb{E}\left[\left(\int_{\Phi(A)} K_h(u-\Phi(X_1))t(u)du\right)^2\right].
\end{aligned}$$

The difference between this computation and the one of $v^{(1)}$ in Section 4.6.6 is the way to deal with the expectation:

$$\begin{aligned}\mathbb{E}\left[\left(\int_{\Phi(A)} K_h(u-\Phi(X_1))t(u)du\right)^2\right] &= \mathbb{E}\left[\left(\check{K}_h \star t\mathbf{1}_{\Phi(A)}\right)^2(\Phi(X_1))\right],\\ &= \int_A \left(\check{K}_h \star t\mathbf{1}_{\Phi(A)}\right)^2(\Phi(x))\, f_X(x)dx,\\ &\leq \left\|\frac{f_X}{\phi}\right\|_{L^\infty(A)} \int_A \left(\check{K}_h \star t\mathbf{1}_{\Phi(A)}\right)^2(\Phi(x))\,\phi(x)dx,\end{aligned}$$

thanks to Assumption ($\mathrm{H1}_{bis}$) (ii). The change of variables $u=\Phi(x)$ leads to

$$\begin{aligned}\mathbb{E}\left[\left(\int_{\Phi(A)} K_h(u-\Phi(X_i))t(u)du\right)^2\right] &\leq \left\|\frac{f_X}{\phi}\right\|_{L^\infty(A)} \int_{\Phi(A)} \left(\check{K}_h \star t\mathbf{1}_{\Phi(A)}\right)^2(u)du,\\ &= \left\|\frac{f_X}{\phi}\right\|_{L^\infty(A)} \left\|\check{K}_h \star t\mathbf{1}_{\Phi(A)}\right\|_{L^2(\mathbb{R})}. \qquad (4.32)\end{aligned}$$

We end the computation in the same way as previously (see $v^{(1)}$, Section 4.6.6), and then $v^{(1b)} = \|f_X/\phi\|_{L^\infty(A)}\|K\|_{L^1(\mathbb{R})}$. Now, applying the Talagrand Inequality is straightforward, and, like in Section 4.6.6, thanks to Assumptions (H2)-(H3), we obtain

$$\sum_{h\in\mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1b)}(t)\right)^2 - 2(1+2\delta)\|K\|^2_{L^2(\mathbb{R})}\mathbb{E}[\theta^2(X_1)]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n}.$$

**Upper-bound for $\nu_{n,h}^{(2,1b)}$**

We also precise the quantities which permit to apply the concentration result. Their computations are also inspired by Section 4.6.6, and precisely by the bound for $\nu_{n,h}^{(2,1)}$. To begin, we choose $M_1^{(2b)} = \kappa_n^b\|K\|_{L^2(\mathbb{R})}\sqrt{h}$. For $H^{(2b)}$, similarly to $H^{(2)}$, we first obtain

$$\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(1)}(t)\right)^2\right] \leq \frac{1}{n}\mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2\int_{\Phi(A)} K_h^2(u-\Phi(X_1))du\right]$$

But, $\int_{\Phi(A)} K_h^2(u-\Phi(X_1))du \leq \|K_h\|^2_{L^2(\mathbb{R})}$, and

$$\mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2\right] = \mathbb{E}\left[\mathrm{Var}(\theta(Y_1)|X_1)\right] \leq \mathbb{E}[\theta^2(Y_1)]. \qquad (4.33)$$

Therefore, we set $(H_1^{(2b)})^2 = \mathbb{E}[\theta^2(Y_1)]\|K\|^2_{L^2(\mathbb{R})}/nh$. Finally,

$$\begin{aligned}
\mathrm{Var}\left(\xi_{t,h}^{(2,1b)}(X_1,Y_1)\right) &\leq \mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2\left(\int_{\Phi(A)} K_h(u-\Phi(X_i))t(u)du\right)^2\right],\\
&= \mathbb{E}\left[\mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2|X_1\right]\left(\int_{\Phi(A)} K_h(u-\Phi(X_i))t(u)du\right)^2\right],\\
&\leq \mathbb{E}\left[\mathbb{E}[\theta^2(Y_1)]\left(\int_{\Phi(A)} K_h(u-\Phi(X_i))t(u)du\right)^2\right],
\end{aligned}$$

by using (4.33). It remains to recall that (4.32) holds, to set $v^{(2b)} = \mathbb{E}[\theta^2(Y_1)]\|K\|_{L^1(\mathbb{R})}$.

Once again, the existence and size of these three quantities justify that we obtain

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,1b)}(t)\right)^2 - 2(1+2\delta)\|K\|^2_{L^2(\mathbb{R})}\mathbb{E}[\theta^2(X_1)]\frac{1}{nh}\right)_+\right] \leq \frac{C}{n}.$$

**Upper-bound for $\nu_{n,h}^{(2,2b)}$**

We exactly follow the same line as for $\nu_{n,h}^{(2,2)}$ (see Section 4.6.6) to write

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\tilde{S}(0,1)}\left(\nu_{n,h}^{(2,2b)}(t)\right)^2\right] &\leq \frac{1}{n}\mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2\mathbf{1}_{\{|\theta(Y_1)-s_0(X_1)|>\kappa_n^b\}}\int_{\Phi(A)}K_h^2\left(u-\Phi(X_1)\right)du\right],\\
&\leq \frac{\|K\|_{L^2(\mathbb{R})}}{nh}\mathbb{E}\left[(\theta(Y_1)-s_0(X_1))^2\mathbf{1}_{\{|\theta(Y_1)-s_0(X_1)|>\kappa_n^b\}}\right],\\
&\leq \|K\|_{L^2(\mathbb{R})}\frac{(\kappa_n^b)^{-p}}{nh}\mathbb{E}\left[|\theta(Y_1)-s_0(X_1)|^{2+p}\right],
\end{aligned}$$

with $p$ defined by (H5). As previously, the way to conclude can be found in Section 4.6.6,

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\sup_{t\in\bar{S}(0,1)}\left(\nu_{n,h}^{(2,2b)}(t)\right)^2\right] \leq \frac{C}{n}.$$

□

## 4.9 Appendix 3: What about a different selection device?

In this chapter, we have chosen an estimator among the collection $(\hat{s}_h)_{h\in\mathcal{H}}$ with a simplified version of the Goldenshluger-Lepski method. In Chapter 3, we have modified this method to adapt it to the model selection framework: see Section 3.2.3, Definitions (3.11) and (3.12). Reciprocally, we propose to study what happens if we choose a bandwidth with a criterion inspired by the one defined in Section 3.2.3.

Keeping in mind that a bandwidth $h$ is "equivalent" to the inverse $1/D_m$ of a model dimension, we set, to replace $A(h)$ (see (4.13)),

$$\tilde{A}(h) = \max_{h' \in \mathcal{H}_n} \left\{ \|\hat{s}_{h \vee h'} - \hat{s}_{h'}\|_\phi^2 - V(h') \right\}_+, \tag{4.34}$$

with $V(h')$ defined by (4.12). We also keep the definition

$$\tilde{s} = \hat{s}_{\tilde{h}} \quad \text{with} \quad \tilde{h} = \arg\min_{h \in \mathcal{H}_n} \left\{ \tilde{A}(h) + V(h) \right\}. \tag{4.35}$$

This estimator also satisfies an oracle-type inequality, similar to Theorem 4.1. Under the same assumptions, but with the regularity condition of Corollary 4.1 additionally, the result is

$$\mathbb{E}\left[ \|\tilde{s} - s\|_\phi^2 \right] \leq c_1 \min_{h \in \mathcal{H}_n} \left\{ h^{2\alpha} + \frac{\mathbb{E}\left[\theta^2(Y_1)\right] \|K\|_{L^2(\mathbb{R})}^2}{nh} \right\} + \frac{c_2}{n}. \tag{4.36}$$

Compare to (4.15): the bias term $\|s - s_h\|_\phi^2$ is replaced by its bound of order $h^{2\alpha}$. This comes from the proof of an upper-bound for $\mathbb{E}[\tilde{A}(h)]$. We refer to Section 4.6.5 for a comparison with the bound of $\mathbb{E}[A(h)]$: instead of $T_b$, the term which appears here is $\|g_{h \vee h'} - g_{h'}\|^2$, and thus, we have to control its maximum over $h' \in \mathcal{H}_n$. This can be done in the following way:

$$\max_{h' \in \mathcal{H}_n} \|g_{h \vee h'} - g_{h'}\|^2 \quad = \quad \max_{\substack{h' \in \mathcal{H}_n \\ h' \leq h}} \|g_h - g_{h'}\|^2 \leq 2\|g_h - g\|^2 + 2 \max_{\substack{h' \in \mathcal{H}_n \\ h' \leq h}} \|g_{h'} - g\|^2.$$

The second term cannot be directly identified as the bias term $\|g_h - g\|^2$. However, it can be bounded as follows, under the regularity condition on $g$:

$$\max_{\substack{h' \in \mathcal{H}_n \\ h' \leq h}} \|g_{h'} - g\|^2 \leq \max_{\substack{h' \in \mathcal{H}_n \\ h' \leq h}} C(\alpha, L)(h')^{2\alpha} \lesssim h^{2\alpha}.$$

And, thus, we can only prove that $\mathbb{E}[\tilde{A}(h)] \lesssim h^{2\alpha}$.

Although analogous, in the considered framework, the selection rule (4.34) is thus less efficient. But we must keep in mind that it can be the only alternative to set up a bandwidth selection method, when the bias cannot be directly expressed as $K_h \star s$: this is the case in a work in progress, with Angelina Roche (Univ. Montpellier) to estimate the conditional c.d.f of a real random variable with respect to a functional covariate.

**Remark 4.4.** To carry on the comparison between the selection devices, let us recall what happens when we apply a strategy inspired by Goldenshluger & Lepski (2011a) for model selection purpose: contrary to the case of kernel estimates, we only have the possibility to define $A(m) = \max_{m' \in \mathcal{M}_n} \{\|\hat{s}_{m \wedge m'} - \hat{s}_{m'}\|_\phi^2 - V(m')\}_+$ (see Paragraph entitled "Stratégie Mixte", Section 1.1.3 of Chapter 1 for the related explanations). But in this case, we can nevertheless obtained $A(m) \leq C\|\Pi_{S_m} s - s\|^2$ (and not only $A(m) \leq C D_m^{-2\alpha}$), by assuming that the models are nested: see e.g. Lemma 3.2 of Chapter 3.

## 4.10 Appendix 4: Extension of the method to warped-bases estimation

In Chapter 3, we have widely detailed the adaptive estimation of a regression function by projection estimators with warped bases (with model selection performed through penalization or through the GL method), where the deformation is the c.d.f.. In this chapter, we have studied different estimation problems, including additive regression (see Examples 1 to 4, Section 4.2.1) with a warped-kernel device. We could also propose the warped-bases strategy for the examples of this chapter, which means that we could extend the results of Chapter 3 to other statistical frameworks. We briefly describe what the warped-bases method looks like in Examples 1 to 4: we present in each case the contrast which is minimized to obtain projection estimator of an auxiliary function.

We have dealt with Example 1 (additive regression) in Chapter 3. To recover $s$, we have used the transformed data $(F_X(X_i), Y_i)$ (like in the warped-kernek strategy of this chapter). Projection estimators are built by the minimization of the contrast $\gamma_n(., F_X)$ (defined by (3.3)) or $\gamma_n(., \hat{F}_n)$. Examples 2-4 can be handled similarly.

In Example 2 (multiplicative regression), we also use the transformed data $(F_X(X_i), Y_i)$. The auxiliary function is still $g = s \circ F_X^{-1}$. The contrast can be written:

$$\gamma_n(t, F_X) := \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i^2 \left(t \circ F_X(X_i)\right).$$

Similarly, with the transformed data $(F_X(X_i), Y_i)$ of Example 3 (interval censoring, case 1), the auxiliary function $g$ is recovered by the minimization of the contrast

$$\gamma_n(t, F_X) := \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i \leq X_i} \left(t \circ F_X(X_i)\right).$$

Finally, in Example 4 (hazard rate estimation from right-censored data), recall that the transformed data are $(\Phi(X_i), Y_i)$, with $\Phi(x) = \int_0^x (1 - F_X(t))dt$. The auxiliary function can be written $g = s \circ \Phi^{-1}$, with $s$ the hazard rate function (see (4.8)) and the constrast

$$\gamma_n(t, \Phi) := \|t\|^2_{L^2(\mathbb{R}_+)} - \frac{2}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i \leq C_i} \left(t \circ \Phi(X_i)\right).$$

These four frameworks can be recapitulated as follows, like in Section 4.2.1: denote by $\Phi$ the c.d.f. $F_X$ in the first three examples, and let $\theta$ be the function $y \mapsto y$ in Examples 1,3,4, and $y \mapsto y^2$ in Example 2. The contrast is thus in any case

$$\gamma_n(t, \Phi) := \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \theta(Y_i) \left(t \circ \Phi(X_i)\right). \tag{4.37}$$

By minimizing it on a model $S_m$, we estimate the function $g = s \circ \Phi^{-1}$ in each of the examples. The estimator of $g$ is thus of form $\sum_{j=1}^{D_m} \hat{a}_j \varphi_j$, and the estimator for $s$ is

| Example | $s$ | $\Phi$ | $\hat{a}_j$ |
|---|---|---|---|
| 1. $Y = s(X) + \varepsilon$ | $s$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j\left(F_X(X_i)\right)$ |
| 2. $Y = \sigma(X)\varepsilon$ | $\sigma^2$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \varphi_j\left(F_X(X_i)\right)$ |
| 3. $(X, \mathbf{1}_{Z\leq X})$ | $F_Z$ | $F_X$ | $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq X_i} \varphi_j\left(F_X(X_i)\right)$ |
| 4. $(X = Z \wedge C, \mathbf{1}_{Z\leq C})$ | $\frac{f_Z}{1-F_Z}$ | $\Phi(x) = \int_0^x (1 - F_X(t))dt$ | $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Z_i\leq C_i} \varphi_j\left(\Phi(X_i)\right)$ |

Table 4.5: Summary of the examples and of the "warping" function used in each case.

$\sum_{j=1}^{D_m} \hat{a}_j \varphi_j \circ \Phi$ (replace $\Phi$ by an empirical version if it is unknown). The values of the coefficients $\hat{a}_j$ are precised in Table 4.5.

These definitions are justified with an equality which amounts to (4.9) of Proposition 4.1: for any squared-integrable function $t$ with support $A$,

$$\mathbb{E}\left[\theta(Y) t \circ \Phi(Y)\right] = \langle t, s \circ \Phi^{-1} \rangle_{\Phi(A)}.$$

Oracle-type inequality and rates of convergence for the risk can also be obtained for each case. The main argument is the control of the supremum of the following empirical process (whose definition extends (3.22)):

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\theta(Y_i) t \circ \Phi(X_i) - \mathbb{E}\left[\theta(Y_i) t \circ \Phi(X_i)\right]\right),$$

for any $t \in L^2(\Phi(A))$. We do not detail it: the proof are similar to Theorem 4.1.

# Chapitre 5

# Méthodes de déformation pour l'estimation de la densité conditionnelle.

**Sommaire**

Une partie de ce chapitre est une version modifiée de l'article *Warped bases for conditional density estimation*, soumis pour publication, et actuellement en révision.

**Résumé.** L'objectif de ce chapitre est d'appliquer la méthode de déformation introduite dans les deux chapitres précédents pour l'estimation d'une fonction d'une seule variable à l'estimation d'une fonction de deux variables, la densité conditionnelle $\pi$ d'une variable réponse $Y$ sachant un prédicteur $X$, de loi continue. Partant toujours de données transformées de type $(F_X(X), Y)$, on propose d'abord un estimateur $\hat{g}$ de la fonction auxiliaire définie cette fois par $g(x, y) = \pi(F_X^{-1}(x), y)$, avant de définir $\hat{\pi}(x, y) = \hat{g}(\hat{F}(x), y)$ ($\hat{F}$ estimant $F_X$). La procédure principalement développée est la sélection de modèles (pour des estimateurs par projection), mise en œuvre par un critère inspiré de Goldenshluger & Lepski (2011a). Le compromis biais-variance est encore une fois réalisé de manière automatique, et la vitesse de convergence du risque est calculée sur des classes de fonctions de régularité anisotropique. Le choix théorique de la règle de sélection est discuté. Des simulations viennent illustrer la méthode. Enfin, on propose une version analogue de la procédure fondée sur des noyaux.

**Abstract.** The aim of this chapter is to apply the "warping" method studied in the previous chapters (for the estimation of functions of one real variable) to a bivariate setting: we consider the problem of estimating the conditional density $\pi$ of a response vector $Y$ given the predictor $X$ (which is assumed to be a continuous variable). Starting with the transformed data $(F_X(X), Y)$, we first recover $\hat{g}$ an adaptive estimate of the auxiliary $g(x, y) = \pi(F_X^{-1}(x), y)$. The resulting estimator is $\hat{\pi}(x, y) = \hat{g}(\hat{F}(x), y)$, where $\hat{F}$ is the empirical distribution function. We mainly tackle the problem with a model selection device for projection estimators, in the spirit of Goldenshluger & Lepski (2011a). The global squared-bias/variance compromise is realized, in a context of anisotropic function classes: non-asymptotic mean-squared integrated risk bounds are established, and risk convergence rates provided. The choice of the selection rule is discussed. Simulation experiments illustrate the method. Finally, we briefly present the analogous "warped"-kernel estimate.

## 5.1 Introduction

### 5.1.1 Motivation

Assume that we observe pairs of real random variables $(X, Y)$ with joint unknown density $f_{(X,Y)}$. The relationship between the predictor $X$ and the response $Y$ is classically described by regression analysis. But this can also be achieved by estimating the entire conditional density, that is

$$\pi(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}, \text{ if } f_X(x) > 0, \tag{5.1}$$

where $f_X$ is the marginal density of the variable $X$, and is assumed not to vanish on the interval of estimation.
The aim of this chapter is to provide a nonparametric strategy to estimate $\pi$, which has to be both adaptive, and simple to compute. Our main ideas are to use warped bases to build projection estimators and to perform model selection in the spirit of Goldenshluger & Lepski (2011a).

### 5.1.2 State of the art

Although nonparametric conditional density estimation has increasingly become a subject of interest since the early 1970s', adaptive estimators, which match the performances of an oracle that knows the regularity of the true function, are still rather scarse. To our knowledge, most of the methods to estimate $\pi$ are based on the principle that it can be seen as a nonparametric weighted regression. This leads mainly to two directions: kernel methods with well-chosen bandwidth(s), which have essentially been studied from an asymptotic point of view, and projection estimators built on regression-type criteria minimised on a well-chosen model.

The literature about the asymptotic properties of kernel estimators is vast. Several adjusted forms of the Nadaraya-Watson estimate have been proposed, for conditional distribution function (Stute, 1986b; Hall *et al.*, 1999) and for the conditional density: "double kernel" estimator (Hyndman *et al.*, 1996; Hyndman & Yao, 2002), generalization using local polynomials (Fan *et al.*, 1996), and reweighted kernel estimate (De Gooijer & Zerom, 2003). Accordingly, data-driven selection rules for the bandwidth are proposed, using methods inspired by Fan & Gijbels (1995), the bootstrap approach (Hall *et al.*, 1999; Bashtannyk & Hyndman, 2001), or cross-validation (Youndjé *et al.*, 1994; Fan & Yim, 2004). All these methods lead to asymptotic results: consistency rates, asymptotic normality, study of the asymptotic variance for example, under various assumptions (such as $\alpha-$ mixing design). A common feature of these estimators is their ratio form. This can be seen as a theoretical difficulty (see Penskaya 1995 for a specific study), which can be bypassed by using quantile regression or the copula function (Carroll *et al.*, 1994; Takeuchi *et al.*, 2009; Faugeras, 2009): the corresponding estimates still satisfy classical asymptotic properties. Another way of avoiding ratio is to consider a transformation of the input data. This strategy, early studied in Stute (1986a) will be detailed in the next section.

Moreover, projection estimators have been developed: the quality criterion which has thus become classical is the mean integrated squared error, or empirical versions of it. Nonasymptotic results, such as oracle inequalities or lower-bounds for the risk are set for estimators based on orthogonal series. For example, a Fourier basis can be used to build a blockwise-shrinkage Efromovich-Pinsker estimator, using characteristic functions to rewrite $\pi$: the regression setting is studied in Efromovich (2007), the general case is the subject of Efromovich (2010b), and multidimensionality is taken into account in Efromovich (2010a). His estimators match the performances of the oracle under the quadratic risk. The oracle-type inequalities stated permit to establish sharp minimaxity over the bivariate anisotropic Sobolev classes. The problem of dimension reduction is also studied in Fan *et al.* (2009), in the spirit of single index results. Model selection theory leads also to adaptation results: by minimizing a least-squares penalised contrast introduced by Lacour (2007), Brunel *et al.* (2007) build an estimator which adapts to an unknown underlying design and is minimax over anisotropic bivariate function classes. But the contrast, considered also by Akakpo & Lacour (2011) to deal with dependent data and inhomogeneous functional classes, does not provide explicit estimator without matrix invertibility requirements (except when using a histogram basis). Moreover the penalty given in Brunel *et al.* (2007) depends on the unknown infinite norm of $\pi$. It can be estimated but it then requires strong regularity assumptions. Notice also that recent works by Cohen & Le Pennec (2012) focus on a penalised maximum likelihood estimator leading to risk bounds for a Jensen-Kullblack-Leibler loss function. The maximisation of the likelihood seems to be difficult without additional assumptions on the shape of the conditional density.

The present work is mainly in the spirit of projection methods (except Appendix 2, Section 5.7. We aim at providing an adaptive estimator, which satisfies non-asymptotic risk bounds, but with a simpler expression, thus avoiding matrix inversion and purifying the penalty function. This goal is achieved by using both a "warping" of the data, like in the works of Stute (1986a,b) and Mehra *et al.* (2000) (no ratio, no matrix inversion), and by applying in a new and original way the Goldenshluger and Lepski method (the key to avoiding nuisance terms in the penalty).

### 5.1.3 Generality about the estimation method

The data are pairs of real random variables $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ (with $n$ a positive integer), independent and identically distributed (i.i.d.) with joint density $f_{(X,Y)}$, supported by a subset $A_1 \times A_2$ of $\mathbb{R}^2$ ($A_2$ a bounded interval). We assume that the marginal density $f_X$ of the $X_i$ does not vanish, and denote by $F_X$ the cumulative distribution function (c.d.f.) of these variables, which consequently admits an inverse.

The fundamental idea to provide a simple explicit estimator is that

$$g(u,y) = \pi\left(F_X^{-1}(u), y\right),\ (u,y) \in (0;1) \times A_2, \tag{5.2}$$

is the joint density of the random pair $(F_X(X_1), Y_1)$. We provide first an estimator of this function $g$, and then an estimator for the target function $\pi$, by using the reverse formula: $\pi(x,y) = g(F_X(x), y)$. This strategy has also been used by Stute (1986a) and Mehra *et al.*

(2000) to build kernel estimates of cumulative conditional distribution function and conditional density respectively, which are shown to be asymptotically normal. More recently, Kerkyacharian & Picard (2004) employed similar ideas to provide wavelet thresholding estimators of a regression function, and Chesneau (2007) used them in a modified Gaussian noise setting.

We adopt here a non-asymptotic point of view, by using model selection, and we aim at adaptive results. Precisely, the assumption that $g$ is squared integrable leads first to projection estimators of the form

$$\forall (u,y) \in (0;1) \times A_2, \;\; \hat{g}_{D_1,D_2}(u,y) = \sum_{j_1=1}^{D_1} \sum_{j_2=1}^{D_2} \hat{a}_{j_1,j_2} \phi_{j_1} \otimes \psi_{j_2}(u,y),$$

with $\phi_{j_1} \otimes \psi_{j_2}(u,y) = \phi_{j_1}(u)\psi_{j_2}(y)$, for different couples $(D_1, D_2)$ with $(\phi_{j_1} \otimes \psi_{j_2})_{j_1,j_2}$ an orthonormal family of functions and $\hat{a}_{j_1,j_2}$ estimated coefficients. Then, we propose an estimator of $\pi$ given by:

$$\forall (x,y) \in A_1 \times A_2, \;\; \hat{\pi}_{D_1,D_2}(x,y) = \hat{g}_{D_1,D_2}\left(\hat{F}_n(x), y\right),$$

with $\hat{F}_n$ an empirical counterpart for $F_X$. To avoid dependency in the proofs, we assume that there exists $(X_{-i})_{i\in\{1,\dots,n\}}$ a sample of variables with the same distribution as the $(X_i)$ and independent of them. Thus, we set

$$\hat{F}_n : x \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_{-i} \leq x}.$$

for the theoretical part. For the practical study, we assume to observe only the pairs $(X_i, Y_i)_i$ and compute successfully the estimator of $\pi$ by using an estimate of $F$ based on the original sample $X_i$.

To sum up, we get a development of $\hat{\pi}_{D_1,D_2}$ in an orthonormal basis, whose first coordinate is warped by $\hat{F}_n$:

$$\forall (x,y) \in A_1 \times A_2, \;\; \hat{\pi}_{D_1,D_2}(x,y) = \sum_{j_1=1}^{D_1} \sum_{j_2=1}^{D_2} \hat{a}_{j_1,j_2} \phi_{j_1} \otimes \psi_{j_2}\left(\hat{F}_n(x), y\right).$$

The procedure is particularly simple and fast to compute, since the coefficients $\hat{a}_{j_1,j_2}$ are just empirical means (they do not involve any matrix inversion). The selection rule of the levels $D_1$ and $D_2$ used in a second step is inspired by recent advances of Goldenshluger & Lepski (2011a) and is particularly well suited to the multidimensional framework.

Our main theorem is an adaptivity non-asymptotic result, an oracle-inequality and permits to deduce asymptotic rates of convergence for the quadratic risk, if the function $g$ belongs to anisotropic functional spaces. We show that adaptation has no price and that the rate corresponds exactly to the best bias-variance compromise, with assumptions stated on the function $g$ instead of $\pi$. Moreover, on the practical examples, the strategy we propose outperforms kernel methods summed up in Fan & Yim (2004) and the penalization device of Brunel *et al.* (2007).

### 5.1.4 Organisation of the chapter

Section 5.2 presents the two warped bases estimators (the one built assuming $F_X$ is known, and the one built in the general case). The performances of each estimator are studied in Section 5.3: the functional spaces are described and global risk bounds and rates of convergence presented. Section 5.4 is devoted to numerical results. The proofs are gathered in Section 5.5. The model selection rule is discussed in Appendix 1 (Section 5.6). Finally, an adaptation of the warping strategy to build a kernel-based estimator is proposed in Appendix 2 (Section 5.7).

## 5.2 Estimation strategy

All the estimators defined in the sequel are projection estimators. Therefore, we begin with the description of the approximation spaces (Section 5.2.1). We then proceed in three steps to estimate the conditional density $\pi$, on $A_1 \times A_2$. First, we define a collection of estimators for the function $g$ (see definition (5.2)), by minimizing a contrast on the models (Section 5.2.2). The second step consists in ensuring the automatic selection of the model, without any knowledge about the regularity of $g$. This leads to a well defined estimator $\hat{g}$ (Section 5.2.3). Finally, we partially warp $\hat{g}$ to estimate $\pi$.

### 5.2.1 Approximation spaces

Our estimation procedure is based on the assumption that the function $g$ belongs to $L^2((0;1) \times A_2)$, the set of square-integrable functions on $(0;1) \times A_2$, which is equipped with its usual Hilbert structure: we denote by $\langle .,. \rangle$ the scalar-product and by $\|.\|$ the norm. Consequently, $g$ can be developed in any orthonormal basis, and can be approximated by its orthogonal projections onto the linear subspaces spanned by the first functions of the basis. For the sake of simplicity, we assume $A_2 = (0;1)$ in the theoretical part. The case of any segment $A_2$ can be easily obtained by making a scaling change, see Section 5.4. Following the example of Efromovich (1999), we choose the Fourier basis $(\varphi_{j_1} \otimes \varphi_{j_2})_{j_1,j_2 \in \mathbb{N}\backslash\{0\}}$ of $L^2((0;1) \times A_2)$, defined for $u, y \in (0;1)$ by

$$\varphi_1(u) = 1, \ \ \forall k \in \mathbb{N}\backslash\{0\}, \ \varphi_{2k}(u) = \sqrt{2}\cos(2\pi k u), \ \varphi_{2k+1}(u) = \sqrt{2}\sin(2\pi k u), \tag{5.3}$$

and $\varphi_{j_1} \otimes \varphi_{j_2}(u,y) = \varphi_{j_1}(u)\varphi_{j_2}(y)$. For an index $l = 1, 2$, we also denote by $S_{m_l}$ the space spanned by $\{\varphi_1, \ldots, \varphi_{D_{m_l}}\}$, for $D_{m_l} = 2m_l + 1$, and $m_l$ an element of the set of indices $\mathcal{I}_n^{(l)} = \{1, \ldots, \lfloor \sqrt{n}/2 \rfloor - 1\}$ ($\lfloor . \rfloor$ is the integer part). The approximation spaces are then $\mathbb{S}_m = S_{m_1} \times S_{m_2}$ for $m = (m_1, m_2) \in \mathcal{M}_n$, with $\mathcal{M}_n = \mathcal{I}_n^{(1)} \times \mathcal{I}_n^{(2)}$. Thus, we have

$$\mathbb{S}_m = S_{m_1} \times S_{m_2} = \mathrm{Span}\left\{\varphi_{j_1} \otimes \varphi_{j_2}, \ j_1 = 1, \ldots, D_{m_1}, \ j_2 = 1, \ldots, D_{m_2}\right\},$$

and the dimension of $\mathbb{S}_m$ is $\mathbb{D}_m = D_{m_1} D_{m_2}$. Notice that for all $m_l \in \mathcal{I}_n^{(l)}$ ($l = 1, 2$), $D_{m_l} \leq \sqrt{n}$ and thus $\mathbb{D}_m \leq n$. This is similar to the dimension constraint required on the models used to estimate a one-variable function in Chapter 3 (see Section 3.2.1).

**Remark 5.1.** – The basis satisfies $\| \sum_{j_1=1}^{D_{m_1}} \sum_{j_2=1}^{D_{m_2}} (\varphi_{j_1} \otimes \varphi_{j_2})^2 \|_{L^\infty((0;1)\times A_2)} \leq \mathbb{D}_m$, where $\|.\|_{L^\infty((0;1)\times A_2)}$ is the supremum of the function on $(0;1)\times A_2$. This is equivalent to the following useful link between the infinite norm and the $L^2$ norm (see Birgé & Massart 1998 for the proof):

$$\forall t \in L^2((0;1)\times A_2),\ \|t\|_{L^\infty((0;1)\times A_2)} \leq \sqrt{D_{m_1} D_{m_2}} \|t\| = \sqrt{\mathbb{D}_m} \|t\|. \tag{5.4}$$

– For each $m_l, m_l' \in \mathcal{I}_n^{(l)}$ $(l = 1, 2)$, we have

$$D_{m_l} \leq D_{m_l'} \implies S_{m_l} \subset S_{m_l'}. \tag{5.5}$$

Notice that other classical models, such as models spanned by regular wavelet basis, histogram basis or dyadic piecewise polynomial basis satisfy similar properties. We refer to Barron *et al.* (1999), and Brunel & Comte (2005) for a precise description. See also Remark 5.2 below about the extension of our results to these models.

### 5.2.2 Estimation on a fixed model

We start with the following criterion

$$\forall t \in L^2((0;1)\times A_2) \mapsto \gamma_n(t, \hat{F}_n) := \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t\left(\hat{F}_n(X_i), Y_i\right).$$

This contrast is new and quite far from the regression and conditional density least-squares criterion. The novelty comes both from the $L^2$ norm which stands in place of the empirical norm used in the classical contrasts (see for example the contrast $\gamma_n^0$ in Brunel *et al.* 2007), and from the presence of the empirical c.d.f $\hat{F}_n$. To justify this choice, plug for a moment the true c.d.f $F_X$ instead of its empirical counterpart, and compute, for $t \in L^2((0;1)\times A_2)$,

$$\begin{aligned}
\mathbb{E}\left[\gamma_n(t, F_X)\right] - \mathbb{E}\left[\gamma_n(g, F_X)\right] &= \|t\|^2 - \|g\|^2 - 2\mathbb{E}\left[(t-g)\left(F_X(X_1), Y_1\right)\right], \\
&= \|t\|^2 - \|g\|^2 - 2\int_{A_1\times A_2} (t-g)(F_X(x), y)\pi(x,y) f_X(x) dxdy, \\
&= \|t\|^2 - \|g\|^2 - 2\int_{(0;1)\times A_2} (t-g)(u,y) g(u,y) dudy, \\
&= \|t\|^2 - \|g\|^2 - 2\langle g, t-g\rangle, \\
&= \|t-g\|^2.
\end{aligned}$$

This quantity is minimal when $t = g$. This shows that $\gamma_n(., F_X)$ (and $\gamma_n(., \hat{F}_n)$ in practice) suits well for the estimation of $g$. We thus set, for each model $\mathbb{S}_m$,

$$\hat{g}_m^{\hat{F}} = \arg\min_{t\in\mathbb{S}_m} \gamma_n(t, \hat{F}_n),\ \hat{g}_m^{F_X} = \arg\min_{t\in\mathbb{S}_m} \gamma_n(t, F_X),$$

or equivalently,

$$\hat{g}_m^{\hat{F}} = \sum_{j_1=1}^{D_{m_1}} \sum_{j_2=1}^{D_{m_2}} \hat{a}_{j_1,j_2}^{\hat{F}} \varphi_{j_1} \otimes \varphi_{j_2},\ \text{ with }\ \hat{a}_{j_1,j_2}^{\hat{F}} = \frac{1}{n} \sum_{i=1}^{n} \varphi_{j_1}(\hat{F}_n(X_i)) \varphi_{j_2}(Y_i),$$

and a similar expression for estimator $\hat{g}_m^{F_X}$ with coefficients $\hat{a}_{j_1,j_2}^{F_X}$. Finally, we set

$$\pi_m^{\hat{F},\hat{F}}(x,y) = \hat{g}_m^{\hat{F}}(\hat{F}_n(x),y) \quad \text{and} \quad \hat{\pi}_m^{F_X,F_X}(x,y) = \hat{g}_m^{F_X}(F_X(x),y),$$

denoted with two super-indexes $\hat{F}$ (or $F_X$) to underline the double dependence of the estimator on this function, through both the coefficients $\hat{a}_{j,k}^{\hat{F}}$ and the composition of the first variable by $F_X$. Notice the advantage of the contrast we define: we get an explicit formula for the estimator. The coefficients are easily computable empirical means. They do not involve a matricial inversion compared to the estimator obtained via the least-squares criterion (see for example Brunel *et al.* 2007). Moreover, in the toy case of known c.d.f. $F_X$, $\hat{g}_m^{F_X}$ is an unbiased estimator of the orthogonal projection of $g$ onto $\mathbb{S}_m$.

### 5.2.3 Model selection

#### Risk on a fixed model

In order to explain which model $\mathbb{S}_m$ we should choose, we first study the quadratic risk of each estimator of the collection $(\pi_m^{F_X,F_X})_{m\in\mathcal{M}_n}$. The loss function naturally associated to our context is the following $L^2-$norm,

$$\forall v \in L^2(A_1\times A_2, f_X),\ \|v\|_{f_X}^2 = \int_{A_1\times A_2} v^2(x,y) f_X(x) dxdy,$$

with $L^2(A_1\times A_2, f_X)$, the space of squared-integrable functions on $A_1\times A_2$ with respect to the Lebesgue measure weighted by the density $f_X$. We denote $\langle .,. \rangle_{f_X}$ the corresponding scalar-product. Notice besides that the following links hold between this norm and the classical norm previously defined: for $t, s \in L^2((0;1)\times A_2)$, we compute, using $F_X' = f_X$,

$$\|t(F_X(.),.)\|_{f_X} = \|t\|,\ \ \langle t(F_X(.),.), s(F_X(.),.)\rangle_{f_X} = \langle t, s\rangle.$$

If we want to bound the classical quadratic risk of the estimator, we can assume that $f_X$ is bounded from below by a strictly positive constant $f_0$: this assumption, which is standard (see for example Assumption $\mathcal{A}_2$ in Brunel *et al.* 2007, or Assumption $(A2)$ in Durot 2008) leads to the inequality $\|v\|_{f_X} \geq f_0\|v\|$, for all $v \in L^2(A_1\times A_2, f_X)$. Thus, it is sufficient to bound the weighted risk.
We can give a simple explanation for the choice of an estimator among the collection, considering first the collection of theoretical estimators $(\hat{\pi}_m^{F_X,F_X})_m$. For the weighted $L^2-$risk which is used in the rest of the chapter, and for each $m \in \mathcal{M}_n$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\pi}_m^{F,F} - \pi\right\|_f^2\right] &= \mathbb{E}\left[\left\|\hat{h}_m^F - h\right\|^2\right] = \|h - h_m\|^2 + \mathbb{E}\left[\left\|h_m - \hat{h}_m^F\right\|^2\right], \\
&= \left\|\pi - \pi_m^F\right\|_f^2 + \mathbb{E}\left[\left\|\pi_m^F - \hat{\pi}_m^{F,F}\right\|_f^2\right], \qquad (5.6)
\end{aligned}$$

where

$$\pi_m^{F_X}(x,y) = g_m(F_X(x),y) \text{ and } g_m \text{ is the orthogonal projection of } g \text{ onto } \mathbb{S}_m.$$

We recover the usual squared-bias/variance decomposition of the risk. The key point is the difference of behaviour of the two terms: they both depend on $\mathbb{D}_m$ but in opposite ways. The first term in the right-hand side of (5.6) decreases when $\mathbb{D}_m$ grows, since $\pi$ is better approximated by its projection when the approximation space grows, while the second term grows with $\mathbb{D}_m$:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\pi}_m^{F_X,F_X}-\pi_m^{F_X}\right\|_{f_X}^2\right] &= \sum_{j_1=1}^{D_{m_1}}\sum_{j_2=1}^{D_{m_2}}\mathrm{Var}\left(\hat{a}_{j_1,j_2}^{F_X}\right),\\ &\leq \frac{1}{n}\sum_{j_1=1}^{D_{m_1}}\sum_{j_2=1}^{D_{m_2}}\mathbb{E}\left[\left(\varphi_{j_1}\left(F_X(X_i)\right)\varphi_{j_2}(Y_i)\right)^2\right]\leq\frac{D_{m_1}D_{m_2}}{n}, \quad (5.7)\end{aligned}$$

using Property (5.4). The best model among the collection is the one which minimises the right-hand side in (5.6), making a trade-off between the squared-bias term and the variance term. However, it is unknown since $g$ and $g_m$ are not observed. Therefore, an adaptive estimator of $\pi$ must make this compromise automatically.

**Selection rule**

The method is the same as in Chapter 3 to define the estimators $\tilde{s}_2^{F_X}$ and $\tilde{s}_2^{\hat{F}}$: see Section 3.2.3. We propose to adapt the scheme proposed by Goldenshluger & Lepski (2011a) for kernel density estimation. The adaptive index is chosen as follows:

$$\hat{m}^{\hat{F}}=\left(\hat{m}_1^{\hat{F}},\hat{m}_2^{\hat{F}}\right)=\arg\min_{m\in\mathcal{M}_n}\left[A(m,\hat{F}_n)+2V^{\hat{F}}(m)\right],$$

where $V^{\hat{F}}$ has the order of the variance term:

$$V^{\hat{F}}: m=(m_1,m_2)\ \mapsto\ c_1\frac{D_{m_1}D_{m_2}}{n}, \quad (5.8)$$

with $c_1>0$ a purely numerical constant. Its theoretical calibration is precised in the proofs (Section 5.5.3), and, in practice, we adjust it by simulations (see details in Section 5.4). The function $A(.,\hat{F}_n)$ is based on the comparison of the estimators built in the first stage:

$$A(m,\hat{F}_n)=\max_{m'\in\mathcal{M}_n}\left(\left\|\hat{g}_{m'}^{\hat{F}}-\hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2-V^{\hat{F}}(m')\right)_+, \quad (5.9)$$

where $x_+=\max(x,0)$, $x\in\mathbb{R}$ and $m\wedge m'=(m_1\wedge m_1',m_2\wedge m_2')$. We prove that $A(m,\hat{F}_n)$ has the order of the bias term (see Inequality (5.18)). Thus we get an estimator, explicitly expressed in a warped basis,

$$\tilde{\pi}(x,y)=\hat{g}_{m^{\hat{F}}}^{\hat{F}}(\hat{F}_n(x),y). \quad (5.10)$$

The $L^2-$norm involved in the definition of $A(.,\hat{m})$ is easy to compute, since the functions $\hat{g}_{m'}^{\hat{F}}$, $m'\in\mathcal{M}_n$ are expressed with a development in an orthonormal basis (see Section 5.4 for details). This advantage must be highlighted compared to other strategies of model selection using the contrast function or to strategies involving bandwith choice for a kernel.
There are several novelties to underline. First, the warping of the basis for the variable $x$

leads to explicit and simple coefficients $\hat{a}_{j_1,j_2}^{\hat{F}}$ for the estimator. The use of a selection device inspired from Goldenshluger & Lepski (2011a) is original in the setting of multidimensional model selection. This choice is discussed in Appendix 2 (Section 5.6). Note also that the specific factor 2 in the definition of $\hat{m}^{\hat{F}}$ plays an important (but technical) role in the proofs. Once the constant $c_1$ is chosen (through simulation experiments), the "penalty" term $V^{\hat{F}}$ is entirely data driven, while the penalty in the regression-type contrast depends, in addition, on the unknown infinite norm of $\pi$ (see Brunel *et al.* 2007, or Lacour 2007).

Finally, let us define an estimator in the theoretical case of known c.d.f. $F_X$:

$$\tilde{\pi}_0(x,y) = \hat{g}_{\hat{m}^{F_X}}^{F_X}(F_X(x),y), \tag{5.11}$$

with $\hat{m}^{F_X}$ selected as the argument-minimum of $A(m,F_X)+V^{F_X}(m)$, where we denote by $V^{F_X}(m) = c_0 D_{m_1} D_{m_2}/n$, $c_0$ a numerical constant, which can be different of $c_1$.

## 5.3 Main results

### 5.3.1 Anisotropic Sobolev spaces

Let us define the functional spaces we consider further for the function $g$ (even if its index of regularity needs not be known). The choice of the trigonometric models leads us to consider spaces of periodic functions, that is Sobolev spaces. We define them directly via Fourier coefficients, keeping in mind that they can also be characterised via weak differentiability (see for example DeVore & Lorentz 1993 and Härdle *et al.* 1998 for functions of one variable, and Adams 1975 for functions of several variables). Precisely, we extend the characterization of Tsybakov (Lemma A.3, p.162, Tsybakov 2009) to functions of two variables. See also Section 2.2.1 of Chapter 2.

Let $t \in L^2((0;1)^2)$. Then there exists a real-valued family $(\theta_{j_1,j_2})_{j_1,j_2\in\mathbb{N}\setminus\{0\}}$ such that, in $L^2$, the following equality holds:

$$t = \sum_{j_1,j_2\in\mathbb{N}\setminus\{0\}} \theta_{j_1,j_2}\varphi_{j_1}\otimes\varphi_{j_2}.$$

Recall that the functions $\varphi_j$ are defined by (5.3). We say that $t$ belongs to the partial ball with radius $L>0$ and regularity $\alpha=(\alpha_1,\alpha_2)$ ($\alpha_l\in\mathbb{N}$, $l=1,2$, $\alpha\neq(0,0)$), if

$$\sum_{j_1,j_2\in\mathbb{N}\setminus\{0\}} \mu_{j_1,\alpha_1}^2\mu_{j_2,\alpha_2}^2\theta_{j_1,j_2}^2 \leq \frac{L^2}{\pi^{2(\alpha_1+\alpha_2)}}, \tag{5.12}$$

with $\mu_{j_l,\alpha_l} = j_l^{\alpha_l}$ for even $j_l$, $\mu_{j_l,\alpha_l} = (j_l-1)^{\alpha_l}$ otherwise. We write $t\in W_2^{\alpha,per}((0;1)^2,L)$, in the spirit of the definition of Tsybakov (2009). These spaces are anisotropic. The function $g$ can thus have different smoothness properties with respect to different directions.

Let us finally give a useful approximation property of this space. We denote by $t_m = t_{(m_1,m_2)}$ the orthogonal projection of the function $t$ onto the subspace $\mathbb{S}_m = \mathbb{S}_{(m_1,m_2)}$. We have the following rate:

$$\|t-t_m\|^2 \leq C(\alpha,L)\left(D_{m_1}^{-2\alpha_1}+D_{m_2}^{-2\alpha_2}\right), \tag{5.13}$$

where $C(\alpha,L)$ is a constant depending on $\alpha$ and $L$. This inequality is proved in Section 2.2.2 of Chapter 2 (see Inequality (2.15)).

### 5.3.2 Nonasymptotic risk bound and consequence

The first theorem provides a non-asymptotic bound for the risk of the estimator $\tilde{\pi}$ (see its definition (5.10)). We recall that the trigonometric models satisfy properties (5.4) and (5.5), and that the dimensions $D_{m_l}$ are bounded by $\sqrt{n}$. Hereafter we denote by $\|.\|_{L^\infty((0;1))}$ the infinite norm of a function over the interval $(0;1)$.

**Theorem 5.1.** *We assume that the function $g$ belongs to the anisotropic Sobolev ball denoted by $W_2^{(1,0),per}((0;1)^2, L)$, for some fixed $L > 0$, is bounded on $(0;1)^2$, and is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$. We also assume that, for some constants $C_a, C_b, C_c$, the trigonometric models satisfy*

$$\forall m = (m_1, m_2) \in \mathcal{M}_n,\ D_{m_1} \leq C_a \left(\frac{n}{\ln^2(n)}\right)^{1/3} \quad \text{and} \quad C_b \ln^5(n) \leq D_{m_2} \leq C_c\sqrt{n}. \tag{5.14}$$

*Then, there exists numerical constants $c$ and $C$ depending on $\|g\|$, $\|\partial_1 g\|$, and $L$, such that*

$$\mathbb{E}\left[\|\tilde{\pi} - \pi\|_{f_X}^2\right] \leq c \min_{m \in \mathcal{M}_n} \left\{ \frac{D_{m_1} D_{m_2}}{n} + \left\|\pi_m^{F_X} - \pi\right\|_{f_X}^2 \right\} + \frac{C}{n}. \tag{5.15}$$

**Remark 5.2.**

- There actually exists an integer $n_0$, depending on the function $g$, such that Inequality (5.15) holds for all $n \geq n_0$ with a purely numerical constant $c$. But the result is non-asymptotic, since the inequality also holds for $n < n_0$, taking a constant $c$ which depends on quantities of the problem.
- Up to this result, the models $S_{m_1}$ and $S_{m_2}$ and their respective dimensions have played the same role. But in the theorem, the dimension constraints (5.14) are not the same in each direction. To be totally rigorous, we should denote by $S_{m_l}^{(l)}$ the models and by $D_{m_l}^{(l)}$ their dimension, for each $l = 1, 2$. For the sake of simplicity, we keep the first notations as there is no possible confusion.
- The assumptions on the model dimensions $D_{m_1}$ and $D_{m_2}$ are not a consequence of the adaptation: they only permit to deal with the deviation between the empirical c.d.f. $\hat{F}_n$ and $F_X$. Let us compare them to the assumption of Theorem 3.2 (which is the analogous of Theorem 5.1, for the warped-basis estimation of a regression function). Up to the logarithm, the constraint on $D_{m_1}$ here is the same as the one on the one-dimensional model $S_m$ in Chapter 3: they are both supposed to be less than $n^{1/3}$. This might be improved by additional technicalities in the proof.
- The function $g$ is supposed to be continuously derivable with respect to its first variable only: this assumption allows us to handle the "warping" (see the proof of Lemma 5.5) which only affects the first variable of the target function (see (5.2) for example.
- If we focus on the simpler situation of known c.d.f., we can derive the same result as Inequality (5.15) for the estimator $\tilde{\pi}_0$ (defined by (5.11)), with few assumptions: we only assume that the function $g$ is bounded on the space $(0;1) \times A_2$, and we have no additional condition on the trigonometric model. In this case, the constant $c$ in (5.15) is purely numerical, and the other constant $C$ depends only on $\|g\|_{L^\infty((0;1)\times A_2)}$. This result holds in a more general setting than trigonometric models. It is sufficient to assume that the models satisfy properties (5.4) and (5.5), and have dimensions bounded

by $\sqrt{n}$. These assumptions are very weak. Since the proof of this result is both short and representative of the method used to prove Theorem 5.1, we begin with it in the proof section (see Section 5.5.2).

The basic outline of model selection (with a method inspired by Goldenshluger & Lepski 2011a) is to estimate the bias-variance sum and to select the model which minimises it. Theorem 5.1 shows that it is a good strategy: the right model (in the sense that it realises the trade-off) has been chosen in a data-driven way and the selected estimator performs as well as the best estimator in the family $\{\pi_m^{\hat{F},\hat{F}},\ m \in \mathcal{M}_n\}$, up to some multiplicative constants and to a negligible residual term of order $1/n$. The constants are given in the proof, which is deferred to Section 5.5.2.

Brunel *et al.* (2007) provide the same kind of oracle inequality for their estimator built by penalization of a regression-type contrast. The assumptions seem first to be slightly less restrictive: it is only assumed that $D_{m_1} \leq n^{1/2}/\ln(n)$. However, the term $V^{\hat{F}}$ does not contain any unknown term and is then entirely computable, contrary to the penalty used in Brunel *et al.* (2007), which depends on $\|\pi\|_{L^\infty(A_1\times A_2)}$. Moreover, replacing this quantity by an estimator requires much more regularity constraints than the one we have, and leads to a semi-asymptotic result (see the appendix of Lacour 2007 for an example of these conditions). Consequently, a model selection strategy in the spirit of Goldenshluger-Lepski applied with warped bases has the advantage of providing an estimator easier to compute than a regression-type estimator and with good theoretical properties under quite weak assumptions.

Theorem 5.1 also enables us to give a rate of convergence for the estimation of $\pi$, under regularity assumptions for function $g$. Recall that the bound of Inequality (5.15) is close to the order of the sum of the variance term and the bias term. The minimization of the left-hand-side of the inequality in the case of regular functions leads to the following corollary, which implies that the obtained rate of convergence is likely to be minimax in most cases (see Section 5.5.4 for the details of the computations).

**Corollary 5.1.** *We assume that the function $g$ belongs to the anisotropic Sobolev ball denoted by $W_2^{\alpha,per}((0;1)^2,L)$, for some fixed $L>0$, and $\alpha=(\alpha_1,\alpha_2)$ ($\alpha_l \in \mathbb{N}$, $l=1,2$, $\alpha \neq (0,0)$) with $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$, and $\alpha_2 - \alpha_1 + 2\alpha_1\alpha_2 > 0$. Then, under the assumptions of Theorem 5.1,*

$$\mathbb{E}\left[\|\tilde{\pi}-\pi\|_{f_X}^2\right] \leq C(\alpha,L)n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}},$$

*with $C(\alpha,L)$ a numerical constant which depends on $\alpha$ and $L$, $\|g\|$, $\|\partial_1 g\|$. The quantity $\bar{\alpha}$ is the harmonic mean of $\alpha_1$ and $\alpha_2$.*

The harmonic mean of $\alpha_1$ and $\alpha_2$ is the real $\bar{\alpha}$ such that $2/\bar{\alpha} = 1/\alpha_1 + 1/\alpha_2$. The corollary signifies that without knowing $\alpha$ and $L$ (depending on the unknown $g$), $\tilde{\pi}$ does as well as the best possible estimator which knows these quantities. It is thus an adaptive estimator.

## 5.4 Simulation study

The aim of this section is to illustrate the behaviour of the estimator $\tilde{\pi}$ and to compare it with the regression-type estimator of Brunel *et al.* (2007) and with kernel estimators (Fan & Yim, 2004).

### 5.4.1 Implementation

The estimate $\tilde{\pi}$ is computed by using simulated sample of data $(X_i, Y_i)_{i=1,\dots,n}$. The empirical c.d.f. function $\hat{F}_n$ is the one of the sample $(X_i)_i$: in practice, we do not use additional observations. We actually experimented it, and a second sample does not improve the results.

For each data sample (that is for each computation of the estimators), we calibrate the set $A_1 \times A_2$ for the estimation, over 90% of the variables $(X_i, Y_i)$: we choose to eliminate the smallest values, and the largest values of the data to avoid edge effects.

To implement each estimator $\tilde{\pi}$, we use warped trigonometric basis. Recall that our procedure is based first on the estimation of the function $g$, which belongs to $L^2((0;1) \times A_2)$. Accordingly, the orthogonal Fourier basis is $(\varphi_{j_1} \otimes \psi_{j_2})_{j_1,j_2}$, with $\varphi_{j_1}$ defined by (5.3), and $\psi_{j_2}$ obtained by:

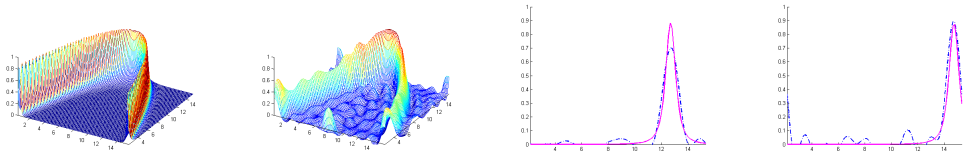
$$\forall y \in (a_2; b_2),\ \psi_{j_2}(y) = \frac{1}{\sqrt{b_2 - a_2}} \varphi_{j_2}\left(\frac{y - a_2}{b_2 - a_2}\right),$$

where $A_2 = (a_2; b_2)$.

We have to compute the sum $A(m, \hat{F}_n) + 2V^{\hat{F}}(m)$ for each $m = (m_1, m_2)$. Notice that the quadratic norm in the definition of $A(m, \hat{F}_n)$ (see (5.9)) is simply equal to a sum of squared-coefficients. For example, if $m \wedge m' = (m_1, m_2')$,

$$\left\| \hat{g}^{\hat{F}}_{m'} - \hat{g}^{\hat{F}}_{m \wedge m'} \right\|^2 = \sum_{j=D_{m_1}+1}^{D_{m_1'}} \sum_{k=1}^{D_{m_2'}} \left(\hat{a}^{\hat{F}}_{j,k}\right)^2.$$

Finally, we calibrate the numerical constant $c_1$ involved in the definition of $V^{\hat{F}}$ (see (5.8)). As in most model selection results, the theoretical value obtained in the proof (Section 5.5.3) is very pessimistic due to rough upper-bounds (for the sake of simplicity of the proof). Thus, specific data-driven calibration has been developed: for example, the so-called "dimension jump" method allows us to apply the slope heuristic (see Baudry *et al.* 2012) to choose the penalty constant of a classical penalised contrast strategy (Barron *et al.*, 1999). But this cannot be applied to the recent method of Goldenshluger and Lepski. Consequently, we adjust $c_1$ prior to the comparison with the other estimates: we look at the quadratic risk with respect to the value of $c_1$, and choose one of the values leading to reasonable risk and complexity of the selected models (recall that in penalty calibration, it is more secure to overpenalise): we set thus $c_1 = 0.05$.



(a) (b) (c) (d)

Figure 5.1: Plots of true function versus estimator $\tilde{\pi}$, Example 1' (ii) with $n = 1000$ observations. (a) true function $\pi$, (b) estimator $\tilde{\pi}$, (c) and (d) plots of $y \mapsto \pi(x, y)$ (full line), $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed $x$.

### 5.4.2 Comparison with kernel estimates

Since kernel methods play a key role in nonparametric conditional density estimation, we begin by comparing the adaptive warped-bases estimate $\tilde{\pi}$ with double-kernel local linear regressions built with a data-driven selection of the bandwidths. Fan & Yim (2004) recall the definition of this estimate. They numerically compare the bandwidths selection rule they propose (a cross-validation strategy based on minimization of the integrated squared error, with three other selection strategies: an ad-hoc empirical bias method (Fan *et al.*, 1996), a bootstrap method (developed first by Hall *et al.* 1999, adapted by Bashtannyk & Hyndman 2001), and a simple rule, based on local parametric models (Hyndman & Yao, 2002). They show that in most cases, the cross-validation method outperforms the other approaches substantially. Thus, we evaluate $\tilde{\pi}$ by comparison with the cross-validation double-kernel estimate, denoted by $\tilde{\pi}_{FY}$.

We consider the examples corresponding to the independent random sample setting in Fan & Yim (2004):

- Example 1': $Y_i = 0.23X_i(16 - X_i) + 0.4\varepsilon_i$, with the $X_i$ uniformly distributed on the interval $(0; 16)$ and the $\varepsilon_i$ are independent, (i) standard normal, (ii) from the Student distribution $t_2$, (iii) from the Student distribution $t_4$ .
- Example 2': $Y_i = 20\cos(\pi X_i/10) + \varepsilon_i$, with the $X_i$ uniformly distributed on the interval $(-20; 20)$ and the $\varepsilon_i$ are standard normal, independent.

For sample size $n = 1000$, we compute the root mean squared error (RMSE) with the same formula as in Section 4 of Fan & Yim (2004), for the estimator $\tilde{\pi}$. Figure Figure 5.1 presents an example of estimate for Example 1' and Table 5.1 summarises the results: we report the risk obtained here for $\tilde{\pi}$ and the risk of $\tilde{\pi}_{FY}$ obtained by Fan & Yim (2004) (see their Tables 1 and 2). We do not mention the median of the RMSE since it is not significantly different from the mean for the estimator $\tilde{\pi}$.

The values are similar for both estimators in most cases, but slightly better for the warped-bases estimator (in bold in Table 5.1). This result has to be put into perspective, since more thorough numerical trials have to be conducted to confirm this. The aim was just here to show that a warped-bases adaptive strategy can compete with kernel methods.

| | Ex 1' (i) | Ex1' (ii) | Ex1' (iii) | Ex2' |
|---|---|---|---|---|
| $\tilde{\pi}$ | **0.7872** | **0.6427** | **0.6911** | **1.0408** |
| $\tilde{\pi}_{FY}$ | 1.0899 | 0.7641 | 1.0143 | 2.7404 |

Table 5.1: Values of RMSE $\times 1000$ averaged over 100 samples of size $n = 1000$ according to Fan & Yim (2004), in Examples 1' and 2' for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{FY}$.

### 5.4.3 Comparison with regression-type estimator

We also compare $\tilde{\pi}$ with another adaptive estimator: the one of Brunel *et al.* (2007) denoted by $\tilde{\pi}_{BCL}$. The estimator $\tilde{\pi}_{BCL}$ of Brunel *et al.* (2007) is defined as a penalised least-squares contrast estimator. The penalty is $\text{pen}(m) = K_0 \|\pi\|_{L^\infty(A_1 \times A_2)} D_{m_1} D_{m_2}/n$. We implement the method with $K_0 = 0.5$ like in Brunel *et al.* (2007) but we do not replace $\|\pi\|_{L^\infty(A_1 \times A_2)}$ by its theoretical value. To have a real data-driven procedure, we estimate it by taking the supremum of the values of a least-squares estimator on a fixed model $\mathbb{S}_m$ on a rough grid, with $m = [(\ln(n) - 1)/2]$.

The aim is to investigate mainly the difference between the classical bases and the warped bases.

We propose to base the simulation study on the following examples: we generate samples $(X_i, Y_i)_{i \in \{1,\dots,n\}}$ such that

- Example 1: $Y_i = b(X_i) + \varepsilon_i$, with the following possibilities. The $X_i$ follow a uniform distribution on the interval $(0;1)$ (denoted by $\mathcal{U}_{(0;1)}$), or on the interval $(-1;1)$ ($\mathcal{U}_{(-1;1)}$), a standard Gaussian distribution ($\mathcal{N}(0,1)$). The $\varepsilon_i$'s are generated following the standard Gaussian distribution, the Gaussian distribution with variance 4 ($\mathcal{N}(0,4)$), or the Gamma distribution ($\Gamma(4,1)$) with parameters 4 and 1 (the 1 is the scale parameter). We denote by $f_\varepsilon$ their density. The sample $(\varepsilon_i)$ is independent of the $(X_i)$. Finally, the regression function $b$ is $b(x) = 2x + 5$ or $b(x) = x^2$. The conditional density $\pi$ is thus given by
$$\pi(x, y) = f_\varepsilon(y - b(x)).$$
- Example 2: $Y_i = b(X_i) + \sigma(X_i)\varepsilon_i$, with a uniform distribution on $(0;1)$ for $X_i$, the previous Gamma distribution for $\varepsilon_i$ (which is independent of $X_i$) and $\sigma(x) = \sqrt{1.3 - |x|}$. Similarly to Example 1, the conditional density is
$$\pi(x, y) = f_\varepsilon(y - b(x)/\sigma(x))/\sigma(x).$$
- Example 3: $X_i$ follows a uniform distribution $\mathcal{U}_{(0;1)}$, and given $X_i = x$, $Y_i$ follows the Gaussian mixture $0.5\mathcal{N}(8 - 4x, 1) + 0.5\mathcal{N}(8 + 4x, 1)$. The function $\pi$ is the density of the mixture.

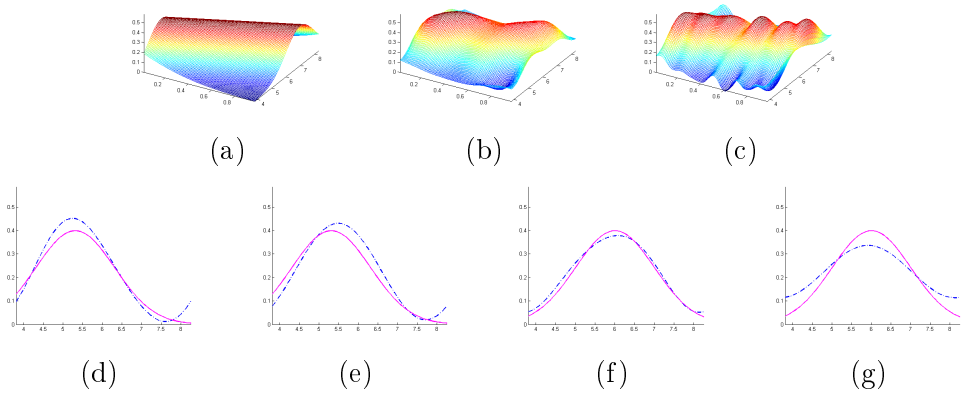


Figure 5.2: Plots of true function versus estimators, Example 1, with $X_i$ i.i.d. $\mathcal{U}_{(0;1)}$, $\varepsilon_i$ i.i.d. $\mathcal{N}(0,1)$, and $b(x) = 2x + 5$ with $n = 2000$ observations. First line: (a) true function $\pi$, (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line, (d) and (f): plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed $x$. Second line, (e) and (g): for two fixed $x$, plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}_{BCL}(x, y)$ (dashed dotted line) .

– Example 4: $X_i$ follows a uniform distribution $\mathcal{U}_{(0;1)}$, and given $X_i = x$, $Y_i$ follows a Gamma distribution with parameters 3 and $x^2+1$. The function $\pi$ is the corresponding Gamma density.

Examples 2 and 3, and some cases of Example 1 have also been studied by Brunel *et al.* (2007), and Example 5 by Fan & Yim (2004) (but no risks for this model are given for the last two examples).

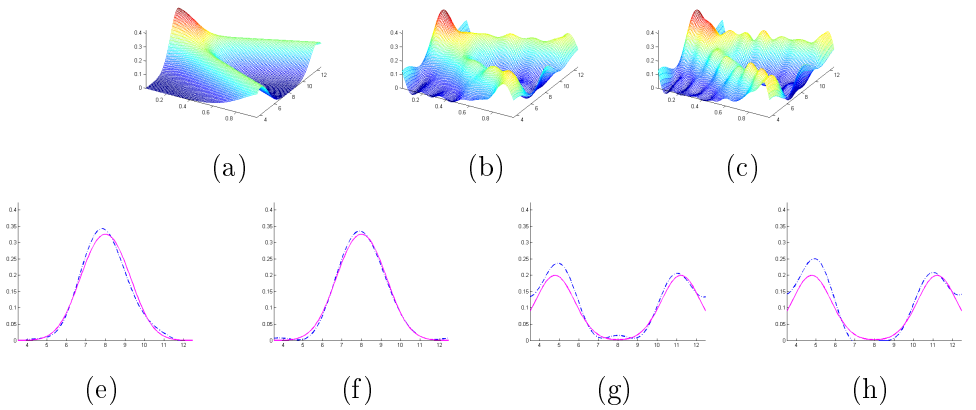
Figures Figure 5.2 and Figure 5.3 illustrate the visual quality of the reconstruction, for a case of Example 1, and for Example 3. We do not observe significant differences between the two estimators, which both behave quite well. However, the computation of $\tilde{\pi}_{BCL}$ requires much more time than the one of $\tilde{\pi}$, probably because of the presence of a matricial inversion, consequence of the least-squares contrast. The warped-bases estimator can thus advantageously be used for estimation problems with large data samples (data deriving from domain such as physics, fluorescence, finance...).

For sample sizes $n = 60, 500$ and 2000, we give in Tables 5.2 and 5.3 the estimated values of the risk $\mathbb{E}[\|\hat{\pi} - \pi\|_2^2]$, with $\|.\|_2$ the quadratic norm on $L^2(A_1 \times A_2)$, and $\hat{\pi} = (\tilde{\pi}_{BCL})_+$ or $(\tilde{\pi})_+$. It is not difficult to see that the choice of the positive part of both estimators can only make their risks decrease. The MISE is computed over $N = 100$ replicated samples, and the quadratic norm is approximated using subdivisions of $A_1$ and $A_2$ (see Brunel *et al.* 2007, Section 5.1, for details about the formula).

The risk of our estimator $\tilde{\pi}$ is often better than the one of the penalised least-squares estimator $\tilde{\pi}_{BCL}$ (in bold in the tables). Precisely, it is always smaller for the sample sizes $n = 500$ and $n = 2000$, which confirms that one can easily use the warped-bases estimator for estimation problems with large samples of data, in spite of its bad performances for very small sample sizes.

| $b(x)$ | $\varepsilon$ | $X$ | $n = 60$ | 500 | 2000 | Method |
|---|---|---|---|---|---|---|
| $2x+5$ | $\mathcal{N}(0,1)$ | $\mathcal{U}_{(0;1)}$ | 19.81 | **2.879** | **1.327** | $\tilde{\pi}$ |
| | | | 7.446 | 2.536 | 1.409 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{U}_{(-1;1)}$ | 20.871 | **4.811** | **2.971** | $\tilde{\pi}$ |
| | | | 9.443 | 6.384 | 5.501 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{N}(0,1)$ | 38.255 | **14.833** | **11.038** | $\tilde{\pi}$ |
| | | | 37.374 | 40.295 | 38.993 | $\tilde{\pi}_{BCL}$ |
| | $\Gamma(4,1)$ | $\mathcal{U}_{(0;1)}$ | 5.628 | **0.969** | **0.479** | $\tilde{\pi}$ |
| | | | 2.361 | 1.417 | 0.715 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{U}_{(-1;1)}$ | 9.224 | **2.255** | **1.376** | $\tilde{\pi}$ |
| | | | 6.666 | 3.569 | 2.314 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{N}(0,1)$ | 20.094 | **8.641** | **6.093** | $\tilde{\pi}$ |
| | | | 24.749 | 20.201 | 22.571 | $\tilde{\pi}_{BCL}$ |
| $x^2$ | $\mathcal{N}(0,2)$ | $\mathcal{U}_{(0;1)}$ | 5.06 | **0.288** | **0.258** | $\tilde{\pi}$ |
| | | | 2.986 | 0.527 | 0.548 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{U}_{(-1;1)}$ | 10.381 | **0.546** | **0.428** | $\tilde{\pi}$ |
| | | | 4.277 | 0.846 | 1.033 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{N}(0,1)$ | 20.113 | **2.658** | **2.379** | $\tilde{\pi}$ |
| | | | 14.442 | 2.754 | 2.395 | $\tilde{\pi}_{BCL}$ |
| | $\Gamma(4,1)$ | $\mathcal{U}_{(0;1)}$ | 5.845 | **0.811** | **0.351** | $\tilde{\pi}$ |
| | | | 2.505 | 0.894 | 0.547 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{U}_{(-1;1)}$ | 10.828 | **0.664** | **0.599** | $\tilde{\pi}$ |
| | | | 4.672 | 1.144 | 0.950 | $\tilde{\pi}_{BCL}$ |
| | | $\mathcal{N}(0,1)$ | 22.337 | **6.277** | **3.551** | $\tilde{\pi}$ |
| | | | 18.723 | 7.367 | 3.792 | $\tilde{\pi}_{BCL}$ |

Table 5.2: Values of MISE $\times 100$ averaged over 100 samples, in Examples 1 (regression models) for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$.



(a) (b) (c)

(e) (f) (g) (h)

Figure 5.3: Plots of true function versus estimators, Example 3 with $n = 2000$ observations. First line: (a) true function $\pi$, (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line, (d) and (f): plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed $x$. Second line, (e) and (g): for two fixed $x$, plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}_{BCL}(x, y)$ (dashed dotted line) .

| Example | $n = 60$ | 500 | 2000 | Method |
|---|---|---|---|---|
| Ex 2 | 7.621 | **1.163** | **0.498** | $\tilde{\pi}$ |
| | 2.739 | 1.178 | 0.657 | $\tilde{\pi}_{BCL}$ |
| Ex 3 | 14.451 | **12.264** | **12.87** | $\tilde{\pi}$ |
| | 13.153 | 12.764 | 13.175 | $\tilde{\pi}_{BCL}$ |
| Ex 4 | 3.617 | **0.907** | **0.557** | $\tilde{\pi}$ |
| | 2.407 | 1.142 | 0.611 | $\tilde{\pi}_{BCL}$ |

Table 5.3: Values of MISE $\times 100$ averaged over 100 samples, in Example 2,3,4,5 for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$.

| | Ex 1' (i) | Ex1' (ii) | Ex1' (iii) | Ex2' |
|---|---|---|---|---|
| $\tilde{\pi}$ | **0.7872** | **0.6427** | **0.6911** | **1.0408** |
| $\tilde{\pi}_{BCL}$ | 0.9167 | 0.7353 | 0.8159 | 1.1046 |
| $\tilde{\pi}_{FY}$ | 1.0899 | 0.7641 | 1.0143 | 2.7404 |

Table 5.4: Values of RMSE $\times 1000$ averaged over 100 samples of size $n = 1000$ according to Fan & Yim (2004), in Examples 1' and 2' for the estimators $\tilde{\pi}$, $\tilde{\pi}_{BCL}$ and $\tilde{\pi}_{FY}$.

**Remark 5.3.** The implementation also permits to compare at the same time the risk of the estimator $\tilde{\pi}_{BCL}$ of Brunel *et al.* (2007) to the estimator $\tilde{\pi}_{FY}$ of Fan & Yim (2004) introduced in Section 5.4.2: the results are presented in Table 5.4 (which thus completes Table 5.1). They first show that our estimator have the smallest risk in every examples: it outperforms both the kernel estimator, what we have already noticed, and the penalized least-squared estimator. Then, the risks of $\tilde{\pi}_{BCL}$ come in between $\tilde{\pi}$ and $\tilde{\pi}_{FY}$: they are smaller than the risks of $\tilde{\pi}_{FY}$.

## 5.5 Proofs

In all the proofs, the letter $C$ denotes a nonnegative real that may change from line to line. We recall that we denote by $\|t\|_{L^\infty(A)}$ the infinite norm of a function $t$ over a set $A$, by $\|t\|_{L^2(A)}$ its Hilbert norm, and by $\langle ., \rangle_A$ the associated scalar product. Notice that the quantities $\|\varphi_2^{(k)}\|_{L^\infty((0;1))}$ $(k = 0, 1, 2, 3)$ appear in the proof: since we consider the trigonometric basis, they can be replaced by $\sqrt{2}$. However, we choose to keep them, to make easier the extension of the results to other orthogonal bases (compactly support wavelets...).

### 5.5.1 Preliminary results

Let us start by setting a result which is the key argument in the proof of the main theorem. We consider the centred empirical process defined by

$$\forall t \in L^2((0;1) \times A_2),\ \nu_n(t) = \frac{1}{n} \sum_{i=1}^{n} t\left(F_X(X_i), Y_i\right) - \mathbb{E}\left[t\left(F_X(X_i), Y_i\right)\right]. \tag{5.16}$$

The aim of the following proposition is to control the deviations of the supremum of this process on the unit sphere of $\mathbb{S}_m$

$$\mathcal{S}(m) = \{t \in \mathbb{S}_m,\ \|t\| = 1\}.$$

**Proposition 5.1.** *If the function $g$ is bounded on $(0;1)\times A_2$, for all $\delta>0$, there exists a constant $C>0$, depending on $\|g\|_{L^\infty((0;1)\times A_2)}$, such that,*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)-2(1+2\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right]\leq\frac{C}{n}.$$

***Proof of Proposition 5.1.*** We first bound the maximum by a sum:

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right]\leq\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right],$$

with the abbreviation $c(\delta)=2(1+2\delta)$ and we apply the Talagrand Inequality (2.1) (stated in Proposition 2.2, Chapter 2). Recall that by using density arguments, we can apply it to the unit sphere of a finite dimensional linear space, that is $\mathcal{S}(m')$, for our problem. We define $r_t:(x,y)\mapsto t(F_X(x),y)$, and compute the constants $M_1$, $H$ and $v$. Notice first that $\|r_t\|_{L^\infty(A_1\times A_2)}\leq\|t\|_{L^\infty((0;1)\times A_2)}$, we deduce from Property (5.4) that we can set $M_1=\sqrt{D_{m'_1}D_{m'_2}}$. If $t\in\mathcal{S}(m')$, it can be written $t=\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}b_{j,k}\varphi_j\otimes\varphi_k$, with $\sum_{j,k}b_{j,k}^2=1$. So, using the linearity of the process, and Cauchy-Schwarz's Inequality, we get $\sup_{t\in\mathcal{S}(m')}\nu_n(t)^2\leq\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}\nu_n^2(\varphi_j\otimes\varphi_k)$. We use anew Property (5.4) to define $H^2$:

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)\right]\leq\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}\frac{1}{n}\mathrm{Var}\left(\varphi_j(F_X(X_1))\varphi_k(Y_1)\right)\leq\frac{D_{m'_1}D_{m'_2}}{n}:=H^2.$$

Finally, $\mathrm{Var}(t(F_X(X_1),Y_1))\leq\mathbb{E}[t^2(F_X(X_1),Y_1)]\leq\|t\|^2\|g\|_{L^\infty((0;1)\times A_2)}=\|g\|_{L^\infty((0;1)\times A_2)}:=v$. We just replace the quantities $M_1,H$ and $v$ by the values derived above in Inequality (2.1) (Proposition 2.2):

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\nu_n(t)^2-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right]$$

$$\leq c_1\left\{\sum_{m'\in\mathcal{M}_n}\frac{1}{n}\exp\left(-c_2D_{m'_1}D_{m'_2}\right)+\sum_{m'\in\mathcal{M}_n}\frac{D_{m'_1}D_{m'_2}}{n^2}\exp\left(-c_3\sqrt{n}\right)\right\}.$$

It remains to remark that the first sum is a constant and that $\sum_{m'\in\mathcal{M}_n}D_{m'_1}D_{m'_2}\leq n^2$ to conclude the proof.

□

We also set the following useful lemma.

**Proposition 5.2.** *Let $\nu:L^2((0;1)\times A_2)\mapsto\mathbb{R}$ be a linear functional. Let also $m=(m_1,m_2)$ be an index of the collection $\mathcal{M}_n$. Then,*

$$\sup_{t\in\mathcal{S}(m)}\nu^2(t)=\sum_{j=1}^{D_{m_1}}\sum_{k=1}^{D_{m_2}}\nu^2(\varphi_j\otimes\varphi_k).$$

***Proof of Proposition 5.2.***
If $t$ belongs to $\mathcal{S}(m)$, it can be written $t = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k} \varphi_j \otimes \varphi_k$, with $\sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k}^2 = 1$. Thus, by the linearity of $\nu$ and the Cauchy-Schwarz Inequality,

$$\nu^2(t) = \left( \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k} \nu(\varphi_j \otimes \varphi_k) \right)^2 \leq \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \nu^2(\varphi_j \otimes \varphi_k).$$

This leads to $\sup_{t \in \mathcal{S}(m)} \nu^2(t) \leq \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \nu^2(\varphi_j \otimes \varphi_k)$. The equality is obtained by choosing $t = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k} \varphi_j \otimes \varphi_k \in L^2((0;1))$, with $b_{j,k} = \nu(\varphi_j \otimes \varphi_k) / (\sum_{j'=1}^{D_{m_1}} \sum_{k'=1}^{D_{m_2}} \nu^2(\varphi_{j'} \otimes \varphi_{k'}))^{1/2}$.

□

### 5.5.2 Proof of Inequality (5.15) in the theoretical case of known c.d.f $F_X$

We first deal with the estimator $\tilde{\pi}_0$ (defined by (5.11)). It satisfies the following inequality:

$$\mathbb{E}\left[ \|\tilde{\pi}_0 - \pi\|_{f_X}^2 \right] \leq c \min_{m \in \mathcal{M}_n} \left\{ \frac{D_{m_1} D_{m_2}}{n} + \left\|\pi_m^{F_X} - \pi\right\|_{f_X}^2 \right\} + \frac{C}{n},$$

and its proof is a simple example of the scheme we will use to prove the main result, Theorem 5.1. For the sake of simplicity, in this section, we denote by $\hat{m}$ the selected index $\hat{m}^{F_X}$, by $V$ the penalty $V^{F_X}$, and by $A$ the quantity $A(., F_X)$. Let $\mathbb{S}_m$ be a fixed model in the collection indexed by $\mathcal{M}_n$.

**Main part of the proof of Inequality (5.15)**

We decompose the loss of the estimator as follows:

$$\|\tilde{\pi}_0 - \pi\|_{f_X}^2 = \left\|\hat{g}_{\hat{m}}^{F_X} - g\right\|^2 \leq 3 \left\|\hat{g}_{\hat{m}}^{F_X} - \hat{g}_{m \wedge \hat{m}}^{F_X}\right\|^2 + 3 \left\|\hat{g}_{m \wedge \hat{m}}^{F_X} - \hat{g}_m^{F_X}\right\|^2 + 3 \left\|\hat{g}_m^{F_X} - g\right\|^2.$$

By definition of $A$,

$$\left\|\hat{g}_{\hat{m}}^{F_X} - g\right\|^2 \leq 3\left(A(m) + V(\hat{m})\right) + 3\left(A(\hat{m}) + V(m)\right) + 3\left\|\hat{g}_m^{F_X} - g\right\|^2,$$

Moreover, by definition of $\hat{m}$, $A(\hat{m}) + V(\hat{m}) \leq A(m) + V(m)$, which leads to

$$\left\|\hat{g}_{\hat{m}}^{F_X} - g\right\|^2 \leq 6\left(A(m) + V(m)\right) + 3\left\|\hat{g}_m^{F_X} - g\right\|^2.$$

We have already bounded the risk of the estimator on a fixed model (see Section 5.2.3, Inequalities (5.6) and (5.7)), therefore, by definition of $V$, we get

$$\mathbb{E}\left[ \left\|\hat{g}_{\hat{m}}^{F_X} - g\right\|^2 \right] \leq 6\mathbb{E}\left[A(m)\right] + (6c_0 + 3)\frac{D_{m_1} D_{m_2}}{n} + 3\|g_m - g\|^2. \tag{5.17}$$

To pursue the proof, we have to control the expectation of $A(m)$. By splitting the norm $\|\hat{g}^{F_X}_{m'} - \hat{g}^{F_X}_{m\wedge m'}\|^2$ for $m, m' \in \mathcal{M}_n$, and using the definition of $A$, we get

$$A(m) \leq 3 \max_{m'\in\mathcal{M}_n} \left[ \left\| \hat{g}^{F_X}_{m'} - g_{m'} \right\|^2 - \frac{V(m')}{6} \right]_+ + 3 \max_{m'\in\mathcal{M}_n} \left[ \left\| g_{m\wedge m'} - \hat{g}^{F_X}_{m\wedge m'} \right\|^2 - \frac{V(m')}{6} \right]_+ + 3 \max_{m'\in\mathcal{M}_n} \|g_{m'} - g_{m\wedge m'}\|^2 .$$

The three terms of the above decomposition are studied in the following lemmas, proved just below.

**Lemma 5.1.** *If the function $g$ is bounded on $(0;1)\times A_2$, there exists a constant $C>0$ such that, for $m \in \mathcal{M}_n$,*

$$(a)\ \mathbb{E}\left[ \max_{m'\in\mathcal{M}_n} \left( \left\| \hat{g}^{F_X}_{m'} - g_{m'} \right\|^2 - \frac{V(m')}{6} \right)_+ \right] \leq \frac{C}{n},$$

$$(b)\ \mathbb{E}\left[ \max_{m'\in\mathcal{M}_n} \left( \left\| g_{m\wedge m'} - \hat{g}^{F_X}_{m\wedge m'} \right\|^2 - \frac{V(m')}{6} \right)_+ \right] \leq \frac{C}{n}.$$

**Lemma 5.2.** *If the function $g$ is bounded on $(0;1)\times A_2$, there exists a constant $C>0$ such that,*

$$\max_{m'\in\mathcal{M}_n} \|g_{m'} - g_{m\wedge m'}\|^2 \leq 4\|g_m - g\|^2.$$

These inequalities imply that

$$\mathbb{E}[A(m)] \leq \frac{C}{n} + 12\|g_m - g\|^2. \tag{5.18}$$

Gathering this with Inequality (5.17) ends the proof of the Theorem.

□

**Proof of Lemma 5.1**

To simplify the notations, we denote by $T_p = \|\hat{g}^{F_X}_p - g_p\|^2$ for $p = m'$ or $p = m\wedge m'$, and by $U_p = (T_p - V(m'))_+$.

*Inequality (a).* Using Proposition 5.2, we first compute,

$$\left\| \hat{g}^{F_X}_{m'} - g_{m'} \right\|^2 = \sum_{j=1}^{D_{m'_1}} \sum_{k=1}^{D_{m'_2}} \left( \hat{a}^{F_X}_{j,k} - a_{j,k} \right)^2 = \sum_{j=1}^{D_{m'_1}} \sum_{k=1}^{D_{m'_2}} \nu_n^2(\varphi_j \otimes \varphi_k) = \sup_{t\in\mathcal{S}(m')} \nu_n^2(t),$$

with $\nu_n$ the empirical process defined by (5.16). Thus,

$$\mathbb{E}\left[ \max_{m'\in\mathcal{M}_n} U_{m'} \right] = \mathbb{E}\left[ \max_{m'\in\mathcal{M}_n} \left( \sup_{t\in\mathcal{S}(m')} \nu_n^2(t) - \frac{V(m')}{6} \right)_+ \right],$$

and Inequality $(a)$ of the lemma is proved by applying Proposition 5.1.

*Inequality (b).* We have to distinguish several cases, depending on the value of $m \wedge m'$: $\max_{m' \in \mathcal{M}_n} U_{m \wedge m'}$

$$\leq \max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m'_2 \leq m_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m_2 \leq m'_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m'_2 \leq m_2}} U_{m \wedge m'}.$$

- *First term:* $m'_1 \leq m_1$ and $m'_2 \leq m_2$. In this case, $m \wedge m' = m'$. Thus, we bound roughly
$$\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m'_2 \leq m_2}} U_{m \wedge m'}\right] \leq \mathbb{E}\left[\max_{m' \in \mathcal{M}_n} U_{m'}\right],$$
and use Inequality $(a)$ to conclude that this term is bounded by $C/n$.
- *Second term:* $m_1 \leq m'_1$ et $m_2 \leq m'_2$. Here, $m \wedge m' = m$. Using $V(m) \leq V(m')$ (because $D_{m_l} \leq D_{m'_l}$, $l = 1, 2$), we have
$$\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} U_{m \wedge m'}\right] \leq \mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} \left(T_m - \frac{V(m)}{6}\right)_+\right] = \mathbb{E}\left[\left(T_m - \frac{V(m)}{6}\right)_+\right],$$
and it can be seen as a consequence of Proposition 5.1 and of the beginning of the proof of Inequality $(a)$ that this last term is bounded by $C/n$.
- *Third term:* $m'_1 \leq m_1$ et $m_2 \leq m'_2$. Here, we have $m \wedge m' = (m'_1, m_2)$. We use $V((m'_1, m_2)) \leq V(m'_1, m'_2)$ to get
$$\begin{aligned}\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m_2 \leq m'_2}} U_{m \wedge m'}\right] &\leq \mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m_2 \leq m'_2}} \left(T_{(m'_1, m_2)} - \frac{V((m'_1, m_2))}{6}\right)_+\right], \\ &\leq \sum_{m'_1 \in \mathcal{I}_n^{(1)}} \mathbb{E}\left[\left(T_{(m'_1, m_2)} - \frac{V((m'_1, m_2))}{6}\right)_+\right].\end{aligned}$$
The last term is also bounded by $C/n$, using a slightly different version of Proposition 5.1 (take the maximum only over $m'_1 \in \mathcal{I}_n^{(1)}$ instead of over $m \in \mathcal{M}_n$, and replace $m$ by $m \wedge m'$).
- *Fourth term:* $m_1 \leq m'_1$ et $m'_2 \leq m_2$. We deal with this case by using the same arguments as for the previous case.

We conclude that $\mathbb{E}[\max_{m' \in \mathcal{M}_n} U_{m \wedge m'}]$ is upper-bounded by $C/n$.

□

**Proof of Lemma 5.2**

Following the same lines as in the proof of Lemma 5.1, we distinguish four cases:
- $m'_1 \leq m_1$ and $m'_2 \leq m_2$. For such couples $(m_1, m_2)$ and $(m'_1, m'_2)$, $\|g_{m'} - g_{m \wedge m'}\|^2 = 0$.

– $m_1 \leq m_1'$ et $m_2 \leq m_2'$. We notice first that $\|g_{m'} - g_{m\wedge m'}\|^2 = \|g_{m'} - g_m\|^2 \leq 2\|g_{m'} - g\|^2 + 2\|g_m - g\|^2$. Since the models are nested in each direction (see Property (5.5)), we have $\mathbb{S}_m = S_{m_1} \times S_{m_2} \subset S_{m_1'} \times S_{m_2'} = \mathbb{S}_{m'}$. Consequently, $g_m \in \mathbb{S}_{m'}$, and by the definition of the orthogonal projection onto $\mathbb{S}_{m'}$, we get $\|g_{m'} - g\| \leq \|g_m - g\|$. This leads to $\|g_{m'} - g_{m\wedge m'}\|^2 \leq 4\|g_m - g\|^2$.

– $m_1' \leq m_1$ et $m_2 \leq m_2'$. To deal with this case, we use first the following remark: if $t$ belongs to $L^2((0;1) \times A_2)$, then for all $u \in (0;1)$, $y \mapsto t(u,y)$ belongs to $L^2(A_2)$ and $y \in A_2$, $u \mapsto t(u,y)$ belongs to $L^2((0;1))$. Moreover, by denoting by $G_1$ (respectively $G_2$) a closed linear subspace of $L^2((0;1))$ (respectively of $L^2(A_2)$), and by $\Pi_G$ the projection operator onto a subspace $G$, the following equality holds:

$$\Pi_{G_1\times G_2} t = \Pi_{G_1\times L^2(A_2)} \left(\Pi_{L^2((0;1))\times G_2} t\right).$$

In our setting, we thus compute

$$\begin{aligned}
\|g_{m'} - g_{m\wedge m'}\|^2 &= \left\|\Pi_{S_{m_1'}\times L^2(A_2)}\left[\Pi_{L^2((0;1))\times S_{m_2'}} g - \Pi_{L^2((0;1))\times S_{m_2}} g\right]\right\|^2, \\
&\leq \left\|\Pi_{L^2((0;1))\times S_{m_2'}} g - \Pi_{L^2((0;1))\times S_{m_2}} g\right\|^2, \\
&\leq 2\left\|\Pi_{L^2((0;1))\times S_{m_2'}} g - g\right\|^2 + 2\left\|\Pi_{L^2((0;1))\times S_{m_2}} g - g\right\|^2, \\
&\leq 4\left\|\Pi_{L^2((0;1))\times S_{m_2}} g - g\right\|^2 \leq 4\|g_m - g\|^2,
\end{aligned}$$

where the inequalities of the last line are obtained by noticing that $S_{m_2} \subset S_{m_2'}$ and that $S_{m_1} \subset L^2((0;1))$, and by using the definition of orthogonal projections.

– $m_1 \leq m_1'$ et $m_2' \leq m_2$. By symmetry, this case can be handled similarly to the latter.

Gathering the bounds of the four cases and taking the maximum of the four upper-bounds lead to the conclusion:

$$\max_{m'\in\mathcal{M}_n} \|g_{m'} - g_{m\wedge m'}\|^2 \leq \max\left\{0, 4\|g_m - g\|^2\right\} = 4\|g_m - g\|^2.$$

□

### 5.5.3 Proof of Theorem 5.1

To simplify the notations, we write in this section $A(m)$ to replace $A(m, \hat{F}_n)$, $V$ for $V^{\hat{F}}$, and $\hat{m}$ instead of $\hat{m}^{\hat{F}}$. The main idea of the proof is to recover the framework of the proof of Section 5.5.2. The computations are more technical, since the estimator $\tilde{\pi} = \hat{g}^{\hat{F}}_{\hat{m}}(\hat{F}(.),.)$ depends in two ways on $\hat{F}$. We denote it by $\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}}$, and coherently, we denote by $\hat{\pi}^{F_X,F_X}_{\hat{m}}$ the estimator previously studied, that is $\tilde{\pi}_0$. We also introduce the following intermediate:

$$\forall (x,y) \in A_1 \times A_2,\ \hat{\pi}^{\hat{F},F_X}(x,y) = \hat{g}^{\hat{F}}_{\hat{m}}(F_X(x), y). \tag{5.19}$$

These notations also suit well for a fixed index $m \in \mathcal{M}_n$. We denote by $\mathbb{E}[.|(X_{-l})]$ the conditional expectation given the sample $(X_{-l})_{l=1,\ldots,n}$ (the conditional variance will be coherently

denoted by $\mathrm{Var}(.|(X_{-l}))$). A key point is the following decomposition which holds for any index $m$: $\|\hat{\pi}_m^{\hat{F},\hat{F}} - \pi\|_{f_X}^2 \leq 6\sum_{l=0}^4 T_l^m$, with

$$\begin{aligned}
T_0^m &= \|\pi - \pi_m^{F_X}\|_{f_X}^2 + \|\pi_m^{F_X} - \hat{\pi}_m^{F_X,F_X}\|_{f_X}^2, \\
T_1^m &= \left\|\hat{\pi}_m^{F_X,F_X} - \hat{\pi}_m^{\hat{F},F_X} - \mathbb{E}\left[\hat{\pi}_m^{F_X,F_X} - \hat{\pi}_m^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_{f_X}^2, \\
T_2^m &= \left\|\hat{\pi}_m^{\hat{F},F_X} - \hat{\pi}_m^{\hat{F},\hat{F}} - \mathbb{E}\left[\hat{\pi}_m^{\hat{F},F_X} - \hat{\pi}_m^{\hat{F},\hat{F}}\,|(X_{-l})_l\right]\right\|_{f_X}^2, \\
T_3^m &= \left\|\mathbb{E}\left[\hat{\pi}_m^{F_X,F_X} - \hat{\pi}_m^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_{f_X}^2, \quad T_4^m = \left\|\mathbb{E}\left[\hat{\pi}_m^{\hat{F},F_X} - \hat{\pi}_m^{\hat{F},\hat{F}}\,|(X_{-l})_l\right]\right\|_{f_X}^2.
\end{aligned} \tag{5.20}$$

Let us remark that $T_0^m$ is the bias-variance decomposition for the risk of an estimator $\hat{\pi}_m^{F_X,F_X}$, and has already been studied (see Section 5.2.3). The sketch of the proof is now to decompose the loss function, using these intermediates and the definition of $A$ and $V$, and then to bound each of the terms by $CD_{m_1}D_{m_2}/n$ or to centre them (so as to show they are negligible).

**Main part of the proof**

We begin by introducing the intermediate estimator defined by (5.19) in the loss of our estimator:

$$\begin{aligned}
\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi\right\|_{f_X}^2 &\leq 3\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{\pi}_{\hat{m}}^{\hat{F},F_X} - \mathbb{E}\left[\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{\pi}_{\hat{m}}^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_{f_X}^2 \\
&\quad +3\left\|\mathbb{E}\left[\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \hat{\pi}_{\hat{m}}^{\hat{F},F_X}\,|(X_{-l})_l\right]\right\|_{f_X}^2 + 3\left\|\hat{\pi}_{\hat{m}}^{\hat{F},F_X} - \pi\right\|_{f_X}^2, \\
&= 3T_2^{\hat{m}} + 3T_4^{\hat{m}} + 3\left\|\hat{g}_{\hat{m}}^{\hat{F}} - g\right\|^2.
\end{aligned}$$

The last term itself can be decomposed, by construction of $A$, $V$, and $\hat{m}$:

$$\begin{aligned}
\left\|\hat{g}_{\hat{m}}^{\hat{F}} - g\right\|^2 &\leq 3\left\|\hat{g}_{\hat{m}}^{\hat{F}} - \hat{g}_{m\wedge\hat{m}}^{\hat{F}}\right\|^2 + 3\left\|\hat{g}_{m\wedge\hat{m}}^{\hat{F}} - \hat{g}_m^{\hat{F}}\right\|^2 + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2, \\
&\leq 3\left(A(m) + V\left(\hat{m}\right)\right) + 3\left(A\left(\hat{m}\right) + V\left(m\right)\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2, \\
&= 3\left(A(m) + 2V\left(m\right)\right) + 3\left(A\left(\hat{m}\right) + 2V\left(\hat{m}\right)\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2 - 3V\left(\hat{m}\right) - 3V\left(m\right), \\
&\leq 6\left(A(m) + 2V(m)\right) - 2V\left(\hat{m}\right) + 3\left\|\hat{g}_m^{\hat{F}} - g\right\|^2.
\end{aligned}$$

Furthermore, $\|\hat{g}_m^{\hat{F}} - g\|^2 = \|\hat{\pi}_m^{\hat{F},F_X} - \pi\|_{f_X}^2 \leq 3T_1^m + 3T_3^m + 6T_0^m$. Consequently,

$$\begin{aligned}
\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi\right\|_{f_X}^2 &\leq 3T_2^{\hat{m}} + 3T_4^{\hat{m}} - 3\times 2V\left(\hat{m}\right) + 3\times 6\left(A(m) + 2V(m)\right) \\
&\quad +3\times 3\times\left(3T_1^m + 3T_3^m + 6T_0^m\right),
\end{aligned} \tag{5.21}$$

where the terms $T_l^m$, $l = 0,\ldots,4$ are defined by (5.20). We split the term $A$, first in a similar way as in Section 5.5.2. Let $(m,m') \in \mathcal{M}_n^2$,

$$\left\|\hat{g}_{m'}^{\hat{F}} - \hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2 \leq 3\left\|\hat{g}_{m'}^{\hat{F}} - g_{m'}\right\|^2 + 3\left\|g_{m'} - g_{m\wedge m'}\right\|^2 + 3\left\|g_{m\wedge m'} - \hat{g}_{m\wedge m'}^{\hat{F}}\right\|^2.$$

But we immediatly try to recover the splitting terms defined by (5.20). By applying Proposition 5.2, we get, for $p = m$ or $p = m \wedge m'$,

$$\left\| g_p - \hat{g}_p^{\hat{F}} \right\|^2 = \sup_{t \in \mathcal{S}(p)} \tilde{\nu}_n^2(t), \quad \tilde{\nu}_n(t) = \frac{1}{n} \sum_{i=1}^{n} t\left(\hat{F}_n(X_i), Y_i\right) - \mathbb{E}\left[t\left(F_X(X_i), Y_i\right)\right],$$

for a function $t \in L^2((0;1) \times A_2)$. We recover the previous empirical process by the decomposition $\tilde{\nu}_n^2(t) \leq 2\nu_n^2(t) + 2R_n^2(t)$, with $R_n(t) = (1/n)\sum_{i=1}^n t(\hat{F}_n(X_i), Y_i) - t(F_X(X_i), Y_i)$. Moreover, if $t$ belongs to $\mathcal{S}(p)$, we have already written $t = \sum_{j=1}^{D_{p_1}} \sum_{k=1}^{D_{p_2}} \theta_{j,k} \varphi_j \otimes \varphi_k$, with $\sum_{j=1}^{D_{p_1}} \sum_{k=1}^{D_{p_2}} \theta_{j,k}^2 = 1$. Using this expression, Cauchy-Schwarz Inequality, and the definition of the coefficients $\hat{a}_{j,k}^{F_X}$ or $\hat{a}_{j,k}^{\hat{F}}$ yield $\sup_{t \in \mathcal{S}(p)} R_n^2(t) = \sum_{j=1}^{D_{p_1}} \sum_{k=1}^{D_{p_2}} (\hat{a}_{j,k}^{\hat{F}} - \hat{a}_{j,k}^{F_X})^2$. The conditional expectation of $\hat{a}_{j,k}^{\hat{F}} - \hat{a}_{j,k}^{F_X}$ is introduced to get $\sup_{t \in \mathcal{S}(p)} R_n^2(t) \leq 2T_1^p + 2T_3^p$. Consequently,

$$\left\| g_p - \hat{g}_p^{\hat{F}} \right\|^2 \leq 2 \sup_{t \in \mathcal{S}(p)} (\nu_n(t))^2 + 4T_1^p + 4T_3^p.$$

By substracting $V(m')$, taking the maximum over $m' \in \mathcal{M}_n$ and integrating give an upper-bound for $\mathbb{E}[A(m)]$. We introduce it into (5.21) to obtain:

$$\mathbb{E}\left[ \left\| \hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi \right\|_{f_X}^2 \right] \tag{5.22}$$

$$\begin{aligned} \leq\ & 36V(m) + 27\mathbb{E}\left[2T_0^m + T_1^m + T_3^m\right] + 18\left\{ 3 \max_{m' \in \mathcal{M}_n} \left\| g_{m \wedge m'} - g_{m'} \right\|^2 \right. \\ & + 3\mathbb{E}\left[ \left( T_2^{\hat{m}} - V(\hat{m}) \right)_+ \right] + 3\mathbb{E}\left[ \left( T_4^{\hat{m}} - V(\hat{m}) \right)_+ \right] \\ & + 6\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( \sup_{t \in \mathcal{S}(m')} (\nu_n(t))^2 - \frac{V(m')}{18 \times 36} \right)_+ \right] \\ & + 6\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( \sup_{t \in \mathcal{S}(m \wedge m')} (\nu_n(t))^2 - \frac{V(m')}{18 \times 36} \right)_+ \right] \\ & + 12\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( T_3^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( T_3^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \\ & \left. + 12\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( T_1^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( T_1^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \right\}. \end{aligned} \tag{5.23}$$

We bound each of these terms. Some of them have already been studied: recall first that

$$\mathbb{E}\left[T_0^m\right] \leq \left\| \pi_m^{F_X} - \pi \right\|^2 + \frac{D_{m_1} D_{m_2}}{n},$$

using (5.6) and (5.7). Moreover, applying twice Proposition 5.1 shows that

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( \sup_{t \in \mathcal{S}(m')} (\nu_n(t))^2 - V_0(m') \right)_+ \right] \leq \frac{C}{n},$$

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_n} \left( \sup_{t \in \mathcal{S}(m \wedge m')} (\nu_n(t))^2 - V_0(m') \right)_+ \right] \leq \frac{C}{n},$$

with $V_0(m') = 2(1+2\delta)D_{m'_1}D_{m'_2}/n$. Choosing $c_1$ (see the definition (5.8)) larger than $2(1+2\delta)$, these inequalities hold with $V$ in place of $V_0$. Finally, we have proved in Lemma 5.2 that $\max_{m'\in\mathcal{M}_n}\|g_{m'} - g_{m\wedge m'}\|^2 \leq 4\|g_m - g\|^2$. Taking into account the previous inequality (5.22) for the risk, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi\right\|_{f_X}^2\right] \leq\ & 36V(m) + 27\times 2\frac{D_{m_1}D_{m_2}}{n} + (12\times 18 + 27\times 2)\|g_m - g\|^2 + \frac{C}{n} \\ & +3\mathbb{E}\left[\left(T_2^{\hat{m}} - V(\hat{m})\right)_+\right] + 3\mathbb{E}\left[\left(T_4^{\hat{m}} - V(\hat{m})\right)_+\right] + 27\mathbb{E}\left[T_1^m + T_3^m\right] \\ & +18\left\{12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'} - \frac{V(m')}{18\times 72}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'} - \frac{V(m')}{18\times 72}\right)_+\right]\right. \\ & \left.+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m'} - \frac{V(m')}{18\times 72}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m\wedge m'} - \frac{V(m')}{18\times 72}\right)_+\right]\right\}.\end{aligned} \tag{5.24}$$

It remains to bound the terms $T_l^m$, $l = 1,2,3,4$ or their centred versions, by quantities of order at most $D_{m_1}D_{m_2}/n$. Let us first notice that, for $l = 2,4$,

$$\mathbb{E}\left[\left(T_l^{\hat{m}} - V(\hat{m})\right)_+\right] \leq \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_l^{m'} - V(m')\right)_+\right],$$

and then use the lemmas just below, whose proofs are deferred to the following sections.

**Lemma 5.3.** *Assuming that the models are trigonometric, there exists a constant $C$ depending only on $\|\varphi_2'\|_{L^\infty((0;1))}$ such that, for $m\in\mathcal{M}_n$,*

$$\mathbb{E}\left[T_1^m\right] \leq C\frac{D_{m_1}^3 D_{m_2}}{n^2}.$$

*Moreover, the following inequality holds, if $D_{m_1} = O(\sqrt{n}/\ln(n))$, for $p_{m'} = m'$ or $p_{m'} = m\wedge m'$, and for a constant $C>0$*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{p_{m'}} - V_1(m')\right)_+\right] \leq \frac{C}{n},$$

*with $V_1(m') = \kappa_1 D_{m'_1}D_{m'_2}/n$, and $\kappa_1$ a constant depending only on $\|\varphi_2'\|_{L^\infty((0;1))}$.*

If $D_{m_1} = O(n^{1/2})$ in particular, the first inequality of Lemma 5.3 leads to $\mathbb{E}[T_1^m] \leq CD_{m_1}D_{m_2}/n$.

**Lemma 5.4.** *Assuming that the models are trigonometric, there exists a constant $C$, which depends on $\|\varphi_2'\|_{L^\infty((0;1))}$, such that*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_2^{m'} - V_2(m')\right)_+\right] \leq C\frac{\ln(n)}{n},$$

*with $V_2(m') = \kappa_2 D_{m'_1}^4 D_{m'_2}\ln^2(n)/n^2$, and $\kappa_2$ a constant depending also on $\|\varphi_2'\|_{L^\infty((0;1))}$.*

Assuming that $D_{m'_1} = O(n^{1/3}/\ln^{2/3}(n))$, we have $V_2(m') \leq V_2^b(m') := \kappa_2' D_{m'_1}D_{m'_2}/n$ ($\kappa_2'$ a constant independent of $g$). The inequality of Lemma 5.4 still holds by replacing $V_2$ by $V_2^b$.

**Lemma 5.5.** *Assuming that the models are trigonometric, and that $g$ is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$, there exists a constant $C$ depending on $\|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\|g\|$ and $\|\partial_1 g\|$ ($\partial_1$ is the derivation operator with respect to the first variable) such that, for $m \in \mathcal{M}_n$,*

$$\mathbb{E}\left[T_3^m\right] \leq C\left(\frac{1}{n} + \frac{D_{m_1}}{n} + \frac{D_{m_1}^4}{n^2} + \frac{D_{m_1}^7}{n^3}\right).$$

*Moreover, the following inequality holds, for $p_{m'} = m'$ or $p_{m'} = m \wedge m'$, for $n \geq n_0(g)$, and assuming $D_{m_1} = O(n^{1/3})$ and $D_{m_2} \geq c\ln^4(n)$ (for a constant $c > 0$) for each $m$,*

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(T_3^{p_{m'}} - V_3(m')\right)_+\right] \leq \frac{C}{n},$$

*with $V_3(m') = \kappa_3 \frac{D_{m_1'} D_{m_2'}}{n}$, $\kappa_3$ a constant independent of $g$, and $n_0(g)$ a nonnegative integer depending on the function $h$.*

If $D_{m_1} = O(n^{1/3})$ in particular, the first inequality of Lemma 5.3 leads to $\mathbb{E}[T_3^m] \leq CD_{m_1}D_{m_2}/n$.

**Lemma 5.6.** *Assuming that the models are trigonometric, that $g$ is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$ and belongs to the anisotropic Sobolev ball denoted by $W_2^{(1,0),per}((0;1)^2, L)$, and that for all $m \in \mathcal{M}_n$, $D_{m_1} = O(n^{1/3}/\ln^{1/3}(n))$ and $D_{m_2} \geq c\ln^5(n)$ (for a constant $c > 0$), there exists a constant $C$, which depends on $\|\varphi_2'\|_{L^\infty((0;1))}$, $\|\varphi_2''\|_{L^\infty((0;1))}$, $\|\varphi_2^{(3)}\|_{L^\infty((0;1))}$, $\|g\|$, $\|\partial_1 g\|$, and $L$ such that, for $n \geq n_1(g)$,*

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(T_4^{m'} - V_4(m')\right)_+\right] \leq C\frac{\ln(n)}{n},$$

*with $V_4(m') = \kappa_4 D_{m_1'} D_{m_2'}/n$, and $\kappa_4$ independent of $g$, and $n_1(g)$ a nonnegative integer depending on the function $g$.*

To conclude the proof, we choose the constant $c_1$ larger than $\kappa_l$ ($l = 1, \ldots, 4$), to have $V(m') \geq V_l(m')$ (or $V_l^b(m')$ for $l = 2$): this enables to apply the inequalities of the lemmas with $V$ and to use it in Inequality (5.24). We then obtain the result of Theorem 5.1.

□

**Technical tools for the proof of Lemmas 5.3 to 5.6**

As usual when using a "warping" method, key arguments for the proof of the lemmas are the properties of the empirical cumulative distribution function $\hat{F}_n$ of the sample $(X_{-l})_l$. First, let $U_{-i} = F_X(X_{-i})$ ($i = 1, \ldots, n$). Recall that it is a uniform variable on $(0;1)$. We denote by $\hat{U}_n$ the empirical c.d.f. associated to the sample $(U_{-i})_{i=1,\ldots,n}$. We get

$$\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}} \,|(X_{-l})_l\right] = \int_{(0;1)\times A_2} \varphi_j \circ \hat{U}_n(u)\varphi_k(y)g(u,y)dudy.$$

More details are given in Section 2.1.2 of Chapter 2. Especially, some inequalities in the spirit of Dvoretzky *et al.* (1956) are stated there. We will use it intensively.

**Proof of Lemma 5.3**

The first part of the lemma is to bound $\mathbb{E}[T_1^m]$. Using the definition of $\hat{\pi}^{F_X,F_X}$ and $\hat{\pi}^{\hat{F},F_X}$ leads to

$$T_1^m = \left\| \hat{g}_m^{F_X} - \hat{g}_m^{\hat{F}} - \mathbb{E}\left[ \hat{g}_m^{F_X} - \hat{g}_m^{\hat{F}} \,|(X_{-l})_l \right] \right\|^2 .$$

The decompositions of the estimators in the orthonormal basis $(\varphi_j \otimes \varphi_k)$ yield $T_1^m = \sum_{j,k} \{ (\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}) - \mathbb{E}[\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}} | (X_{-l})] \}^2$. Thus,

$$\mathbb{E}\left[T_1^m \,|(X_{-l})_l\right] = \sum_{j,k} \mathrm{Var}\left( \hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}} \,|(X_{-l})_l \right).$$

We work out the conditional variance for any couple $(j,k)$:

$$\begin{aligned} \mathrm{Var}\left( \hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}} \,|(X_{-l})_l \right) &= \frac{1}{n} \mathrm{Var}\left( \varphi_j\left(F_X(X_1)\right) \varphi_k(Y_1) - \varphi_j\left(\hat{F}_n(X_1)\right) \varphi_k(Y_1) \,|(X_{-l})_l \right), \\ &\leq \frac{1}{n} \mathbb{E}\left[ \varphi_k^2(Y_1) \left\{ \varphi_j\left(F_X(X_1)\right) - \varphi_j\left(\hat{F}_n(X_1)\right) \right\}^2 |(X_{-l})_l \right]. \end{aligned}$$

We apply the mean value theorem, sum over the indices $j$ and $k$, and remark $\|\varphi_j'\|_{L^\infty((0;1))} \leq D_{m_1} \|\varphi_2'\|_{L^\infty((0;1))}$ (property of the trigonometric basis):

$$\begin{aligned} \mathbb{E}\left[T_1^m \,|(X_{-l})_l\right] &\leq \frac{1}{n} \left\| \sum_{k=1}^{D_{m_2}} \varphi_k^2 \right\|_{L^\infty((0;1))} \sum_{j=1}^{D_{m_1}} \|\varphi_j'\|_{L^\infty((0;1))}^2 \left\| F_X - \hat{F}_n \right\|_{L^\infty(A_1)}^2, \\ &\leq \|\varphi_2'\|_{L^\infty((0;1))}^2 \frac{D_{m_1}^3 D_{m_2}}{n} \left\| F_X - \hat{F}_n \right\|_{L^\infty(A_1)}^2 . \end{aligned}$$

It remains to use Proposition 2.4 of Chapter 2 with $p = 2$ to bound the expectation:

$$\mathbb{E}\left[T_1^m\right] \leq C \|\varphi_2'\|_{L^\infty((0;1))}^2 \frac{D_{m_1}^3 D_{m_2}}{n^2}$$

This completes the proof of the first inequality. For the second, let us begin with $V_1(p_{m'}) \leq V_1(m')$. Therefore $\mathbb{E}[\max_{m' \in \mathcal{M}_n} (T_1^{p_{m'}} - V_1(m'))_+] \leq \mathbb{E}[\max_{m' \in \mathcal{M}_n} (T_1^{p_{m'}} - V_1(p_{m'}))_+]$. In the sequel, we simplify the notations by setting $p = p_{m'}$. We apply Proposition 5.2, which leads to $T_1^p = \sup_{t \in \mathcal{S}(p)} \left(\nu_n^a(t)\right)^2$ with

$$\nu_n^a(t) = \frac{1}{n} \sum_{i=1}^n \left( t(F_X(X_i), Y_i) - t(\hat{F}_n(X_i), Y_i) \right) - \mathbb{E}\left[ \left( t(F_X(X_i), Y_i) - t(\hat{F}_n(X_i), Y_i) \right) |(X_{-l})_l \right],$$

a process which is centred conditionally to the sample $(X_{-l})_l$. Thus we apply Talagrand's inequality (2.1) (Proposition 2.2), as in the proof of Proposition 5.1, but conditionally to $(X_{-l})_l$. In this setting the key quantities are such that

$$\sup_{t \in \mathcal{S}(p)} \|r_t\|_{L^\infty((0;1) \times A_2)} \leq M_{1,a}, \; \mathbb{E}\left[ \sup_{t \in \mathcal{S}(p)} |\nu_n^a(t)| \,|(X_{-l})_l \right] \leq H_{a,p},$$

$$\text{and} \sup_{t \in \mathcal{S}(p)} \frac{1}{n} \sum_{i=1}^n \mathrm{Var}\left( r_t(X_i, Y_i) \,|(X_{-l})_l \right) \leq v_a.$$

We compute

$$M_{1,a} = \|\varphi_2'\|_{L^\infty((0;1))} D_{p_1}^{3/2} D_{p_2}^{1/2} \left\|\hat{F}_n - F_X\right\|_{L^\infty(A_1)},$$
$$H_{a,p}^2 = \tfrac{1}{n} \|\varphi_2'\|_{L^\infty((0;1))}^2 D_{p_1}^3 D_{p_2} \left\|\hat{F}_n - F_X\right\|_{L^\infty(A_1)}^2, \quad v_a = nH_{a,p}^2,$$

and thus obtain for $\delta > 0$, $\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(p)} (\nu_n^a(t))^2 - 2(1+2\delta)H_{a,p}^2\right)_+ |(X_{-l})_l\right]$

$$\leq C_0 \left\{ H_{a,p}^2 \exp(-C\delta) + \frac{H_{a,p}^2}{C^2(\delta)n} \exp\left(-C\sqrt{\delta}\sqrt{n}\right)\right\}.$$

Here, $C_0$ is a random constant, which depends on $\|F_X - \hat{F}_n\|_{L^\infty(A_1)}$, and $C$ is purely numerical. But $C_0$ can be also bounded by a fixed quantity, since the infinite norm is smaller than 1. Thus we write anew $C$ in the sequel. We choose $\delta = \kappa \ln(n)$ ($\kappa > 0$), so that $C(\delta) = 1$. We now put $p = m'$ (The case $p = m \wedge m'$ can be handled similarly). We thus have $\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(T_1^{m'} - 2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+ |(X_{-l})_l\right]$

$$\begin{aligned}
&\leq \sum_{m'\in\mathcal{M}_n} \mathbb{E}\left[\left(T_1^{m'} - 2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+ |(X_{-l})_l\right], \\
&\leq C\left\{ n^{-C\kappa} \sum_{m'\in\mathcal{M}_n} \frac{D_{m_1'}^3 D_{m_2'}}{n} + \exp\left(-C\sqrt{n}\right) \sum_{m'\in\mathcal{M}_n} \frac{D_{m_1'}^3 D_{m_2'}}{n^2}\right\}.
\end{aligned}$$

Moreover, we use $D_{m_l} = O(\sqrt{n})$ ($l = 1, 2$), and remark that the cardinal of $\mathcal{M}_n$ is smaller than $n$, to get $\sum_{m'\in\mathcal{M}_n} D_{m_1'}^3 D_{m_2'}/n \leq \sum_{m'\in\mathcal{M}_n} Cn^{3/2}n^{1/2}/n \leq Cn^2$. Thus, if we choose $\kappa$ large enough,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(T_1^{m'} - 2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+ |(X_{-l})_l\right] \leq C\left\{n^{2-C\kappa} + n\exp\left(-C\sqrt{n}\right)\right\} \leq Cn^{-1}.$$

We then notice that, for any $\alpha_n > 0$

$$\begin{aligned}
2(1+2\kappa\ln(n))H_{a,m'}^2 &\leq 6\kappa \left\|\varphi_2'\right\|_{L^\infty((0;1))}^2 \frac{D_{m_1'}^3 D_{m_2'} \ln(n)}{n} \left\|\hat{F}_n - F_X\right\|_{L^\infty(A_1)}^2, \\
&\leq 6\kappa \left\|\varphi_2'\right\|_{L^\infty((0;1))}^2 \frac{D_{m_1'}^3 D_{m_2'} \ln(n)}{n} \left(\alpha_n^2 + \mathbf{1}_{\|\hat{F}_n - F_X\|_{L^\infty(A_1)} \geq \alpha_n}\right).
\end{aligned}$$

Choosing $\alpha_n = \sqrt{3\ln(n)/n}$, and using $D_{m_1'} = O(\sqrt{n}/\ln(n))$,

$$\begin{aligned}
2(1+2\kappa\ln(n))H_{a,m'}^2 &\leq 18\kappa \left\|\varphi_2'\right\|_{L^\infty((0;1))}^2 \frac{D_{m_1'} D_{m_2'}}{n} + C\frac{n}{\ln^2(n)} \mathbf{1}_{\|\hat{F}_n - F_X\|_{L^\infty(A_1)} \geq \alpha_n}, \\
&= V_1(m') + C\mathbf{1}_{\|\hat{F}_n - F_X\|_{L^\infty(A_1)}^2 \geq \alpha_n},
\end{aligned}$$

Besides,

$$\begin{aligned}\mathbb{E}\left[\left(T_1^{m'}-V_1(m')\right)_+\right] &\leq \mathbb{E}\left[\left(T_1^{m'}-2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+\right]+\mathbb{E}\left[C\frac{n}{\ln^2(n)}\mathbf{1}_{\|\hat{F}_n-F_X\|_{L^\infty(A_1)}\geq\alpha_n}\right],\\ &\leq \mathbb{E}\left[\left(T_1^{m'}-2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+\right]+Cn^{-2}\ln^{-1}(n),\end{aligned}$$

with the inequality of Proposition 2.3 (Chapter 2). To conclude, $\sum_{m'\in\mathcal{M}_n}\mathbb{E}[(T_1^{m'}-V_1(m'))_+]\leq C/n$.

□

**Proof of Lemma 5.4**

For convenience, the constant $\kappa_2$ in the definition of $V_2$ is split into two parts, that is $\kappa_2=\kappa\kappa'$. The first step is to write $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_2^{m'}-V_2(m'))_+]\leq\sum_{m'\in\mathcal{M}_n}\mathbb{E}[(T_2^{m'}-V_2(m'))_+]$. Then it is enough to bound this quantity for each index $m'$. We write in a shortened form the sum "$\sum_{j=1}^{D_{m_1'}}$": "$\sum_j$" (and the analogous for $\sum_{k=1}^{D_{m_2'}}$). We compute
$T_2^{m'}$

$$\begin{aligned}&= \int_{A_1\times A_2}\left(\hat{g}_{m'}^{\hat{F}}\left(F_X(x),y\right)-\hat{g}_{m'}^{\hat{F}}\left(\hat{F}_n(x),y\right)\right.\\ &\quad\left.-\mathbb{E}\left[\hat{g}_{m'}^{\hat{F}}\left(F_X(x),y\right)-\hat{g}_{m'}^{\hat{F}}\left(\hat{F}_n(x),y\right)|(X_{-l})_l\right]\right)^2 f_X(x)dxdy,\\ &= \int_{A_1}\sum_{j,j'}\sum_{k,k'}\left(\hat{a}_{j,k}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right]\right)\left(\hat{a}_{j',k'}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j',k'}^{\hat{F}}|(X_{-l})_l\right]\right)\\ &\quad\times\left(\varphi_j\circ F_X(x)-\varphi_j\circ\hat{F}_n(x)\right)\left(\varphi_{j'}\circ F_X(x)-\varphi_{j'}\circ\hat{F}_n(x)\right)\int_{A_2}\varphi_k(y)\varphi_{k'}(y)dy f_X(x)dx,\\ &= \int_{(0;1)}\sum_{k=1}^{D_{m_2'}}\left\{\sum_{j=1}^{D_{m_1'}}\left(\hat{a}_{j,k}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right]\right)\left(\varphi_j(u)-\varphi_j\circ\hat{U}_n(u)\right)\right\}^2 du,\end{aligned}$$

By the Cauchy-Schwarz Inequality, and the mean value theorem,

$$T_2^{m'} \leq \left\|\varphi_2'\right\|_{\infty,(0;1)}^2 D_{m_1'}^3\left\|\hat{U}_n-id\right\|_{L^\infty((0;1))}^2\sum_{k=1}^{D_{m_2'}}\sum_{j=1}^{D_{m_1'}}\left(\hat{a}_{j,k}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right]\right)^2.$$

Thus, $\mathbb{E}[(T_2^{m'}-V_2(m'))_+]\leq T_{2,a}^{m'}+T_{2,b}^{m'}$, with

$$T_{2,a}^{m'}=D_{m_1'}^3\left\|\varphi_2'\right\|_{L^\infty((0;1))}^2\mathbb{E}\left[\sum_{j,k}\left(\hat{a}_{j,k}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right]\right)^2\left(\left\|\hat{U}_n-id\right\|_{L^\infty((0;1))}^2-\kappa'\frac{\ln(n)}{n}\right)_+\right],$$

$$T_{2,b}^{m'}=D_{m_1'}^3\left\|\varphi_2'\right\|_{L^\infty((0;1))}^2\kappa'\frac{\ln(n)}{n}\mathbb{E}\left[\left(\sum_{j,k}\left(\hat{a}_{j,k}^{\hat{F}}-\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right]\right)^2-\frac{\kappa}{\|\varphi_2'\|_{L^\infty((0;1))}^2}\frac{D_{m_1'}D_{m_2'}}{n}\ln(n)\right)_+\right].$$

Bounding roughly $\sum_{j,k}(\hat{a}_{j,k}^{\hat{F}} - \mathbb{E}[\hat{a}_{j,k}^{\hat{F}}(X_{-l})_l])^2$ leads to

$$\begin{aligned} T_{2,a}^{m'} &\leq 2D_{m_1'}^4 D_{m_2'} \left\|\varphi_2'\right\|_{L^\infty((0;1))}^2 \mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2 - \kappa' \frac{\ln(n)}{n}\right)_+\right], \\ &\leq Cn^{4/2} n^{1/2} \left(\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2 - \kappa' \frac{\ln(n)}{n}\right)^2\right]\right)^{1/2} \text{ using } D_{m_l} \leq \sqrt{n}\,(l=1,2), \\ &\leq Cn^{5/2} n^{-1-\kappa'} = Cn^{-3/2-\kappa'} \text{ (Inequality (2.6) of Corollary 2.1)} \end{aligned}$$

Thus, choosing $\kappa' \geq 7/2$, $\sum_{m' \in \mathcal{M}_n} T_{2,a}^{m'} \leq C/n$. For the second term $T_{2,b}^{m'}$, we notice first that $\sum_{j,k}(\hat{a}_{j,k}^{\hat{F}} - \mathbb{E}[\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l])^2 = \sup_{t \in \mathcal{S}(m')}(\nu_n^b)^2(t)$ (Proposition 5.2) with

$$\nu_n^b(t) = \frac{1}{n}\sum_{i=1}^n t\left(\hat{F}_n(X_i), Y_i\right) - \mathbb{E}\left[t\left(\hat{F}_n(X_i), Y_i\right)|(X_{-l})_l\right].$$

We now bound the deviations of this empirical process, centred conditonally to $(X_{-l})$, exactly as we bound $\nu_n^a$ in the proof of Lemma 5.3: they are controlled by the Talagrand Inequality (2.1). We finally obtain $\sum_{m' \in \mathcal{M}n} T_{2,b}^{m'} \leq C\ln(n)/n$, which ends the proof, by gathering this bound with the one of $\sum_{m' \in \mathcal{M}n} T_{2,a}^{m'}$.

□

**Proof of Lemma 5.5**

To compute a bound for $\mathbb{E}[T_3^m]$, let us begin with the definition of the estimators and their coefficients, to get $T_3^m = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \{\langle \varphi_k, \Lambda_j(y)\rangle_{A_2}\}^2$ with $\Lambda_j(y) = \int_{A_1}(\varphi_j(\hat{F}_n(x)) - \varphi_j(F_X(x)))f_{(X,Y)}(x,y)dx$. Thus we can write $T_3^m = \sum_{j=1}^{D_{m_1}} \|\Pi_{S_{m_2}}\Lambda_j\|_{A_2}^2 \leq \sum_{j=1}^{D_{m_1}} \|\Lambda_j\|_{A_2}^2$, which can be developed as

$$T_3^m \leq \sum_{j=1}^{D_{m_1}} \int_{A_2} \left(\int_{(0;1)} (\varphi_j(\hat{U}_n(u)) - \varphi_j(u))g(u,y)du\right)^2 dy := \int_{A_2} T_3'^m(y)dy.$$

We apply Taylor's formula with the Lagrange form for the remainder: there exists a random number depending on $j$, $\hat{\alpha}_{j,n,u}$, such that the following splitting holds:

$$\mathbb{E}\left[T_3'^m(y)\right] \leq 3\mathbb{E}\left[T_{3,1}^m(y)\right] + 3\mathbb{E}\left[T_{3,2}^m(y)\right] + 3\mathbb{E}\left[T_{3,3}^m(y)\right],$$

with notations

$$\begin{aligned} T_{3,1}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{\int_0^1 g(u,y)\left(\hat{U}_n(u) - u\right)\varphi_j'(u)du\right\}^2, \\ T_{3,2}^m(y) &= (1/4)\sum_{j=1}^{D_{m_1}} \left\{\int_0^1 g(u,y)\left(\hat{U}_n(u) - u\right)^2\varphi_j''(u)du\right\}^2, \\ T_{3,3}^m(y) &= (1/36)\sum_{j=1}^{D_{m_1}} \left\{\int_0^1 g(u,y)\left(\hat{U}_n(u) - u\right)^3\varphi_j^{(3)}(\hat{\alpha}_{j,n,u})du\right\}^2. \end{aligned}$$

Writing the definition of $\hat{U}_n(u)$, and noting that $u = \mathbb{E}[\mathbf{1}_{U_i \leq u}]$ $(i = 1, \ldots, n)$, we get for the first term

$$\mathbb{E}\left[T_{3,1}^m(y)\right] = \mathbb{E}\left[\sum_{j=1}^{D_{m_1}} \left(\frac{1}{n}\sum_{i=1}^{n} A_{i,j}(y) - \mathbb{E}[A_{i,j}(y)]\right)^2\right], \quad \text{with } A_{i,j}(y) = \int_{U_i}^{1} g(u,y)\varphi_j'(u)du.$$

We integrate by parts in $A_{i,j}$ ($g$ is assumed to be $\mathcal{C}^1$ with respect to its first variable). This leads to another splitting, for each $y \in A_2$:

$$\mathbb{E}\left[T_{3,1}^m(y)\right] \leq 2\mathbb{E}\left[T_{3,1,1}^m(y)\right] + 2\mathbb{E}\left[T_{3,1,2}^m(y)\right],$$

where

$$\begin{aligned} T_{3,1,1}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{\frac{1}{n}\sum_{i=1}^{n} g(U_i,y)\varphi_j(U_i) - \mathbb{E}\left[g(U_i,y)\varphi_j(U_i)\right]\right\}^2, \\ T_{3,1,2}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{\int_0^1 \partial_1 g(u,y)\left(\hat{U}_n(u) - u\right)\varphi_j(u)du\right\}^2. \end{aligned} \tag{5.25}$$

In the spirit of the bound given for $T_1^m$, the first term is controlled as follows:

$$\mathbb{E}\left[T_{3,1,1}^m(y)\right] \leq \frac{1}{n}\sum_{j=1}^{D_{m_1}} \mathbb{E}\left[(g(U_1,y)\varphi_j(U_1))^2\right] \leq \frac{D_{m_1}}{n}\int_0^1 g(u,y)^2 du.$$

Thus, $\int_{A_2} \mathbb{E}[T_{3,1,1}^m(y)]dy \leq \|g\|^2 D_{m_1}/n$. Then, by definition and properties of the orthogonal projection on $\mathbb{S}_m$,

$$\mathbb{E}\left[T_{3,1,2}^m(y)\right] = \mathbb{E}\left[\sum_{j=1}^{D_{m_1}} \left(\langle \partial_1 g(.,y)(\hat{U}_n - id), \varphi_j\rangle_{(0;1)}\right)^2\right] \leq \mathbb{E}\left[\left\|\partial_1 g(.,y)(\hat{U}_n - id)\right\|_{(0;1)}^2\right].$$

Finally, $T_{3,1,2}^m(y) \leq C\|\partial_1 g(.,y)\|_{L^2((0;1))}^2/n$ by Proposition 2.4, and thus, by gathering the bounds for $T_{3,1,1}^m(y)$ and $T_{3,1,2}^m(y)$,

$$\int_{A_2} \mathbb{E}\left[T_{3,1}^m(y)\right] dy \leq C\left(\frac{1}{n} + \frac{D_{m_1}}{n}\right).$$

As regards $T_{3,2}^m(y)$, we remark first that for $j \geq 2$, $\varphi_j'' = -(\pi\mu_j)^2\varphi_j$, with $\mu_j = j$ for even $j$, and $\mu_j = j - 1$ otherwise, so that $\mu_j$ is bounded by $D_{m_1}$. Hence

$$\begin{aligned} \mathbb{E}\left[T_{3,2}^m(y)\right] &\leq (\pi^4/4)D_{m_1}^4 \mathbb{E}\left[\sum_{j=2}^{D_{m_1}} \left\{\int_0^1 g(u,y)\left(\hat{U}_n(u) - u\right)^2 \varphi_j(u)du\right\}^2\right], \\ &\leq (\pi^4/4)D_{m_1}^4 \mathbb{E}\left[\left\|g(.,y)\left(\hat{U}_n - id\right)^2\right\|_{L^2((0;1))}^2\right] \leq C\int_{(0;1)} g^2(u,y)du\frac{D_{m_1}^4}{n^2}, \end{aligned}$$

by proceeding with the previous arguments (properties of orthogonal projection and Proposition 2.4). So we prove $\int_{A_2} \mathbb{E}[T_{3,2}^m(y)]dy \leq CD_{m_1}^4/n^2$. The computations for the last term are less technical:

$$\mathbb{E}\left[T_{3,3}^m(y)\right] \leq (1/36)\sum_{j=1}^{D_{m_1}} \left\|\varphi_j^{(3)}\right\|_{L^\infty((0;1))}^2 \|g(.,y)\|_{L^2((0;1))}^2 \mathbb{E}\left[\left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^6\right],$$

thus $\int_{A_2} \mathbb{E}[T^m_{3,3}(y)]dy \leq CD^7_{m_1}/n^3$. This completes the proof of the first inequality of Lemma 5.5.

With regard to the second inequality, it is enough to bound $\mathbb{E}[\max_{m' \in \mathcal{M}_n}(T^p_1 - V_1(p))_+]$, like for the second part of Lemma 5.3 ($p = m'$ or $p = m \wedge m'$). As previously, we get the splitting

$$T^p_3 \leq 6\int_{A_2} T^p_{3,1,1}(y)dy + 6\int_{A_2} T^p_{3,1,2}(y)dy + 3\int_{A_2} T^p_{3,2}(y)dy + 3\int_{A_2} T^p_{3,3}(y)dy, \tag{5.26}$$

and

$$\begin{aligned}\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(T^{p,b}_3 - V_3(p)\right)_+\right] \leq\ & \mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(6\int_{A_2} T^p_{3,1,1}(y)dy - V_3(p)/3\right)_+\right] \\ & + \mathbb{E}\left[\max_{m' \in \mathcal{M}_n} 6\int_{A_2} T^p_{3,1,2}(y)dy\right] \\ & + \mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(3\int_{A_2} T^p_{3,2}(y)dy - V_3(p)/3\right)_+\right] \\ & + \mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(3\int_{A_2} T^p_{3,3}(y)dy - V_3(p)/3\right)_+\right].\end{aligned}$$

The term which is not centred is direcly negligible : denoting by $m_{\max}$ the largest couple of index (maximum is taken term by term) in the collection $\mathcal{M}_n$, we remark that $T^p_{3,1,2} \leq T^{m_{\max}}_{3,1,2}$ (by (5.25)). Hence, $\mathbb{E}[\max_{m' \in \mathcal{M}_n} 6\int_{A_2} T^p_{3,1,2}(y)dy] \leq C/n$. Let us briefly study each of the other terms: first $T^p_{3,1,1}(y) = \sup_{s \in S_{p_1}, \|s\|_{L^2((0;1))}=1} \nu^2_{n,y}(s)$ (Proposition 5.2), with

$$\nu_{n,y}(s) = \frac{1}{n}\sum_{i=1}^n \pi(X_i, y)s \circ F_X(X_i) - \mathbb{E}\left[\pi(X_i, y)s \circ F_X(X_i)\right].$$

Using once more time Talagrand's Inequality (2.1) leads to

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(6\int_{A_2} T^p_{3,1,1}(y)dy - V_{3,1,1}(p)\right)_+\right] \leq \frac{C}{n}, \tag{5.27}$$

with $V_{3,1,1}(p) = 6 \times 2(1+2\delta)\|g\|^2_{L^\infty((0;1)\times A_2)} D_{p_1}/n$, ($\delta > 0$). Besides, for $n \geq n_0 = \exp(\|g\|^2_{L^\infty((0;1)\times A_2)})$,

$$V_{3,1,1}(p) \leq 12(1+2\delta)\ln(n)\frac{D_{p_1}}{n} \leq C\frac{D_{p_1}D_{p_2}}{n} := V^b_{3,1,1}(p),$$

since $D_{p_2} \geq c\ln(n)$ ($c > 0$). Inequality (5.27) holds with $V^b_{3,1,1}$. The last two terms, involving $T^m_{3,2}(y)$ and $T^m_{3,3}(y)$ can be computed with the same strategy: use the proof of the first inequality of Lemma 5.5 to bound $\int_{A_2} T^m_{3,l}(y)dy$ ($l = 2, 3$) by quantity of the form $C\|\hat{U}_n - id\|^k_{L^\infty((0;1))}$, and then apply Inequality (2.5) of Corollary 2.1. The conclusion is that

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(3\int_{A_2} T^p_{3,l}(y)dy - V_{3,l}(p)\right)_+\right] \leq C\frac{\ln(n)}{n}, \tag{5.28}$$

for $l = 2, 3$, with $V_{3,2}(p) = CD_{p_1}^4 \ln^2(n)/n^2$, and $V_{3,3}(p) = CD_{p_1}^7 \ln^3(n)/n^3$. Assuming both $n \geq n_1 = \exp(\|g\|^2)$, and $D_{p_1} = O(n^{1/3})$, $D_{p_2} \geq c \ln^3(n)$, we have

$$V_{3,2}(p) \leq C \frac{D_{p_1} D_{p_2}}{n} := V_{3,2}^b(p).$$

With the more restrictive low bound $D_{p_2} \geq c \ln^4(n)$, we also get $V_{3,3}(p) \leq CD_{p_1} D_{p_2}/n := V_{3,3}^b(p)$. As usual, Inequalities (5.28) still hold with $V_{3,l}^b$ instead of $V_{3,l}$. The proof is complete if we gather all these bounds and if we choose the constant $\kappa_3$, such that $V_3 \geq 3V_{3,1,1}^b$, $V_3 \geq 3V_{3,2}^b$, et $V_3 \geq 3V_{3,3}^b$.

□

**Proof of Lemma 5.6**

Let us first split the term $T_4^{m'}$ in several parts. Similarly to the bound obtained for $T_3^m$, we use the definitions of the estimators and their coefficients, and the fact that the basis $(\varphi_k)_k$ is orthonormal: hence

$$T_4^{m'} = \int_{A_1} \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \left( \sum_{j=1}^{D_{m_1'}} \hat{a}_{j,k}^{\hat{F}} \left( \varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x) \right) \right)^2 |(X_{-l})_l \right] f_X(x) dx.$$

We write it $T_4^{m'} \leq 2T_{4,1}^{m'} + 2T_{4,2}^{m'}$, with

$$T_{4,1}^{m'} = \int_{A_1} \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \left( \sum_{j=1}^{D_{m_1'}} \left( \hat{a}_{j,k}^{\hat{F}} - a_{j,k} \right) \left( \varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x) \right) \right)^2 |(X_{-l})_l \right] f_X(x) dx,$$

$$T_{4,2}^{m'} = \int_{A_1} \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \left( \sum_{j=1}^{D_{m_1'}} a_{j,k} \left( \varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x) \right) \right)^2 |(X_{-l})_l \right] f_X(x) dx,$$

where we denote by $a_{j,k} = \langle g, \varphi_j \otimes \varphi_k \rangle$, the Fourier coefficients of the function $g$. Then we have also $T_{4,1}^{m'} \leq 2T_{4,1,1}^{m'} + 2T_{4,1,2}^{m'}$ with the notations

$$T_{4,1,1}^{m'} = \int_{(0;1)} \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \left\{ \sum_{j=1}^{D_{m_1'}} \left( \hat{a}_{j,k}^{\hat{F}} - \mathbb{E}\left[ \hat{a}_{j,k}^{\hat{F}} |(X_{-l})_l \right] \right)^2 \right\} \left\{ \sum_{j=1}^{D_{m_1'}} \left( \varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right)^2 \right\} |(X_{-l})_l \right] du,$$

$$T_{4,1,2}^{m'} = \int_{(0;1)} \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \left\{ \sum_{j=1}^{D_{m_1'}} \left( \mathbb{E}\left[ \hat{a}_{j,k}^{\hat{F}} |(X_{-l})_l \right] - a_{j,k} \right)^2 \right\} \left\{ \sum_{j=1}^{D_{m_1'}} \left( \varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right)^2 \right\} |(X_{-l})_l \right] du.$$

As

$$\mathbb{E}\left[ T_{4,2}^{m'} \right] = \mathbb{E}\left[ \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k} a_{j',k} \int_0^1 \left( \varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right) \left( \varphi_{j'}(u) - \varphi_{j'} \circ \hat{U}_n(u) \right) du \right],$$

a Taylor formula yields $\mathbb{E}[T_{4,2}^{m'}] = \mathbb{E}[T_{4,2,1}^{m'} + T_{4,2,2}^{m'} + T_{4,2,3}^{m'}]$, where

$$T_{4,2,1}^{m'} = \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^2 \varphi_j'(u)\varphi_{j'}'(u)du,$$
$$T_{4,2,2}^{m'} = (1/4) \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^4 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_{j'}''(\hat{\alpha}_{j',n,u})du,$$
$$T_{4,2,3}^{m'} = \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^3 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_{j'}'(u)du.$$

Hence, the decomposition of the studied term is $T_4^{m'} \leq 4T_{4,1,1}^{m'} + 4T_{4,1,2}^{m'} + 2T_{4,2,1}^{m'} + 2T_{4,2,2}^{m'} + 2T_{4,2,3}^{m'}$, and consequently

$$\begin{aligned}
\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(T_4^{m'} - V_4(m')\right)_+\right] \leq\ & \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(4T_{4,1,1}^{m'} - V_4(m')/3\right)_+\right] \\
& + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(4T_{4,1,2}^{m'} - V_4(m')/3\right)_+\right] \\
& + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} \left(2T_{4,2,3}^{m'} - V_4(m')/3\right)_+\right] \\
& + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} 2T_{4,2,1}^{m'}\right] + \mathbb{E}\left[\max_{m'\in\mathcal{M}_n} 2T_{4,2,2}^{m'}\right].
\end{aligned}$$

The methods use to bound each of these terms have already been detailed for other terms: with regard to the two quantities which are not centred, we bound it to show that they are negligible (that is of order at most $C/n$). For the others, we first bound each $T_{4,l}^{m'}$ by a quantity of the form $C\|\hat{U}_n - id\|_{L^\infty((0;1))}$, and we finally apply Inequality (2.5) (Corollary 2.1), as we have already done for $T_{2,a}^m$ for example. That is why we only give the bounds for each $T_{4,l}^{m'}$. To begin, the term $T_{4,1,1}^{m'}$ can be written

$$T_{4,1,1}^{m'} = \sum_{k=1}^{D_{m_2'}} \sum_{j=1}^{D_{m_1'}} \mathrm{Var}\left(\hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l\right) \int_{(0;1)} \sum_{j=1}^{D_{m_1'}} \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u)\right)^2 du. \tag{5.29}$$

The conditional variance is

$$\begin{aligned}
\mathrm{Var}\left(\hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l\right) &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n \varphi_k(Y_i)\varphi_j \circ \hat{F}_n(X_i) | (X_{-l})_l\right\}, \\
&\leq \frac{1}{n}\mathbb{E}\left[\varphi_k(Y_1)^2 \left(\varphi_j \circ \hat{F}_n(X_1)\right)^2 | (X_{-l})_l\right].
\end{aligned}$$

By Property (5.4) applied to the sum over $j,k$ of the last quantity, $\sum_{j,k} \mathrm{Var}(\hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l) \leq D_{m_1'} D_{m_2'}/n$. Besides, we use the mean value theorem to bound the integral of (5.29) so that

$$T_{4,1,1}^{m'} \leq \frac{D_{m_1'} D_{m_2'}}{n} \times D_{m_1'}^3 \|\varphi_2'\|_{L^\infty((0;1))}^2 \left\|\hat{U}_n - id\right\|_{L^\infty((0;1))}^2,$$

which allows us to control $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(4T^{m'}_{4,1,1}-V_4(m')/3)_+]$ as explained previously. Furthermore,

$$T^{m'}_{4,1,2}=T^{m'}_3\int_{(0;1)}\sum_{j=1}^{D_{m'_1}}\left(\varphi_j(u)-\varphi_j\circ\hat{U}_n(u)\right)^2du,$$

which leads to $T^{m'}_{4,1,2}\leq T^{m'}_3D^3_{m'_1}\|\varphi'_2\|^2_{L^\infty((0;1))}\|\hat{U}_n-id\|^2_{L^\infty((0;1))}$. The term $T^{m'}_3$ is replaced by its detailed upper-bound (5.26), and as a result, $T^{m'}_{4,1,2}\leq\sum_{l=1}^4T^{m'}_{4,1,2,l}$. Roughly speaking, we get $T^{m'}_{4,1,2,l}\leq C\|\hat{U}_n-id\|_{L^\infty((0;1))}$ and apply the previous strategy for each $l=1,\dots,4$. Let us consider now the terms $T^{m'}_{4,2,1}$ and $T^{m'}_{4,2,2}$ which do not require to be centred. It is usefull to remark that the Fourier coefficients of $g$ can be written

$$a_{j,k}=\langle\xi_k,\varphi_j\rangle_{(0;1)}=\int_{(0;1)}\xi_k(u)\varphi_j(u)du,\quad\text{with }\xi_k(u)=\int_{A_2}g(u,y)\varphi_k(y)dy.$$

Since the term $T^{m'}_{4,2,1}$ involves the derivative of the projection of $\xi_k$ onto $S_{m'_1}$, we use a specific property of the trigonometric basis: $\sum_{j=1}^{D_{m'_1}}a_{j,k}\varphi'_j=\left(\Pi_{S_{m'_1}}(\xi_k)\right)'=\Pi_{S_{m'_1}}(\xi'_k)$, so

$$T^{m'}_{4,2,1}\quad\leq\quad\left\|\hat{U}_n-id\right\|^2_{L^\infty((0;1))}\sum_{k=1}^{D_{m'_2}}\left\|\xi'_k\right\|^2_{L^2((0;1))}.$$

Let us compute then the derivative of $\xi_k$ to bound roughly

$$\sum_{k=1}^{D_{m'_2}}\left\|\xi'_k\right\|^2_{L^2((0;1))}\quad=\quad\sum_{k=1}^{D_{m'_2}}\int_{(0;1)}\left(\int_{A_2}\partial_1g(u,y)\varphi_k(y)dy\right)^2du\leq\int_{(0;1)}\|\partial_1g(u,.)\|^2_{A_2}du=\|\partial_1g\|^2.$$

We thus have $\mathbb{E}[\max_{m'\in\mathcal{M}_n}T^{m'}_{4,2,1}]\leq\|\partial_1g\|^2\mathbb{E}[\|\hat{U}_n-id\|^2_{L^\infty((0;1))}]\leq C/n$ with Proposition 2.4. Recall now that

$$T^{m'}_{4,2,2}=(1/4)\sum_{k=1}^{D_{m'_2}}\int_0^1(u-\hat{U}_n(u))^4\left(\sum_{j=1}^{D_{m'_1}}a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)^2du.$$

We introduce $\mu_j=j$ for even $j$ and $\mu_j=j-1$ for odd $j$. Since $g$ belongs to $W^2_{per}((0;1)^2,L,(1,0))$ and according to (5.12)

$$\begin{aligned}\sum_{k=1}^{D_{m'_2}}\left(\sum_{j=1}^{D_{m'_1}}a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)^2\quad&\leq\quad\left\|\varphi''_2\right\|^2_{L^\infty((0;1))}\sum_{k=1}^{D_{m'_2}}\left(\sum_{j=1}^{D_{m'_1}}|a_{j,k}|\,\mu_j^2\right)^2,\\&\leq\quad\left\|\varphi''_2\right\|^2_{L^\infty((0;1))}\sum_{k=1}^{D_{m'_2}}\sum_{j=1}^{D_{m'_1}}a^2_{j,k}\mu_j^2\sum_{j=1}^{D_{m'_1}}\mu_j^2,\\&\leq\quad\left\|\varphi''_2\right\|^2_{L^\infty((0;1))}\frac{L^2}{\pi^2}D^3_{m'_1}\leq CD^3_{m_{1,\max}}.\end{aligned}$$

Hence, $\mathbb{E}[\max_{m' \in \mathcal{M}_n} T^{m'}_{4,2,2}] \leq \mathbb{E}[\|\hat{U}_n - id\|^4_{L^\infty((0;1))}] C D^3_{m_{1,\max}} \leq C D^3_{m_{1,\max}}/n^2 \leq C/n$ as soon as $D_{m_{1,\max}} \leq n^{1/3}$ (we denote by $D_{m_{1,\max}}$ the largest index on the collection $(D_{m_1})$). Following the same sketch for the last term, we write

$$T^{m'}_{4,2,3} = \int_{(0;1)} (u - \hat{U}_n)^3 \sum_{k=1}^{D_{m'_2}} \left( \sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi''_j(\hat{\alpha}_{j,n,u}) \right) \left( \sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi'_j(u) \right).$$

and compute like the term $T^{m'}_{4,2,2}$:

$$\sum_{k=1}^{D_{m'_2}} \left( \sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi''_j(\hat{\alpha}_{j,n,u}) \right)^2 \leq \|\varphi''_2\|^2_{L^\infty((0;1))} \frac{L^2}{\pi^2} D^3_{m'_1}, \quad \sum_{k=1}^{D_{m'_2}} \left( \sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi'_j(u) \right)^2 \leq \|\varphi'_2\|^2_{L^\infty((0;1))} \frac{L^2}{\pi^2} D_{m'_1}.$$

This leads to

$$T^{m'}_{4,2,3} \leq \|\varphi'_2\|_{L^\infty((0;1))} \|\varphi''_2\|_{L^\infty((0;1))} \frac{L^2}{\pi^2} D_{m'_1} \left\| \hat{U}_n - id \right\|^3_{L^\infty((0;1))},$$

and we apply tools already used to complete the proof.

□

### 5.5.4 Proof of Corollary 5.1

Start with Inequality (5.15). The bias term which appears in the right-hand side is bounded with Inequality (5.13). We have to compute the quantity $\min_{m_1,m_2}\{D^{-2\alpha_1}_{m_1} + D^{-2\alpha_2}_{m_2} + D_{m_1} D_{m_2}/n\} + C/n$. Define the bivariate function

$$\psi : (x, y) \mapsto x^{-2\alpha_1} + y^{-2\alpha_2} + \frac{xy}{n},$$

which is $\mathcal{C}^\infty$ on $(\mathbb{R}^*_+)^2$. Precisely, compute the gradient:

$$\nabla\psi(x, y) = {}^t\left( -2\alpha_1 x^{-2\alpha_1 - 1} + \frac{y}{n}; -2\alpha_2 y^{-2\alpha_2 - 1} + \frac{x}{n} \right).$$

The only point of $(\mathbb{R}^*_+)^2$ such that $\nabla\psi(x_0, y_0) = 0$ is

$$(x_0, y_0) = \left( C_\alpha n^{\frac{\alpha_2}{\alpha_1 + \alpha_2 + 2\alpha_1\alpha_2}}; 2\alpha_1 C_\alpha^{-2\alpha_1 - 1} n^{\frac{\alpha_1}{\alpha_1 + \alpha_2 + 2\alpha_1\alpha_2}} \right),$$

with $C_\alpha = [(2\alpha_1)^{2\alpha_2 + 1}/(2\alpha_2)]^{1/(2\alpha_1 + 2\alpha_2 + 4\alpha_1\alpha_2)}$. We now check that there is a minimum for $\psi$ on this point. It is sufficient to show that the Hessian matrix is positive. For any $(x, y) \in (\mathbb{R}^*_+)^2$, the matrix can be written

$$H\psi(x, y) = \begin{pmatrix} 2\alpha_1(2\alpha_1 + 1)x^{-2\alpha_1 - 2} & \frac{1}{n} \\ \frac{1}{n} & 2\alpha_2(2\alpha_2 + 1)y^{-2\alpha_2 - 2} \end{pmatrix},$$

and thus its determinant is $\det(H\varphi(x, y)) = 4\alpha_1\alpha_2(2\alpha_1 + 1)(2\alpha_2 + 1)x^{-2\alpha_1 - 2} y^{-2\alpha_2 - 2}$. Especially,

$$\det(H\psi(x_0, y_0)) = \frac{1}{n^2}\left[(2\alpha_1 + 1)(2\alpha_2 + 1) - 1\right] > 0,$$

which justifies $\psi$ is minimum on $(x_0, y_0)$, with order

$$\varphi(x_0, y_0) \propto n^{\frac{-2\alpha_1\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}}.$$

□

## 5.6 Appendix 1: What about a penalization strategy?

In this section we would like to discuss the way we perform the model selection device to choose an index $m \in \mathcal{M}_n$, and thus an estimator in the collection $(\hat{\pi}_m^{\hat{F},\hat{F}})_{m\in\mathcal{M}_n}$ in Section 5.2.3. The method we use is inspired by Goldenshluger & Lepski (2011a). We could have also chosen the index with a penalization of the contrast, in a similar way than the one use in Chapter 3 to define $\tilde{s}_1^{F_X}$ and $\tilde{s}_1^{\hat{F}}$. For the estimation settings of Chapter 3, we have shown that the two methods are equivalent. The same oracle-type inequality was proved for the two estimators (see Theorem 3.2).

However, to estimate the conditional density, our feeling is that the classical model selection device (penalized contrast) may not completely amount to the Goldenshluger-Lepski method. We did not succeed in proving the main theoretical result (Theorem 5.1) for a penalized contrast, with the same assumptions. A similar non-asymptotic risk bound can be proven, but with more restrictive conditions.

Define $\tilde{m} = \text{argmin}_{m\in\mathcal{M}_n}\{\gamma_n(\hat{g}_m^{\hat{F}}, \hat{F}_n) + \text{pen}(m)\}$, with, for a numerical constant $c(\|g\|_{L^\infty((0;1)\times A_2)})$ which depends on $\|g\|_{L^\infty((0;1)\times A_2)}$,

$$\text{pen} : m = (m_1, m_2) \in \mathcal{M}_n \mapsto c(\|g\|_{L^\infty((0;1)\times A_2)})\frac{D_{m_1}D_{m_2}}{n}. \tag{5.30}$$

The penalized contrast estimator of the conditional density is $\tilde{\tilde{\pi}}(x, y) = \hat{g}_{\tilde{m}}^{\hat{F}}(\hat{F}_n(x), y)$. We have $\gamma_n(\hat{g}_m^{\hat{F}}, \hat{F}_n) = -\sum_{j=1}^{D_m}(\hat{a}_j^{\hat{F}})^2$, thus $\tilde{m}$ can be easily computed.

Under the assumptions of Theorem 5.1, but with the restrictive conditions

$$\forall m = (m_1, m_2) \in \mathcal{M}_n,\ D_{m_1} \leq C_a n^{1/6} \text{ and } C_b \ln^5(n) \leq D_{m_2} \leq C_c\sqrt{n}, \tag{5.31}$$

instead of (5.14), the estimator $\tilde{\tilde{\pi}}$ also satisfies Inequality (5.15). However, notice that the penalty pen defined by (5.30) depends on the unknown $\|g\|_{L^\infty((0;1)\times A_2)}$. It can be replaced by an estimator (see Section 3.7 of Chapter 3 above and especially Theorem 6.2 of Chapter 6), but the price to pay is additional assumptions.

Let us explain why we have these differences between the two methods. As we have seen in the proofs of the oracle-type inequalities for the model selection estimators (built with penalization or with the Goldenshluger-Lepski method) of this thesis, the main challenge is always to find upper-bounds for the supremum of some empirical processes over some unit balls of finite dimensional subspaces of $L^2$: see the proofs of Theorems 3.1 and 3.2 for the

estimation of a regression function (Chapter 3) or the proof of Theorem 5.1 for the conditional density estimation. The quantity has the form $\sup_{t\in\mathcal{S}} \nu_n(t)$ (where $\nu_n$ is the centred empirical process, and $\mathcal{S}$ the unit ball).

The difference between the two selection methods is the following. We both use the notations of this Chapter (see Section 5.2.1) and the notations of Chapter 3 (Section 3.2.1).

- In the penalization device, $\mathcal{S}$ is the unit ball of a space which has the schematic form $S_{\hat{m}} + S_m$. When we deal with function of one variable, such as the regression function of Chapter 3, we have $S_{\hat{m}} + S_m \subset S_{m\vee\hat{m}}$, by assuming that the models are nested, with

$$\dim(S_m + S_{\hat{m}}) = D_{m\vee\hat{m}} \leq D_m + D_{m'}.$$

  The argument is used in Section 3.5.2. When we deal with a multivariate problem, such as the conditional density estimation in this chapter, the sum $\mathbb{S}_m + \mathbb{S}_{\hat{m}}$ can also be written

$$\begin{aligned} \mathbb{S}_m + \mathbb{S}_{\hat{m}} &= S_{m_1} \times S_{m_2} + S_{\hat{m}_1} \times S_{\hat{m}_2}, \\ &\subset (S_{m_1} + S_{\hat{m}_1}) \times (S_{m_2} + S_{\hat{m}_2}), \\ &= S_{m_1\vee\hat{m}_1} \times S_{m_2\vee\hat{m}_2} = \mathbb{S}_{m\vee\hat{m}}, \end{aligned}$$

  if the models are nested in each of the directions. However, the dimension of this last model is

$$\dim(\mathbb{S}_{m\vee\hat{m}}) = D_{m_1\vee\hat{m}_1} D_{m_2\vee\hat{m}_2},$$

  which cannot be bounded by $D_{m_1} D_{m_2} + D_{\hat{m}_1} D_{\hat{m}_2}$, even if we can nevertheless bound

$$\dim(\mathbb{S}_m + \mathbb{S}_{\hat{m}}) \leq \dim(\mathbb{S}_m) + \dim(\mathbb{S}_{\hat{m}}) = D_{m_1} D_{m_2} + D_{\hat{m}_1} D_{\hat{m}_2}.$$

  Thus, the extension of the one-dimension scheme to the multivariate setting is not straightforward, and requires other tricks (see below).
- In the Goldenshluger-Lepski device, $\mathcal{S}$ is either the unit ball of a space of the collection ($S_{m'}$ in Chapter 3 for the functions of one variable or $\mathbb{S}_{m'} = S_{m'_1} \times S_{m'_2}$ for the conditional density estimation) or the unit ball of a space of form $S_{m\wedge m'}$ (one variable), or $\mathbb{S}_{m\wedge m'}$ (two-variables): in this last case, even with product spaces, the problem can also be reduced to bound the deviations of the empirical process over ONE space (see the proof of Inequality (b) of Lemma 5.1 above).

Therefore, even if similar results can be obtained with the Goldenshluger-Lepski device and with the penalization strategy for one-dimension model selection, the problem is different in the multivariate setting of conditional density estimation. As an example, we develop why the penalty depends on the norm $\|g\|_{L^\infty((0;1)\times A_2)}$, whereas its Lepski's equivalent, namely $V$ (see (5.8)) is free of it. To that aim, assume for a moment that $F_X$ is known (the replacement by the empirical counterpart does not change anything in the problem, here). The penalization device applied with the scheme of Section 3.5.2 requires to bound

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{T}_{m,m'}} \nu_n^2(t) - p(m,m')\right)_+\right],$$

with $\nu_n$ defined by (5.16), $\mathcal{T}_{m,m'} = \{t \in \mathbb{S}_m + \mathbb{S}_{m'},\ \|t\| = 1\}$, and $p(m,m')$ less than $\text{pen}(m) + \text{pen}(m')$. The definition of pen is thus imposed by the application of the Talagrand inequality (Proposition 2.2). This is similar but quite different to the proof of Lemma 5.1 above (see Section 5.5.1). Let us see what differs, by computing $M_1$, $H$ and $v$ to apply the concentration inequality (2.1). First, notice that the space $\mathbb{S}_m + \mathbb{S}_{m'}$ does not belong to the collection $(\mathbb{S}_m)_m$ but is a subspace of $L^2((0;1) \times A_2))$. Denote by $(\chi_{j,k})_{j,k \in \mathcal{I}_{m,m'}}$ an orthonormal basis of this space. The cardinality of $\mathcal{I}_{m,m'}$ is the dimension of $\mathbb{S}_m + \mathbb{S}_{m'}$, and is thus bounded by $D_{m_1} D_{m_2} + D_{m_1'} D_{m_2'}$. But we also have $\mathbb{S}_m + \mathbb{S}_{m'} \subset \mathbb{S}_{m \vee m'}$, with dimension $D_{m_1 \vee m_1'} D_{m_2 \vee m_2'}$. First, the quantity $v$ is the same as in Section 5.5.1. Then, the computation of $M_1$ is similar to Section 5.5.1:

$$\|t(F_X(.),.)\|_{L^\infty(A_1 \times A_2)} \quad \leq \quad \|t\|_{L^\infty((0;1)\times A_2)} \leq \sqrt{D_{m_1 \vee m_1'} D_{m_2 \vee m_2'}},$$

by using the norm connexion property (5.4). With $D_{m_l} \leq \sqrt{n}$ $(l = 1, 2)$, we set $M_1 = \sqrt{n}$. The main changes comes from $H^2$: if a function $t$ belongs to $\mathcal{T}_{m,m'}$, we write it

$$t = \sum_{(j,k) \in I_{m,m'}} \theta_{j,k} \chi_{j,k}, \quad \text{with} \sum_{(j,k) \in I_{m,m'}} \theta_{j,k}^2 = 1.$$

We thus have $\nu_n^2(t) \leq \sum_{(j,k) \in I_{m,m'}} \nu_n^2(\chi_{j,k})$. Then,

$$\begin{aligned}
\mathbb{E}\left[\nu_n^2\left(\chi_{j,k}\right)\right] &= \frac{1}{n} \text{Var}\left(\chi_{j,k}\left(F_X(X_1), Y_1\right)\right) \leq \frac{1}{n} \mathbb{E}\left[\chi_{j,k}^2\left(F_X(X_1), Y_1\right)\right], \\
&= \frac{1}{n} \int_{(0;1)\times A_2} \chi_{j,k}(u,y)^2 \pi\left(F_X^{-1}(u), y\right) du dy, \\
&\leq \frac{\|g\|_{L^\infty((0;1)\times A_2)}}{n} \|\chi_{j,k}\|^2 = \frac{\|g\|_{L^\infty((0;1)\times A_2)}}{n}.
\end{aligned}$$

Thus $\mathbb{E}[\sup_{t \in \mathcal{T}_{m,m'}} \nu_n^2(t)] \leq \text{card}(I_{m,m'}) \|g\|_{L^\infty((0;1)\times A_2)}/n$, and

$$H^2 = \|g\|_{L^\infty((0;1)\times A_2)} \frac{D_{m_1} D_{m_2} + D_{m_1'} D_{m_2'}}{n}.$$

Therefore, we see that we cannot apply Property (5.4) contrary to what we have done in Section 5.5.1. An infinite norm appears here, which explains the definition (5.30), and thus the advantage of the Goldenshluger-Lepski method in this setting (recall also that the infinite norm is also involved in the procedure of Brunel *et al.* 2007).

## 5.7 Appendix 2: Warped kernel for conditional density estimation

### 5.7.1 Goal of this section

The aim of this section is to show that the "warping" device can be also applied to recover the conditional density with kernel estimator. We thus consider an extension of the warped-kernel method developed in Chapter 4 to estimate a bivariate function. Assume again that

we observe pairs $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ and try again to estimate the conditional density $\pi$ defined by (5.1).

Kernel estimators for $\pi$ have been widely studied, as we said in Section 5.1: typically, Nadaraya-Watson type estimates, such as "double-kernel" ratio estimators, with cross-validation methods to select bandwidths are studied from the asymptotic point of view (convergence rates and asymptotic normality are shown): among others, see Hyndman *et al.* (1996), Hyndman & Yao (2002), De Gooijer & Zerom (2003) or Fan & Yim (2004).

We propose to build a warped-kernel estimate for the conditional density $\pi$, which is the analogous version of the warped-bases estimator $\tilde{\pi}$. To adapt the method of Chapter 4, we only warp the first coordinate of $\pi$, by using the c.d.f. $F_X$ of the design $X$. The auxiliary function is still $g$, defined by (5.2). For the sake of clarity, we describe the procedure by assuming $F_X$ known, and build an estimate which is a kernel version of the estimator $\tilde{\pi}_0$ defined by (5.11). However, the substitution of $\hat{F}_n$ to $F_X$ can be done as usual, in a second step.

Let us introduce some notations. We consider two kernel functions $K^{(1)}$ and $K^{(2)}$, which are supposed to be squared-integrable on $\mathbb{R}$. From this point, $\mathcal{H}_n^{(cd)}$ denotes a set of bandwidth couples $(h_1, h_2) \in (\mathbb{R}_+^*)^2$. We set again $K_{h_l}^{(l)} : x \mapsto K^{(l)}(x/h_l)/h_l$ $(l = 1, 2)$, and denote by $\mathbb{K}(u, y) = K_{h_1}^{(1)} \otimes K_{h_2}^{(2)}(u, y)$ the product $K_{h_1}^{(1)}(u)K_{h_2}^{(2)}(y)$, for all real numbers $u$ and $y$. In this bivariate framework, Assumptions (H2) and (H3) of Section 4.3.2 become:

(H2') There exists $\alpha_0 > 0$ such that $\displaystyle\sum_{h\in\mathcal{H}_n} \frac{1}{h_1 h_2} \leq k_0 n^{\alpha_0}$, for a constant $k_0 \geq 0$.

(H3') For all $\kappa_0 > 0$, there exists $C_0 > 0$, such that $\displaystyle\sum_{(h_1,h_2)\in\mathcal{H}_n} \exp\left(-\frac{\kappa_0}{h_1 h_2}\right) \leq C_0$.

### 5.7.2 Estimation and performance

The cornerstone of the method is to remark that the auxiliary $g$ defined by (5.2) is the density of the transformed data $(F_X(X), Y)$. Thus, a collection of kernel estimators for $g$ is

$$\hat{g}_{h_1,h_2}^{(cd)} : (u, y) \mapsto \frac{1}{n}\sum_{i=1}^{n} K_{h_1}^{(1)}\left(u - F_X(X_i)\right) K_{h_2}^{(2)}\left(y - Y_i\right), \quad (h_1, h_2) \in \mathcal{H}_n^{(cd)},$$

and the analogous collection for $\pi$ is $(\hat{\pi}_{h_1,h_2})_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}}$, with $\hat{\pi}_{h_1,h_2}(x, y) = \hat{g}_{h_1,h_2}^{(cd)}(F_X(x), y)$. Stute (1986a) studied similar estimators for a conditional distribution function. More recently, this collection has already been considered by Mehra *et al.* (2000), who compared it asymptotically to classical Nadaraya-Watson estimates.

The novelty which must be underlined here is the Goldenshluger-Lepski selection of the best bandwidths $(\hat{h}_1, \hat{h}_2)$: we set, in the same way as (4.12) and (4.13) (Chapter 4), and for any $(h_1, h_2) \in \mathcal{H}_n^{(cd)}$,

$$\begin{cases} V^{(cd)}(h_1, h_2) = \delta'(1 + \|\mathbb{K}\|_{L^1(\mathbb{R})}^2)\dfrac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}}{n h_1 h_2} \\ A^{(cd)}(h_1, h_2) = \displaystyle\max_{(h_1',h_2')\in\mathcal{H}_n} \left\{\|\hat{g}_{(h_1,h_2),(h_1',h_2')} - \hat{g}_{h_1',h_2'}\|^2 - V^{(cd)}(h_1', h_2')\right\}_+ \end{cases}$$

with

$$\hat{g}_{(h_1,h_2),(h_1',h_2')} \ : \ (u,y) \mapsto K_{h_1'}^{(1)} \otimes K_{h_2'}^{(2)} \star \left(\hat{g}_{h_1,h_2} \mathbf{1}_{(0;1)\times A_2}\right)(u,y).$$

To realize the bias-variance compromize, we define:

$$(\hat{h}_1, \hat{h}_2) = \arg \min_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}} \{A^{(cd)}(h_1,h_2) + V^{(cd)}(h_1,h_2)\}.$$

We now set the oracle-type inequality, concerning the selected estimator $\hat{\pi}_{\hat{h}_1,\hat{h}_2}$, which is the analogous of $\tilde{\pi}_0$ defined by (5.11).

**Theorem 5.2.** *Assume that $\pi$ is bounded, and that Assumptions* (H2') *and* (H3') *holds for the collection $\mathcal{H}_n^{(cd)}$. Then, there exists three constants $\kappa_l$, $l = 1,2,3$ such that*

$$\mathbb{E}\left[\left\|\hat{\pi}_{\hat{h}_1,\hat{h}_2} - \pi\right\|_{f_X}^2\right] \leq \min_{(h_1,h_2)\in\mathcal{H}_n^{(cd)}} \left\{ c_1 \left\|g - g_{h_1,h_2}^{(cd)}\right\|_{f_X}^2 + c_2 \frac{\|K^{(1)} \times K^{(2)}\|_{L^2(\mathbb{R}^2)}^2}{nh_1h_2} \right\} + \frac{c_3}{n}, \tag{5.32}$$

*where $g_{h_1,h_2}^{(cd)} = (K_{h_1}^{(1)} \otimes K_{h_2}) \star (g\mathbf{1}_{(0;1)\times A_2})$, and where the $\kappa_l$ depend on $\|K^{(1)} \times K^{(2)}\|_{L^1(\mathbb{R}^2)}$ ($l = 1,2,3$) and $\kappa_3$ additionally depends on $\|g\|_{L^\infty((0;1)\times A_2)}$.*

Therefore, the warped kernel strategy could be successfully adapted to a bivariate framework. The crucial choice of the bandwidth is automatically performed: thanks to the Goldenshluger-Lepski method, the optimal trade-off is reached. Therefore, we extend the results of Mehra *et al.* (2000) about warped kernel conditional density estimator.

Moreover, it should be mentioned that the estimator admits a simple expression, and can be consequently implemented with low complexity, like in the four univariate examples of Chapter 4.

Finally, Inequality (5.32) can be used for derivation of adaptive optimal rates for conditional density estimation. For that purpose, we need to assume that the kernels $K^{(1)}$ and $K^{(2)}$ have vanishing moment property (such as Assumption $(K_l)$), and the convergence rate is established over anisotropic Hölder classes (defined for example in Section 2.4 of Comte & Lacour 2013). We do not detail this, since the main goal of this section is to show the adaptation of warped-bases strategy to establish non-asymptotic bound for conditional density estimation.

### 5.7.3 Proof of Theorem 5.2

The arguments and the sketch of the proof are exactly the same as those used to prove Theorem 4.1 of Chapter 4. It follows thus exactly the sketch of Section 4.6.4. We first obtain an inequality equivalent to (4.18):

$$\mathbb{E}\left[\left\|\hat{\pi}_{\hat{h}_1,\hat{h}_2} - \pi\right\|_{f_X}^2\right] \leq 6\mathbb{E}\left[A^{(cd)}(h_1,h_2)\right] + 6V^{(cd)}(h_1,h_2) + \frac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}^2}{nh_1h_2} + 3\|g_{h_1,h_2}^{(cd)} - g\|^2.$$

Similarly, we must bound $A^{(cd)}$: we use the splitting (4.20). Then, bounding $A^{(cd)}$ amounts to control the deviation of the following centered empirical process:

$$\nu^{(cd)}_{n,h_1,h_2}(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_{(0;1)\times A_2} K^{(1)}_{h_1}\left(u - F_X(X_i)\right) K^{(2)}_{h_2}\left(y - Y_i\right) t(u,y)dudy - \mathbb{E}\left[\int_{(0;1)\times A_2} K^{(1)}_{h_1}\left(u - F_X(X_i)\right) K^{(2)}_{h_2}\left(y - Y_i\right) t(u,y)dudy\right]\right\}.$$

Precisely, we apply the Talagrand Inequality (Proposition 2.2, Chapter 2) with the following quantity:

$$M_1 = \frac{\|\mathbb{K}\|_{L^2(\mathbb{R}^2)}}{\sqrt{h_1 h_2}}, \quad H^2 = \frac{\|\mathbb{K}\|^2_{L^2(\mathbb{R}^2)}}{nh_1h_2}, \quad v = \|g\|_{L^\infty((0;1)\times A_2)}\|\mathbb{K}\|^2_{L^1(\mathbb{R}^2)}.$$

This proves that

$$\sum_{(h_1,h_2)\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}^{(cd)}(0,1)}\left(\nu^{(cd)}_{n,h_1,h_2}(t)\right)^2 - V^{(cd)}(h_1,h_2)\right)_+\right] \leq \frac{C}{n},$$

which is the key point of the proof.

□

Deuxième partie

# Comparaison de la loi de deux échantillons : estimation de la densité relative

# Chapitre 6

# Sélection de modèles pour l'estimation de la densité relative.

**Sommaire**



Ce chapitre est une version modifiée de l'article *Model selection for relative density estimation*, (bientôt) soumis pour publication.

**Résumé.** L'objectif de ce travail est de proposer un estimateur adaptatif pour une fonction récemment utilisée dans les problèmes à deux échantillons, la densité relative. Cette fonction, outil pour la comparaison des distributions de deux variables $X$ et $X_0$, est définie comme la densité de $F_0(X)$, où $F_0$ est la fonction de répartition de $X_0$. La technique d'estimation choisie s'inspire de la sélection de modèles : un estimateur est sélectionné automatiquement sur la bases des observations par un critère adapté des travaux de Goldenshluger et Lepski (2011), à partir d'une collection d'estimateurs par projection. Le compromis biais-variance est réalisé, et une borne non asymptotique pour le risque quadratique intégré établie. Des vitesses de convergence sont également démontrées. La méthode est illustrée par des simulations.

**Abstract.** The aim of this work is to propose an adaptive estimator for the relative density, a function recently used in two-sample problems. This function is useful to compare the distributions of two variables $X$ and $X_0$. It is defined as the density of the variable $F_0(X)$, where $F_0$ is the cumulative distribution function of $X_0$. The estimation device is inspired by model selection: an estimator is automatically selected from the data, with a criterion adapted from the work of Goldenshluger and Lepski (2011), in a collection of projection estimators. The global squared-bias/variance compromize is realized, and a non-asymptotic risk bound for the mean integrated squared error is proved. Convergence rates are also deduced. Short simulation experiments illustrate the method.

## 6.1 Introduction

The study of differences among groups is the main challenge of two-sample problems, and statistical methods are required to do this in various fields (biology or social research for example). Nonparametric inference procedures are well-developed for comparing samples coming from two populations, modeled by two real random variables $X_0$ and $X$. Most of the methods are based on the comparison of the cumulative distribution functions (c.d.f. in the sequel) $F_0$ and $F$ of $X_0$ and $X$ respectively. The study of the relative density $r$ of $X$ with respect to $X_0$ is quite recent. Assume that $f_0$, the density of $X_0$, does not vanish on its support $A_0$, and denote by $F_0^{-1}$ the inverse of $F_0$. The relative density is defined as the density of the variable $F_0(X)$ and can be expressed as

$$r(x) = \frac{f \circ F_0^{-1}(x)}{f_0 \circ F_0^{-1}(x)}, \quad x \in F_0(A), \tag{6.1}$$

where $\circ$ is the composition symbol, $f$ is a density of $X$, and $A \subset \mathbb{R}$ its support. In the present work, we focus on the adaptive estimation of this function, from two independent samples $(X_i)_{i\in\{1,\dots,n\}}$ and $(X_{i_0})_{i_0\in\{1,\dots,n_0\}}$ of variables $X$ and $X_0$.

The most classical nonparametric methods to tackle the initial issue of the comparison of $F$ and $F_0$ are statistical tests such as Kolmogorov and Smirnov, Wilcoxon, or Mann and Whitney tests, which all propose to check the null hypothesis of equal c.d.f.. Another usual tool is the Receiving Operating Characteristic (ROC) curve, which can be defined as the c.d.f. of the variable $1 - F_0(X)$ and is well-known in fields such as signal detection and diagnostic test for example. Estimators for this function based on the empirical c.d.f. of $X$ and $X_0$ have been proposed (see Hsieh & Turnbull 1996b,a and references therein), as well as kernel smoothers (see among all Lloyd 1998, Lloyd & Yong 1999, Hall & Hyndman 2003). The relative distribution approach has been developed at the same time: the estimation of the relative distribution function, the c.d.f. of $F_0(X)$ has first been exploited. The same kind of estimators (kernel and empirical) were studied, in the context of applications in social sciences (Gastwirth, 1968; Hsieh, 1995; Handcock & Morris, 1999). The relative density defined by (6.1) is the derivative of this function. A discussion about its advantages, compared to the ROC curve can be found in Molanes-López & Cao (2008b) (page 694). Early references for its definition are Bell & Doksum (1966) and Silverman (1978), who approached the problem with the maximum likelihood point of view. A kernel estimate was first proposed by Ćwik & Mielniczuk (1993), and modified by Molanes-López & Cao (2008a) who proved asymptotic developments for the Mean Integrated Squared Error (MISE), under the assumption that $r$ is twice continuously derivable (see also Handcock & Janssen 2002). The problem of bandwidth selection is also adressed, but few theoretical results are proved for the estimators with the selected parameters. The question has also been studied in a semiparametric setting (see Cheng & Chu 2004 and references therein). Our work is the first to study a nonparametric projection estimators and provide a theoretical study of the adaptive estimator.

We build in this chapter a new estimator of the relative density, based on the minimization of a contrast inspired by the classical density contrast (Massart, 2007). A collection of projection estimators on linear models is built in Section 6.2, and the quadratic risk is studied: the upper-bound is non trivial, and requires non straightforward splittings. We obtain a bias-variance decomposition which permits to understand what we can expect at best from adaptive estimation, which is the subject of Section 6.3: the model selection is automatically performed in the spirit of the Goldenshluger-Lepski method in a data-driven way (Goldenshluger & Lepski, 2011a). Both non-asymptotic and asymptotic results are derived: an oracle-type inequality shows that adaptation has no cost, and rates of convergence are deduced, for functions $r$ belonging to Besov balls. Such results are new for this estimation problem: especially, no assumption about a link between the sample sizes $n$ and $n_0$ is required, and the regularity assumptions are not restrictive. Section 6.4 provides brief illustrations. Finally, the proofs are gathered in Section 6.5.

## 6.2 The collection of projection estimators

For the sake of clarity, we assume that the variables $X$ and $X_0$ has the same support: $A = A_0$. Hence, $F_0(A) = (0;1)$ is the estimation interval. This assumption is natural to compare the distribution of $X$ to the one of $X_0$.

### 6.2.1 Approximation spaces

We denote by $L^2((0;1))$, the space of square integrable functions on $(0;1)$, equipped with its usual Hilbert structure: $\langle .,. \rangle$ is the scalar product, and $\|.\|$ the associated norm. The relative density $r$, defined by (6.1) and estimated on its definition set $(0;1)$ is assumed to belong to $L^2((0;1))$. Our estimation procedure is based on this device: we consider a family $S_m$, $m \in \mathcal{M}_{n,n_0}$ of finite dimensional subspaces of $L^2((0;1))$ and compute a collection of estimators $(\hat{r}_m)_{m\in\mathcal{M}_{n,n_0}}$, where, for all $m$, $\hat{r}_m$ belongs to $S_m$. In a second step a data driven procedure chooses among the collection the final estimator $\hat{r}_{\hat{m}}$.

Here, simple projection spaces are considered, namely trigonometric spaces: the set $S_m$ is linearly spanned by $\varphi_1, \ldots, \varphi_{2m+1}$, with $\varphi_1(x) = 1$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$, $\varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx)$, for $x \in (0;1)$. This basis is also used for many nonparametric estimation problems, by several authors (see e.g. Efromovich 1999 among all). We set $D_m = 2m+1$, the dimension of $S_m$, and $\mathcal{M}_{n,n_0} = \{1, 2, \ldots, [(n \wedge n_0)]/2 - 1\}$, the collection of indices. The largest space in the collection has maximal dimension $D_{m_{\max}}$, which is subject to constraints appearing later. Note that, for all $x \in (0;1)$, $\sum_{j=1}^{D_m} \varphi_j^2(x) = D_m$. Thus, for any function $t \in S_m$, $\|t\|_{L^\infty((0;1))} := \sup_{x\in(0;1)} |t(x)| \leq \sqrt{D_m}\|t\|$.

### 6.2.2 Estimation on a fixed model

For each index $m \in \mathcal{M}_{n,n_0}$, and for a function $t \in S_m$, the following contrast function is introduced

$$\gamma_n(t, \hat{F}_{0,n_0}) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t \circ \hat{F}_{0,n_0}(X_i), \tag{6.2}$$

where $\hat{F}_{0,n_0}$ is the empirical c.d.f. of the sample $(X_{0,i_0})_{i_0=1,\dots,n_0}$, that is

$$\hat{F}_{0,n_0} : x \mapsto \frac{1}{n_0}\sum_{i_0=1}^{n_0} \mathbf{1}_{X_{0,i_0}\leq x}.$$

It is built as an empirical version of the $L^2-$distance between a function $t \in S_m$ and the function of interest $r$. To see this, compute

$$\mathbb{E}\left[\gamma_n(t, F_0)\right] = \|t\|^2 - 2\mathbb{E}\left[t \circ F_0(x)\right] = \|t\|^2 - 2\int_A t \circ F_0(x) f(x) dx.$$

Make the change of variables $u = F_0(x)$ and recall that $F_0(A) = (0;1)$, to obtain

$$\int_A t \circ F_0(x) f(x) dx = \int_{F_0(A)} t(u) \frac{f \circ F_0^{-1}(u)}{f_0 \circ F_0^{-1}(u)} du = \langle t, r\rangle, \tag{6.3}$$

and thus $\mathbb{E}[\gamma_n(t, F_0)] = \|t - r\|^2 - \|r\|^2$. This illustrates that minimizing the criterion $\gamma_n(., \hat{F}_{0,n_0})$, is likely to provide a function $t$ that minimizes in mean $\|t-r\|^2$ and thus estimates $r$ on the space $S_m$. Consequently, let, for $m \in \mathcal{M}_{n,n_0}$,

$$\hat{r}_m(., \hat{F}_{0,n_0}) = \arg\inf_{t \in S_m} \gamma_n(t, \hat{F}_{0,n_0}). \tag{6.4}$$

An explicit and unique expression follows from this definition by using the orthogonal basis $(\varphi_j)_j$ of $S_m$ described above: for $x \in (0;1)$,

$$\hat{r}_m(x, \hat{F}_{0,n_0}) = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{F}_{0,n_0}} \varphi_j(x), \quad \text{with } \hat{a}_j^{\hat{F}_{0,n_0}} = \frac{1}{n}\sum_{i=1}^{n} \varphi_j\left(\hat{F}_{0,n_0}(X_i)\right). \tag{6.5}$$

Note that in the "toy" case of known c.d.f. $F_0$, the procedure amounts to estimate a density: $\gamma_n(., F_0)$ is the density contrast defined in Massart (2007) (adapted to the estimation of the density of $F_0(X)$), and $\hat{r}_m(., F_0)$ is the density projection estimator.

### 6.2.3 Risk of a projection estimator

The global squared error is the natural criterion associated to the projection estimation procedure. We define $r_m$ as the orthogonal projection of $r$ onto the model $S_m$. It can be written $r_m = \sum_{j=1}^{D_m} \varphi_j$ with $a_j = \langle r, \varphi_j\rangle$. In the toy case of known c.d.f. $F_0$, computations similar to (6.3) give $\mathbb{E}[\hat{a}_j^{F_0}] = a_j$, and thus, the Pythagoras theorem simply leads to the classical bias-variance decomposition:

$$\|r - \hat{r}_m(., F_0)\|^2 = \|r - r_m\|^2 + \|\hat{r}_m(., F_0) - r_m\|^2. \tag{6.6}$$

Moreover, the variance term can be easily bounded:

$$\mathbb{E}\left[\|\hat{r}_m(., F_0) - r_m\|^2\right] = \sum_{j=1}^{D_m} \mathrm{Var}\left(\hat{a}_j^{F_0}\right) \leq \frac{1}{n}\sum_{j=1}^{D_m} \mathbb{E}\left[\varphi_j^2(F_0(X_1))\right] = \frac{D_m}{n}. \tag{6.7}$$

The challenge in the general case comes from the plug-in of the empirical $\hat{F}_{0,n_0}$: it seems natural but involves non straightforward computations. This is why the proof of the following upper-bound for the risk is postponed to Section 6.5.

**Proposition 6.1.** *Assume that the relative density $r$ is continuously derivable on $(0;1)$. Assume also that $D_m \leq \kappa n_0^{1/3}$, for a constant $\kappa > 0$. Then, there exist two constants $c_1$ and $c_2$ such that,*

$$\mathbb{E}\left[\left\|\hat{r}_m(.,\hat{F}_{0,n_0}) - r\right\|^2\right] \leq 3\left\|r - r_m\right\|^2 + \left(3\frac{D_m}{n} + c_1\|r\|^2\frac{D_m}{n_0}\right) + c_2\left(\frac{1}{n} + \frac{1}{n_0}\right). \quad (6.8)$$

*The constants $c_1$ and $c_2$ do not depend on $n$, $n_0$ and $m$. Moreover, $c_1$ also does not depend on $r$.*

Proposition 6.1 shows that the risk is divided into three terms: a squared-bias term, a variance term (proportional to $D_m(n^{-1} + n_0^{-1})$) and a remainder (proportional to $(n^{-1} + n_0^{-1})$). The upper bound of (6.1) is non trivial, and the proof requires tricky approximations (see Section 6.5.1 e.g.). The assumption on the model dimension $D_m$ comes from the control of the deviation between $\hat{F}_{n_0}$ and $F_0$. It can be compared to what is required to obtain the warped-estimation results: in Proposition 3.1 (regression estimation with warped-bases) or in Theorem 5.1 (conditional density estimation with warped-bases), the dimension of the model (which is warped) is also supposed to be bounded by the power 1/3 of the number of observations used to estimate the empirical counterpart for the c.d.f.

### 6.2.4 Rates of convergence on Besov balls

The result (6.8) gives also the asymptotic rate for an estimator if we consider that $r$ belongs to a Besov ball $B^{\alpha}_{p,\infty}((0;1),L)$ of radius $L$ ($p$ a nonnegative integer, $L > 0$, $\alpha > 0$), for the Besov norm $\|.\|_{\alpha,p}$ on the Besov space $\mathcal{B}^{\alpha}_{p,\infty}((0;1))$. For a precise definition of those notions, we refer to DeVore & Lorentz (1993), Chapter 2, Section 7, where it is also proved that $\mathcal{B}^{\alpha}_{p,\infty}((0;1)) \subset \mathcal{B}^{\alpha}_{2,\infty}((0;1))$ for $p \geq 2$. This justifies that we now restrict to $\mathcal{B}^{\alpha}_{2,\infty}((0;1))$. The following rate is obtained:

**Corollary 6.1.** *Assume that the relative density $r$ belongs to the Besov ball $B^{\alpha}_{2,\infty}((0;1),L)$, for $L > 0$, and $\alpha \geq 1$. Choose a model $m_{n,n_0}$ such that $D_{m_{n,n_0}} = C(n^{-1} + n_0^{-1})^{-1/(2\alpha+1)}$, for $C > 0$. Then, under the assumptions of Proposition 6.1, there exists a numerical constant $C'$ such that*

$$\mathbb{E}\left[\left\|\hat{r}_{m_{n,n_0}}(.,\hat{F}_{0,n_0}) - r\right\|^2\right] \leq C'\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

This inequality is a straightforward consequence of the result of DeVore & Lorentz (1993) and of Lemma 12 of Barron *et al.* (1999), which imply that the bias term $\|r - r_m\|^2$ is of order $D_m^{-2\alpha}$. The minimum of the right-hand side term of (6.8) can thus be computed, leading to Corollary 6.1. Heuristically, the rate is $[\min(n, n_0)]^{2\alpha/(2\alpha+1)}$. Nevertheless, it is worth noticing that it depends on the two sample sizes $n$ and $n_0$. The rate we obtain is new in nonparametric estimation, but it is not surprising. Actually, it looks like the Kolmogorov-Smirnov two-sample test convergence result: it is well-known that the test statistic rate is $\sqrt{nn_0/(n+n_0)}$ (see for example Doob 1949).

**Remark 6.1.** The regularity condition $\alpha \geq 1$ ensures that there exists a dimension $D_{m_{n,n_0}}$ which satisfies $D_m \leq Cn_0^{1/3}$ while being of order $(n^{-1} + n_0^{-1})^{-1/(2\alpha+1)}$. When $\alpha < 1$, this

choice remains possible and the convergence rate is preserved under the additional assumption $n \leq n_0/(n_0^{(2-2\alpha)/3} - 1)$. Roughly, this condition means that $n \leq n_0^{(2\alpha+1)/3} < n_0$, and thus $n$ and $n_0$ must be put in order to handle this case.

It follows from Corollary 6.1 that the optimal dimension depends on the unknown regularity $\alpha$ of the function to be estimated. The aim is to perform an adaptive selection only based on the data.

## 6.3 Adaptive estimation

### 6.3.1 Model selection

Consider the collection $(S_m)_{m\in\mathcal{M}_{n,n_0}}$ of models defined in Section 6.2.1 and the collection $(\hat{r}_m)_{m\in\mathcal{M}_{n,n_0}}$ of estimators defined by (6.4). The aim is to propose a data driven choice of $m$ leading to an estimator with risk near of the squared-bias/variance compromize (see (6.8)). The selection combines two strategies: the model selection device performed with a penalization of the contrast (see e.g. Barron *et al.* 1999), and the recent Goldenshluger-Lepski method (Goldenshluger & Lepski, 2011a). A similar device has already been used in Chapter 3 and by Comte & Johannes (2012). We set, for every index $m$,

$$\begin{aligned} V(m) &= c_0 \left( \frac{D_m}{n} + \|r\|^2 \frac{D_m}{n_0} \right), \\ A(m, \hat{F}_{0,n_0}) &= \max_{m' \in \mathcal{M}_{n,n_0}} \left( \left\| \hat{r}_{m'}(.,\hat{F}_{0,n_0}) - \hat{r}_{m\wedge m'}(.,\hat{F}_{0,n_0}) \right\|^2 - V(m') \right)_+, \end{aligned} \tag{6.9}$$

where $m \wedge m'$ is the minimum between $m$ and $m'$, and $(x)_+$ the maximum between $x$ and 0 (for a real number $x$). The quantity $V$ must be understood as a penalty term, and $A$ is an approximation of the squared-bias term (see Lemma 6.3). The estimator of $r$ is now given by $\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0})$, with

$$\hat{m} = \text{argmin}_{m\in\mathcal{M}_{n,n_0}} \{A(m, \hat{F}_{0,n_0}) + V(m)\}.$$

By construction, the choice of the index $m$, and hence the estimator $\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0})$ do not depend on the regularity assumption on the relative density $r$.

### 6.3.2 Main results

A non-asymptotic upper-bound is derived for the risk of the estimator $\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0})$.

**Theorem 6.1.** *Assume that the relative density $r$ is continuously derivable on $(0;1)$. Assume also that $D_m \leq \kappa n_0^{1/3}/\ln^{2/3}(n_0)$, for a constant $\kappa > 0$. Then, there exist two constants $c$ and $C$ such that,*

$$\mathbb{E}\left[ \left\| \hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0}) - r \right\|^2 \right] \leq c \min_{m\in\mathcal{M}_{n,n_0}} \left\{ \left( \frac{D_m}{n} + \|r\|^2 \frac{D_m}{n_0} \right) + \|r_m - r\|^2 \right\} + C\left( \frac{1}{n} + \frac{1}{n_0} \right).$$

*The constant $c$ is purely numerical, while $C$ depends on $r$, but neither on $n$ nor $n_0$.*

For every index $m \in \mathcal{M}_{n,n_0}$, $\{(D_m/n + \|r\|^2 D_m/n_0) + \|r_m - r\|^2\}$ has the same order as $\mathbb{E}\left[\|\hat{r}_m(.,\hat{F}_{0,n_0}) - r\|^2\right]$ (see Proposition 6.1). Thus Theorem 6.1 indicates that up to a multiplicative constant, the estimator $\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0})$ converges as fast as the best estimator in the collection. The convergence result is stated in the following corollary.

**Corollary 6.2.** *Assume that the relative density $r$ belongs to $B^{\alpha}_{2,\infty}((0;1),L)$, for $L>0$, and $\alpha \geq 1$. Under the assumptions of Theorem 6.1,*

$$\mathbb{E}\left[\left\|\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0}) - r\right\|^2\right] \leq C\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

It is worth noticing that the optimal rate of convergence (that is the one of the best estimator among the collection) is automatically achieved by the estimator $\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0})$.

The penalty term $V$ given in (6.9) cannot be used in practice, since it depends on the unknown quantity $\|r\|^2$. A solution is to replace it by an estimator, and to prove that the estimator of $r$ built with this random penalty keeps the adaptation property. To that aim, set, for an index $m^* \in \mathcal{M}_{n,n_0}$,

$$\begin{aligned}
\tilde{V}(m) &= c_0\left(\frac{D_m}{n} + 4\|\hat{r}_{m^*}(.,\hat{F}_{0,n_0})\|^2\frac{D_m}{n_0}\right), \\
\tilde{A}(m,\hat{F}_{0,n_0}) &= \max_{m'\in\mathcal{M}_{n,n_0}}\left(\left\|\hat{r}_{m'}(.,\hat{F}_{0,n_0}) - \hat{r}_{m\wedge m'}(.,\hat{F}_{0,n_0})\right\|^2 - \tilde{V}(m')\right)_+,
\end{aligned} \tag{6.10}$$

and $\tilde{m} = \operatorname{argmin}_{m\in\mathcal{M}_{n,n_0}}\{\tilde{A}(m,\hat{F}_{0,n_0}) + \tilde{V}(m)\}$. The result for $\hat{r}_{\tilde{m}}$ is described in the following theorem.

**Theorem 6.2.** *Assume that the assumptions of Theorem 6.1 are satisfied, and that $r$ belongs to $B^{\alpha}_{2,\infty}((0;1),L)$, for $L>0$, and $\alpha \geq 1$. Choose $m^*$ in the definition of $\tilde{V}$ such that $D_{m^*} \geq \ln(n_0)$ and $D_{m^*} = O(n^{1/4}/\ln^{1/4}(n))$. Then, for $n_0$ large enough, there exist two constants $c$ and $C$ such that,*

$$\mathbb{E}\left[\left\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\right\|^2\right] \leq c\min_{m\in\mathcal{M}_{n,n_0}}\left\{\left(\frac{D_m}{n} + \|r\|^2\frac{D_m}{n_0}\right) + \|r_m - r\|^2\right\} + C\left(\frac{1}{n} + \frac{1}{n_0}\right).$$

## 6.4 Simulation

In this section, we briefly present the performance of the estimator $\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})$ on simulated data.

### 6.4.1 Implementation

The implementation of the estimator is very simple, and follows the steps below.

- For each $m \in \mathcal{M}_{n,n_0}$, compute $(\hat{r}_m(x_k,\hat{F}_{0,n_0}))_{k=1,\dots,K}$ defined by (6.5) for grid points $(x_k)_{k=1,\dots,K}$ evenly distributed across $(0;1)$, with $K=50$.
- For each $m \in \mathcal{M}_{n,n_0}$, compute $\tilde{V}(m)$ and $\tilde{A}(m)$, defined by (6.10).

- For $\tilde{V}(h)$: the calibration of the constant $c_0$ involved in the definition of $\tilde{V}$ has been done before the study. We choose $c_0 = 1.5$, but the estimation results seem quite robust to slight changes. The index $m^*$ of the estimator $\hat{r}_{m^*}(.,\hat{F}_{0,n_0})$ used in $\tilde{V}$ is the smallest integer greater than $\ln(n_0) - 1$.
- For $\tilde{A}(h)$: we approximate the $L^2$ norms by the corresponding Riemann sums computed over the grid points $(x_k)_k$:

$$\left\|\hat{r}_{m'}(.,\hat{F}_{0,n_0}) - \hat{r}_{m\wedge m'}(.,\hat{F}_{0,n_0})\right\|^2 \approx \frac{1}{K}\sum_{k=1}^{K}\left(\hat{r}_{m'}(x_k,\hat{F}_{0,n_0}) - \hat{r}_{m\wedge m'}(x_k,\hat{F}_{0,n_0})\right)^2.$$

- Select the argmin $\tilde{m}$ of $\tilde{A}(m) + \tilde{V}(m)$, and choose $\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})$.

The risks $\mathbb{E}[\|(\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}))_+ - r\|^2]$ are computed: we average the integrated squared error (ISE) computed with $N = 500$ replications of the samples $(X_{0,i_0})_{i_0}$ and $(X_i)_i$ to obtain the mean ISE (MISE). Notice that the grid size ($K = 50$), and the number of replications ($N = 500$) are the same as Ćwik & Mielniczuk (1993). It is not difficult to see that the choice of the positive part of the estimator can only make its risk decreases. The procedure is applied for different sample sizes $n$ and $n_0$ going from 50 to 500, like in Molanes-López & Cao (2008a).

### 6.4.2 Examples

The estimation procedure is evaluated by generating two samples $(X_{0,i_0})_{i_0=1,\ldots,n_0}$ and $(X_i)_{i=1,\ldots,n}$ coming from random variables $X_0$ and $X$ respectively, with probability distributions described below.

1. The variable $X_0$ is chosen to have a uniform distribution in the set $(0;1)$. The variable $X$ fits one of the following models:
   (a1) $(1/4)(U_1 + U_2 + U_3 + U_4)$ where $U_j$, $j = 1,\ldots,4$ are independent and uniform on $(0;1)$,
   (b1) a mixture of $V_1$ with probability $1/2$ and $V_2$ with probability $1/2$, where $V_1 = V/2$ and $V_2 = (V+1)/2$, and $V$ as for model (a),
   (c1) a beta distribution with parameters 4 and 5 (denoted by $\mathcal{B}(4,5)$),
   (d1) a mixture of $X_j$, $j = 1,2,3$ with probability $1/3$, where the $X_j$ have respective distributions $\mathcal{B}(10,5)$, $\mathcal{B}(7,7)$ and $\mathcal{B}(5,10)$,
   (e1) a mixture of $X_1$ with probability $1/2$ and $X_2$ with probability $1/2$, where $X_1$ and $X_2$ have respective distributions $\mathcal{B}(15,4)$ and $\mathcal{B}(5,11)$.

   Hence the investigated relative densities are defined as the densities of $X$, in these five examples.
2. The variable $X_0$ is from the Weibull distribution with parameters (2,3). We denote by $W$ the corresponding c.d.f.. The variable $X$ is built such that $X = W^{-1}(S)$, with $S$ chosen from one of the three following distribution:
   (a2) a beta distribution $\mathcal{B}(14,17)$,
   (b2) a mixture of $V_1$ with probability $4/5$ and $V_2$ with probability $1/5$, where $V_1$ is from $\mathcal{B}(14,37)$ and $V_2$ is from $\mathcal{B}(14,20)$,

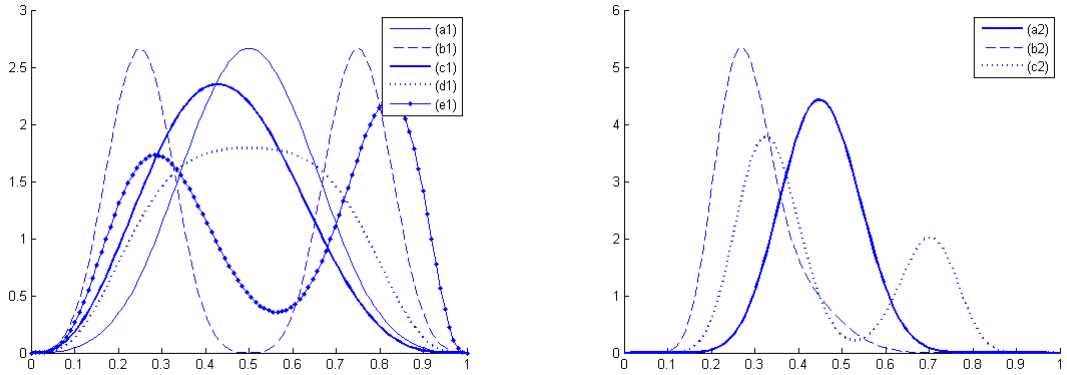



Figure 6.1: Plot of the differents investigated relative densities of Examples (1) and (2)

(c2) a mixture of $V_1$ with probability $1/3$ and $V_2$ with probability $2/3$, where $V_1$ is from $\mathcal{B}(34, 15)$ and $V_2$ is from $\mathcal{B}(15, 30)$.

In these three cases, the relative density $r$ to recover is the density of the variable $S$.

3. The variables $X_0$ and $X$ have the same distribution. Consequently, $r$ equals to 1 on $(0; 1)$. The common distribution can be: (a3) a uniform distribution in the set $(0; 1)$, (b3) a beta distribution $\mathcal{B}(2, 5)$, (c3) a gaussian distribution with mean 0 and variance 1, (d3) an exponential distribution with mean $1/5$.
4. The variable $X_0$ is from the uniform distribution on $(0; 1)$, and the variable $X$ has the density $f(x) = c_1\mathbf{1}_{(0;0.5)}(x) + (2 - c_1)\mathbf{1}_{(0.5;1)}(x)$, with $c_1 \in \{1.01, 1.05, 1.1, 1.5\}$ (the case $c_1 = 1$ is the case of the uniform distribution on $(0; 1)$).
5. The variable $X_0$ is from the beta distribution $\mathcal{B}(2, 5)$, and the variable $X$ from a beta distribution $\mathcal{B}(a, 5)$ with $a \in \{2.01, 2.05, 2.1, 2.5\}$ (the case $c_1 = 1$ is the case of an uniform distribution on $(0; 1)$). For this example, the risks are computed over a regular grid of the interval $[F_0(0.01); F_0(0.99)]$.

Recall that the further $r$ is from the uniform density on $(0; 1)$, the more different $X$ and $X_0$ are from each other. The five sets of examples above are chosen to illustrate how the estimator performs when $X$ is close to $X_0$ as well as when $X$ strays from it. The first set of examples is borrowed from Ćwik & Mielniczuk (1993) and the second from Molanes-López & Cao (2008a). The true relative densities associated to each framework of these two sets are plotted in Figure Figure 6.1: they are quite far from the uniform distribution, since the distributions of $X$ and $X_0$ are not similar. The third set of examples permits to check that the estimation procedure easily handles the case of identical probability distribution for $X$ and $X_0$. Finally, we illustrate with Examples (4) and (5) what happens when $X$ is close to $X$ but slightly different. Figure Figure 6.2 shows the relative densities of these last two examples.

Figure Figure 6.3 illustrates the stability of the method and shows beams of estimates $\hat{r}_{\tilde{m}}(., \hat{F}_{0,n_0})$: 10 estimators built from i.i.d. samples of data are plotted together with the true functions.

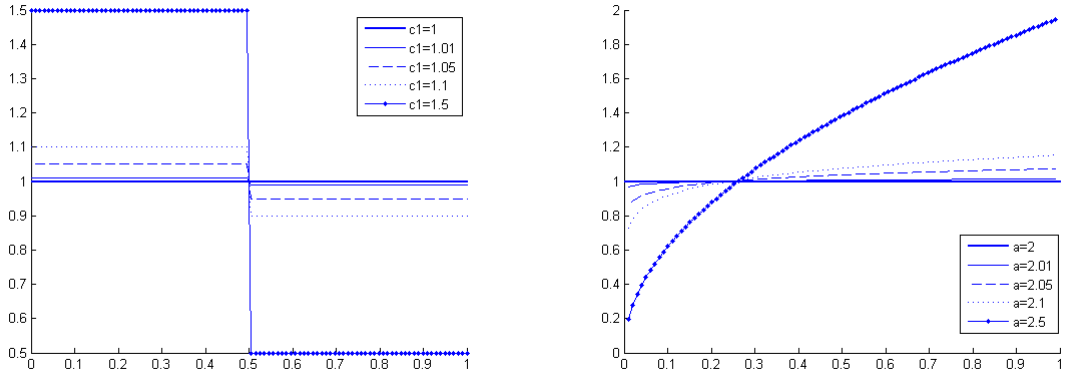



Figure 6.2: Plot of the differents investigated relative densities of Examples (4) and (5)

The risks for Examples (1) and (2) are displayed in Table 6.1. As expected, the values of MISE get smaller when the sample sizes $n$ and $n_0$ increase. But the MISE is not symmetric with respect to $n$ and $n_0$: increasing $n_0$ for fixed $n$ seems to improve more substantially the risk than the other way round. We provide in Figure Figure 6.4 a visual illustration of this fact (the improvement when $n_0$ increases appears horizontally). Notice that the values we find are of the same order as the ones of Ćwik & Mielniczuk (1993) and of Molanes-López & Cao (2008a).

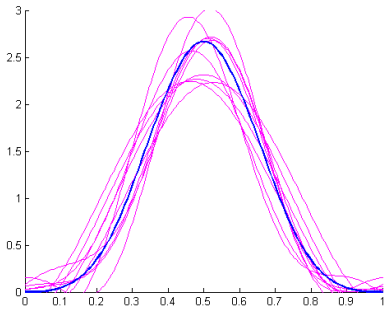
If $X$ and $X_0$ have the same probability distribution, the estimator is expected to be constant equal to 1: the selected index $m$ must thus be 0. This is the case for most of the samples we simulate: for example, only 1% of the 500 estimators computed with 50 i.i.d. Gaussian pairs $(X, X_0)$ are not identically equal to 1. For all the examples of the third set, the medians of the ISE over the 500 replicated samples are always equal to 0, whatever the distribution of $X$ and $X_0$, chosen in Examples (3) (uniform, beta, Gaussian, or exponential). The MISE are dispayed in Table 6.2: they are much more smaller than the MISE obtained with two different distributions for $X$ and $X_0$ (see Table 6.1). Thus the procedure suits well to identify two identical distributions.

The next question is what happens when the distribution of $X$ gets progressively further from the one of $X$ gradually. The MISEs in Examples (4) and (5), given in Table 6.3 illustrate this situation. The larger $c_1$ (resp. $a$), the further $X$ from $X_0$ the larger the MISE in Example (4) (resp. (5)).

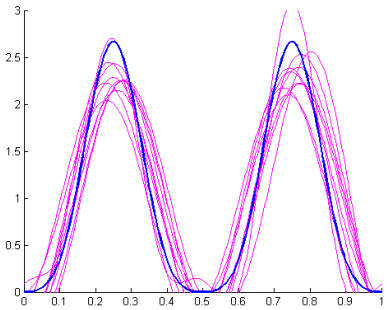
As a conclusion, we can see that the estimator $\hat{r}_{\tilde{m}}(., \hat{F}_{0,n_0})$ performs very well and gives account of what has to be detected in a satisfactory way: similarity or not between two distributions.

## 6.5 Proofs

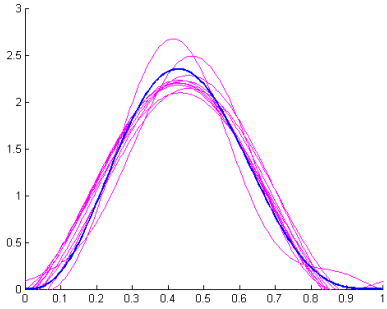
We need additional notations in this section. First, for any measurable function $t$ defined on a Borel set $B$, we denote by $\|t\|_{L^\infty(B)}$ the quantity $\sup_{x\in B} |t(x)|$. Then, we set $U_{0,i_0} = F_0(X_{0,i_0})$ $(i_0 = 1, \ldots, n_0)$, and let $\hat{U}_{n_0}$ be the empirical c.d.f. associated to the sample



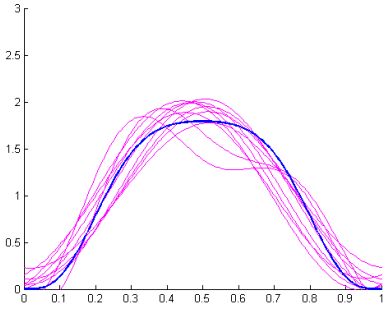
Example (a1)



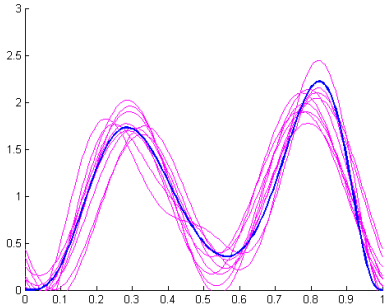
Example (b1)



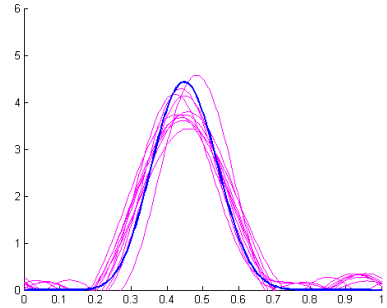
Example (c1)



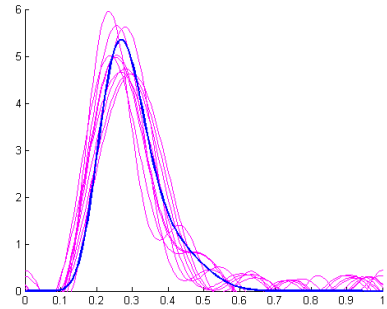
Example (d1)



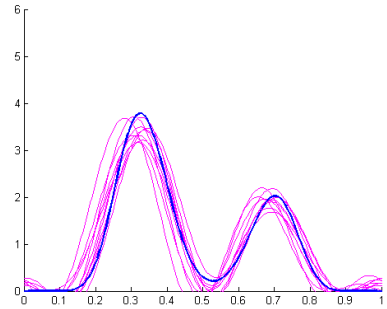
Example (e1)



Example (a2)

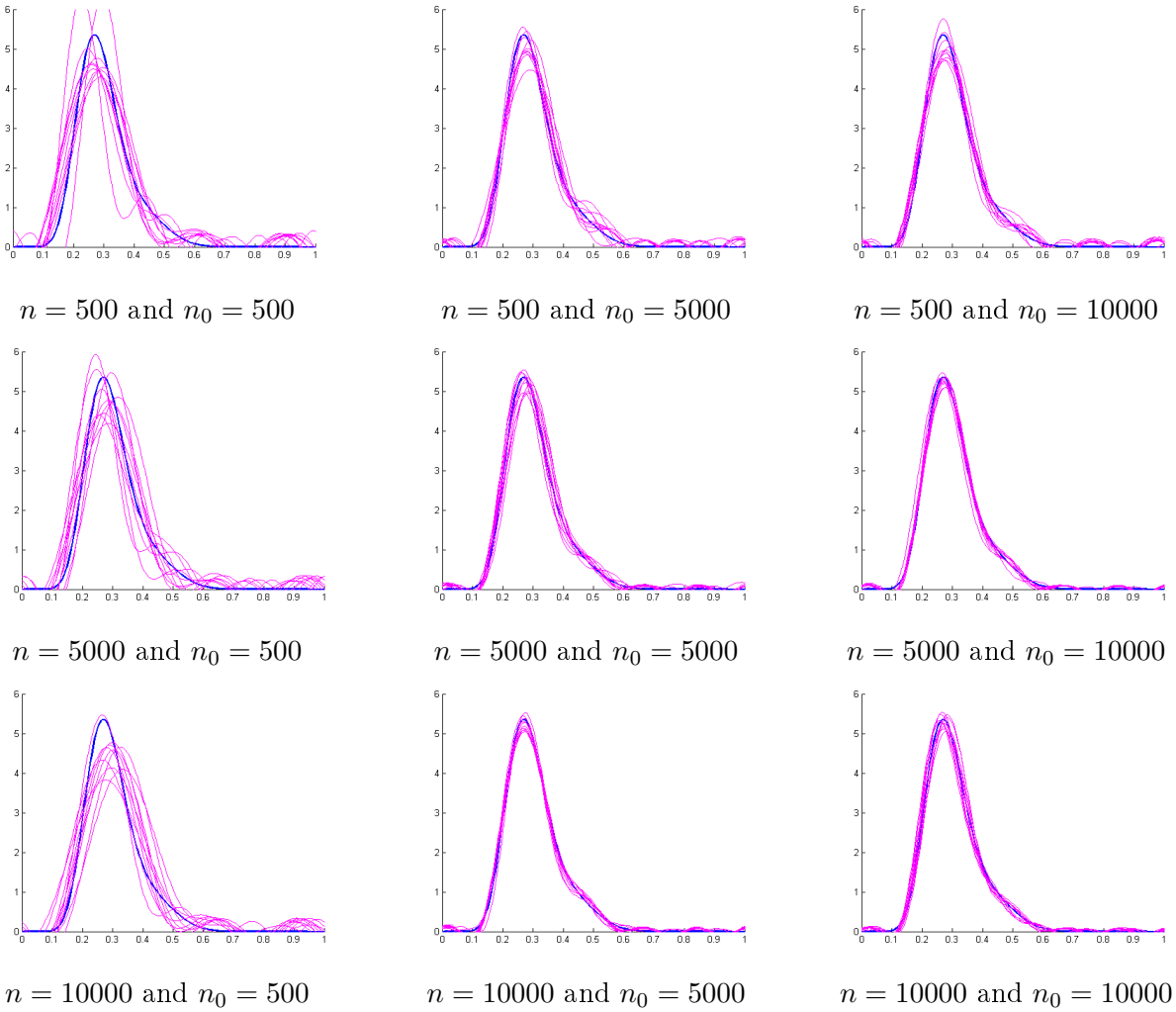


Example (b2)



Example (c2)

Figure 6.3: Beams of 10 estimators built from i.i.d. samples of size $n = n_0 = 500$ (thin lines) versus true function (thick line) in Examples (1) and (2).



$n = 500$ and $n_0 = 500$ | $n = 500$ and $n_0 = 5000$ | $n = 500$ and $n_0 = 10000$

$n = 5000$ and $n_0 = 500$ | $n = 5000$ and $n_0 = 5000$ | $n = 5000$ and $n_0 = 10000$

$n = 10000$ and $n_0 = 500$ | $n = 10000$ and $n_0 = 5000$ | $n = 10000$ and $n_0 = 10000$

Figure 6.4: Beams of 10 estimators built from i.i.d. samples of various sizes $(n; n_0)$ (thin lines) versus true function (thick line) in Example (2)(b2) .

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.2379 | 0.1503 | 0.1094 | 0.0935 |
| 100 | 0.2303 | 0.1453 | 0.0955 | 0.0752 |
| 200 | 0.2298 | 0.1409 | 0.1032 | 0.0685 |
| 400 | 0.1903 | 0.1372 | 0.0944 | 0.0634 |

Example (a1)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.7616 | 0.3514 | 0.2053 | 0.1470 |
| 100 | 0.6133 | 0.2649 | 0.1637 | 0.1295 |
| 200 | 0.5324 | 0.2739 | 0.1539 | 0.1138 |
| 400 | 0.4977 | 0.2445 | 0.1567 | 0.1080 |

Example (b1)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.1989 | 0.1082 | 0.0661 | 0.0573 |
| 100 | 0.1684 | 0.0920 | 0.0643 | 0.0456 |
| 200 | 0.1581 | 0.0950 | 0.0572 | 0.0417 |
| 400 | 0.1541 | 0.0893 | 0.0586 | 0.0381 |

Example (c1)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.2860 | 0.1475 | 0.0820 | 0.0625 |
| 100 | 0.2009 | 0.0901 | 0.0584 | 0.0454 |
| 200 | 0.1699 | 0.0761 | 0.0496 | 0.0358 |
| 400 | 0.1465 | 0.0759 | 0.0446 | 0.0330 |

Example (d1)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.4559 | 0.3911 | 0.2976 | 0.2287 |
| 100 | 0.4432 | 0.2884 | 0.1521 | 0.0899 |
| 200 | 0.4267 | 0.2026 | 0.1005 | 0.0751 |
| 400 | 0.4129 | 0.1731 | 0.0955 | 0.0660 |

Example (e1)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.9094 | 0.6008 | 0.3727 | 0.2642 |
| 100 | 0.8884 | 0.5738 | 0.2896 | 0.2108 |
| 200 | 0.8875 | 0.4877 | 0.2837 | 0.1878 |
| 400 | 0.9225 | 0.4848 | 0.2860 | 0.1817 |

Example (a2)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 1.2313 | 0.8564 | 0.5358 | 0.4151 |
| 100 | 1.1737 | 0.7826 | 0.4892 | 0.3613 |
| 200 | 1.0919 | 0.7586 | 0.4379 | 0.2876 |
| 400 | 1.1359 | 0.7180 | 0.4486 | 0.2823 |

Example (b2)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.9589 | 0.6272 | 0.4593 | 0.3251 |
| 100 | 0.8717 | 0.5665 | 0.3287 | 0.2035 |
| 200 | 0.7746 | 0.5055 | 0.2848 | 0.1625 |
| 400 | 0.7908 | 0.4752 | 0.2555 | 0.1501 |

Example (c2)

Table 6.1: Values of MISE averaged over 500 samples for the estimator $\hat{r}_{\tilde{m}}(., \hat{F}_{0,n_0})$, in Examples (1) ((a1) to (e1)) and (2) ((a2) to (c2)).

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0018 | 0.0056 | 0.0073 | 0.0046 |
| 100 | 0.0011 | 0.0032 | 0.0025 | 0.0017 |
| 200 | 0.0011 | 0.0011 | 0.0014 | 0.0006 |
| 400 | 0 | 0.0010 | 0.0007 | 0.0004 |

Example (a3)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0034 | 0.0042 | 0.0018 | 0.0035 |
| 100 | 0.0005 | 0.0020 | 0.0011 | 0.0015 |
| 200 | 0.0009 | 0.0008 | 0.0011 | 0.0022 |
| 400 | 0.0005 | 0.0009 | 0.0013 | 0.0009 |

Example (b3)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0021 | 0.0018 | 0.0035 | 0.0047 |
| 100 | 0.0015 | 0.0014 | 0.0016 | 0.0010 |
| 200 | 0.0004 | 0.0018 | 0.0016 | 0.0019 |
| 400 | 0.0023 | 0.0023 | 0.0005 | 0.0013 |

Example (c3)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0038 | 0.0015 | 0.0041 | 0 |
| 100 | 0.0008 | 0.0028 | 0.0023 | 0.0016 |
| 200 | 0.0017 | 0.0012 | 0.0009 | 0.0012 |
| 400 | 0.0012 | 0.0011 | 0.0009 | 0.0008 |

Example (d3)

Table 6.2: Values of MISE averaged over 500 samples for the estimator $\hat{r}_{\tilde{m}}(., \hat{F}_{0,n_0})$, in Examples (3) ((a3) to (c3)).

Example (4)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0022 | 0.0021 | 0.0027 | 0.0023 |
| 100 | 0.0026 | 0.0026 | 0.0023 | 0.0016 |
| 200 | 0.0021 | 0.0024 | 0.0017 | 0.0021 |
| 400 | 0.0021 | 0.0020 | 0.0016 | 0.0004 |

$c_1 = 1.01$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0038 | 0.0040 | 0.0061 | 0.0100 |
| 100 | 0.0053 | 0.0062 | 0.0053 | 0.0036 |
| 200 | 0.0025 | 0.0054 | 0.0039 | 0.0037 |
| 400 | 0.0028 | 0.0041 | 0.0041 | 0.0036 |

$c_1 = 1.05$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0109 | 0.0159 | 0.0166 | 0.0151 |
| 100 | 0.0121 | 0.0138 | 0.0125 | 0.0151 |
| 200 | 0.0125 | 0.0129 | 0.0130 | 0.0127 |
| 400 | 0.0121 | 0.0129 | 0.0120 | 0.0117 |

$c_1 = 1.1$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.2406 | 0.2086 | 0.1779 | 0.1491 |
| 100 | 0.2202 | 0.1573 | 0.1112 | 0.0971 |
| 200 | 0.2030 | 0.1203 | 0.0829 | 0.0712 |
| 400 | 0.2013 | 0.1090 | 0.0744 | 0.0657 |

$c_1 = 1.5$

Example (5)

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0041 | 0.0080 | 0.0052 | 0.0042 |
| 100 | 0.0032 | 0.0049 | 0.0018 | 0.0015 |
| 200 | 0.0048 | 0.0036 | 0.0018 | 0.0014 |
| 400 | 0.0033 | 0.0012 | 0.0012 | 0.0014 |

$a = 2.01$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0131 | 0.008 | 0.0058 | 0.0092 |
| 100 | 0.0067 | 0.0093 | 0.0056 | 0.0056 |
| 200 | 0.0059 | 0.0071 | 0.0034 | 0.0041 |
| 400 | 0.0067 | 0.0042 | 0.0051 | 0.0035 |

$a = 2.05$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.0144 | 0.0165 | 0.0159 | 0.0134 |
| 100 | 0.0175 | 0.0154 | 0.0133 | 0.0138 |
| 200 | 0.0160 | 0.0152 | 0.0134 | 0.0128 |
| 400 | 0.0149 | 0.0135 | 0.0129 | 0.0117 |

$a = 2.1$

| $n$ \ $n_0$ | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| 50 | 0.3124 | 0.3100 | 0.3091 | 0.3110 |
| 100 | 0.3128 | 0.3104 | 0.3101 | 0.3027 |
| 200 | 0.3167 | 0.3090 | 0.2964 | 0.2933 |
| 400 | 0.3087 | 0.3078 | 0.3031 | 0.2857 |

$a = 2.5$

Table 6.3: Values of MISE averaged over 500 samples for the estimator $\hat{r}_{\tilde{m}}(., \hat{F}_{0,n_0})$, in Examples (4) and (5).

$(U_{0,i_0})_{i_0=1,\dots,n_0}$. Finally $id$ is the function such that $u \mapsto u$, on the interval $(0;1)$. We also denote by $\mathbb{E}[.|(X_0)]$ the conditional expectation given the sample $(X_{0,i_0})_{i_0=1,\dots,n_0}$ (the conditional variance will be coherently denoted by $\mathrm{Var}(.|(X_0))$). We will use the tools of Chapter 2, especially the concentration inequalities of Section 2.1.1 and the inequalities of Dvoretzky *et al.* (1956) of Section 2.1.2.

### 6.5.1 Proof of Proposition 6.1

A key point is the following decomposition which holds for any index $m$

$$\left\|\hat{r}_m(.,\hat{F}_{0,n_0}) - r\right\|^2 \leq 3T_1^m + 3T_2^m + 3\left\|\hat{r}_m(.,F_0) - r\right\|^2,$$

with

$$\begin{aligned} T_1^m &= \left\|\hat{r}_m(.,\hat{F}_{0,n_0}) - \hat{r}_m(.,F_0) - \mathbb{E}\left[\hat{r}_m(.,\hat{F}_{0,n_0}) - \hat{r}_m(.,F_0)\,|(X_0)\right]\right\|^2, \\ T_2^m &= \left\|\mathbb{E}\left[\hat{r}_m(.,\hat{F}_{0,n_0}) - \hat{r}_m(.,F_0)\,|(X_0)\right]\right\|^2. \end{aligned} \tag{6.11}$$

We have already proved (see (6.6) and (6.7)) that

$$\left\|\hat{r}_m(.,F_0) - r\right\|^2 \leq \frac{D_m}{n} + \left\|\Pi_{S_m} r - r\right\|^2.$$

Therefore, it remains to apply the two following lemmas, proved in the two following sections.

**Lemma 6.1.** *Under the assumptions of Proposition 6.1,*

$$\mathbb{E}\left[T_1^m\right] \leq \|\varphi_2'\|^2_{L^\infty((0;1))} \frac{D_m^3}{4nn_0}$$

The result of Lemma 6.1 is also satisfied if we only assume that the models $(S_m)_m$ are spanned by continuously derivable functions $\varphi_j$, $j = 1, \dots, D_m$.

**Lemma 6.2.** *Under the assumptions of Proposition 6.1, ,*

$$\mathbb{E}\left[T_2^m\right] \leq 3\|r\|^2 \frac{D_m}{n_0} + 3\frac{\pi^4}{4} C_4 \|r\|^2 \frac{D_m^4}{n_0^2} + \frac{C_6}{12}\|\varphi_2^{(3)}\|^2_{L^\infty((0;1))} \frac{D_m^7}{n_0^3} + 3\frac{\|r'\|^2}{n_0}.$$

The result follows if $D_m \leq \kappa n_0^{1/3}$.

□

**Proof of Lemma 6.1**

The decompositions of the estimator in the orthogonal basis $(\varphi_j)_j$ yields

$$T_1^m = \sum_{j=1}^{D_m} \left(\hat{a}_j^{\hat{F}_{0,n_0}} - \hat{a}_j^{F_0} - \mathbb{E}\left[\hat{a}_j^{\hat{F}_{0,n_0}} - \hat{a}_j^{F_0}\,|(X_0)\right]\right)^2,$$

and therefore, $\mathbb{E}[T_1^m|(X_0)] = \sum_{j=1}^{D_m} \mathrm{Var}(\hat{a}_j^{\hat{F}_{0,n_0}} - \hat{a}_j^{F_0}|(X_0))$. Moreover, for any index $j$,

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_j^{\hat{F}_{0,n_0}} - \hat{a}_j^{F_0}|(X_0)\right) &\leq \frac{1}{n}\mathbb{E}\left[\left(\varphi_j \circ \hat{F}_{0,n_0}(X_1) - \varphi_j \circ F_0(X_1)\right)^2 |(X_0)\right], \\ &\leq \frac{1}{n}\left\|\varphi_j'\right\|^2_{L^\infty((0;1))} \int_A \left(\hat{F}_{0,n_0}(x) - F_0(x)\right)^2 f(x)dx,\end{aligned}$$

by using the mean-value theorem. Since $\|\varphi_j'\|^2_{L^\infty((0;1))} \leq D_m^2 \|\varphi_2'\|^2_{L^\infty((0;1))}$ in the Fourier basis, this leads to

$$\mathbb{E}\left[T_1^m\right] \leq \|\varphi_2'\|^2_{L^\infty((0;1))} \frac{1}{n} D_m^3 \int_A \mathbb{E}\left[\left(\hat{F}_{0,n_0}(x) - F_0(x)\right)^2\right] f(x)dx.$$

Notice finally that $\mathbb{E}[(\hat{F}_{0,n_0}(x) - F_0(x))^2] = \mathrm{Var}(\hat{F}_{0,n_0}(x)) = (F_0(x)(1-F_0(x)))/n_0 \leq 1/(4n_0)$. This permits to conclude the proof of Lemma 6.1.

□

**Proof of Lemma 6.2**

Arguing as in the beginning of the proof of Lemma 6.1 yields

$$T_2^m = \sum_{j=1}^{D_m} \left(\int_A \left(\varphi_j \circ \hat{F}_{0,n_0}(x) - \varphi_j \circ F_0(x)\right) f(x)dx\right)^2. \tag{6.12}$$

We apply the Taylor formula, with the Lagrange form for the remainder to the function $\varphi_j$. There exists a random number $\hat{\alpha}_{j,n_0,x}$ such that the following decomposition holds: $T_2^m \leq 3T_{2,1}^m + 3T_{2,2}^m + 3T_{2,3}^m$, where

$$\begin{aligned}T_{2,1}^m &= \sum_{j=1}^{D_m} \left(\int_A \varphi_j'(F_0(x)) \left(\hat{F}_{0,n_0}(x) - F_0(x)\right) f(x)dx\right)^2, \\ T_{2,2}^m &= \sum_{j=1}^{D_m} \left(\int_A \varphi_j''(F_0(x)) \frac{\left(\hat{F}_{0,n_0}(x) - F_0(x)\right)^2}{2} f(x)dx\right)^2, \\ T_{2,3}^m &= \sum_{j=1}^{D_m} \left(\int_A \varphi_j^{(3)}(\hat{\alpha}_{j,n_0,x}) \frac{\left(\hat{F}_{0,n_0}(x) - F_0(x)\right)^3}{6} f(x)dx\right)^2.\end{aligned}$$

We now bound each of these three terms. Let us begin with $T_{2,1}^m$. The change of variables $u = F_0(x)$ permits to obtain first

$$T_{2,1}^m = \sum_{j=1}^{D_m} \left(\int_{(0;1)} \varphi_j'(u) \left(\hat{U}_{n_0}(u) - u\right) r(u)du\right)^2,$$

and, with the definition of $\hat{U}_{n_0}(u)$, we get

$$T_{2,1}^m = \sum_{j=1}^{D_m} \left(\frac{1}{n_0}\sum_{i=1}^{n_0} B_{i,j} - \mathbb{E}[B_{i,j}]\right)^2, \quad \text{with } B_{i,j} = \int_{U_{0,i}}^1 r(u)\varphi_j'(u)du.$$

An integration by parts for $B_{i,j}$ leads to another splitting $T^m_{2,1} \leq 2T^m_{2,1,1} + 2T^m_{2,1,2}$, with notations

$$T^m_{2,1,1} = \sum_{j=1}^{D_m} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} r(U_{0,i})\varphi_j(U_{0,i}) - \mathbb{E}\left[r(U_{0,i})\varphi_j(U_{0,i})\right] \right\}^2,$$
$$T^m_{2,1,2} = \sum_{j=1}^{D_m} \left\{ \int_{(0;1)} r'(u) \left(\hat{U}_{n_0}(u) - u\right) \varphi_j(u)du \right\}^2.$$

The expectation of the first term is a variance and is bounded as follows:

$$\mathbb{E}\left[T^m_{2,1,1}\right] \leq \frac{1}{n_0} \sum_{j=1}^{D_m} \mathbb{E}\left[\left(r(U_{0,1})\varphi_j(U_{0,1})\right)^2\right] \leq \int_0^1 r(u)^2 du \frac{D_m}{n_0}.$$

For $T^m_{2,1,2}$, we use the definitions and properties of the orthogonal projection operator $\Pi_{S_m}$ on the space $S_m$:

$$\begin{aligned} T^m_{2,1,2} &= \sum_{j=1}^{D_m} \left(\langle r'(\hat{U}_{n_0} - id), \varphi_j\rangle_{(0;1)}\right)^2 = \left\|\Pi_{S_m}(r'(\hat{U}_{n_0} - id))\right\|^2, \\ &\leq \left\|r'(\hat{U}_{n_0} - id)\right\|^2 \leq \|r'\|^2 \|\hat{U}_{n_0} - id\|^2_{L^\infty((0;1))}. \end{aligned}$$

Applying Proposition 2.4 (Chapter 2) proves that $\mathbb{E}[T^m_{2,1,2}] \leq C_2\|r'\|^2/n_0$. Therefore,

$$\mathbb{E}\left[T^m_{2,1}\right] \leq \|r\|^2 \frac{D_m}{n_0} + C_2\|r'\|^2 \frac{1}{n_0}. \tag{6.13}$$

Consider now $T^m_{2,2}$. The trigonometric basis satisfies $\varphi_j'' = -(\pi\mu_j)^2\varphi_j$, with $\mu_j = j$ for even $j \geq 2$, and $\mu_j = j-1$ for odd $j \geq 2$. We thus have,

$$\begin{aligned} \mathbb{E}\left[T^m_{2,2}\right] &= (\pi^4/4)\mathbb{E}\left[\sum_{j=1}^{D_m} \left\{ \int_{(0;1)} r(u) \left(\hat{U}_{n_0}(u) - u\right)^2 \mu_j^2 \varphi_j(u)du \right\}^2\right], \\ &\leq (\pi^4/4) D_m^4 \mathbb{E}\left[\sum_{j=1}^{D_m} \left\{ \langle r\left(\hat{U}_{n_0} - id\right)^2, \varphi_j\rangle_{(0;1)} \right\}^2\right], \\ &\leq (\pi^4/4) D_m^4 \mathbb{E}\left[\left\| r\left(\hat{U}_{n_0} - id\right)^2 \right\|^2\right] \leq (\pi^4/4) D_m^4 \mathbb{E}\left[\left\|\hat{U}_{n_0} - id\right\|^4_{L^\infty((0;1))}\right] \int_{(0;1)} r^2(u)du, \end{aligned}$$

Thanks to Proposition 2.4, we obtain

$$\mathbb{E}\left[T^m_{2,1}\right] \leq C_4(\pi^4/4)\|r\|^2 \frac{D_m^4}{n_0^2}. \tag{6.14}$$

The last term is then easily controlled, using also Proposition 2.4:

$$\begin{aligned} \mathbb{E}\left[T^m_{2,3}\right] &\leq \frac{1}{6} \sum_{j=1}^{D_m} \left\|\varphi_j^{(3)}\right\|^2_{L^\infty((0;1))} \|r\|^2 \mathbb{E}\left[\left\|\hat{U}_{n_0} - id\right\|^6_{L^\infty((0;1))}\right], \qquad (6.15) \\ &\leq \frac{C_6}{6} \left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;1))} \|r\|^2 \frac{D_m^7}{n_0^3}. \end{aligned}$$

Lemma 6.2 is proved by gathering (6.13), (6.14) and (6.15).

□

### 6.5.2 Proof of Theorem 6.1

In the proof, $C$ is a constant which may change from line to line, and is independent of all $m \in \mathcal{M}_{n,n_0}$, $n$, and $n_0$. Let $m \in \mathcal{M}_{n,n_0}$ be fixed. The following decomposition holds:

$$\begin{aligned}\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2 &\leq 3\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-\hat{r}_{m\wedge\hat{m}}\left(.,\hat{F}_{0,n_0}\right)\right\|^2 \\ &\quad +3\left\|\hat{r}_{m\wedge\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-\hat{r}_{m}\left(.,\hat{F}_{0,n_0}\right)\right\|^2+3\left\|\hat{r}_{m}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2.\end{aligned}$$

We use successively the definition of $A(\hat{m},\hat{F}_{0,n_0})$, $A(m,\hat{F}_{0,n_0})$, and $\hat{m}$ to obtain

$$\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2 \leq 6\left(A(m,\hat{F}_{0,n_0})+V(m)\right)+3\left\|\hat{r}_{m}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2.$$

Keeping in mind that we can split $\|\hat{r}_m(.,\hat{F}_{0,n_0})-r\|^2 \leq 3T_1^m+3T_2^m+3\|\hat{r}_m(.,F_0)-r\|^2$ with the notations of Section 6.5.1, we derive from (6.6) and (6.7):

$$\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2 \leq 6\left(A(m,\hat{F}_{0,n_0})+V(m)\right)+9T_1^m+9T_2^m+9\frac{D_m}{n}+9\left\|r_m-r\right\|^2.$$

We also apply Lemmas 6.1 and 6.2. Taking into account that $D_m \leq \kappa n_0^{1/3}$, we thus have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_{0,n_0}\right)-r\right\|^2\right] &\leq 6\mathbb{E}\left[A(m,\hat{F}_{0,n_0})\right]+6V(m)+C\frac{D_m}{n}+C\|r\|^2\frac{D_m}{n_0} \\ &\quad +9\left\|r_m-r\right\|^2+\frac{C}{n_0}.\end{aligned}$$

Therefore, the conclusion of Theorem 6.1 is the result of the following lemma.

**Lemma 6.3.** *Under the assumptions of Theorem 6.1, there exists a constant $C>0$ such that, for any $m \in \mathcal{M}_{n,n_0}$,*

$$\mathbb{E}\left[A(m,\hat{F}_{0,n_0})\right] \leq C\frac{1}{n}+C\frac{1}{n_0}+12\left\|r_m-r\right\|^2.$$

□

**Proof of Lemma 6.3**

To study $A(m,\hat{F}_{0,n_0})$, we write, for $m' \in \mathcal{M}_{n,n_0}$.

$$\begin{aligned}\left\|\hat{r}_{m'}\left(.,\hat{F}_{0,n_0}\right)-\hat{r}_{m\wedge m'}\left(.,\hat{F}_{0,n_0}\right)\right\|^2 &\leq 3\left\|\hat{r}_{m'}\left(.,\hat{F}_{0,n_0}\right)-r_{m'}\right\|^2+3\left\|r_{m'}-r_{m\wedge m'}\right\|^2 \\ &\quad +3\left\|r_{m\wedge m'}-\hat{r}_{m\wedge m'}\left(.,\hat{F}_{0,n_0}\right)\right\|^2.\end{aligned}$$

Let $\mathcal{S}(p_{m'})$ be the set $\{t \in S_{p_{m'}},\ \|t\|=1\}$, for $p_{m'}=m'$ or $p_{m'}=m\wedge m'$ . We note that

$$\left\|r_{p_{m'}}-\hat{r}_{p_{m'}}(.,\hat{F}_{0,n_0})\right\|^2=\sum_{j=1}^{D_{p_{m'}}}\left(\tilde{\nu}_n(\varphi_j)\right)^2=\sup_{t\in\mathcal{S}(p_{m'})}\tilde{\nu}_n(t)^2, \tag{6.16}$$

with

$$\tilde{\nu}_n(t) = \frac{1}{n}\sum_{i=1}^{n} t\circ\hat{F}_{0,n_0}(X_i) - \mathbb{E}\left[t\circ F_0(X_i)\right].$$

Since the empirical process $\tilde{\nu}_n$ is not centred, we consider the following splitting: $(\tilde{\nu}_n(t))^2 \leq 2\nu_n^2(t) + 2((1/n)\sum_{i=1}^n (t\circ\hat{F}_{0,n_0}(X_i) - t\circ F_0(X_i)))^2$, with

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(t\circ F_0(X_i) - \mathbb{E}\left[t\circ F_0(X_i)\right]\right). \tag{6.17}$$

But we also have

$$\sup_{t\in\mathcal{S}(p_{m'})}\left(\frac{1}{n}\sum_{i=1}^{n}\left(t\circ\hat{F}_{0,n_0}(X_i) - t\circ F_0(X_i)\right)\right)^2 \;=\; \sum_{j=1}^{D_{p_{m'}}}\left(\hat{a}_j^{\hat{F}_0} - \hat{a}_j^{F_0}\right)^2 \leq 2T_1^{p_{m'}} + 2T_2^{p_{m'}},$$

with the notations of Section 6.5.1. This shows that

$$\left\|r_{p_{m'}} - \hat{r}_{p_{m'}}(.,\hat{F}_{0,n_0})\right\|^2 \leq 2\sup_{t\in\mathcal{S}(p_{m'})}(\nu_n(t))^2 + 4T_1^{p_{m'}} + 4T_2^{p_{m'}}. \tag{6.18}$$

We thus have

$$\begin{aligned}\left\|\hat{r}_{m'}\left(.,\hat{F}_{0,n_0}\right) - \hat{r}_{m\wedge m'}\left(.,\hat{F}_{0,n_0}\right)\right\|^2 \;\leq\; & 6\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2 + 6\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2 + 12T_2^{m'} \\ & + 12T_2^{m\wedge m'} + 12T_1^{m'} + 12T_1^{m\wedge m'} + 3\left\|r_{m'} - r_{m\wedge m'}\right\|^2.\end{aligned}$$

We get back to the definition of $A(m,\hat{F}_{0,n_0})$. To do so, we subtract $V(m')$. For convenience, we split it into two terms: $V(m') = V^{\hat{F}_0,1}(m') + V^{\hat{F}_0,2}(m')$, with $V^{\hat{F}_0,1}(m') = c_0 D_m/n$, and $V^{\hat{F}_0,2}(m') = c_0\|r\|^2 D_m/n_0$. Thus,

$$\begin{aligned}\mathbb{E}\left[A\left(m,\hat{F}_{0,n_0}\right)\right] \;\leq\; & 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2 - \frac{V^{\hat{F}_0,1}(m')}{12}\right)_+\right] \\ & + 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2 - \frac{V^{\hat{F}_0,1}(m')}{12}\right)_+\right] \\ & + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(T_2^{m'} - \frac{V^{\hat{F}_0,2}(m')}{24}\right)_+\right] \\ & + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(T_2^{m\wedge m'} - \frac{V^{\hat{F}_0,2}(m')}{24}\right)_+\right] \\ & + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}} T_1^{m'}\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}} T_1^{m\wedge m'}\right] + 3\max_{m'\in\mathcal{M}_{n,n_0}}\left\|r_{m'} - r_{m\wedge m'}\right\|^2.\end{aligned}$$

For the deterministic term, we notice that

$$\max_{m'\in\mathcal{M}_{n,n_0}}\left\|r_{m'} - r_{m\wedge m'}\right\|^2 \;\leq\; 2\max_{\substack{m'\in\mathcal{M}_{n,n_0}\\ m\leq m'}}\left\|r_{m'} - r\right\|^2 + 2\left\|r - r_m\right\|^2.$$

If $m \leq m'$, the spaces are nested $S_m \subset S_{m'}$, thus the orthogonal projections $r_m$ and $r_{m'}$ of $r$ onto $S_m$ and $S'_m$ respectively satisfy $\|r_{m'} - r\|^2 \leq \|r_m - r\|^2$. Thus,

$$\max_{m' \in \mathcal{M}_{n,n_0}} \|r_{m'} - r_{m \wedge m'}\|^2 \leq 4 \|r_m - r\|^2 . \tag{6.19}$$

Moreover, for $p_{m'} = m'$ or $p_{m'} = m \wedge m'$, $T_1^{p_{m'}} \leq T_1^{m_{\max}}$ (recall that $m_{\max}$ is the largest index in the collection $\mathcal{M}_{n,n_0}$. Therefore,

$$12\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} T_1^{m'}\right] + 12\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} T_1^{m \wedge m'}\right] \leq 24\mathbb{E}\left[T_1^{m_{\max}}\right] \leq \frac{D^3_{m_{\max}}}{nn_0} \leq \frac{C}{n}.$$

Consequently, we have at this stage

$$\begin{aligned}
\mathbb{E}\left[A\left(m, \hat{F}_{0,n_0}\right)\right] &\leq \frac{C}{n} + 12 \|r_m - r\|^2 \\
&\leq 6\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(\sup_{t \in \mathcal{S}(m')} (\nu_n(t))^2 - \frac{V^{\hat{F}_{0,1}}(m')}{12}\right)_+\right] \\
&\quad +6\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(\sup_{t \in \mathcal{S}(m \wedge m')} (\nu_n(t))^2 - \frac{V^{\hat{F}_{0,1}}(m')}{12}\right)_+\right] \\
&\quad +12\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(T_2^{m'} - \frac{V^{\hat{F}_{0,2}}(m')}{48}\right)_+\right] \\
&\quad +12\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(T_2^{m \wedge m'} - \frac{V^{\hat{F}_{0,2}}(m')}{48}\right)_+\right].
\end{aligned}$$

Since $V^{\hat{F}_{0,l}}(m') \geq V^{\hat{F}_{0,l}}(m' \wedge m)$ it remains to bound the two following terms:

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(\sup_{t \in \mathcal{S}(p_{m'})} (\nu_n(t))^2 - \frac{V^{\hat{F}_{0,1}}(p_{m'})}{12}\right)_+\right] \text{ and } \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(T_2^{p_{m'}} - \frac{V^{\hat{F}_{0,2}}(p_{m'})}{48}\right)_+\right]$$

We use the two following lemmas, which are proved below.

**Lemma 6.4.** *Assume that $r$ is bounded on $(0;1)$. The deviations of the empirical process $\nu_n$ defined by (6.17) can be controlled as follows,*

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left\{\sup_{t \in \mathcal{S}(p_{m'})} \nu_n^2(t) - \tilde{V}(p_{m'})\right\}_+\right] \leq \frac{C}{n},$$

*where $\tilde{V}(p_{m'}) = 2(1+2\delta)D_{p_{m'}}/n$, and $\delta$ a purely numerical constant.*

We choose $c_0$ in the definition of $V$ (see (6.9)) large enough to have $V^{\hat{F}_{0,1}}(p_{m'})/12 \geq \tilde{V}(p_{m'})$, for every $m'$. The inequality of Lemma 6.4 with $V^{\hat{F}_{0,1}}(p_{m'})$ as a replacement for $\tilde{V}(p_{m'})$.

**Lemma 6.5.** *Under the assumptions of Theorem 6.1,*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(T_2^{p_{m'}}-V_2(p_{m'})\right)_+\right]\leq\frac{C}{n_0},$$

*with* $V_2(p_{m'})=c_2\|r\|^2\frac{D_{p'_m}}{n_0}$, *$c_2$ a positive constant large enough, and $C$ depending on the basis, on $r$, and on the constants $C_p$ of Proposition 2.4.*

We choose $c_0$ in the definition of $V$ (see (6.9)) large enough to have $V^{\hat{F}_0,2}(p_{m'})/24\geq V_2(p_{m'})$, for every $m'$. This enables to apply Lemma 6.5 with $V^{\hat{F}_0,2}(p_{m'})$ as a replacement for $V_2(p_{m'})$.

The proof of Lemma 6.3 is completed.

□

**Proof of Lemma 6.4**

We roughly bound

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left\{\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-\tilde{V}(p_{m'})\right\}_+\right]\leq\sum_{m'\in\mathcal{M}_{n,n_0}}\mathbb{E}\left[\left\{\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-\tilde{V}(p_{m'})\right\}_+\right].$$

We apply the Talagrand Inequality recalled in Proposition 2.2 (Chapter 2). To this aim, we compute $M_1$, $H^2$ and $v$. Write for a moment $\nu_n(t)=(1/n)\sum_{i=1}^n\psi_t(X_i)-\mathbb{E}[\psi_t(X_i)]$, with $\psi_t(x)=t\circ F_0(x)$.

- First, for $t\in\mathcal{S}(p_{m'})$,
$$\|\psi_t\|_{L^\infty(A)}\leq\|t\|_{L^\infty((0;1))}\leq\sqrt{D_{p_{m'}}}\|t\|=\sqrt{D_{p_{m'}}}=:M_1.$$
- Next, we develop $t\in\mathcal{S}(p_{m'})$ in the orthogonal basis $(\varphi_j)_{j=1,\dots,D_{p_{m'}}}$. This leads to
$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)\right]\leq\sum_{j=1}^{D_{p_{m'}}}\mathbb{E}\left[\nu_n(\varphi_j^2)\right]=\sum_{j=1}^{D_{p_{m'}}}\mathbb{E}\left[\left(\hat{a}_j^{F_0}-a_j\right)^2\right]\leq\frac{D_{p_{m'}}}{n}=:H^2,$$
thanks to the upper-bound for the variance term (see (6.7).
- Last, for $t\in\mathcal{S}(p_{m'})$,
$$\begin{aligned}\mathrm{Var}\left(\psi_t(X_1)\right)&\leq\int_A t^2\left(F_0(x)\right)f(x)dx=\int_{(0;1)}t^2\left(u\right)r(u)du,\\&\leq\|r\|_{L^\infty((0;1))}\|t\|^2=\|r\|_{L^\infty((0;1))}=:v.\end{aligned}$$

Inequality (2.1) gives, for $\delta>0$,

$$\begin{aligned}\sum_{m'\in\mathcal{M}_{n,n_0}}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-c(\delta)H^2\right)_+\right]&\leq c_1\sum_{m'\in\mathcal{M}_{n,n_0}}\left\{\frac{1}{n}\exp\left(-c_2\delta D_{p_{m'}}\right)\right.\\&\left.+\frac{D_{p_{m'}}}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\sqrt{n}\right)\right\},\end{aligned}$$

where $c_l$, $l=1,2,3$ are three constants. Now, it is sufficient to use that $D_{p'_m}=2p_{m'}+1$, and that the cardinal of $\mathcal{M}_{n,n_0}$ is bounded by $n$ to end the proof of Lemma 6.4.

**Proof of Lemma 6.5**

The proof is based on the proof of Lemma 6.2, Section 6.5.1. Let us abbreviate $p_{m'}$ by $p$. We proceed as in this proof and obtain $T_2^p \leq 6T_{2,1,1}^p + 6T_{2,1,2}^p + 3T_{2,2}^p + 3T_{2,3}^p$. Thus,

$$\begin{aligned}\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}(T_2^p - V_2(p))_+\right] \leq\ & \mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(6T_{2,1,1}^p - V_2(p)/3\right)_+\right] \\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}} 6T_{2,1,2}^p\right] + \mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - V_2(p)/3\right)_+\right] \\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,3}^p - V_2(p)/3\right)_+\right].\end{aligned} \tag{6.20}$$

We do not subtract $V(p)$ to one of the term, since we immediately derive from Section 6.5.1

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}} 6T_{2,1,2}^p\right] \leq \mathbb{E}\left[6T_{2,1,2}^{m_{\max}}\right] \leq 6C_2\|r'\|^2/n_0. \tag{6.21}$$

For the term depending on $T_{2,1,1}^p$, note that $T_{2,1,1}^p = \sum_{j=1}^{D_p}(\nu_{n_0}^b(\varphi_j))^2$, with

$$\nu_{n_0}^b(t) = \frac{1}{n_0}\sum_{i=1}^{n_0}\psi_t(X_{0,i}) - \mathbb{E}\left[\psi_t(X_{0,i})\right], \text{ avec } \psi_t(X_{0,i}) = r(F_0(X_{0,i}))t(F_0(X_{0,i})).$$

We proceed as in (6.16) to write $T_{2,1,1}^p = \sup_{t\in\mathcal{S}(p)}(\nu_{n_0}^b(t))^2$. We anew apply the Talagrand Inequality (Proposition 2.2). We compute easily $M_1 = \|r\|_{L^\infty((0;1))}\sqrt{D_p}$, and $v = \|r\|^2_{L^\infty((0;1))}$. For $H^2$, the same computations as in Lemma 6.2 give

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(p)}\left(\nu_{n_0}^b(t)\right)^2\right] = \mathbb{E}\left[T_{2,1,1}^p\right] \leq \|r\|^2\frac{D_p}{n_0} := H^2.$$

The result is the following, with $V_{2,1,1}(p) = 6\times 2(1+2\delta)\|r\|^2 D_p/n_0$, $\delta > 0$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(6T_{2,1,1}^p - V_{2,1,1}(p)\right)_+\right] \leq \frac{C}{n_0}. \tag{6.22}$$

For the term in which $T_{2,2}^p$ is involved, we begin with

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - \frac{V_2(p)}{3}\right)_+\right] \leq \sum_{m'\in\mathcal{M}_{n,n_0}}\mathbb{E}\left[\left(3T_{2,2}^p - \frac{V_2(p)}{3}\right)_+\right],$$

and compute the right-hand side, for a fixed index $p$. We prove in Section 6.5.1 that $T_{2,2}^p \leq (\pi^4/4)\|r\|^2 D_p^4\|\hat{U}_{n_0} - id\|^4_{L^\infty((0;1))}$. Inequality (2.5) of Corollary 2.1 with $p = 4$ (Chapter 2) gives, for all $\kappa > 0$,

$$\mathbb{E}\left[\left(\left\|\hat{U}_{n_0} - id\right\|^4_{L^\infty((0;1))} - \frac{\kappa}{(3\pi^4/4)\|r\|^2}\frac{\ln^2(n_0)}{n_0^2}\right)_+\right] \leq Cn_0^{-\frac{1}{\sqrt{2}}\left(\frac{\kappa}{(3\pi^4/4)\|r\|^2}\right)^{1/2}}.$$

Therefore, denoting by $V_{2,2}(p) = (3\pi^4/4)\|r\|^2\kappa\frac{D_p^4\ln^2(n_0)}{n_0^2}$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - V_{2,2}(p)\right)_+\right] \leq C\sum_{m'\in\mathcal{M}_{n,n_0}} D_p^4 n_0^{-\frac{1}{\sqrt{2}}\left(\frac{\kappa}{(3\pi^4/4)\|r\|^2}\right)^{1/2}}. \tag{6.23}$$

But we roughly bound $\sum_{m'\in\mathcal{M}_{n,n_0}} D_p^4 \leq D_{m_{\max}}^5 \leq n_0^5$. The right-hand side of (6.23) is thus bounded by $Cn_0^{5-c\sqrt{\kappa}}$, with $c$ a constant, and this last bound is smaller than $C/n_0$ if $\kappa$ is large enough. Since we assume $D_p \leq n_0^{1/3}/\ln^{2/3}(n)$, we have

$$V_{2,2}(p) \leq V_{2,2}^{bis}(p) = (3\pi^4/4)\|r\|^2\kappa\frac{D_p}{n_0},$$

and (6.23) is still true with $V_{2,2}$ replaced by $V_{2,2}^{bis}$.

We proceed similarly for the term which depends on $T_{2,3}^p$. We see in Section 6.5.1 that $T_{2,3}^p \leq (1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2 D_p^7\|\hat{U}_{n_0} - id\|_{L^\infty((0;1))}^6$, and thanks to Inequality (2.5) of Corollary 2.1 with $p=6$

$$\mathbb{E}\left[\left(\left\|\hat{U}_{n_0} - id\right\|_{L^\infty((0;1))}^6 - \frac{\kappa}{(1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2}\frac{\ln^3(n_0)}{n_0^3}\right)_+\right] \leq Cn_0^{-\frac{1}{2^{1/3}}\left(\frac{\kappa}{(1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2}\right)^{2/3}}.$$

Thus, for $V_{2,3}(p) = (1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2\kappa D_p^7\ln^3(n_0)/n_0^3$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,3}^p - V_{2,3}(p)\right)_+\right] \leq C\sum_{m'\in\mathcal{M}_{n,n_0}} D_p^4 n_0^{-\frac{1}{2^{1/3}}\left(\frac{\kappa}{(1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2}\right)^{2/3}}. \tag{6.24}$$

If $\kappa$ is large enough, the right-hand side of (6.24) is bounded by $C/n_0$, and $V_{2,3}$ can be replaced by an upper-bound, if $D_p \leq n_0^{1/3}/\ln^{1/2}(n)$: $V_{2,3}(p) \leq V_{2,3}^{bis}(p) = (1/6)\|\varphi_2^{(3)}\|_{L^\infty((0;1))}^2\|r\|^2\kappa\frac{D_p}{n_0}$.

We gather (6.21), (6.22), (6.23), and (6.24) in Inequality (6.20), and choose $V_2(p)$ with form $c_2\|r\|^2\frac{D_p}{n_0}$ for $c_2$ large enough.

□

### 6.5.3 Proof of Theorem 6.2

We introduce the set

$$\Lambda = \left\{\left|\frac{\|\hat{r}_{m^*}(.,\hat{F}_{0,n_0})\|}{\|r\|} - 1\right| < \frac{1}{2}\right\},$$

and split

$$\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\|^2\right] = \mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\|^2\mathbf{1}_\Lambda\right] + \mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\|^2\mathbf{1}_{\Lambda^c}\right].$$

We show in the sequel that the first term give the order of the upper-bound of Theorem 6.2, and that the probability of the set $\Lambda^c$ is negligible compared to $1/n + 1/n_0$.

• **Upper-bound for** $\mathbb{E}[\|\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0}) - r\|^2 \mathbf{1}_\Lambda]$. Arguing as in Section 6.5.2, we first obtain, for $m \in \mathcal{M}_{n,n_0}$

$$\left\|\hat{r}_{\tilde{m}}\left(.,\hat{F}_{0,n_0}\right) - r\right\|^2 \leq 6\left(\tilde{A}(m,\hat{F}_{0,n_0}) + \tilde{V}(m)\right) + 3\left\|\hat{r}_m\left(.,\hat{F}_{0,n_0}\right) - r\right\|^2.$$

Moreover, $\tilde{A}(m,\hat{F}_{0,n_0}) \leq A(m,\hat{F}_{0,n_0}) + \max_{m'\in\mathcal{M}_{n,n_0}}(V(m') - \tilde{V}(m'))_+$. Thus

$$\begin{aligned}\left\|\hat{r}_{\tilde{m}}\left(.,\hat{F}_{0,n_0}\right) - r\right\|^2 \leq\ & 6\left(A(m,\hat{F}_{0,n_0}) + V(m)\right) + 3\left\|\hat{r}_m\left(.,\hat{F}_{0,n_0}\right) - r\right\|^2, \\ & + \max_{m'\in\mathcal{M}_{n,n_0}}\left(V(m') - \tilde{V}(m')\right)_+ + 6\left(\tilde{V}(m) - V(m)\right).\end{aligned}$$

For every $m \in \mathcal{M}_{n,n_0}$,

$$\tilde{V}(m) - V(m) = c_0\frac{D_m}{n_0}\left(4\|\hat{r}_{m^*}(.,\hat{F}_{0,n_0})\|^2 - \|r\|^2\right).$$

On the set $\Lambda$, since $\|r\| < 2\|\hat{r}_{m*}(.,\hat{F}_{0,n_0})\|$, we thus have $(V(m') - \tilde{V}(m'))_+ = 0$. On this set, we also have : $\|\hat{r}_{m*}(.,\hat{F}_{0,n_0})\| \leq (3/2)\|r\|$,

$$\left(\hat{V}(m) - V(m)\right) \leq c\frac{D_m}{n_0}\left(4 \times \frac{9}{4}\|r\|^2 - \|r\|^2\right) = 8c\|r\|^2\frac{D_m}{n_0}.$$

Using also Lemma 6.3 enables to conclude

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\right\|^2 \mathbf{1}_\Lambda\right] \leq\ & \delta \min_{m\in\mathcal{M}_{n,n_0}}\left\{\frac{D_m}{n} + \|r\|^2\frac{D_m}{n_0} + 15\,\|r_m - r\|^2\right\} \\ & + \frac{C}{n} + \frac{C}{n_0}.\end{aligned}$$

• **Upper-bound for** $\mathbb{E}[\|\hat{r}_{\hat{m}}(.,\hat{F}_{0,n_0}) - r\|^2 \mathbf{1}_{\Lambda^c}]$. First, $\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\|^2 \leq 2\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})\|^2 + 2\|r\|^2$, and

$$\left\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})\right\|^2 = \sum_{j=1}^{D_{\tilde{m}}}\left(\hat{a}_j^{\hat{F}_0}\right)^2 \leq \left\|\sum_{j=1}^{D_{\tilde{m}}}\varphi_j^2\right\| \leq D_{\tilde{m}} \leq n \wedge n_0.$$

Thus

$$\mathbb{E}\left[\|\tilde{r}_{\tilde{m}}(.,\hat{F}_{0,n_0}) - r\|^2 \mathbf{1}_{\Lambda^c}\right] \leq \mathbb{P}\left(\Lambda^c\right) 2\left(n \wedge n_0 + 2\|r\|^2\right).$$

It remains to bound $\mathbb{P}(\Lambda^c)$. We split

$$\mathbb{P}(\Lambda^c) \leq \mathbf{1}_{\left\{\|r - r_{m^*}\| \geq \frac{\|r\|}{4}\right\}} + \mathbb{P}\left(\|r_{m^*} - \hat{r}_{m^*}(.,\hat{F}_{0,n_0})\| \geq \frac{\|r\|}{4}\right).$$

Recall that $r$ belongs to the Besov ball $B^\alpha_{2,\infty}((0;1),L)$, and that $D_{m*} \geq \ln(n_0)$. Hence, $\|r - r_{m^*}\| \leq D_{m^*}^{-\alpha} \leq (\ln(n_0))^{-\alpha}$. This quantity goes to 0 when $n_0$ goes to $+\infty$. Therefore, $(\ln(n_0))^{-\alpha} \leq \|r\|/4$ for $n_0$ large enough. Consequently,

$$\mathbf{1}_{\left\{\|r - r_{m^*}\| \geq \frac{\|r\|}{4}\right\}} = 0. \tag{6.25}$$

Thanks to (6.18), we also have

$$\mathbb{P}\left(\|r_{m^*} - \hat{r}_{m^*}(.,\hat{F}_{0,n_0})\|^2 \geq \frac{\|r\|^2}{16}\right) \leq \mathbb{P}\left(\sup_{t\in\mathcal{S}(m^*)} (\nu_n(t))^2 \geq \frac{\|r\|^2}{3\times 16}\right) + \mathbb{P}\left(4T_1^{m^*} \geq \frac{\|r\|^2}{3\times 16}\right) + \mathbb{P}\left(4T_2^{m^*} \geq \frac{\|r\|^2}{3\times 16}\right),$$

with $\nu_n$ defined by (6.17) and $T_1^{m^*}$, $T_2^{m^*}$ by (6.11). We use (6.12), and the mean-value theorem to obtain, $T_2^{m^*} \leq D_{m^*}^3 \|\varphi_2'\|^2_{L^\infty((0;1))} \|\hat{U}_{n_0} - id\|^2_{L^\infty((0;1))}$. Thus

$$\begin{aligned}\mathbb{P}\left(4T_2^{m^*} \geq \frac{\|r\|^2}{3\times 16}\right) &\leq \mathbb{P}\left(\left\|\hat{U}_{n_0} - id\right\|_\infty^2 \geq \frac{\|r\|^2}{192 D_{m^*}^3 \|\varphi_2'\|^2_{L^\infty((0;1))}}\right), \\ &\leq C\exp\left(-2n_0 \frac{\|r\|^2}{192 D_{m^*}^3 \|\varphi_2'\|^2_{L^\infty((0;1))}}\right).\end{aligned}$$

by applying Proposition 2.3. Since $D_{m^*}^3 \leq n_0/\ln^2(n_0)$, we have

$$\mathbb{P}\left(4T_2^{m^*} \geq \frac{\|r\|^2}{3\times 16}\right) \leq C\exp\left(-2\frac{\|r\|^2}{192\|\varphi_2'\|^2_{L^\infty((0;1))}} (\ln(n_0))^2\right). \tag{6.26}$$

The same arguments permit to bound the term in which $T_1^{m^*}$ is involved. We first note that $T_1^{m^*} \leq 4D_{m^*}^3 \|\varphi_2'\|^2_{L^\infty((0;1))} \|\hat{F}_{n_0} - F_0\|^2_{L^\infty(\mathbb{R})}$ and conclude with Proposition 2.3:

$$\mathbb{P}\left(4T_1^{m^*} \geq \frac{\|r\|^2}{48}\right) \leq C\exp\left(-\frac{\|r\|^2}{2\times 192\|\varphi_2'\|^2_{L^\infty((0;1))}} (\ln(n_0))^2\right). \tag{6.27}$$

We go back to the term involving the empirical process:

$$\mathbb{P}\left(\sup_{t\in\mathcal{S}(m^*)} (\nu_n(t))^2 \geq \frac{\|r\|^2}{48}\right) \leq \sum_{j=1}^{D_{m^*}} \mathbb{P}\left(\nu_n^2(\varphi_j) \geq \frac{\|r\|^2}{48 D_{m^*}}\right).$$

Writing $\nu_n(\varphi_j) = (1/n)\sum_{i=1}^n Z_i^j - \mathbb{E}[Z_i^j]$ with $Z_i^j = \varphi_j(F_0(X_i))$ (see (6.17)) allows to apply the Bernstein Inequality recalled in Proposition 2.1 (Chapter 2). We compute $b = \sqrt{D_{m^*}}$, and $v = n\|r\|_{L^\infty((0;1))}$. This leads, for $u > 0$

$$\mathbb{P}\left(\nu_n^2(\varphi_j) \geq \sqrt{2\|r\|_{L^\infty((0;1))} u} + u\frac{1}{3}\sqrt{D_{m^*}}\right) \leq e^{-nu}.$$

Choosing $u = a/D_{m^*}^4$, for a constant $a$ and using $D_{m^*} \geq \ln(n_0)$, we can obtain $\sqrt{2\|r\|_{L^\infty((0;1))} u} + u\sqrt{D_{m^*}}/3 \leq \|r\|^2/48 D_{m^*}$, for $n_0$ large enough, and

$$\sum_{j=1}^{D_{m^*}} \mathbb{P}\left(\nu_n^2(\varphi_j) \geq \frac{\|r\|^2}{48 D_{m^*}}\right) \leq D_{m^*}\exp\left(-n\frac{a}{D_{m^*}^4}\right). \tag{6.28}$$

Putting together (6.25), (6.27), (6.26) and (6.28), we have proved

$$\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})-r\|^2\mathbf{1}_{\Lambda^c}\right] \leq 2\left(n\wedge n_0+2\|r\|^2\right)\left\{C\exp\left(-\frac{\|r\|^2}{2\times 192\|\varphi_2'\|^2_{L^\infty((0;1))}}\left(\ln(n_0)\right)^2\right)\right.$$
$$\left.+C\exp\left(-2\frac{\|r\|^2}{192\|\varphi_2'\|^2_{L^\infty((0;1))}}\left(\ln(n_0)\right)^2\right)+D_{m^*}\exp\left(-\frac{na}{D^4_{m^*}}\right)\right\}.$$

Recall that $D_{m^*}\leq C(n/\ln(n))^{1/4}$. The last term of this upper-bound is thus negligible compared to $1/n$ (if $a$ is large enough). The two other terms have the order $n_0\exp(-C(\ln(n_0))^2)$, and are thus smaller than $C/n_0$. Finally,

$$\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_{0,n_0})-r\|^2\mathbf{1}_{\Lambda^c}\right] \leq \frac{C}{n_0}+\frac{C}{n}.$$

□

# Chapitre 7

# Estimation de la densité relative à partir de données censurées à droite.

**Sommaire**



Ce chapitre présente un travail en cours.

**Résumé.** L'objectif est d'étendre la problématique du chapitre précédent, à savoir l'estimation de la densité relative, au cas où les observations sont incomplètes. Précisément, on suppose que les données peuvent être censurées aléatoirement à droite, ce qui intervient fréquemment dans les applications : un seul, ou alors les deux échantillons pourront être seulement partiellement observés. Deux stratégies d'estimation sont proposées. La première consiste en une adaptation de la procédure du chapitre précédent, faisant appel à des méthodes classiques en analyse de survie (estimateur de Kaplan-Meier et correction de censure). La seconde, plus originale, repose sur le principe de déformation, développé dans la première partie de cette thèse. Dans les deux cas, des bornes supérieures pour le risque intégré de collections d'estimateurs de type projection sont démontrées, et une procédure de sélection de modèles mise en place.

**Abstract.** The aim is to extend the problem of relative density estimation, which is the subject of the previous chapter, to partially observed data. We assume that a random right censoring may occur for one, or both of the data samples. Two estimation strategies are proposed. The guiding idea for the first method is to adapt the estimator of the previous chapter, thanks to technicalities which are classical in survival analysis (Kaplan-Meier estimator, correction of censoring). The second one is more original. It is based on the deformation device, developed in the first part of this thesis. In both cases, upper-bounds for the global squared-errors of projection-types estimator are proved, and a model selection method is set.

## 7.1 Introduction

La comparaison d'un groupe d'individus à une population de référence au travers de la comparaison de deux fonctions de répartition $F$ et $F_0$ de variables réelles $X$ et $X_0$ est un objectif important en statistique et les applications dans des domaines variés comme la recherche médicale (comparaison d'un groupe de malades à un groupe sain) ou les sciences sociales sont nombreuses. Un cas fréquent en pratique est celui où les observations sont incomplètes, car possiblement censurées de manière aléatoire à droite. On peut par exemple être intéressé par la comparaison de l'effet de deux traitements sur des groupes de patients qui ne sont pas nécessairement observés jusqu'à la date prévue de fin d'étude. Ceci motive l'estimation de la densité relative (6.1) de $X_0$ sachant $X$, introduite au chapitre précédent, à partir d'observations censurées.

Les données ne sont alors plus composées d'échantillons des variables $X_0$ et $X$, mais d'échantillons de couples $(Z_0, \delta_0)$ et $(Z, \delta)$, avec $Z_0 = X_0 \wedge C_0$, $\delta_0 = \mathbf{1}_{X_0 \leq C_0}$, $Z = X \wedge C$, et $\delta = \mathbf{1}_{X \leq C}$. Dans un tel modèle, les variables d'intérêt $X$ et $X_0$ représentent typiquement des temps aléatoires de survie, de supports respectifs $(0; \tau_X)$ et $(0; \tau_{X_0})$ $(0 < \tau_X, \tau_{X_0} \leq \infty)$, et $C$ et $C_0$ désignent des instants aléatoires de censure (de supports respectifs $(0; \tau_C)$ et $(0; \tau_{C_0})$, $0 < \tau_C, \tau_{C_0} \leq \infty$). Les variables $\delta$ et $\delta_0$ sont alors des indicateurs de non censure. La fonction d'intérêt est toujours la densité relative de $X$ par rapport à $X_0$, dont on rappelle la définition :

$$r(x) = \frac{f \circ F_0^{-1}(x)}{f_0 \circ F_0^{-1}(x)}, \quad x \in F_0((0; \tau_X)),$$

où $f$ et $f_0$ sont des densités respectives de $X$ et $X_0$, et $F_0^{-1}$ l'inverse de $F_0$, supposé exister. On supposera également que les temps de censure $C$ et $C_0$ admettent des densités qui seront notées $f_C$ et $f_{C_0}$ ; leurs fonctions de répartition seront notées $G$ et $G_0$.

L'extension des méthodes non paramétriques utilisées dans les problèmes à deux échantillons et brièvement décrites à la Section 6.1 (tests statistiques, courbes ROC...) à des données partiellement observées fait l'objet de nombreuses études. Citons par exemple les travaux de Andersen & Rønn (1995) (données censurées à droite ou à gauche) et Zhang *et al.* (2003) (données censurées par intervalle) concernant la construction de tests de comparaison des lois de deux échantillons, tests fondés sur des estimateurs de type maximum de vraisemblance pour le premier, et de type empiriques pour le second. Dans le domaine de l'estimation, les travaux dans ce cadre s'appuient souvent sur l'estimateur désormais classique de la fonction de survie, l'estimateur de Kaplan & Meier (1958), dont la définition ainsi que quelques propriétés sont rappelées au Chapitre 2, Section 2.1.3. Wang *et al.* (2009) ont cherché par exemple à estimer l'aire sous la courbe ROC, à partir d'estimateurs non paramétriques de celle-ci, alors que Zhou & Liang (2005) se sont intéressés, avec une stratégie semi-paramétrique à différents problèmes de comparaison de deux échantillons, l'un de loi paramétrique et l'autre de loi non paramétrique. L'estimation de la densité relative à partir d'observations censurées a été envisagée par Cao *et al.* (2000) : des estimateurs à noyaux sont bâtis, et une représentation asymptotique de la loi limite démontrée. Un choix algorithmique de la fenêtre est proposé ensuite dans Cao *et al.* (2001), et les méthodes sont étendues à la fonction de risque ins-

tantané relative dans Cao *et al.* (2005). Pons (2007) s'intéresse également à la convergence en probabilité d'estimateurs à noyaux dans ce cadre, tout en envisageant également des modèles semi-paramétriques. Enfin, Molanes-López & Cao (2008b) traitent le cas de données censurées à la fois à droite et à gauche, et proposent un développement asymptotique du risque ponctuel et intégré (biais et variance) d'un estimateur à noyau de la densité relative. La sélection de la fenêtre se fait sous condition de régularité sur la fonction estimée, à l'aide d'une règle de type "*rule of thumb*" (Silverman, 1986).

Ce travail, encore en cours, présente deux stratégies possibles pour estimer la densité relative de $X$ par rapport à $X_0$ par projection, dans le cas où l'un des deux échantillons d'observations au moins est censuré à droite. La première stratégie fait l'objet de la Section 7.2 : partant de la construction d'estimateurs par projection dans le cas non-censuré faite au Chapitre 6, nous nous proposons d'adapter la définition d'un contraste au cas de la censure pour construire une collection d'estimateurs par projection de $r$ sur des modèles trigonométriques. Les outils classiques d'analyse de survie de type correction de censure sont utilisés. La sélection du modèle se fait comme au chapitre précédent par une méthode inspirée de Goldenshluger & Lepski (2011a). On obtient une borne non-asymptotique pour le risque de l'estimateur sélectionné, et l'adaptation entraîne l'obtention de la vitesse attendue, sur un intervalle d'estimation de la forme $(0;\tau)$ avec $\tau$ strictement inférieur aux extrémités des supports de toutes les variables. La seconde stratégie, décrite à la Section 7.3, permet l'estimation sur un intervalle de longueur légèrement supérieure, mais ne permet pour l'instant que d'envisager le cas où un seul des deux échantillons est censuré. Elle est fondée sur la méthode de déformation détaillée dans la Partie A de la thèse (voir particulièrement le Chapitre 3) : on introduit une fonction originale de déformation, et on débute par l'estimation d'une fonction auxiliaire via la minimisation d'un contraste. On obtient à nouveau une collection d'estimateurs par projection, que l'on déforme ensuite pour estimer la fonction cible. Enfin, les preuves sont rassemblées dans la Section 7.4.

## 7.2 Estimation de la densité relative avec censure, sans déformation

Dans cette partie, nous présentons une première méthode d'estimation de la densité relative à partir d'un double échantillon de variables aléatoires sujettes à une censure aléatoire à droite.

Les observations sont constituées des deux échantillons indépendants suivants : un premier échantillon $(Z_{0,i_0},\delta_{0,i_0})_{i_0=1,\dots,n_0}$ de variables aléatoires $Z_{0,i_0}=X_{0,i_0}\wedge C_{0,i_0}$, et $\delta_{0,i_0}=\mathbf{1}_{X_{0,i_0}\leq C_{0,i_0}}$, distribuées comme $(Z_0,\delta_0)$ définies ci-dessus (voir Section 7.1) ; et un second échantillon $(Z_i,\delta_i)_{i=1,\dots,n_0}$ de variables aléatoires $Z_i=X_i\wedge C_i$, et $\delta_i=\mathbf{1}_{X_i\leq C_i}$, distribuées comme $(Z,\delta)$.

L'objectif est d'estimer la densité relative de $X$ par rapport à $X_0$ à partir de ces échantillons, sur l'intervalle $(0, F_0(\tau))$, pour un réel strictement positif $\tau$ tel que

$$\tau < \min(\tau_X, \tau_C, \tau_{X_0}, \tau_{C_0}),$$

à l'aide d'estimateurs par projection sur des modèles engendrés par la base trigonométrique $S_m = \text{Vect}\{\varphi_1, \ldots, \varphi_{D_m}\}$, $m \in \mathcal{M}_{n,n_0}$, modèles analogues à ceux définis à la Section 6.2.1 du chapitre précédent, mais translatés pour obtenir une base orthonormée de $L^2((0; F_0(\tau)))$. On note $\|.\|$ la norme $L^2$ sur $(0; F_0(\tau))$ et $\langle ., . \rangle_{(0;F_0(\tau))}$ le produit scalaire associé. Dans un premier temps, la minimisation d'un contraste permet de définir une collection d'estimateurs. Pour chacun d'entre eux, le risque quadratique intégré admet une décomposition de type biais-variance classique, suggérant ainsi la mise en place dans un second temps d'une procédure de sélection réalisant automatiquement le compromis.

### 7.2.1 Collection d'estimateurs par projection

#### Contraste

Nous adaptons la définition du contraste (6.2) défini au Chapitre 6 pour maintenant retrouver la fonction $r$ à l'aide d'échantillons de variables éventuellement censurées. La fonction de répartition empirique $\hat{F}_{0,n_0}$ des observations $(X_{0,i_0})_{i_0}$ est donc tout d'abord remplacé par l'estimateur de Kaplan-Meier modifié $\check{F}_{0,n_0}$ défini par (2.8) (Chapitre 2), construit à partir de l'échantillon de référence $((Z_{0,i_0}, \delta_{0,i_0}))_{i_0}$. On introduit également la correction de censure de Koul *et al.* (1981), pour tenir compte de la censure dont les variables $(X_i)_i$ font l'objet. Finalement, la définition est donc la suivante, pour $t \in L^2((0; F_0(\tau_Z)))$,

$$\gamma_n^c(t, \check{F}_{0,n_0}, \check{\bar{G}}_n) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} \frac{\delta_i}{\check{\bar{G}}_n(Z_i)} t \circ \check{F}_{0,n_0}(Z_i) \mathbf{1}_{Z_i \leq \tau}, \tag{7.1}$$

où $\check{\bar{G}}_n$ est l'estimateur de Kaplan-Meier modifié de la survie $\bar{G} = 1 - G$ de la variable $C$ (voir toujours (2.8)). Pour chaque indice $m \in \mathcal{M}_{n,n_0}$, posons

$$\hat{r}_m(., \check{F}_{0,n_0}, \check{\bar{G}}_n) = \arg\min_{t \in \mathcal{S}_m} \gamma_n^c(t, \check{F}_{0,n_0}, \check{\bar{G}}_n).$$

Le minimum est défini de façon unique, et la fonction en lequel il est atteint s'écrit

$$\hat{r}_m(., \check{F}_{0,n_0}, \check{\bar{G}}_n) = \sum_{j=1}^{D_m} \hat{a}_j^{\check{F}_{0,n_0}, \check{\bar{G}}_n} \varphi_j, \quad \text{avec} \quad \hat{a}_j^{\check{F}_{0,n_0}, \check{\bar{G}}_n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\check{\bar{G}}_n(Z_i)} \varphi_j \left( \check{F}_{0,n_0}(Z_i) \right) \mathbf{1}_{Z_i \leq \tau}. \tag{7.2}$$

La pertinence de l'introduction du contraste est illustrée par le calcul du biais du pseudo-estimateur $\hat{r}_m(., F_0, \bar{G})$, obtenu en substituant les "vraies" fonctions $F_0$ et $\bar{G}$ à leurs contreparties empiriques $\check{F}_{0,n_0}$, $\check{\bar{G}}_n$. Calculons tout d'abord, à l'aide d'un calcul détaillé au Lemme

7.1 (voir Section 7.4.1),

$$\begin{aligned}\mathbb{E}\left[\hat{a}_j^{F_0,\bar{G}}\right] &= \mathbb{E}\left[\frac{\mathbf{1}_{X\leq C}}{\bar{G}(X)}\varphi_j\circ F_0(X)\mathbf{1}_{X\leq\tau}\right], \qquad (7.3)\\ &= \int_{(0;\tau_Z)}\frac{\mathbf{1}_{x\leq\tau}}{\bar{G}(x)}\varphi_j\circ F_0(x)f(x)\bar{G}(x)dx,\\ &= \int_{(0;\tau)}\varphi_j\circ F_0(x)f(x)dx = \int_{(0;F_0(\tau))}\varphi_j(u)r(u)du = \langle\varphi_j,r\rangle_{(0;F_0(\tau))}. \qquad (7.4)\end{aligned}$$

Comme $\mathbb{E}[\hat{r}_m(.,F_0,\bar{G})] = \sum_{j=1}^{D_m}\mathbb{E}[\hat{a}_j^{F_0,\bar{G}}]\varphi_j = \sum_{j=1}^{D_m}\langle\varphi_j,r\rangle_{(0;F_0(\tau))}\varphi_j$, la fonction $\hat{r}_m(.,F_0,\bar{G})$ estime sans biais la projection orthogonale de la cible $r$ sur le modèle $S_m$.

**Remarque 7.1. Complément pour la justification de la correction de censure.** L'idée du contraste dans le cas sans censure était de remplacer $\int_0^{\tau_X} t(F_0(x))f(x)dx = \langle t,r\rangle_{(0;F_0(\tau_X))} = \mathbb{E}[t\circ F_0(X)]$ par son équivalent empirique : $n^{-1}\sum_{i=1}^n t\circ F_0(X_i)$ (dans le cas où $F_0$ est connue). Heuristiquement, $f(x)dx$ était estimé par la variable $X_i$ (et on moyennait sur tout l'échantillon). Dans le cas où $X_i$ est censurée, il faut une autre solution pour estimer $f(x)dx$. Pour cela, on constate, par le Lemme 7.1, que

$$\mathbb{E}\left[\mathbf{1}_{Z\leq x,\delta=1}\right] = \mathbb{E}\left[\mathbf{1}_{Z\leq x}\delta\right] = \int_0^x f(x')\bar{G}(x')dx'.$$

Ainsi, $f(x)dx$ est estimé par une moyenne empirique des "$d\mathbf{1}_{Z\leq x,\delta=1}/\bar{G}(x)$" en remplaçant également $\bar{G}$ par son estimateur de Kaplan-Meier. Donc, en notant

$$\bar{N}(x) = \sum_{i=1}^n \mathbf{1}_{Z_i\leq x,\delta_i=1},$$

on estime $f(x)dx$ par $d\bar{N}(x)/(n\check{\bar{G}}_n(x))$. Comme

$$\frac{1}{n}\int_0^\tau t\circ F_0(x)\frac{d\bar{N}(x)}{\bar{G}(x)} = \frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\bar{G}(Z_i)}t\circ F_0(Z_i)\mathbf{1}_{Z_i\leq\tau},$$

où l'on a également introduit la restriction à l'intervalle $(0;\tau)$, le contraste annoncé ci-dessus est ainsi justifié.

### 7.2.2 Risque d'un estimateur par projection

Le calcul (7.3) prouve que dans le cas simplifié où $F_0$ et $\bar{G}$ sont connues, le risque quadratique va se décomposer en la somme d'un terme de biais et d'un terme de variance. La proposition suivante donne l'équivalent dans le cas général. Sa preuve est détaillée à la Section 7.4.1.

**Proposition 7.1.** *La fonction $r$ est supposée de classe $\mathcal{C}^1$ sur $[0;F_0(\tau)]$. Pour chaque modèle $S_m$, $m\in\mathcal{M}_{n,n_0}$, tel que $D_m\leq\kappa n_0^{1/3}$ (pour une constante $\kappa>0$), l'estimateur $\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)$ défini par (7.2) vérifie l'inégalité suivante*

$$\mathbb{E}\left[\left\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2\right] \leq 4\left\|r-r_m\right\|^2 + c_1'\left(\frac{D_m}{n}+\frac{D_m}{n_0}\right)+c_2'\left(\frac{1}{n}+\frac{1}{n_0}\right).$$

*où $c_1'$ et $c_2'$ sont deux constantes indépendantes de $m$, $n$, et $n_0$.*

La Proposition 7.1 est l'analogue, dans le cas d'observations censurées, de la Proposition 6.1 du chapitre précédent. La censure ne dégrade donc pas le résultat obtenu dans le cas où les données sont complètes. Les hypothèses sont les mêmes, seul l'intervalle d'estimation a dû être restreint (estimation sur $(0; F_0(\tau))$ au lieu de $(0; F_0(\tau_X))$ dans le cas sans censure). Le risque se décompose donc à nouveau en trois termes : le premier est le terme de biais, le second la variance, et le dernier un reste négligeable. Si l'on suppose que la fonction $r$ appartient à une boule d'un espace de Besov d'indice de régularité $\alpha$ connu, l'indice $m$ minimisant le membre de droite de l'inégalité de la Proposition 7.1 (c'est-à-dire réalisant le compromis biais-variance) peut-être calculé, et l'on obtient la vitesse classique de convergence du risque, adaptée aux problèmes à deux échantillons : $\mathbb{E}[\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n) - r\|^2] = O([\max(1/n, 1/n_0)]^{(2\alpha/(2\alpha+1))})$, comme obtenu au Corollaire 6.1 du Chapitre 6. Le problème est maintenant de construire un estimateur atteignant cette même vitesse tout en étant construit sans supposer $\alpha$ connu.

### 7.2.3 Estimation adaptative

#### Critère de sélection de modèles

On part de la collection d'estimateurs $(\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n))_{m\in\mathcal{M}_{n,n_0}}$, et l'on souhaite en sélectionner un. On utilise toujours la méthode inspirée de celle de Goldenshluger & Lepski (2011a) pour la sélection de fenêtres.

On définit les deux quantités suivantes, quel que soit $m \in \mathcal{M}_{n,n_0}$, et pour une constante $c > 0$ suffisamment grande (voir preuves) :

$$\begin{aligned} V(m) &= c\left\{a_1\frac{D_m}{n} + (a_2+a_3)\frac{D_m}{n_0}\right\}, \\ A(m) &= \max_{m'\in\mathcal{M}_{n,n_0}}\left(\left\|\hat{r}_{m'}(.,\check{F}_{0,n_0},\check{\bar{G}}_n) - \hat{r}_{m\wedge m'}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)\right\|^2 - V(m')\right)_+, \end{aligned} \tag{7.5}$$

où $a_1$, $a_2$, $a_3$ sont définies comme suit ($\bar{c}_2$ est la constante de l'Inégalité de Bitouzé *et al.* 1999, voir Proposition 2.6) :

$$a_1 = \mathbb{E}\left[\frac{\delta \mathbf{1}_{Z\le\tau}}{\bar{G}^2(Z)}\right], \quad a_2 = 3\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^6(x)} f_0(x)dx, \quad a_3 = \frac{\bar{c}_2 r^2(F_0(\tau))}{\bar{G}_0^2(\tau)}.$$

Ces quantités inconnues intervenant dans la définition de $V$ peuvent être remplacées par des estimateurs, de manière similaire à ce qui a été fait au Théorème 6.2 du chapitre précédent. Enfin, posons

$$\hat{m} = \mathrm{argmin}_{m\in\mathcal{M}_{n,n_0}}\{A(m) + V(m)\}.$$

L'estimateur sélectionné ainsi est donc $\hat{r}_{\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)$.

### 7.2.4 Borne non asymptotique pour le risque

Le résultat principal que l'on obtient est le suivant, démontré à la Section 7.4.1.

**Théorème 7.1.** *On suppose que la fonction $r$ est de classe $\mathcal{C}^1$ sur $[0; F_0(\tau)]$, $\sup_{x\in(0;\tau)} f(x) \leq f_\infty < \infty$ et $\sup_{x\in(0;\tau)} f_0(x) \leq f_{0,\infty} < \infty$. On suppose également, pour une constante $\kappa > 0$,*

$$\forall m \in \mathcal{M}_{n,n_0} \;\; D_m \leq \kappa \frac{n_0^{1/3}}{\ln^{2/3}(n_0)}, \quad D_m \leq \frac{\sqrt{n_0}}{4f_{0,\infty}}, \quad D_m \leq \frac{\sqrt{n}}{4f_\infty}. \tag{7.6}$$

*L'estimateur $\hat{r}_{\hat{m}}(., \check{F}_{0,n_0}, \check{\check{G}}_n)$ vérifie l'inégalité de type oracle suivante :*

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{r}_{\hat{m}}(., \check{F}_{0,n_0}, \check{\check{G}}_n) - r\right\|^2\right] \;\leq\; & \delta \min_{m\in\mathcal{M}_{n,n_0}} \left\{\frac{D_m}{n} + \frac{D_m}{n_0} + \|r_m - r\|^2\right\} \\ & + C\left(\frac{\ln(n)}{n} + \frac{\ln(n_0)}{n_0}\right),\end{aligned}$$

*avec $\delta$ et $C$ deux constantes dépendant de la fonction $r$, mais ni de $n$, ni de $n_0$.*

**Remarque 7.2.** Si $n_0$ est assez grand, la seconde contrainte $D_m \leq \sqrt{n_0}/(4f_{0,\infty})$ est entraînée par la première $D_m \leq \kappa n_0^{1/3}/(\ln^{2/3}(n_0))$.

L'adaptation se fait donc de nouveau sans perte : le compromis biais-variance est automatiquement réalisé, et la vitesse de décroissance du risque est donc celle que l'on attendait, suite à l'étude du risque sur un modèle fixé. En supposant que la fonction $r$ appartient à une boule d'un espace de Besov de régularité $\alpha$, on en déduit en effet

$$\mathbb{E}\left[\left\|\hat{r}_{\hat{m}}(., \check{F}_{0,n_0}, \check{\check{G}}_n) - r\right\|^2\right] \leq C\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

## 7.3 Estimation de la densité relative avec censure, méthode de déformation

Dans cette partie, nous présentons une méthode d'estimation de la densité relative à partir d'un double échantillon de variables aléatoires dont un seul est sujet à censure à droite, à l'aide d'une méthode plus originale que celle de la Section 7.2 ci-dessus, fondée sur une déformation, au sens introduit dans la première partie de cette thèse (Chapitres 3 à 5). Notre but est de construire un estimateur ayant des propriétés analogues à l'estimateur $\hat{r}_{\hat{m}}(., \check{F}_{0,n_0}, \check{\check{G}}_n)$ de la Section 7.3. La déformation va permettre d'éviter la présence d'un estimateur de Kaplan-Meier au dénominateur (voir (7.2) et (7.1)) et donc d'augmenter la taille de l'intervalle d'estimation. En contrepartie, nous supposons pour l'instant qu'un seul des deux échantillons est susceptible d'être censuré. Les variables de l'échantillon défini comme étant celui de référence $(X_{0,i_0})_{i_0}$ sont supposées complètement observées.

Les observations sont donc d'abord constituées des deux échantillons indépendants suivants : un premier échantillon $(X_{0,i_0})_{i_0=1,\ldots,n_0}$ de variables aléatoires, distribuées comme $X_0$, et un second échantillon $(Z_i, \delta_i)_{i=1,\ldots,n_0}$ de variables aléatoires $Z_i = X_i \wedge C_i$, et $\delta_i = \mathbf{1}_{X_i \leq C_i}$, distribuées comme $(Z, \delta)$. Enfin, et comme pour les méthodes de déformation introduites aux Chapitres 3 et 5 par exemple, on suppose observer un échantillon indépendant des observations

de départ $(Z_{-i}, \delta_{-i})_{i=1,\dots n}$, distribué comme $(Z, \delta)$. Ceci revient une nouvelle fois uniquement à supposer l'échantillon de départ de taille $2n$ et à le couper en deux morceaux. Par contre, il est important de noter que l'on ne suppose pas ici disposer d'un second échantillon pour la variable $X_0$.

L'objectif est d'estimer la densité relative de $X$ par rapport à $X_0$ à partir de ces échantillons, sur l'intervalle $(0, F_0(\tau_X))$. On supposera dans la suite que les inégalités suivantes portant sur les extrémités des supports des variables sont vérifiées :

$$\tau_{X_0} < \tau_X \leq \tau_C. \tag{7.7}$$

On a donc $\tau_Z = \tau_X \wedge \tau_C = \tau_X$.

### 7.3.1 Présentation de la stratégie d'estimation

Comme dans la méthode générale de déformation introduite au Chapitre 4, on va définir une fonction de déformation $\Phi$, commencer par estimer la fonction auxiliaire $g = r \circ \Phi^{-1}$ par une collection d'estimateurs par projection $(\hat{g}_m)_m$ sur des sous-espaces $S_m$, puis définir une collection d'estimateurs pour la fonction cible $r$ en posant $\hat{r}_m = \hat{g}_m \circ \Phi$ ou $\hat{r}_m = \hat{g}_m \circ \hat{\Phi}$, selon que $\Phi$ est supposée connue ou non ($\hat{\Phi}$ désignant un équivalent empirique de $\Phi$).

Précisément, on suppose que $F_0$ admet un inverse sur $(0; \tau_{X_0})$. On pose

$$\Phi(x) = \int_0^x \bar{G} \circ F_0^{-1}(u) du, \quad x \in (0; 1). \tag{7.8}$$

La transformation $\Phi$ est dérivable, et $\Phi'(x) = \bar{G} \circ F_0^{-1}(x)$. Cette quantité est strictement positive pour $0 < x < F_0(\tau_C)$. Donc $\Phi$ est bijective sur $(0, F_0(\tau_C))$.

On définit la fonction auxiliaire

$$g : x \in (0; \Phi \circ F_0(\tau_Z)) \ \mapsto \ r \circ \Phi^{-1}(x). \tag{7.9}$$

L'introduction de cette fonction sera justifiée à la section suivante, lors de la définition du contraste. La fonction suivante va également jouer un rôle important : on pose, pour $x \in (0; \tau_{X_0})$,

$$B(x) = \Phi \circ F_0(x). \tag{7.10}$$

On a

$$g \circ B = g \circ \Phi \circ F_0 = r \circ F_0 = \frac{f}{f_0},$$

et aussi, $B'(x) = \Phi' \circ F_0(x) f_0(x) = \bar{G}(x) f_0(x)$. Enfin, remarquons que

$$B(x) = \int_0^{F_0(x)} \bar{G} \circ F_0^{-1}(u) du = \int_0^x \bar{G}(v) f_0(v) dv = \mathbb{E}\left[\bar{G}(X_0) \mathbf{1}_{X_0 \leq x}\right], \tag{7.11}$$

ce qui permettra de substituer facilement à $B$ un estimateur.

### 7.3.2 Estimation de la fonction auxiliaire

Dans cette section, on se propose d'estimer la fonction auxiliaire $g = r \circ \Phi^{-1}$ sur $(0; \Phi \circ F_0(\tau_Z)) = (0; B(\tau_Z))$. Nous construisons toujours des estimateurs par projection sur des modèles trigonométriques $S_m = \text{Vect}\{\varphi_1, \ldots, \varphi_{D_m}\}$, $m \in \mathcal{M}_{n,n_0}$, où $(\varphi_j)_j$ forme cette fois une base orthonormée de $L^2((0; B(\tau_Z)))$. On note $\|.\|_B$ la norme $L^2$ sur $(0; B(\tau_Z))$ et $\langle ., . \rangle_{(0;B(\tau_Z))}$ le produit scalaire associé.

#### Construction d'une collection d'estimateurs par projection

Plaçons nous un instant dans le cas simplifié où les fonctions $\Phi$, et $F_0$ (et donc aussi $B$) seraient connues. On introduit, pour $t \in L^2((0; B(\tau_Z)))$,

$$\gamma_n(t, B) = \|t\|_B^2 - \frac{2}{n} \sum_{i=1}^{n} \delta_i t \circ B(Z_i).$$

Alors,

$$\begin{aligned} \mathbb{E}\left[\gamma_n(t, B)\right] - \mathbb{E}\left[\gamma_n(g, B)\right] &= \|t\|_B^2 - \|r \circ \Phi^{-1}\|_B^2 \\ &\quad -2\mathbb{E}\left[\delta_1 (t - r \circ \Phi^{-1}) \circ \Phi \circ F_0(X_1)\right], \end{aligned}$$

avec

$$\mathbb{E}\left[\delta_1 (t - g) \circ \Phi \circ F_0(X_1)\right] = \int_{(0;\tau_Z)} \bar{G}(x)(t - r \circ \Phi^{-1}) \circ \Phi \circ F_0(x) f(x) dx,$$

par le Lemme 7.1. On pose ensuite le changement de variables $u = B(x)$. Donc, $du = f_0(x)\bar{G}(x)dx$, et donc $\bar{G}(x)dx = du/f_0 \circ F_0^{-1} \circ \Phi^{-1}(u)$ :

$$\begin{aligned} \mathbb{E}\left[\delta_1 (t - r \circ \Phi^{-1}) \circ \Phi \circ F_0(X_1)\right] &= \int_{(0;B(\tau_Z))} (t - r \circ \Phi^{-1})(u) \frac{f \circ F_0^{-1} \circ \Phi^{-1}(u)}{f_0 \circ F_0^{-1} \circ \Phi^{-1}(u)} du, \\ &= \int_{(0;B(\tau_Z))} (t - r \circ \Phi^{-1})(u) r \circ \Phi^{-1}(u) du. \end{aligned}$$

Puis,

$$\begin{aligned} \mathbb{E}\left[\gamma_n(t, B)\right] - \mathbb{E}\left[\gamma_n(g, B)\right] &= \|t\|_B^2 - \|r \circ \Phi^{-1}\|_B^2 \\ &\quad -2\langle t - r \circ \Phi^{-1}, r \circ \Phi^{-1} \rangle_{(0;B(\tau_Z))}, \\ &= \|t - r \circ \Phi^{-1}\|_B^2 = \|t - g\|_B^2. \end{aligned}$$

Ainsi, minimiser $\gamma_n(., B)$ sur un sous espace vectoriel $S_m$ conduit à un estimateur raisonable pour la fonction $g$.

Cependant, la fonction $B$ n'est a priori pas connue. On lui substitue donc un estimateur. Comme dans les autres problèmes de déformation, celui-ci va être calculé à partir de l'échantillon indépendant des observations de départ $(Z_{-i}, \delta_{-i})_{i=1,\ldots n}$. L'expression de $B$ sous forme

d'espérance (cf (7.11)) nous permet de proposer un estimateur de type moment :

$$\hat{B} : x \mapsto \frac{1}{n_0} \sum_{i_0=1}^{n_0} \check{\bar{G}}_n^{(-)}(X_{0,i_0}) \mathbf{1}_{X_{0,i_0} \leq x}, \tag{7.12}$$

où l'on note $\check{\bar{G}}_n^{(-)}$ est l'estimateur de Kaplan-Meier de la survie de la variable $C$, calculé à partir de l'échantillon $(X_{-i}, C_{-i})_i$. On pose alors, pour $m \in \mathcal{M}_{n,n_0}$ :

$$\forall m \in \mathcal{M}_{n,n_0},\ \hat{g}_m(.,\hat{B}) = \arg \inf_{t \in S_m} \gamma_n(t, \hat{B}).$$

On obtient l'expression suivante :

$$\hat{g}_m(.,\hat{B}) = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{B}} \varphi_j, \quad \text{avec } \hat{a}_j^{\hat{B}} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \varphi_j \circ \hat{B}(Z_i). \tag{7.13}$$

L'estimateur $\hat{g}_m(.,\hat{B})$ admet donc une expression simple : le calcul de ses coefficients dans la base $(\varphi_j)$ ne fait pas intervenir de quotient. Notons aussi que la fonction $\hat{g}_m(.,B)$ estime sans biais la projection $\Pi_{S_m} g$ de $g$ sur le modèle $S_m$ : en effet, par un calcul similaire à celui effectué ci-dessus,

$$\mathbb{E}\left[\hat{a}_j^B\right] \quad = \quad \mathbb{E}\left[\delta_1 \varphi_j \circ \Phi \circ F_0(Z_1)\right] = \langle r \circ \Phi^{-1}, \varphi_j \rangle_{(0;B(\tau_Z))}.$$

**Etude du risque**

La borne que nous obtenons pour le risque est la suivante :

**Proposition 7.2.** *L'hypothèse (7.7) est supposée vérifiée. On suppose aussi $g$ bornée et les trois intégrales suivantes finies : $\int_0^{\tau_Z} (g \circ B)^2(x) \bar{G}(x) f_0(x) dx$, $\int_0^{\tau_Z} (g \circ B)^2(x) dx$ et $\int_0^{\tau_Z} \bar{G}^2(x) f_0^2(x) dx$. Enfin, la dimension du modèle $S_m$, $m \in \mathcal{M}_{n,n_0}$, sur lequel on se place est supposée majorée :*

$$D_m \leq \kappa n^{1/3} \quad et \quad D_m \leq \kappa n_0^{1/3},$$

*pour $\kappa > 0$. Alors, l'estimateur $\hat{g}_m(.,\hat{B})$ satisfait l'inégalité suivante, pour des constantes $C_1$ et $C_2$ indépendantes de $m, n$ et $n_0$,*

$$\mathbb{E}\left[\left\|\hat{g}_m(.,\hat{B}) - g\right\|_B^2\right] \quad \leq \quad C_1 \frac{D_m}{n} + C_2 \frac{D_m}{n_0} + 3 \left\|\Pi_{S_m} g - g\right\|_B^2.$$

Comme dans tous les problèmes faisant intervenir une déformation, un argument important de la preuve est le contrôle de l'écart entre $\hat{B}$ et $B$. C'est l'objet du premier paragraphe de la Section 7.4.2 regroupant les démonstrations. Une fois de telles propriétés démontrées, la preuve de la Proposition 7.2 fait appel à des arguments déjà utilisés dans cette thèse. On en donnera donc uniquement les grandes lignes.

### 7.3.3 Estimation de la fonction cible

Dans une première étape, on s'est intéressé à l'estimation de $g = r \circ \Phi^{-1}$, sur $(0; B(\tau_Z))$. On a construit des estimateurs $\hat{g}_m(., \hat{B})$, $m \in \mathcal{M}_{n,n_0}$. Pour estimer $r$, on définit donc naturellement

$$\hat{r}_m(., \hat{B}, \hat{\Phi}) = \hat{g}_m(., \hat{B}) \circ \hat{\Phi}.$$

où $\hat{\Phi}$ est un estimateur de $\Phi$, défini sur $(0; 1)$ par,

$$\hat{\Phi} : x \mapsto \int_0^x \check{\bar{G}}_n^{(-)} \circ \hat{F}_{0,n_0}^{-1}(u) du, \tag{7.14}$$

où $\hat{F}_{0,n_0}^{-1}$ est l'inverse généralisé de la fonction de répartition empirique de l'échantillon $(X_{0,i_0})_{i_0}$, et où l'on rappelle que $\check{\bar{G}}_n^{(-)}$ est l'estimateur de Kaplan-Meier de la survie $\bar{G}$ calculé à partir de l'échantillon supplémentaire $(X_{-i}, C_{-i})_i$.

Le risque quadratique intégré d'un estimateur pour la fonction cible $r$ est calculé sur $(0; F_0(\tau_Z))$ et pondéré par la dérivée de la déformation $\Phi'$ : on note $\|.\|_{\Phi'}$ la fonction de perte correspondante. Nous obtenons le résultat suivant.

**Proposition 7.3.** *L'hypothèse (7.7) est supposée vérifiée. On suppose aussi que $g$ est de classe $\mathcal{C}^1$, et que $g(0) = g \circ B(\tau_Z)$. De plus, les trois intégrales suivantes sont supposées finies : $\int_0^{\tau_Z} (g \circ B)^2(x) \bar{G}(x) f_0(x) dx$, $\int_0^{\tau_Z} (g \circ B)^2(x) dx$ et $\int_0^{\tau_Z} \bar{G}^2(x) f_0^2(x) dx$. Enfin, la dimension du modèle $S_m$, $m \in \mathcal{M}_{n,n_0}$, sur lequel on se place est supposée majorée :*

$$D_m \leq \kappa \frac{n^{1/3}}{\ln^{1/3}(n)} \quad et \quad D_m \leq \kappa \frac{n_0^{1/3}}{\ln^{1/3}(n_0)},$$

*pour $\kappa > 0$. L'estimateur $\hat{r}_m(., \hat{B}, \hat{\Phi})$ satisfait l'inégalité suivante, pour deux constantes $C_1$ et $C_2$, indépendantes de $m, n$ et $n_0$,*

$$\mathbb{E}\left[ \left\| \hat{r}_m(., \hat{B}, \hat{\Phi}) - r \right\|_{\Phi'}^2 \right] \leq C_1 \frac{D_m}{n} + C_2 \frac{D_m}{n_0} + 9 \left\| \Pi_{S_m} g - g \right\|_B^2 .$$

### 7.3.4 Eléments pour l'estimation adaptative : perspectives

Nous disposons d'une collection d'estimateurs de la fonction $r$, construits par projection sur les sous-espaces $(S_m)_{m \in \mathcal{M}_{n,n_0}}$. La question qui se pose est donc, classiquement, la sélection d'un modèle dans cette collection. La procédure de sélection inspirée de Goldenshluger & Lepski (2011a) et proposée déjà dans d'autres problèmes de cette thèse (voir Section 7.2.3 ci-dessus, Section 6.3 au chapitre précédent, ou Section 3.2.3 du Chapitre 3 pour la régression par exemple) peut à nouveau être mise en place.

En posant

$$\begin{aligned} V^{(d)}(m) &= c \left( \frac{D_m}{n} + \frac{D_m}{n_0} \right), \\ A^{(d)}(m) &= \max_{m' \in \mathcal{M}_{n,n_0}} \left( \left\| \hat{r}_{m'}(., \hat{B}, \hat{\Phi}) - \hat{r}_{m \wedge m'}(., \hat{B}, \hat{\Phi}) \right\|_{\Phi'}^2 - V^{(d)}(m') \right)_+ , \end{aligned} \tag{7.15}$$

et

$$\hat{m}^{(d)} = \text{argmin}_{m\in\mathcal{M}_{n,n_0}}\{A^{(d)}(m) + V^{(d)}(m)\},$$

il est probable que l'estimateur $\hat{r}_{\hat{m}^{(d)}}(.,\hat{B},\hat{\Phi})$ vérifie une inégalité de type-oracle.

En effet, dans le cas "jouet" où les fonctions $\Phi$ et $B$ sont supposées connues, la démonstration d'une telle inégalité repose sur le contrôle du supremum d'un processus empirique, qui serait ici défini pour $t \in L^2((0; B(\tau_Z)))$, par

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} \psi_t(X_i, C_i) - \mathbb{E}\left[\psi_t(X_i, C_i)\right], \text{ avec } \psi_t(x,c) = \mathbf{1}_{x\leq c} t \circ B(x \wedge c).$$

Or, l'Inégalité de Talagrand (Proposition 2.2) s'applique avec le calcul des quantités suivantes :

$$M_1 = \sqrt{D_m}, \;\; H^2 = \frac{D_m}{n}, \;\; v = \|r\|_{L^\infty((0;F_0(\tau_Z)))}.$$

Dans ce cas simple, une borne non-asymptotique est donc facilement obtenue pour le risque. Comme le montre les exemples détaillés de l'estimation de la fonction de régression, ou de la densité conditionnelle (Chapitres 3 et 5), tout laisse donc à penser qu'un résultat similaire peut également être obtenu dans le cas général.

## 7.4 Preuves

### 7.4.1 Preuves relatives à la Section 7.2

#### Résultats techniques

Nous énonçons ici deux résultats utilisés dans les preuves ci-dessus. Le premier est un calcul simple d'espérance, classique dans le cas de données censurées à droite.

**Lemme 7.1.** *Soit une fonction $\psi : \mathbb{R} \to \mathbb{R}$, telle que la variable $\psi(Z)$ est intégrable. Alors,*

$$\mathbb{E}\left[\delta\psi(Z)\right] = \mathbb{E}\left[\delta\psi(X)\right] = \int_0^{\tau_Z} \bar{G}(x)\psi(x)f(x)dx.$$

*En particulier, $\mathbb{E}[\delta] = \int_0^{\tau_Z} \bar{G}(x)f(x)dx$.*

**Preuve du Lemme 7.1** On calcule

$$\begin{aligned}
\mathbb{E}\left[\delta\psi(Z)\right] &= \mathbb{E}\left[\mathbf{1}_{X\leq C}\psi(X)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{X\leq C}|X\right]\psi(X)\right], \\
&= \mathbb{E}\left[\int_{(0;\tau_C)} \mathbf{1}_{X\leq c} f_C(c)dx\psi(X)\right] = \mathbb{E}\left[\int_{(0;\tau_C)} \mathbf{1}_{X\leq c} f_C(c)dx\mathbf{1}_{X\leq\tau_C}\psi(X)\right], \\
&= \mathbb{E}\left[\bar{G}(X)\mathbf{1}_{X\leq\tau_C}\psi(X)\right] = \int_{(0;\tau_Z)} \bar{G}(x)\psi(x)f(x)dx.
\end{aligned}$$

□

Le second est plus technique, mais essentiel pour les majorations du risque de l'estimateur adaptatif.

**Lemme 7.2.** *On rappelle que $\tau < \tau_X \wedge \tau_C$. On a,*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i \mathbf{1}_{Z_i\leq\tau}\left|\frac{1}{\check{\bar{G}}_n(Z_i)}-\frac{1}{\bar{G}(Z_i)}\right|\right)^2\right]\leq\frac{\kappa}{n},$$

*pour une constante $\kappa$ dépendant de $\bar{F}(\tau)$, $\bar{G}(\tau)$, et de $\bar{c}_2$ et $\bar{c}_4$, les constantes de l'inégalité de la Proposition 2.6 (Chapitre 2).*

Lors de l'utilisation de ce lemme avec les échantillons des variables $Z_0$ et $\delta_0$, nous noterons de façon analogue $\kappa_0$ la constante correspondante.

**Preuve du Lemme 7.2.** Notons $T$ le terme à majorer. En réduisant au même dénominateur, et en minorant $\bar{G}(Z_i)$ par $\bar{G}(\tau)$ (ce qui est valable si $Z_i \leq \tau$), on obtient tout d'abord

$$\begin{aligned}
T &\leq \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\check{\bar{G}}_n(Z_i)}\right)^2\frac{1}{\bar{G}^2(\tau)}\sup_{x\in(0;\tau)}\left|\check{\bar{G}}_n(x)-\bar{G}(x)\right|^2,\\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\check{\bar{G}}_n(Z_i)}\right)^2\frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)}\sup_{x\in(0;\tau)}\left|\bar{F}(x)\left(\check{G}_n(x)-G(x)\right)\right|^2.
\end{aligned}$$

Pour appliquer l'inégalité de la Proposition 2.6, il faut minorer $\check{\bar{G}}_n(Z_i)$. Sans autre information, cette quantité est seulement minorée par $1/(n+1)$, ce qui ne suffit pas ici. On introduit donc l'évènement

$$\Omega_n=\left\{\sup_{x\in(0;\tau)}(\check{\bar{G}}_n(x)-\bar{G}(x))\geq-\frac{\bar{G}(\tau)}{2}\right\},$$

de telle sorte que $T \leq T_a + T_b$, avec

$$\begin{aligned}
T_a &= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\check{\bar{G}}_n(Z_i)}\right)^2\frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)}\sup_{x\in(0;\tau)}\left|\bar{F}(x)\left(\check{G}_n(x)-G(x)\right)\right|^2\mathbf{1}_{\Omega_n},\\
T_b &= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\check{\bar{G}}_n(Z_i)}\right)^2\frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)}\sup_{x\in(0;\tau)}\left|\bar{F}(x)\left(\check{G}_n(x)-G(x)\right)\right|^2\mathbf{1}_{\Omega_n^c}.
\end{aligned}$$

**Majoration de $T_a$.** Sur l'évènement $\Omega_n$,

$$\check{\bar{G}}_n(Z_i)\geq-\frac{\bar{G}(\tau)}{2}+\bar{G}(Z_i)\geq-\frac{\bar{G}(\tau)}{2}+\bar{G}(\tau)=\frac{\bar{G}(\tau)}{2},$$

ce qui entraîne

$$T_a\leq\frac{4}{\bar{F}^2(\tau)\bar{G}^4(\tau)}\sup_{x\in(0;\tau)}\left|\bar{F}(x)\left(\check{G}_n(x)-G(x)\right)\right|^2\mathbf{1}_{\Omega_n},$$

et donc, d'après la Proposition 2.6,

$$\mathbb{E}[T_a]\leq\frac{4}{\bar{F}^2(\tau)\bar{G}^4(\tau)}\mathbb{E}\left[\left\|\bar{F}\left(\check{G}_n-G\right)\right\|^2_{L^\infty((0;\tau))}\right]\leq\frac{4\bar{c}_2}{\bar{F}^2(\tau)\bar{G}^4(\tau)}\frac{1}{n}.$$

**Majoration de $T_b$.** Sur l'évènement $\Omega_n^c$, on a seulement $\check{\tilde{G}}_n(Z_i) \geq 1/(n+1)$, et donc $(\sum_{i=1}^n \check{\tilde{G}}_n^{-1}(Z_i)/n)^2 \leq (n+1)^2 \leq 4n^2$. Ainsi,

$$\begin{aligned}
\mathbb{E}[T_b] &\leq 4n^2 \frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)} \mathbb{E}\left[\left\|\bar{F}\left(\check{G}_n - G\right)\right\|^2_{L^\infty((0;\tau))} \mathbf{1}_{\Omega_n^c}\right], \\
&\leq 4n^2 \frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)} \left(\mathbb{E}\left[\left\|\bar{F}\left(\check{G}_n - G\right)\right\|^4_{L^\infty((0;\tau))}\right] \mathbb{P}\left(\Omega_n^c\right)\right)^{1/2}, \\
&= 4n^2 \frac{1}{\bar{F}^2(\tau)\bar{G}^2(\tau)} \frac{\sqrt{\bar{c}_4}}{n} \mathbb{P}\left(\Omega_n^c\right)^{1/2}, \\
&= n \frac{4\sqrt{\bar{c}_4}}{\bar{F}^2(\tau)\bar{G}^2(\tau)} \mathbb{P}\left(\Omega_n^c\right)^{1/2},
\end{aligned}$$

par l'Inégalité de Cauchy-Schwarz, puis la Proposition 2.6. Enfin,

$$\begin{aligned}
\mathbb{P}\left(\Omega_n^c\right) &= \mathbb{P}\left(\sup_{x\in(0;\tau)} \left(\check{G}_n(x) - G(x)\right) < -\frac{\bar{G}(\tau)}{2}\right), \\
&\leq \mathbb{P}\left(\sup_{x\in(0;\tau)} \left\{\bar{F}(x)\left(\check{G}_n(x) - G(x)\right)\right\} < -\frac{\bar{G}(\tau)}{2}\right), \\
&\leq \mathbb{P}\left(\sup_{x\in(0;\tau)} \left|\bar{F}(x)\left(\check{G}_n(x) - G(x)\right)\right| > \frac{\bar{G}(\tau)}{2}\right), \\
&\leq 2.5 \exp\left(-2n\frac{\bar{G}^2(\tau)}{4} + C\frac{\bar{G}(\tau)}{2}\sqrt{n}\right),
\end{aligned}$$

par la Proposition 2.5 (Chapitre 2). Donc,

$$\begin{aligned}
\mathbb{E}[T_b] &\leq \frac{4\sqrt{2.5}\sqrt{\bar{c}_4}}{\bar{F}^2(\tau)\bar{G}^2(\tau)} n \exp\left(-n\frac{\bar{G}^2(\tau)}{4} + C\frac{\bar{G}(\tau)}{4}\sqrt{n}\right), \\
&\leq \frac{C}{n},
\end{aligned}$$

pour une certaine constante $C$, puisque la suite $n \mapsto n^2 \exp(-c_1 n + c_2\sqrt{n})$ tend vers 0 quand $n$ tend vers l'infini, quels que soient $c_1$ et $c_2$ strictement positifs.
On obtient donc le résultat en sommant cette majoration et celle de l'espérance de $T_a$.

□

**Preuve de la Proposition 7.1**

**Partie principale de la preuve** La clé de la preuve est la décomposition suivante de la perte de l'estimateur en quatre termes :

$$\left\|\hat{r}_m(., \check{F}_{0,n_0}, \check{\tilde{G}}_n) - r\right\|^2 \leq 4 \sum_{l=0}^{3} T_l^{c,m},$$

avec

$$
\begin{aligned}
T_0^{c,m} &= \left\|\hat{r}_m(.,F_0,\bar{G}) - r\right\|^2, \\
T_1^{c,m} &= \left\|\hat{r}_m(.,\check{F}_{0,n_0},\bar{G}) - \hat{r}_m(.,F_0,\bar{G}) - \mathbb{E}\left[\hat{r}_m(.,\check{F}_{0,n_0},\bar{G}) - \hat{r}_m(.,F_0,\bar{G})\,|(X_0,C_0)\right]\right\|^2, \\
T_2^{c,m} &= \left\|\mathbb{E}\left[\hat{r}_m(.,\check{F}_{0,n_0},\bar{G}) - \hat{r}_m(.,F_0,\bar{G})\,|(X_0,C_0)\right]\right\|^2, \\
T_3^{c,m} &= \left\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n) - \hat{r}_m(.,\check{F}_{0,n_0},\bar{G})\right\|^2,
\end{aligned}
$$

où $\mathbb{E}[.|(X_0,C_0)]$ désigne l'espérance conditionnellement aux échantillons $(X_{0,i_0})_{i_0=1,\ldots,n_0}$ et $(C_{0,i_0})_{i_0=1,\ldots,n_0}$. Par ailleurs, $T_0^{c,m}$ est le risque de l'estimateur quand on suppose $F_0$ et $\bar{G}$ connues les termes $T_1^{c,m}$ et $T_2^{c,m}$ sont relatifs au cas où $\bar{G}$ est connu et pas $F_0$, et le dernier terme est celui qui contient donc "le plus" d'aléas. L'enjeu est donc de majorer chacun de ces termes. Les résultats sont résumés dans les lemmes suivants, dont les démonstrations, qui font l'objet des paragraphes ci-dessous, sont fondées sur les propriétés de l'estimateur de Kaplan-Meier, rappelées à la Section 2.1.3 du Chapitre 2, en particulier sur les inégalités inspirées de Bitouzé *et al.* (1999).

**Lemme 7.3.** *(Majoration du risque de l'estimateur lorsque $F_0$ et $\bar{G}$ sont connus) L'espérance du terme $T_0^{c,m}$ admet le majorant suivant :*

$$
\mathbb{E}\left[T_0^{c,m}\right] \leq \mathbb{E}\left[\frac{\delta \mathbf{1}_{Z\leq\tau}}{\bar{G}^2(Z)}\right]\frac{D_m}{n} + \left\|r_m - r\right\|^2.
$$

L'espérance $\mathbb{E}[\delta \mathbf{1}_{Z\leq\tau}/\bar{G}^2(Z)]$ peut encore éventuellement être majorée par $1/\bar{G}^2(\tau)$.

**Lemme 7.4.** *L'espérance du terme $T_1^{c,m}$ admet le majorant suivant :*

$$
\mathbb{E}\left[T_1^{c,m}\right] \leq \left\|\varphi_2'\right\|^2_{L^\infty((0;F_0(\tau)))}\mathbb{E}\left[\frac{\delta \mathbf{1}_{Z\leq\tau}}{\bar{G}_0^2(Z)\bar{G}^2(Z)}\right]\frac{D_m^3}{nn_0}.
$$

L'espérance $\mathbb{E}[\delta \mathbf{1}_{Z\leq\tau}/(\bar{G}_0^2(Z)\bar{G}^2(Z))]$ peut éventuellement être majorée par $1/(\bar{G}_0^2(\tau)\bar{G}^2(\tau))$.

**Lemme 7.5.** *L'espérance du terme $T_2^{c,m}$ admet le majorant suivant :*

$$
\begin{aligned}
\mathbb{E}\left[T_2^{c,m}\right] \quad \leq \quad & \left(18\|r\|^2_{L^\infty((0;F_0(\tau)))}\kappa_0\frac{D_m}{n}\right. \\
& \left.+9\left(2\int_0^\tau (r\circ F_0(x))^2 f_0(x)dx + \frac{\bar{c}_2}{\bar{G}_0^2(\tau)}\|r\|^2_{L^\infty((0;F_0(\tau)))}\right)\frac{D_m}{n_0}\right) \\
& +3\frac{\pi^4}{4}\bar{c}_4\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)}f_0(x)dx\frac{D_m^4}{n_0^2} \\
& +3\frac{\bar{c}_6}{6}\left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;F_0(\tau)))}\int_0^\tau \frac{(r\circ F_0)^2(x)}{\bar{G}_0^6(x)}f_0(x)dx\frac{D_m^7}{n_0^3} \\
& +9\bar{c}_2\int_{(0;\tau)}\frac{\left(r'\circ F_0(x)\right)^2}{\bar{G}_0(\tau)}f_0(x)dx\frac{1}{n_0},
\end{aligned}
$$

*où la constante $\kappa_0$ est celle du Lemme 7.2, les constantes $\bar{c}_l$ ($l = 2,4,6$) sont celles de la Proposition 2.6.*

Les intégrales apparaissant dans l'énoncé du lemme sont finies, puisque $r$ et $r'$ sont bornées, et $\bar{G}_0$ est minorée sur l'intervalle d'intégration par $\bar{G}_0(\tau)$.

**Lemme 7.6.** *L'espérance du terme $T_3^{c,m}$ admet le majorant suivant :*

$$\mathbb{E}\left[T_3^{c,m}\right] \leq \kappa \frac{D_m}{n},$$

*où la constante $\kappa$ est celle du Lemme 7.2.*

En rassemblant les résultats de ces lemmes, et en tenant compte de la contrainte $D_m \leq \kappa n_0^{1/3}$, la Proposition 7.1 est prouvée.

□

**Preuve du Lemme 7.3** On a vu que $\hat{r}_m(.,F_0,\bar{G})$ estimait sans biais la projection $r_m$ de $r$ sur $S_m$. On a donc

$$\mathbb{E}\left[\left\|\hat{r}_m(.,F_0,\bar{G}) - r_m\right\|^2\right] = \mathbb{E}\left[\left\|\hat{r}_m(.,F_0,\bar{G}) - r_m\right\|^2\right] + \left\|r_m - r\right\|^2.$$

On majore maintenant le terme de variance :

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{r}_m(.,F_0,\bar{G}) - r_m\right\|^2\right] &= \frac{1}{n}\sum_{j=1}^{D_m} \mathrm{Var}\left(\frac{\delta}{\bar{G}(Z)}\varphi_j\left(F_0(Z)\right)\mathbf{1}_{Z\leq\tau}\right), \\
&\leq \frac{1}{n}\sum_{j=1}^{D_m} \mathbb{E}\left[\frac{\delta}{\bar{G}(Z)^2}\varphi_j^2\left(F_0(Z)\right)\mathbf{1}_{Z\leq\tau}\right], \\
&= \frac{1}{n}\sum_{j=1}^{D_m} \int_{(0;\tau)} \varphi_j^2(F_0(x)) f(x) \frac{1}{\bar{G}(x)} dx,
\end{aligned}$$

par le Lemme 7.1. Ainsi, par connexion de normes,

$$\mathbb{E}\left[\left\|\hat{r}_m(.,F_0,\bar{G}) - r_m\right\|^2\right] \leq \frac{D_m}{n}\int_{(0;\tau)} f(x)\frac{dx}{\bar{G}(x)}$$

Cette dernière intégrale se réécrit comme l'espérance de $\delta\mathbf{1}_{Z\leq\tau}/\bar{G}^2(Z)$. En effet, par le Lemme 7.1,

$$\mathbb{E}\left[\frac{\delta}{\bar{G}^2(Z)}\mathbf{1}_{Z\leq\tau}\right] = \int_{(0;\tau_Z)} \mathbf{1}_{x\leq\tau}\frac{1}{\bar{G}(x)} f(x) dx.$$

□

**Preuve du Lemme 7.4** L'espérance conditionnelle de ce terme s'écrit

$$\mathbb{E}\left[T_1^{c,m}\,|(X_0,C_0)\right] = \sum_{j=1}^{D_m} \mathrm{Var}\left(\hat{a}_j^{\check{F}_{0,n_0},\bar{G}} - \hat{a}_j^{F_0,\bar{G}}\,|(X_0,C_0)\right).$$

Or, quel que soit $j$,

$$\begin{aligned}
\mathrm{Var}\left(\hat{a}_j^{\check{F}_{0,n_0},\bar{G}} - \hat{a}_j^{F_0,\bar{G}} \,|(X_0,C_0)\right) &= \frac{1}{n}\mathrm{Var}\left(\frac{\delta}{\bar{G}(X)}\mathbf{1}_{Z\leq\tau}\left(\varphi_j\circ\check{F}_{0,n_0} - \varphi_j\circ F_0\right)(X_1)\,|(X_0,C_0)\right),\\
&\leq \frac{1}{n}\mathbb{E}\left[\frac{\delta}{\bar{G}^2(X)}\mathbf{1}_{Z\leq\tau}\left(\varphi_j\circ\check{F}_{0,n_0} - \varphi_j\circ F_0\right)^2(X)\,|(X_0,C_0)\right],\\
&= \frac{1}{n}\int_{(0;\tau)}\frac{1}{\bar{G}(x)}\left(\varphi_j\circ\check{F}_{0,n_0} - \varphi_j\circ F_0\right)^2(x)f(x)dx,\\
&\leq \frac{1}{n}\left\|\varphi_j'\right\|^2_{L^\infty((0;F_0(\tau)))}\int_{(0;\tau)}\left(\check{F}_{0,n_0} - F_0\right)^2(x)\frac{1}{\bar{G}(x)}f(x)dx,
\end{aligned}$$

en utilisant le Lemme 7.1 et l'inégalité des accroissements finis. Ainsi, comme $\|\varphi_j'\|^2_{L^\infty((0;F_0(\tau)))} \leq D_m^2\|\varphi_2'\|^2_{L^\infty((0;F_0(\tau)))}$, on a

$$\mathbb{E}\left[T_1^{c,m}\,|(X_0,C_0)\right] \leq \frac{D_m^3}{n}\left\|\varphi_2'\right\|^2_{L^\infty((0;F_0(\tau)))}\int_{(0;\tau)}\left(\check{F}_{0,n_0} - F_0\right)^2(x)\frac{1}{\bar{G}(x)}f(x)dx,$$

ce qui aboutit encore à

$$\begin{aligned}
\mathbb{E}\left[T_1^{c,m}\right] &\leq \frac{D_m^3}{n}\left\|\varphi_2'\right\|^2_{L^\infty((0;F_0(\tau)))}\int_{(0;\tau)}\mathbb{E}\left[\left(\check{F}_{0,n_0} - F_0\right)^2(x)\right]\frac{1}{\bar{G}(x)}f(x)dx,\\
&\leq \frac{D_m^3}{n}\left\|\varphi_2'\right\|^2_{L^\infty((0;F_0(\tau)))}\int_{(0;\tau)}\mathbb{E}\left[\sup_{(0;\tau)}\left|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right|^2\right]\frac{1}{\bar{G}_0^2(x)\bar{G}(x)}f(x)dx.
\end{aligned}$$

Il reste à appliquer la Proposition 2.6 pour conclure.

□

**Preuve du Lemme 7.5** De même,

$$\begin{aligned}
T_2^{c,m} &= \sum_{j=1}^{D_m}\left(\mathbb{E}\left[\frac{\mathbf{1}_{X\leq C}\mathbf{1}_{X\leq\tau}}{\bar{G}(X)}\left(\varphi_j\circ\check{F}_{0,n_0} - \varphi_j\circ F_0\right)(X)\,|(X_0,C_0)\right]\right)^2,\\
&= \sum_{j=1}^{D_m}\left(\int_{(0;\tau)}\left(\varphi_j\circ\check{F}_{0,n_0} - \varphi_j\circ F_0\right)(x)f(x)dx\right)^2
\end{aligned}$$

On fait un développement de Taylor-Lagrange à l'ordre 3, et on a $T_2^{c,m} \leq 3(T_{2,1}^{c,m}+T_{2,2}^{c,m}+T_{2,3}^{c,m})$, avec pour $\hat{\alpha}_{j,n_0,x}$ un réel aléatoire,

$$\begin{aligned}
T_{2,1}^{c,m} &= \sum_{j=1}^{D_m}\left(\int_{(0;\tau)}\varphi_j'(F_0(x))\left(\check{F}_{0,n_0}(x) - F_0(x)\right)f(x)dx\right)^2,\\
T_{2,2}^{c,m} &= \sum_{j=1}^{D_m}\left(\int_{(0;\tau)}\varphi_j''(F_0(x))\frac{\left(\check{F}_{0,n_0}(x) - F_0(x)\right)^2}{2}f(x)dx\right)^2,\\
T_{2,3}^{c,m} &= \sum_{j=1}^{D_m}\left(\int_{(0;\tau)}\varphi_j^{(3)}(\hat{\alpha}_{j,n_0,x})\frac{\left(\check{F}_{0,n_0}(x) - F_0(x)\right)^3}{6}f(x)dx\right)^2.
\end{aligned}$$

On majore chacun des termes.

• **Majoration de $T_{2,1}^{c,m}$.** Le premier terme est le plus complexe à majorer. Après changement de variables $u = F_0(x)$, il est égal à

$$T_{2,1}^{c,m} = \sum_{j=1}^{D_m} \left( \int_0^{F_0(\tau)} \varphi_j'(u) \left( \check{F}_{0,n_0}\left(F_0^{-1}(u)\right) - u \right) r(u) du \right)^2 .$$

Rappelons l'expression (2.10) de l'estimateur de Kaplan-Meier modifié, prouvée à la Section 2.1.3 du Chapitre 2 :

$$\check{F}_{0,n_0}(y) = \frac{1}{n_0+1} \sum_{i_0=1}^{n_0} \frac{\delta_{0,i_0}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})} \mathbf{1}_{Z_{0,i_0} \leq y} = \frac{1}{n_0+1} \sum_{i_0=1}^{n_0} \frac{\delta_{0,i_0}}{\check{\bar{G}}_{0,n_0}(X_{0,i_0})} \mathbf{1}_{X_{0,i_0} \leq y}.$$

Comme on évalue l'estimateur en $F_0^{-1}(u)$, l'indicatrice qui apparait est

$$\mathbf{1}_{X_{0,i_0} \leq F_0^{-1}(u)} = \mathbf{1}_{F_0(X_{0,i_0}) \leq u} := \mathbf{1}_{U_{0,i_0} \leq u}.$$

On a donc,

$$T_{2,1}^{c,m} = \sum_{j=1}^{D_m} \left( \left\{ \frac{1}{n_0+1} \sum_{i_0=1}^{n_0} \frac{\delta_{0,i_0} \mathbf{1}_{Z_{0,i_0} \leq \tau}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})} \int_{U_{0,i_0}}^{F_0(\tau)} \varphi_j'(u) r(u) du \right\} - \int_0^{F_0(\tau)} \varphi_j'(u) u r(u) du \right)^2 .$$

L'indicatrice de l'évènement $\{Z_{0,i_0} \leq \tau\}$ a été conservée, car l'intégrale de $U_{0,i_0} = F_0(X_{0,i_0})$ à $F_0(\tau)$ est nulle si $\tau < X_{0,i_0}$, c'est-à-dire si $\tau < Z_{0,i_0}$ (grâce à la présence du $\delta_{0,i_0}$). Par intégration par parties,

$$\int_{U_{0,i_0}}^{F_0(\tau)} \varphi_j'(u) r(u) du = \varphi_j(F_0(\tau)) r(F_0(\tau)) - \varphi_j(U_{0,i_0}) r(U_{0,i_0}) - \int_0^{F_0(\tau)} \varphi_j(u) r'(u) \mathbf{1}_{U_{0,i_0} \leq u} du.$$

De même,

$$\int_0^{F_0(\tau)} \varphi_j'(u) u r(u) du = \varphi_j(F_0(\tau)) r(F_0(\tau)) F_0(\tau) - \int_0^{F_0(\tau)} \varphi_j(u) \left( r(u) + u r'(u) \right) du.$$

Donc,

$$\begin{aligned} T_{2,1}^{c,m} = \sum_{j=1}^{D_m} \Bigg\{ & \left( -\frac{1}{n_0+1} \sum_{i_0=1}^{n_0} \frac{\delta_{0,i_0} \mathbf{1}_{Z_{0,i_0} \leq \tau}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})} \varphi_j(F_0(Z_{0,i_0})) r(F_0(Z_{0,i_0})) - \int_0^{F_0(\tau)} \varphi_j(u) r(u) du \right) \\ & + \left( - \int_0^{F_0(\tau)} \left( \check{F}_{0,n_0}\left(F_0^{-1}(u)\right) - u \right) \varphi_j(u) r'(u) du \right) \\ & + \left( \varphi_j(F_0(\tau)) r(F_0(\tau)) \left\{ \frac{1}{n_0+1} \sum_{i_0=1}^{n_0} \frac{\delta_{0,i_0} \mathbf{1}_{Z_{0,i_0} \leq \tau}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})} - F_0(\tau) \right\} \right) \Bigg\}^2 \end{aligned}$$

Les deux lignes ci-dessus montrent qu'on peut décomposer $T_{2,1}^{c,m} \leq 3T_{2,1,1}^{c,m} + 3T_{2,1,2}^{c,m} + 3T_{2,1,3}^{c,m}$. Majorons chaque terme.

– *Majoration de* $T_{2,1,1}^{c,m}$. Le premier terme est :

$$T_{2,1,1}^{c,m} = \sum_{j=1}^{D_m}\left(-\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\frac{\delta_{0,i_0}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))+\int_0^{F_0(\tau)}\varphi_j(u)r(u)du\right)^2.$$

On remarque alors, en utilisant le changement de variables $u=F_0(x)$ et le Lemme 7.1,

$$\int_0^{F_0(\tau)}\varphi_j(u)r(u)du=\int_0^{\tau}\varphi_j\circ F_0(x)r\circ F_0(x)f_0(x)dx=\mathbb{E}\left[\frac{\delta_0\mathbf{1}_{Z_0\le\tau}}{\bar{G}_0(Z_0)}\varphi_j(F_0(Z_0))r(F_0(Z_0))\right].$$

Ainsi,

$$\begin{aligned}T_{2,1,1}^{c,m} &= \sum_{j=1}^{D_m}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\left\{\frac{\delta_{0,i_0}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\mathbf{1}_{Z_{0,i_0}\le\tau}\right.\right.\\ &\qquad\left.\left.-\mathbb{E}\left[\frac{\delta_{0,i_0}\mathbf{1}_{Z_{0,i_0}\le\tau}}{\bar{G}_0(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\right]\right\}\right)^2.\end{aligned}$$

On décompose à nouveau en deux termes : $T_{2,1,1}^{c,m}\le 2(T_{2,1,1,1}^{c,m}+T_{2,1,1,2}^{c,m})$, avec

$$\begin{aligned}T_{2,1,1,1}^{c,m} &= \sum_{j=1}^{D_m}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\delta_{0,i_0}\left\{\frac{1}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}-\frac{1}{\bar{G}_0(Z_{0,i_0})}\right\}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\mathbf{1}_{Z_{0,i_0}\le\tau}\right)^2,\\ T_{2,1,1,2}^{c,m} &= \sum_{j=1}^{D_m}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\left\{\frac{\delta_{0,i_0}}{\bar{G}_0(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\mathbf{1}_{Z_{0,i_0}\le\tau}\right.\right.\\ &\qquad\left.\left.-\mathbb{E}\left[\frac{\delta_{0,i_0}\mathbf{1}_{Z_{0,i_0}\le\tau}}{\bar{G}_0(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\right]\right\}\right)^2.\end{aligned}$$

Pour le premier terme, la connexion de normes entraîne

$$T_{2,1,1,1}^{c,m}\le D_m\|r\|^2_{L^\infty((0;F_0(\tau)))}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\delta_{0,i_0}\left|\frac{1}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}-\frac{1}{\bar{G}_0(Z_{0,i_0})}\right|\mathbf{1}_{Z_{0,i_0}\le\tau}\right)^2.$$

Il reste alors à appliquer le Lemme 7.2, pour obtenir $\mathbb{E}[T_{2,1,1,1}^{c,m}]\le\|r\|^2_{L^\infty((0;F_0(\tau)))}\kappa_0 D_m/n$.

Pour le second, on constate que

$$\begin{aligned}\mathbb{E}\left[T_{2,1,1,2}^{c,m}\right] &= \sum_{j=1}^{D_m}\frac{1}{(n_0+1)^2}\mathrm{Var}\left(\sum_{i_0=1}^{n_0}\frac{\delta_{0,i_0}}{\bar{G}_0(Z_{0,i_0})}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\mathbf{1}_{Z_{0,i_0}\le\tau}\right),\\ &\le \sum_{j=1}^{D_m}\frac{n_0}{(n_0+1)^2}\mathbb{E}\left[\frac{\delta_0}{\bar{G}_0^2(Z_0)}\varphi_j^2(F_0(Z_0))r^2(F_0(Z_0))\mathbf{1}_{Z_0\le\tau}\right],\\ &\le \frac{1}{n_0}D_m\int_0^{\tau}(r\circ F_0(x))^2f_0(x)dx,\end{aligned}$$

en utilisant à nouveau la connexion de normes.

Par conséquent, on a montré que

$$\mathbb{E}\left[T_{2,1,1}^{c,m}\right] \leq 2\left(\|r\|^2_{L^\infty((0;F_0(\tau)))}\kappa_0\frac{D_m}{n} + \int_0^\tau (r\circ F_0(x))^2 f_0(x)dx\frac{D_m}{n_0}\right).$$

– *Majoration de* $T_{2,1,2}^{c,m}$. Ce second terme est

$$T_{2,1,2}^{c,m} = \sum_{j=1}^{D_m}\left(\int_0^{F_0(\tau)}\left(\check{F}_{0,n_0}\left(F_0^{-1}(u)\right)-u\right)\varphi_j(u)r'(u)du\right)^2.$$

On le majore de la façon suivante :

$$\begin{aligned}
T_{2,1,2}^{c,m} &= \sum_{j=1}^{D_m}\left(\langle\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)r',\varphi_j\rangle\right)^2,\\
&= \left\|\Pi_{S_m}\left\{\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)r'\right\}\right\|^2,\\
&\leq \left\|\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)r'\right\|^2,\\
&= \int_{(0;F_0(\tau))}\left(r'(u)\right)^2\left(\check{F}_{0,n_0}\circ F_0^{-1}(u)-u\right)^2du,\\
&= \int_{(0;\tau)}\left(r'\circ F_0(x)\right)^2\left(\check{F}_{0,n_0}(x)-F_0(x)\right)^2f_0(x)dx,
\end{aligned}$$

en posant à nouveau $u=F_0(x)$. Pour la suite, on introduit $\bar{G}_0$ :

$$T_{2,1,2}^{c,m} \leq \int_{(0;\tau)}\frac{\left(r'\circ F_0(x)\right)^2}{\bar{G}_0(x)}f_0(x)dx\left\|\bar{G}_0\left(\check{F}_{0,n_0}-F_0\right)\right\|^2_{L^\infty((0;\tau))},$$

et on applique directement la Proposition 2.6 :

$$\mathbb{E}\left[T_{2,1,2}^{c,m}\right] \leq \bar{c}_2\frac{1}{n_0}\int_{(0;\tau)}\frac{\left(r'\circ F_0(x)\right)^2}{\bar{G}_0(\tau)}f_0(x)dx.$$

– *Majoration de* $T_{2,1,3}^{c,m}$. On majore d'abord le troisième terme en utilisant la connexion de normes :

$$\begin{aligned}
T_{2,1,3}^{c,m} &= \sum_{j=1}^{D_m}\varphi_j^2(F_0(\tau))r^2(F_0(\tau))\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\frac{\delta_{0,i_0}\mathbf{1}_{Z_{0,i_0}\leq\tau}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}-F_0(\tau)\right)^2,\\
&\leq D_m\|r\|^2_{L^\infty((0;F_0(\tau)))}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\frac{\delta_{0,i_0}\mathbf{1}_{Z_{0,i_0}\leq\tau}}{\check{\bar{G}}_{0,n_0}(Z_{0,i_0})}-F_0(\tau)\right)^2,\\
&= D_m\|r\|^2_{L^\infty((0;F_0(\tau)))}\left(\check{F}_{0,n_0}(\tau)-F_0(\tau)\right)^2,\\
&\leq \frac{1}{\bar{G}_0^2(\tau)}D_m\|r\|^2_{L^\infty((0;F_0(\tau)))}\left\|\bar{G}_0\left(\check{F}_{0,n_0}-F_0\right)\right\|^2_{L^\infty(0;\tau)},
\end{aligned}$$

en utilisant la définition sous forme de somme de l'estimateur de Kaplan-Meier (voir (2.10)). En prenant l'espérance, et en appliquant la Proposition 2.6, on obtient

$$\mathbb{E}\left[T_{2,1,3}^{c,m}\right] \leq \frac{1}{\bar{G}_0^2(\tau)}\|r\|^2_{L^\infty((0;F_0(\tau)))}\bar{c}_2\frac{D_m}{n_0}.$$

Finalement, on a obtenu,

$$\begin{aligned}\mathbb{E}\left[T_{2,1}^{c,m}\right] &\leq 3\left\{2\left(\|r\|^2_{L^\infty((0;F_0(\tau)))}\kappa_0\frac{D_m}{n}+\int_0^\tau (r\circ F_0(x))^2 f_0(x)dx\frac{D_m}{n_0}\right)\right.\\ &\left.+\bar{c}_2\frac{1}{\bar{G}_0^2(\tau)}\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{D_m}{n_0}+\bar{c}_2\int_{(0;\tau)}\frac{(r'\circ F_0(x))^2}{\bar{G}_0(\tau)}f_0(x)dx\frac{1}{n_0}\right\}. \quad (7.16)\end{aligned}$$

• **Majoration de $T_{2,2}^{c,m}$**. On utilise les propriétés de la base trigonométrique pour obtenir

$$\mathbb{E}\left[T_{2,2}^{c,m}\right] = (\pi^4/4)\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_0^{F_0(\tau)} r(u)\left(\check{F}_{0,n_0}\circ F_0^{-1}(u)-u\right)^2\mu_j^2\varphi_j(u)du\right\}^2\right],$$

avec $\mu_j = j$ ou $j-1$ selon que $j$ est pair ou non. Ainsi,

$$\begin{aligned}\mathbb{E}\left[T_{2,2}^{c,m}\right] &\leq (\pi^4/4)D_m^4\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_0^{F_0(\tau)} r(u)\left(\check{F}_{0,n_0}\circ F_0^{-1}(u)-u\right)^2\varphi_j(u)du\right\}^2\right],\\ &= (\pi^4/4)D_m^4\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\langle r\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)^2,\varphi_j\rangle\right\}^2\right],\\ &= (\pi^4/4)D_m^4\mathbb{E}\left[\left\|\Pi_{S_m}\left\{r\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)^2\right\}\right\|^2\right],\\ &\leq (\pi^4/4)D_m^4\mathbb{E}\left[\left\|r\left(\check{F}_{0,n_0}\circ F_0^{-1}-id\right)^2\right\|\right],\\ &= (\pi^4/4)D_m^4\mathbb{E}\left[\int_0^{F_0(\tau)} r^2(u)\left(\check{F}_{0,n_0}\circ F_0^{-1}(u)-u\right)^4du\right],\end{aligned}$$

puisque la norme d'une projection est majorée par la norme de la fonction projetée. Puis, par définition de la norme, et changement de variables,

$$\begin{aligned}\mathbb{E}\left[T_{2,2}^{c,m}\right] &= (\pi^4/4)D_m^4\mathbb{E}\left[\int_0^\tau (r\circ F_0(x))^2\left(\check{F}_{0,n_0}(x)-F_0(x)\right)^4 f_0(x)dx\right],\\ &\leq (\pi^4/4)D_m^4\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)}f_0(x)dx\mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0}-F_0\right)\right\|^4_{L^\infty((0;\tau))}\right].\end{aligned}$$

Et enfin, par la Proposition 2.6.

$$\mathbb{E}\left[T_{2,2}^{c,m}\right] \leq (\pi^4/4)\bar{c}_4\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)}f_0(x)dx\frac{D_m^4}{n_0^2}. \quad (7.17)$$

• **Majoration de $T_{2,3}^{c,m}$.** Le terme $T_{2,3}^{c,m}$ peut-être appréhendé de manière similaire :

$$\begin{aligned}
\mathbb{E}\left[T_{2,3}^{c,m}\right] &= (1/6)\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_0^{F_0(\tau)} r(u)\left(\check{F}_{0,n_0}\circ {F_0}^{-1}(u)-u\right)^3\varphi_j^{(3)}(\hat{\alpha}_{j,n_0,F_0^{-1}(u)})du\right\}^2\right],\\
&\leq (1/6)\sum_{j=1}^{D_m}\left\|\varphi_j^{(3)}\right\|^2_{L^\infty((0;F_0(\tau)))}\mathbb{E}\left[\left\{\int_0^{F_0(\tau)}\left|r(u)\left(\check{F}_{0,n_0}\circ {F_0}^{-1}(u)-u\right)^3\right|du\right\}^2\right],\\
&\leq (1/6)\left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;F_0(\tau)))}D_m^7\mathbb{E}\left[\int_0^{F_0(\tau)} r^2(u)\left(\check{F}_{0,n_0}\circ {F_0}^{-1}(u)-u\right)^6du\right],\\
&\leq (1/6)\left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;F_0(\tau)))}D_m^7\mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0}-F_0\right)\right\|^6_{L^\infty((0;\tau))}\right]\int_0^\tau\frac{(r\circ F_0)^2(x)}{\bar{G}_0^6(x)}f_0(x)dx,
\end{aligned}$$

en appliquant l'Inégalité de Cauchy-Schwarz... Puis, par la Proposition 2.6,

$$\mathbb{E}\left[T_{2,3}^{c,m}\right] \leq (\bar{c}_6/6)\left\|\varphi_2^{(3)}\right\|^2_{L^\infty((0;F_0(\tau)))}\frac{D_m^7}{n_0^3}\int_0^\tau\frac{(r\circ F_0)^2(x)}{\bar{G}_0^6(x)}f_0(x)dx. \tag{7.18}$$

Le résultat du Lemme 7.5 est obtenu en regroupant les bornes des trois termes, à savoir (7.16), (7.17), et (7.18).

□

**Preuve du Lemme 7.6** On calcule tout d'abord,

$$\begin{aligned}
T_3^{c,m} &= \left\|\hat{r}_m(,\check{F}_{0,n_0},\check{\bar{G}}_n)-\hat{r}_m(,\check{F}_{0,n_0},\bar{G})\right\|^2\\
&= \sum_{j=1}^{D_m}\left(\hat{a}_j^{\check{F}_{0,n_0},\check{\bar{G}}_n}-\hat{a}_j^{\check{F}_{0,n_0},\bar{G}}\right)^2,\\
&= \sum_{j=1}^{D_m}\left(\frac{1}{n}\sum_{i=1}^n\left\{\frac{\delta_i}{\check{\bar{G}}_n(Z_i)}-\frac{\delta_i}{\bar{G}_(Z_i)}\right\}\varphi_j\left(\check{F}_{0,n_0}(Z_i)\right)\mathbf{1}_{Z_i\leq\tau}\right)^2,\\
&\leq D_m\left(\frac{1}{n}\sum_{i=1}^n\delta_i\mathbf{1}_{Z_i\leq\tau}\left|\frac{1}{\check{\bar{G}}_n(Z_i)}-\frac{1}{\bar{G}(Z_i)}\right|\right)^2,
\end{aligned}$$

puis on applique le Lemme 7.2, qui permet de conclure.

□

**Preuve du Théorème 7.1**

La lettre $C$ désigne une constante indépendante de $m$, $n$ et $n_0$, qui pourra changer de ligne en ligne. On décompose la perte de l'estimateur, de manière à utiliser ensuite successivement les définitions de $\Phi(m)$, $\Phi(\hat{m})$, puis $\hat{m}$ :

$$\begin{aligned}
\left\|\hat{r}_{\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2 &\leq 3\left\|\hat{r}_{\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-\hat{r}_{m\wedge\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)\right\|^2 \\
&\quad +3\left\|\hat{r}_{m\wedge\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)\right\|^2 \\
&\quad +3\left\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2, \\
&\leq 3\left(A(m)+V(\hat{m})\right)+3\left(A(\hat{m})+V(m)\right)+3\left\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2, \\
&\leq 6\left(A(m)+V(m)\right)+3\left\|\hat{r}_m(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2.
\end{aligned}$$

Par la Proposition 7.1, on obtient donc

$$\mathbb{E}\left[\left\|\hat{r}_{\hat{m}}(.,\check{F}_{0,n_0},\check{\bar{G}}_n)-r\right\|^2\right] \leq \mathbb{E}\left[A(m)\right]+C\left\{\frac{D_m}{n}+\frac{D_m}{n_0}\right\}+12\left\|r_m-r\right\|^2.$$

Il reste à appliquer le lemme ci-dessous, dont la démonstration fait l'objet du paragraphe suivant.

**Lemme 7.7.** *Sous les hypothèses du Théorème 7.1, il existe une constante $C$, dépendant de la fonction $r$, mais ni de $n$, ni de $n_0$, telle que*

$$\mathbb{E}\left[A(m)\right]\leq C\frac{\ln(n)}{n}+C\frac{\ln(n_0)}{n_0}+12\left\|r_m-r\right\|^2.$$

□

**Preuve du Lemme 7.7**

Soit $m'\in\mathcal{M}_{n,n_0}$. On décompose,

$$\begin{aligned}
\left\|\hat{r}_{m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)-\hat{r}_{m\wedge m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)\right\|^2 &\leq 3\left\|\hat{r}_{m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)-r_{m'}\right\|^2+3\left\|r_{m'}-r_{m\wedge m'}\right\|^2 \\
&\quad +3\left\|r_{m\wedge m'}-\hat{r}_{m\wedge m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)\right\|^2.
\end{aligned}$$

On note ensuite que, pour $p=m'$ ou $p=m\wedge m'$

$$\left\|r_p-\hat{r}_p\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)\right\|^2=\sum_{j=1}^{D_p}\left(\bar{\nu}_n(\varphi_j)\right)^2$$

et où l'on a défini

$$\bar{\nu}_n(\varphi_j)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i}{\check{\bar{G}}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\varphi_j\circ\check{F}_{0,n_0}(Z_i)-\mathbb{E}\left[\frac{\delta_i}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\varphi_j\circ F_0(Z_i)\right]\right).$$

On cherche ensuite à faire apparaître un processus empirique centré. A cet effet, on décompose de nouveau, pour $\varphi_j$ une fonction de base,

$$\begin{aligned}
\bar{\nu}_n(\varphi_j) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\frac{1}{\check{\bar{G}}(Z_i)}-\frac{1}{\bar{G}(Z_i)}\right)\mathbf{1}_{Z_i\leq\tau}\varphi_j\circ\check{F}_{0,n_0}(Z_i)\\
&+\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{1}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\left(\varphi_j\circ\check{F}_{0,n_0}(Z_i)-\varphi_j\circ F_0(Z_i)\right)\\
&\quad-\mathbb{E}\left[\delta_i\frac{1}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\left(\varphi_j\circ\check{F}_{0,n_0}(Z_i)-\varphi_j\circ F_0(Z_i)\right)|(X_0,C_0)\right],\\
&+\mathbb{E}\left[\delta_i\frac{1}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\left(\varphi_j\circ\check{F}_{0,n_0}(Z_i)-\varphi_j\circ F_0(Z_i)\right)|(X_0,C_0)\right],\\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\varphi_j\circ F_0(Z_i)-\mathbb{E}\left[\frac{\delta_i}{\bar{G}(Z_i)}\mathbf{1}_{Z_i\leq\tau}\varphi_j\circ F_0(Z_i)\right].
\end{aligned}$$

En notant $\nu_n(\varphi_j)$ le processus de la dernière ligne, et en utilisant les notations de la preuve de la Proposition 7.1, on obtient

$$\sum_{j=1}^{p}\bar{\nu}_n^2(\varphi_j)\leq 4\left(T_3^{c,p}+T_1^{c,p}+T_2^{c,p}+\sum_{j=1}^{p}\nu_n^2(\varphi_j)\right).$$

Comme $\sum_{j=1}^{p}\nu_n^2(\varphi_j)=\sup_{t\in\mathcal{S}(p)}\nu_n^2(t)$, où $\mathcal{S}(p)$ désigne la sphère unité de $S_p$ (voir par exemple le Lemme 3.1 du Chapitre 3), il vient

$$\begin{aligned}
\left\|\hat{r}_{m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)-\hat{r}_{m\wedge m'}\left(.,\check{F}_{0,n_0},\check{\bar{G}}_n\right)\right\|^2 &\leq 12T_3^{c,m}+12T_3^{c,m\wedge m'}+12T_2^{c,m}+12T_2^{c,m\wedge m'}\\
&+12T_1^{c,m}+12T_1^{c,m\wedge m'}+12\sup_{t\in\mathcal{S}(m)}\nu_n^2(t)\\
&+12\sup_{t\in\mathcal{S}(m\wedge m')}\nu_n^2(t)+3\left\|r_{m'}-r_{m\wedge m'}\right\|^2.
\end{aligned}$$

Le terme non aléatoire est le même que celui étudié dans la preuve du Lemme 6.3 du Chapitre 6. Par l'Inégalité (6.19) ,

$$\max_{m'\in\mathcal{M}_{n,n_0}}\left\|r_{m'}-r_{m\wedge m'}\right\|^2\leq 4\left\|r_m-r\right\|^2.$$

Pour calculer $\Phi(m)$, on retranche maintenant la quantité $V(m')$. On la sépare en plusieurs parties, pour recentrer correctement chacun des termes : $V(m')=V_n(m')+V_{0,n_0}(m')$, avec

$$\begin{aligned}
V_n(m') &= c\mathbb{E}\left[\frac{\delta\mathbf{1}_{Z\leq\tau}}{\bar{G}^2(Z)}\right]\frac{D_{m'}}{n},\\
V_{0,n_0}(m') &= c\left\{3\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^6(x)}f_0(x)dx\frac{D_{m'}}{n_0}+\frac{\bar{c}_2r^2(F_0(\tau))}{\bar{G}_0^4(\tau)}\frac{D_{m'}}{n_0}\right\}.
\end{aligned}\tag{7.19}$$

On a ainsi, comme de plus $V_k(m') \geq V_k(m \wedge m')$ pour $k \in \{n, \{0, n_0\}\}$,

$$\begin{aligned} A(m) \leq\ & 12\,\|r_m - r\|^2 + 12 \max_{m' \in \mathcal{M}_{n,n_0}} T_1^{c,m'} + \max_{m' \in \mathcal{M}_{n,n_0}} T_1^{c,m \wedge m'} \\ & + 12 \max_{m' \in \mathcal{M}_{n,n_0}} \left( \sup_{t \in \mathcal{S}(m)} \nu_n^2(t) - \frac{V_n(m')}{24} \right)_+ + 12 \max_{m' \in \mathcal{M}_{n,n_0}} \left( \sup_{t \in \mathcal{S}(m \wedge m')} \nu_n^2(t) - \frac{V_n(m' \wedge m)}{24} \right)_+ \\ & + 12 \max_{m' \in \mathcal{M}_{n,n_0}} T_3^{c,m'} + 12 \max_{m' \in \mathcal{M}_{n,n_0}} T_3^{c,m' \wedge m}, \\ & + 12 \max_{m' \in \mathcal{M}_{n,n_0}} \left( T_2^{c,m'} - \frac{V_{0,n_0}(m')}{24} \right)_+ + 12 \max_{m' \in \mathcal{M}_{n,n_0}} \left( T_2^{c,m' \wedge m} - \frac{V_{0,n_0}(m' \wedge m)}{24} \right)_+ . \end{aligned}$$

Les termes $T_1^{c,m'}$ et $T_1^{c,m \wedge m'}$ n'ont pas été recentrés. Notons $p_{m'} = m'$ ou $p_{m'} = m \wedge m'$. La majoration suivante provient du Lemme 7.4 et de sa preuve :

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_{n,n_0}} T_1^{c,p_{m'}} \right] \leq \mathbb{E}\left[ T_1^{c,m_{\max}} \right] \leq \left\| \varphi_2' \right\|^2_{L^\infty((0;F_0(\tau)))} \mathbb{E}\left[ \frac{\delta \mathbf{1}_{Z \leq \tau}}{\bar{G}_0^2(Z) \bar{G}^2(Z)} \right] \frac{D_{m_{\max}}^3}{n n_0} \leq \frac{C}{n},$$

puisque $D_{m_{\max}} \leq C n_0^{1/3}$. Ensuite les lemmes suivant, prouvés aux paragraphes qui suivent, sont utiles :

**Lemme 7.8.** *Sous les hypothèses du Théorème 7.1, il existe une constante $C$, dépendant de la fonction $r$, mais ni de $n$, ni de $n_0$, telle que, pour $p_{m'} = m'$ ou $p_{m'} = m \wedge m'$,*

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_{n,n_0}} \left( \sup_{t \in \mathcal{S}(p_{m'})} \nu_n^2(t) - V_\nu(p_{m'}) \right)_+ \right] \leq \frac{C}{n},$$

*avec $V_\nu(m') = 2(1 + 2\kappa_\nu)\mathbb{E}\left[ \delta \mathbf{1}_{Z \leq \tau} / \bar{G}^2(Z) \right] D_{m'}/n$, pour $\kappa_\nu > 0$.*

On choisit alors la constante $c$ dans la définition de $V_n$ de telle sorte que $V_n(p_{m'})/24 > V_\nu(p_{m'})$. Le résultat du Lemme 7.8 est alors encore valable en substituant $V_n/24$ à $V_\nu$.

**Lemme 7.9.** *Sous les hypothèses du Théorème 7.1, il existe une constante $C$, dépendant de la fonction $r$, mais ni de $n$, ni de $n_0$, telle que, pour $p_{m'} = m'$ ou $p_{m'} = m \wedge m'$,*

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_{n,n_0}} \left( T_2^{c,p_{m'}} - V_{0,n_0}(p_{m'})/24 \right)_+ \right] \leq C \frac{\ln(n_0)}{n_0}.$$

**Lemme 7.10.** *Sous les hypothèses du Théorème 7.1, il existe une constante $C$, dépendant de la fonction $r$, mais ni de $n$, ni de $n_0$, telle que, pour $p_{m'} = m'$ ou $p_{m'} = m \wedge m'$,*

$$\mathbb{E}\left[ \max_{m' \in \mathcal{M}_{n,n_0}} T_3^{c,p_{m'}} \right] \leq C \frac{\ln(n)}{n}.$$

Ceci termine la majoration de l'espérance de $\Phi(m)$.

□

**Preuve du Lemme 7.8** On commence par majorer grossièrement

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-V_\nu(p_{m'})\right)_+\right]\leq\sum_{m'\in\mathcal{M}_{n,n_0}}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-V_\nu(p_{m'})\right)_+\right].$$

Ecrivons $\nu_n(t)=\frac{1}{n}\sum_{i=1}^n\psi_t(Z_i,\delta_i)-\mathbb{E}[\psi_t(Z_i,\delta_i)]$, avec la notation $\psi_t(z,\delta)=t\circ F_0(z)\delta\mathbf{1}_{Z\leq\tau}/\bar{G}(z)$. Le processus $\nu_n$ est centré, et controlable par l'Inégalité de Talagrand (Proposition 2.2, Chapitre 2). Calculons les quantités $M_1$, $H$ et $v$ qui sont nécessaires pour l'appliquer.

- Calcul de $M_1$. On a, pour $t\in\mathcal{S}(p_{m'})$, $z\in(0;\tau)$, et $\delta\in\{0,1\}$,

$$|\psi_t(z,\delta)|\leq\bar{G}(\tau)^{-1}\|t\|_{L^\infty((0;F_0(\tau)))}\leq\Phi_0\sqrt{D_{p'_m}}\bar{G}(\tau)^{-1}\|t\|=\bar{G}(\tau)^{-1}\Phi_0\sqrt{D_{p_{m'}}}:=M_1.$$

- Calcul de $H^2$. En décomposant $t\in\mathcal{S}(p_{m'})$ dans la base orthonormée $(\varphi_j)_{j=1,\ldots,D_{p_{m'}}}$, on a

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)\right] &= \sum_{j=1}^{D_{p_{m'}}}\mathbb{E}\left[\nu_n(\varphi_j^2)\right],\\ &= \sum_{j=1}^{D_{p_{m'}}}\mathrm{Var}\left(\hat{a}_j^{F_0,\bar{G}}\right)\leq\mathbb{E}\left[\frac{\delta\mathbf{1}_{Z\leq\tau}}{\bar{G}^2(Z)}\right]\frac{D_{p_{m'}}}{n}:=H^2,\end{aligned}$$

puisqu'il s'agit de la majoration du terme de variance en non adaptatif, obtenue au Lemme 7.3.

- Calcul de $v$. Pour $t\in\mathcal{S}(p_{m'})$, d'après le Lemme 7.1

$$\begin{aligned}\mathrm{Var}\left(\psi_t(X_1)\right) &\leq \mathbb{E}\left[\frac{\delta\mathbf{1}_{Z\leq\tau}}{\bar{G}^2(Z)}t^2\left(F_0(Z_1)\right)\right],\\ &= \int_0^\tau\frac{t^2\left(F_0(x)\right)}{\bar{G}(x)}f(x)dx\leq\frac{1}{\bar{G}(\tau)}\int_0^{F_0(\tau)}t^2\left(u\right)r(u)du,\\ &\leq \frac{1}{\bar{G}(\tau)}\|r\|_{L^\infty((0;F_0(\tau)))}\|t\|^2=\frac{1}{\bar{G}(\tau)}\|r\|_{L^\infty((0;F_0(\tau)))}:=v.\end{aligned}$$

L'application de l'Inégalité (2.1) conduit ensuite au résultat cherché (voir par exemple la preuve du Lemme 6.5.2 au Chapitre 6).

□

**Preuve du Lemme 7.9** On note $p=p_{m'}$ pour simplifier les notations dans la suite. La décomposition effectuée dans la preuve du Lemme 7.5 est

$$T_2^{c,p}\leq 18T_{2,1,1,1}^{c,p}+18T_{2,1,1,2}^{c,p}+9T_{2,1,2}^{c,p}+9T_{2,1,3}^{c,p}+3T_{2,2}^{c,p}+3T_{2,3}^{c,p}.$$

On sépare $V_{0,n_0}$ de façon adaptée, pour recentrer chaque terme non négligeable. À cet effet, on minore l'intégrale intervenant dans la définition de $V_{0,n_0}$ (voir (7.19)) de la façon artificielle suivante :

$$\begin{aligned}3\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^6(x)}f_0(x)dx &\geq \int_0^\tau(r\circ F_0(x))^2f_0(x)dx+\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)}f_0(x)dx\\ &\quad+\int_0^\tau\frac{(r\circ F_0(x))^2}{\bar{G}_0^6(x)}f_0(x)dx.\end{aligned}$$

Ceci permet le découpage $V_{0,n_0}(p)/24 \geq V^b_{2,1,1}(p) + V^b_{2,1,3}(p) + V^b_{2,2}(p) + V^b_{2,3}(p)$, avec, pour $c' = c/24$

$$
\begin{aligned}
V^b_{2,1,1}(p) &= c' \int_0^\tau (r \circ F_0(x))^2 f_0(x) dx \frac{D_p}{n_0}, \\
V^b_{2,1,3}(p) &= c' \frac{\bar{c}_2 r^2(F_0(\tau))}{\bar{G}_0^2(\tau)} \frac{D_p}{n_0}, \\
V^b_{2,2}(p) &= c' \int_0^\tau \frac{(r \circ F_0(x))^2}{\bar{G}_0^2(x)} f_0(x) dx \frac{D_p}{n_0}, \\
V^b_{2,3}(p) &= c' \int_0^\tau \frac{(r \circ F_0(x))^2}{\bar{G}_0^6(x)} f_0(x) dx \frac{D_p}{n_0}.
\end{aligned}
$$

On peut donc écrire

$$
\begin{aligned}
&\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} (T_2^{c,p} - V_{0,n_0}(p)/24)_+\right] \\
&\leq \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} 18 T^{c,p}_{2,1,1,1}\right] + \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(18 T^{c,p}_{2,1,1,2} - V^b_{2,1,1}(p)\right)_+\right] \\
&\quad + \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} 9 T^{c,p}_{2,1,2}\right] + \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(9 T^{c,p}_{2,1,3} - V^b_{2,1,3}(p)\right)_+\right] \\
&\quad + \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(3 T^{c,p}_{2,2} - V^b_{2,2}(p)\right)_+\right] + \mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} \left(3 T^{c,p}_{2,3} - V^b_{2,3}(p)\right)_+\right].
\end{aligned}
$$

Globalement, tous ces termes se majorent avec des techniques déjà utilisées dans les Chapitres 3 et 5 par exemple, excepté celui faisant intervenir $T^{c,p}_{2,1,1,1}$. Commençons par les termes les plus classiques.

• **Terme faisant intervenir** $T^{c,p}_{2,1,2}$. Ce terme n'a pas été recentré, puisque l'on a directement (cf. preuve du Lemme 7.5),

$$
\mathbb{E}\left[\max_{m' \in \mathcal{M}_{n,n_0}} 9 T^{c,p}_{2,1,2}\right] \leq \mathbb{E}\left[9 T^{c,m_{\max}}_{2,1,2}\right] \leq \frac{C}{n_0}. \tag{7.20}
$$

• **Terme faisant intervenir** $T^{c,p}_{2,1,1,2}$. Ce terme se réécrit, en majorant $(n_0+1)^{-1}$ par $n_0^{-1}$, comme le supremum du carré d'un processus empirique centré borné : $T^{c,p}_{2,1,1,2} \leq \sum_{j=1}^{D_p} (\nu^b_{n_0}(\varphi_j))^2 = \sup_{t \in \mathcal{S}(p)} (\nu^b_{n_0}(t))^2$, avec

$$
\begin{aligned}
\nu^b_{n_0}(t) &= \frac{1}{n_0} \sum_{i_0=1}^{n_0} \left\{ \frac{\delta_{0,i_0}}{\bar{G}_0(Z_{0,i_0})} \varphi_j(F_0(Z_{0,i_0})) r(F_0(Z_{0,i_0})) \mathbf{1}_{Z_{0,i_0} \leq \tau} \right. \\
&\quad \left. - \mathbb{E}\left[\frac{\delta_{0,i_0} \mathbf{1}_{Z_{0,i_0} \leq \tau}}{\bar{G}_0(Z_{0,i_0})} \varphi_j(F_0(Z_{0,i_0})) r(F_0(Z_{0,i_0}))\right] \right\}.
\end{aligned}
$$

On lui applique l'Inégalité de Talagrand (Proposition 2.2). Les calculs des quantités impliquées sont analogues à celles calculées pour démontrer le Lemme 7.8 :

$$
\begin{aligned}
M_1^b &= \|r\|_{L^\infty((0;F_0(\tau)))} \Phi_0 \sqrt{D_p}/\bar{G}_0(\tau), \ v^b = \|r\|_{L^\infty((0;F_0(\tau)))}/\bar{G}_0(\tau), \\
(H^b)^2 &= \int_0^\tau (r \circ F_0(x))^2 f_0(x) dx D_p/n_0.
\end{aligned}
$$

On prouve ainsi que, pour tout $\kappa_1 > 0$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(\sup_{t\in\mathcal{S}(p_{m'})}\left(\nu_n^b(t)\right) - 2(1+2\kappa_1)\int_0^\tau (r\circ F_0(x))^2 f_0(x)dx\frac{D_p}{n_0}\right)_+\right] \leq \frac{C}{n_0}.$$

On peut choisir la constante $c'$ dans la définition de $V_{2,1}^b$ de telle sorte que

$$V_{2,1,1}^b(p) \geq 2(1+2\kappa_1)\int_0^\tau (r\circ F_0(x))^2 f_0(x)dx\frac{D_p}{n_0},$$

et donc on a finalement montré que

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(18T_{2,1,1,2}^{c,p} - V_{2,1,1}^b(p)\right)_+\right] \leq \frac{C}{n_0}. \tag{7.21}$$

• **Terme faisant intervenir** $T_{2,2}^{c,p}$. Pour ce terme et les deux suivants, l'idée générale est d'appliquer le Corollaire 2.2 (Chapitre 2). Pour cela, on va utiliser des majorations intermédiaires obtenues dans la preuve du Lemme 7.5. On commence comme toujours par

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^{c,p} - V_{2,2}^b(p)\right)_+\right] \leq \sum_{m'\in\mathcal{M}_{n,n_0}}\mathbb{E}\left[\left(3T_{2,2}^{c,p} - V_{2,2}^b(p)\right)_+\right],$$

et on raisonne pour $p$ fixé. On a montré en non adaptatif que

$$T_{2,2}^{c,p} \leq (\pi^4/4)\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)} f_0(x)dxD_p^4 \left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty((0;\tau))}^4.$$

Posons, $V_{2,2}(p) = \kappa_2(3\pi^4/4)\int_0^\tau (r\circ F_0(x))^2/\bar{G}_0^4(x)f_0(x)dxD_p^4\ln^2(n_0)/n_0^2$, pour $\kappa_2 > 0$. On a alors,

$$\begin{aligned}
\mathbb{E}\left[\left(3T_{2,2}^p - V_{2,2}(p)\right)_+\right] &\leq (3\pi^4/4)\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)} f_0(x)dxD_p^4\mathbb{E}\left[\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty((0;\tau))}^4\right.\right. \\
&\qquad \left.\left. - \frac{V_{2,2}(p)}{(3\pi^4/4)\int_0^\tau (r\circ F_0(x))^2/\bar{G}_0^4(x)f_0(x)dxD_p^4}\right)_+\right], \\
&= (3\pi^4/4)\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)} f_0(x)dxD_p^4\mathbb{E}\left[\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty((0;\tau))}^4\right.\right. \\
&\qquad \left.\left. - \kappa_2\frac{\ln^2(n_0)}{n_0^2}\right)_+\right], \\
&\leq CD_p^4 n_0^{-2^{-3/2}\kappa_2^{1/2}}
\end{aligned}$$

en utilisant le Corollaire 2.2 avec $p = 4$. Donc,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - V_{2,2}(p)\right)_+\right] \leq C\sum_{m'\in\mathcal{M}_{n,n_0}} D_p^4 n_0^{-2^{-3/2}\kappa_2^{1/2}}, \tag{7.22}$$

On peut majorer $\sum_{m'\in\mathcal{M}_{n,n_0}} D_p^4 \leq D_{m_{\max}}^5 \leq cn_0^{5/3}$ avec $c$ une constante. Le membre de droite de (7.22) est alors inférieur ou égal à une quantité de l'ordre de $n_0^{5/3-2^{-3/2}\sqrt{\kappa_2}}$. Celle-ci sera bornée par $C/n_0$ si $\kappa_2$ est assez grand.

On note de plus que, sous la contrainte $D_p \leq n_0^{1/3}/\ln^{2/3}(n_0)$,

$$V_{2,2}(p) \leq V_{2,2}^{bis}(p) = \kappa_2(3\pi^4/4)\int_0^\tau \frac{(r\circ F_0(x))^2}{\bar{G}_0^4(x)} f_0(x)dx\frac{D_p}{n_0},$$

et l'égalité (7.22) est encore valide en remplaçant $V_{2,2}$ par $V_{2,2}^{bis}$ : ainsi

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - V_{2,2}^{bis}(p)\right)_+\right] \leq \frac{C}{n_0}.$$

Enfin, on choisit $c'$ dans la définition de $V_{2,2}^b$ de telle sorte que $V_{2,2}^b(p) > V_{2,2}^{bis}(p)$, pour finalement conclure que l'on a aussi

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T_{2,2}^p - V_{2,2}^{b}(p)\right)_+\right] \leq \frac{C}{n_0}. \tag{7.23}$$

• **Terme faisant intervenir** $T_{2,1,3}^{c,p}$. On a montré à la preuve du Lemme 7.5 :

$$T_{2,1,3}^{c,p} \leq \frac{1}{\bar{G}_0^2(\tau)} D_p r^2(F_0(\tau)) \left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2.$$

$$T_{2,1,3}^{c,p} \leq \frac{1}{\bar{G}_0^2(\tau)} D_p r^2(F_0(\tau)) \left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2.$$

En posant $V_{2,1,3}(p) = (9r^2(F_0(\tau))\bar{c}_2/\bar{G}_0^2(\tau))D_p/n_0$, on a

$$\begin{aligned}\mathbb{E}\left[\left(9T_{2,1,3}^p - V_{2,1,3}(p)\right)_+\right] &\leq 9\frac{r^2(F_0(\tau))}{\bar{G}_0^2(\tau)} D_p \mathbb{E}\left[\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2 - \frac{\bar{c}_2}{n_0}\right)_+\right], \\ &\leq C(A_1 + A_2).\end{aligned}$$

avec la décomposition

$$\begin{aligned}A_1 &= \mathbb{E}\left[\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2 - \mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2\right]\right)_+\right], \\ A_2 &= \left(\mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2\right] - \frac{\bar{c}_2}{n_0}\right)_+.\end{aligned}$$

Mais, par la Proposition 2.6, le terme $A_2$ est nul. Quant au terme $A_1$, il vaut

$$\begin{aligned}A_1 &= \mathbb{E}\left[\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2 - \mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2\right]\right)\right. \\ &\qquad \left.\times \mathbf{1}_{\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2 \geq \mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2\right]}\right], \\ &\leq 2\mathbb{P}\left(\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)} \geq \sqrt{\mathbb{E}\left[\left\|\bar{G}_0\left(\check{F}_{0,n_0} - F_0\right)\right\|_{L^\infty(0;\tau)}^2\right]}\right), \\ &\leq 2.5\exp\left(-2n_0a^2 + Ca\sqrt{n_0}\right),\end{aligned}$$

d'après la Proposition 2.5, avec $a = \sqrt{\mathbb{E}[\|\bar{G}_0(\check{F}_{0,n_0} - F_0)\|_{L^\infty(0;\tau)}^2]}$. Cette quantité est inférieure ou égale à $C/n_0^l$, pour tout entier positif $l$, dès que $a > 0$, ce que l'on peut supposer,

sans quoi le terme $T^{c,p}_{2,1,3}$ est p.s. nul. Enfin, le choix d'une constante $c'$ dans la définition de $V^b_{2,1,3}$ satisfaisant $V^b_{2,1,3}(p) > V_{2,1,3}(p)$, entraîne, en sommant sur $m'$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(9T^p_{2,1,3}-V^b_{2,2}(p)\right)_+\right]\leq\frac{C}{n_0}. \tag{7.24}$$

• **Terme faisant intervenir** $T^{c,p}_{2,3}$. Le raisonnement est exactement le même que pour le terme précédent. On sait, toujours par la preuve du Lemme 7.5, que

$$T^{c,p}_{2,3}\leq(1/6)\left\|\varphi^{(3)}_2\right\|^2_{L^\infty((0;F_0(\tau)))}\int_0^\tau\frac{(r\circ F_0)^2(x)}{\bar G^6_0(x)}f_0(x)dxD_p^7\left\|\bar G_0\left(\check F_{0,n_0}-F_0\right)\right\|^6_{L^\infty((0;\tau))}.$$

En posant, pour $\kappa_3>0$,

$$V_{2,3}(p)=\kappa_3(1/2)\left\|\varphi^{(3)}_2\right\|^2_{L^\infty((0;F_0(\tau)))}\int_0^\tau\frac{(r\circ F_0)^2(x)}{\bar G^6_0(x)}f_0(x)dxD_p^7\frac{\ln^3(n_0)}{n_0^3},$$

on a de façon analogue, en appliquant le Corollaire 2.2 avec $p=6$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T^p_{2,3}-V_{2,3}(p)\right)_+\right]\leq C\sum_{m'\in\mathcal{M}_{n,n_0}}D_p^7n_0^{-2^{-5/3}\kappa_3^{1/3}}, \tag{7.25}$$

puis, si $D_p\leq n_0^{1/3}/\ln^{1/2}(n_0)$,

$$V_{2,3}(p)\leq V^{bis}_{2,3}(p)=(1/2)\left\|\varphi^{(3)}_2\right\|^2_{L^\infty((0;F_0(\tau)))}\int_0^\tau\frac{(r\circ F_0)^2(x)}{\bar G^6_0(x)}f_0(x)dx\kappa_3\frac{D_p}{n_0},$$

mais aussi $V^b_{2,3}(p)>V^{bis}_{2,3}(p)$ à condition que la constante $c'$ soit suffisamment grande ; on en déduit que (7.25) est vrai en substituant $V^b_{2,3}(p)$ à $V_{2,3}(p)$ :

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}\left(3T^p_{2,3}-V^b_{2,3}(p)\right)_+\right]\leq\frac{C}{n_0}. \tag{7.26}$$

• **Terme faisant intervenir** $T^{c,p}_{2,1,1,1}$. C'est le terme le plus "nouveau". Rappelons que

$$T^{c,p}_{2,1,1,1}=\sum_{j=1}^{D_p}\left(\frac{1}{n_0+1}\sum_{i_0=1}^{n_0}\delta_{0,i_0}\left\{\frac{1}{\check{\bar G}_{0,n_0}(Z_{0,i_0})}-\frac{1}{\bar G_0(Z_{0,i_0})}\right\}\varphi_j(F_0(Z_{0,i_0}))r(F_0(Z_{0,i_0}))\mathbf{1}_{Z_{0,i_0}\leq\tau}\right)^2,$$

On a donc $\mathbb{E}[\max_{m'\in\mathcal{M}_{n,n_0}}T^{c,p}_{2,1,1,1}]\leq\mathbb{E}[T^{c,m_{\max}}_{2,1,1,1}]$, et, en utilisant en particulier l'Inégalité de Cauchy-Schwarz, $T^{c,m_{\max}}_{2,1,1,1}\leq T^{c,m_{\max}}$, avec

$$\begin{aligned}T^{c,m_{\max}} &= \frac{\|r\|^2_{L^\infty((0;F_0(\tau)))}}{\bar F_0^2(\tau)\bar G_0^2(\tau)}\left(\frac{1}{n_0}\sum_{i_0=1}^{n_0}\frac{1}{(\check{\bar G}_{0,n_0}(X_{0,i_0}))^2}\right)\\ &\quad\times\left\|\bar F_0\left(\check G_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\sum_{j=1}^{D_{m_{\max}}}\left(\frac{1}{n_0}\sum_{i_0=1}^{n_0}\varphi_j^2\left(F_0(X_{i_0})\right)\mathbf{1}_{X_{0,i_0}\leq\tau}\right).\end{aligned}$$

On introduit l'évènement $\Omega_{n_0} = \{\sup_{x\in(0;\tau)}(\check{G}_{0,n_0}(x) - G_0(x)) \geq -\bar{G}_0(\tau)/2\}$, comme dans la preuve du Lemme 7.2 : $T^{c,p}_{2,1,1,1} \leq T^{c,p}(\mathbf{1}_{\Omega_{n_0}} + \mathbf{1}_{\Omega^c_{n_0}})$. Ainsi,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}} 18T^{c,p}_{2,1,1,1}\right] \leq \mathbb{E}\left[18T^{c,m_{\max}}\mathbf{1}_{\Omega_{n_0}} + 18T^{c,m_{\max}}\mathbf{1}_{\Omega^c_{n_0}}\right].$$

Sur $\Omega^c_{n_0}$, on effectue les majorations effectuées dans la preuve du Lemme 7.2 pour majorer le terme $T_b$ :

$$\begin{aligned}
\mathbb{E}\left[18T^{c,m_{\max}}\mathbf{1}_{\Omega^c_{n_0}}\right] &\leq 18\sum_{m'\in\mathcal{M}_{n,n_0}} D_{m_{\max}}\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{1}{\bar{F}_0^2(\tau)\bar{G}_0^2(\tau)}\times 4n_0^2 \\
&\quad\times\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^4_{L^\infty((0;\tau))}\right]^{1/2}\mathbb{P}\left(\Omega^c_{n_0}\right)^{1/2}, \\
&\leq 18\sum_{m'\in\mathcal{M}_{n,n_0}} D_{m_{\max}}\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{1}{\bar{F}_0^2(\tau)\bar{G}_0^2(\tau)}\times 4n_0^2\frac{\sqrt{\bar{c}_4}}{n_0}\mathbb{P}\left(\Omega^c_{n_0}\right)^{1/2}, \\
&\leq 18\times 4\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{\sqrt{\bar{c}_4}}{\bar{F}_0^2(\tau)\bar{G}_0^2(\tau)}n_0^4\exp\left(-n_0\frac{\bar{G}_0^2(\tau)}{4}+C\frac{\bar{G}_0(\tau)}{4}\sqrt{n_0}\right), \\
&\leq \frac{C}{n_0}.
\end{aligned}$$

Sur $\Omega_{n_0}$ maintenant, on commence par majorer $(1/n_0)\sum_{i_0=1}^{n_0}(\check{\bar{G}}_{0,n_0}(X_{i_0}))^{-2}$ par $4/\bar{G}_0^2(\tau)$, comme pour le terme $T_a$ de la preuve du Lemme 7.2, puis on remarque que

$$\sum_{j=1}^{D_{m_{\max}}}\left(\frac{1}{n_0}\sum_{i_0=1}^{n_0}\varphi_j^2\left(F_0(X_{0,i_0})\right)\mathbf{1}_{X_{0,i_0}\leq\tau}\right) = \sup_{t\in\mathcal{S}(m_{\max})}\frac{1}{n_0}\sum_{i_0=1}^{n_0}t^2\left(F_0(X_{0,i_0})\right)\mathbf{1}_{X_{0,i_0}\leq\tau},$$

de telle sorte que,

$$\begin{aligned}
\mathbb{E}\left[18T^{c,m_{\max}}\mathbf{1}_{\Omega_{n_0}}\right] &\leq 18\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{4}{\bar{F}_0^2(\tau)\bar{G}_0^4(\tau)}\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\right. \\
&\quad\left.\times\sup_{t\in\mathcal{S}(m_{\max})}\frac{1}{n_0}\sum_{i_0=1}^{n_0}t^2\left(F_0(Z_{i_0})\right)\mathbf{1}_{X_{0,i_0}\leq\tau}\right], \\
&\leq 18\|r\|^2_{L^\infty((0;F_0(\tau)))}\frac{4}{\bar{F}_0^2(\tau)\bar{G}_0^4(\tau)}\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\right. \\
&\quad\left.\times\left\{\sup_{t\in\mathcal{S}(m_{\max})}\mathbb{E}\left[t^2\left(F_0(X_0)\right)\mathbf{1}_{X_0\leq\tau}\right]+\sup_{t\in\mathcal{S}(m_{\max})}\nu^c_{n_0}(t^2)\right\}\right],
\end{aligned}$$

avec $\nu^c_{n_0}(t^2) = (1/n_0)\sum_{i_0=1}^{n_0}t^2(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq\tau} - \mathbb{E}[t^2(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq\tau}]$. On note ensuite, par le Lemme 7.1, pour $t\in\mathcal{S}(m_{\max})$,

$$\mathbb{E}\left[t^2\left(F_0(X_0)\right)\mathbf{1}_{X_0\leq\tau}\right] = \int_0^\tau t^2(F_0(x))\bar{G}_0(x)f_0(x)dx \leq f_{0,\infty}\|t\|^2 = f_{0,\infty},$$

puisque la densité $f$ est supposée majorée par $f_{0,\infty}$ sur $(0;\tau)$. Par la Proposition 2.6,

$$\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\left(\sup_{t\in\mathcal{S}(m_{\max})}\mathbb{E}\left[t^2\left(F_0(X_0)\right)\mathbf{1}_{X_0\leq\tau}\right]\right)\right]$$

$$\leq \quad f_{0,\infty}\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\right]\leq\frac{\bar{c}_2 f_{0,\infty}}{n_0}.$$

Par ailleurs, par l'Inégalité de Cauchy-Schwarz et toujours la Proposition 2.6,

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\left(\sup_{t\in\mathcal{S}(m_{\max})}\nu^c_{n_0}(t^2)\right)\right] &\leq \mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^4_{L^\infty((0;\tau))}\right]^{1/2}\\ &\quad\times\mathbb{E}\left[\sup_{t\in\mathcal{S}(m_{\max})}\left(\nu^c_{n_0}(t^2)\right)^2\right]^{1/2},\\ &\leq \frac{\sqrt{\bar{c}_4}}{n_0}\mathbb{E}\left[\sup_{t\in\mathcal{S}(m_{\max})}\left(\nu^c_{n_0}(t^2)\right)^2\right]^{1/2}.\end{aligned}$$

Le supremum du processus $\nu^c_{n_0}$ est majoré dans Brunel & Comte (2005) (p.469), grâce aux travaux de Baraud (2002) : si $D_{m_{\max}}\leq\sqrt{n_0}/(4f_{0,\infty})$, alors

$$\mathbb{E}\left[\left\|\bar{F}_0\left(\check{G}_{0,n_0}-G_0\right)\right\|^2_{L^\infty((0;\tau))}\sup_{t\in\mathcal{S}(m_{\max})}\nu^c_{n_0}(t^2)\right] \quad\leq\quad \frac{\sqrt{\bar{c}_4}2\ln(n_0)}{n_0}.$$

Finalement, $\mathbb{E}[18T^{c,m_{\max}}\mathbf{1}_{\Omega_{n_0}}]\leq C\ln(n_0)/n_0$, et par suite,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_{n,n_0}}9T^{c,p}_{2,1,2}\right]\leq\mathbb{E}\left[9T^{c,m_{\max}}_{2,1,2}\right]\leq\frac{C\ln(n_0)}{n_0}. \tag{7.27}$$

La preuve du Lemme 7.9 est complète en regroupant (7.20), (7.21), (7.23), (7.24), (7.26) et (7.27).

□

**Preuve du Lemme 7.10** On a d'abord $\max_{m'\in\mathcal{M}_{n,n_0}}T^{c,p}_3\leq T^{c,m_{\max}}_3$, avec

$$\begin{aligned}T^{c,m_{\max}}_3 &= \sum_{j=1}^{D_{m_{\max}}}\left(\frac{1}{n}\sum_{i=1}^n\left\{\frac{\delta_i}{\check{\bar{G}}_n(Z_i)}-\frac{\delta_i}{\bar{G}_(Z_i)}\right\}\varphi_j\left(\check{F}_{0,n_0}(Z_i)\right)\mathbf{1}_{Z_i\leq\tau}\right)^2,\\ &\leq \frac{1}{\bar{G}^2(\tau)\bar{F}^2(\tau)}\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{\check{\bar{G}}^2_n(X_i)}\right)\left\|\bar{F}\left(\check{G}_n-G\right)\right\|^2_{L^\infty((0;\tau))}\\ &\quad\times\sum_{j=1}^{D_{m_{\max}}}\left(\frac{1}{n}\sum_{i=1}^n\varphi^2_j(\check{F}_{0,n_0}(X_i))\mathbf{1}_{X_i\leq\tau}\right).\end{aligned}$$

La stratégie est ensuite exactement la même que pour le terme $T^{c,p}_{2,1,1,1}$ de la preuve du Lemme 7.9 ci-dessus : on introduit l'évènement $\Omega_n$ de la preuve du Lemme 7.2 : sur $\Omega^c_n$, l'argument

principal est que la probabilité $\mathbb{P}(\Omega_n^c)$ est plus petite à constante près que n'importe quelle puissance de $n$. Sur $\Omega_n$, on fait apparaître le processus

$$\nu_n^d(t^2) = \frac{1}{n}\sum_{i=1}^{n} t^2(\check{F}_{0,n_0}(X_i))\mathbf{1}_{X_i\leq\tau} - \mathbb{E}\left[t^2(\check{F}_{0,n_0}(X_i))\mathbf{1}_{X_i\leq\tau}\,|(X_0,C_0)\right],$$

et tout fonctionne exactement de la même façon, en utilisant $\sup_{(0;\tau)} f \leq f_\infty$, et $D_m \leq \sqrt{n}/(4f_\infty)$.

□

### 7.4.2 Eléments pour les preuves relatives à la Section 7.3

Comme pour les autres méthodes de déformation introduites dans cette thèse (voir par exemple Section 4.4 du Chapitre 4), les preuves reposent sur le contrôle de l'écart entre la fonction de déformation et son estimateur. Nous introduisons donc ci-dessous, pour les fonctions $\Phi$ et $B$, des propriétés similaires aux inégalités de type Dvoretzky *et al.* (1956) pour la répartition empirique, ou Bitouzé *et al.* (1999) pour le Kaplan-Meier, introduites au Chapitre 2 (Section 2.1.2 et Section 2.1.3). Lorsque l'on dispose de ces outils, les preuves sont ensuite extrêmement similaires à des preuves déjà détaillées dans cette thèse. On se contentera donc d'en donner les grandes lignes.

#### Propriétés des estimateurs des fonctions de déformation

Rappelons que la fonction $B$ est définie par (7.10), et son estimateur $\hat{B}$ par (7.12). On introduit également le pseudo-estimateur

$$\tilde{B} : x \mapsto \frac{1}{n_0}\sum_{i_0=1}^{n_0} \bar{G}(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq x}.$$

On obtient les inégalités intégrées suivantes :

**Proposition 7.4.** *Soit $x$ un réel positif. Soit $k$ un entier supérieur ou égal à 2.*

*(i) On a,*

$$\mathbb{E}\left[\left|\tilde{B}(x) - B(x)\right|^k\right] \leq \tilde{c}_k \frac{1}{n_0^{k/2}},$$

*avec $\tilde{c}_2 = 1/4$, et $\tilde{c}_k = 2\bar{\bar{c}}_k$ si $k > 2$, où $\bar{\bar{c}}_k$ est la constante de la Proposition 2.7.*

*(ii) Si l'on suppose $\tau_X > \tau_{X_0}$,*

$$\mathbb{E}\left[\left|\hat{B}(x) - \tilde{B}(x)\right|^k\right] \leq \frac{\bar{c}_k}{\bar{F}_X^k(\tau_{X_0})}\frac{1}{n^{k/2}},$$

*avec $\bar{c}_k$ la constante de l'Inégalité (2.6).*

*(iii) On a donc aussi,*

$$\mathbb{E}\left[\left|\hat{B}(x) - B(x)\right|^k\right] \leq 2^{k-1}\left(\frac{\bar{c}_k}{\bar{F}_X^k(\tau_{X_0})}\frac{1}{n^{k/2}} + \tilde{c}_k\frac{1}{n_0^{k/2}}\right).$$

De même, rappelons que $\Phi$ est définie par (7.8), et son estimateur $\hat{\Phi}$ par (7.14). On note également

$$\tilde{\Phi} : x \mapsto \int_0^x \bar{G} \circ \hat{F}_{0,n_0}^{-1}(u)du.$$

On obtient les inégalités suivantes :

**Proposition 7.5.** *Soit $k$ un entier supérieur ou égal à 2.*

*(i) On suppose $\tau_{X_0} < \infty$, et la densité $f_C$ bornée. Alors, pour tout $x \in (0;1)$,*

$$\mathbb{E}\left[\left|\tilde{\Phi}(x) - \Phi(x)\right|^k\right] \leq C_k \|f_C\|^k_{L^\infty((0;\tau_C))} \tau^k_{X_0} \frac{1}{n_0^{k/2}},$$

*où $C_k$ est la constante de l'Inégalité (2.4).*

*(ii) On suppose $\tau_{X_0} < \tau_X$. Alors, pour tout $x \in (0;1)$,*

$$\mathbb{E}\left[\left|\hat{\Phi}(x) - \tilde{\Phi}(x)\right|^k\right] \leq \frac{\bar{c}_k}{\bar{F}^k_X(\tau_{X_0})} \frac{1}{n^{k/2}},$$

*où $\bar{c}_k$ est la constante de l'Inégalité (2.6).*

**Preuve de la Proposition 7.4**

– Pour l'Inégalité (i), on commence par le cas particulier $k = 2$ : dans ce cas, on a le calcul exact suivant,

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{B}(x) - B(x)\right|^2\right] &= \mathrm{Var}\left(\frac{1}{n_0}\sum_{i_0=1}^{n_0} \bar{G}(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq x}\right), \\ &= \frac{1}{n_0}\mathrm{Var}\left(\bar{G}(X_{0,1})\mathbf{1}_{X_{0,1}\leq x}\right) := \frac{1}{n_0}\mathrm{Var}(T_0),\end{aligned}$$

avec $T_0 \in (0;1)$. On utilise maintenant

$$\mathrm{Var}(T_0) = \mathbb{E}[T_0^2] - (\mathbb{E}[T_0])^2 \leq \mathbb{E}[T_0] - (\mathbb{E}[T_0])^2 \leq \frac{1}{4},$$

ce qui donne le résultat cherché.

Si $k > 2$, on a

$$\mathbb{E}\left[\left|\tilde{B}(x) - B(x)\right|^k\right] = \frac{1}{n_0^k}\mathbb{E}\left[\left|\sum_{i_0=1}^{n_0} \xi_{i_0}\right|^k\right],$$

avec $\xi_{i_0} = \bar{G}(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq x} - \mathbb{E}[\bar{G}(X_{0,i_0})\mathbf{1}_{X_{0,i_0}\leq x}]$, variable centrée bornée. On peut donc appliquer l'Inégalité de Rosenthal (Proposition 2.7). Ici, $\mathbb{E}|\xi_{i_0}|^k \leq 1$, et $\mathbb{E}[\xi_{i_0}^2] = \mathrm{Var}(\xi_{i_0}) \leq \mathbb{E}[\bar{G}^2(X_{0,i_0})] \leq 1$. Donc,

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{B}(x) - B(x)\right|^k\right] &= \frac{1}{n_0^k}\bar{\bar{c}}_k\left\{\sum_{i_0=1}^{n_0} 1 + \left(\sum_{i_0=1}^{n_0} 1\right)^{k/2}\right\}, \\ &\leq \frac{1}{n_0^k}\bar{\bar{c}}_k\left\{n_0 + n_0^{k/2}\right\}, \\ &\leq 2\bar{\bar{c}}_k \frac{1}{n_0^{k/2}}.\end{aligned}$$

– Pour l'Inégalité (ii),

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{B}(x)-\tilde{B}(x)\right|^k\right] &\leq \mathbb{E}\left[\left|\frac{1}{n_0}\sum_{i_0=1}^{n_0}\left(\breve{\bar{G}}_n(X_{0,i_0})-\bar{G}(X_{0,i_0})\right)\mathbf{1}_{X_{0,i_0}\leq x}\right|^k\right], \\
&\leq \mathbb{E}\left[\left\|\breve{\bar{G}}_n-\bar{G}\right\|^k_{L^\infty((0;\tau_{X_0}))}\right], \\
&\leq \frac{1}{\bar{F}^k_X(\tau_{X_0})}\mathbb{E}\left[\left\|\bar{F}_X\left(\breve{\bar{G}}_n-\bar{G}\right)\right\|^k_{L^\infty((0;\tau_{X_0}))}\right],
\end{aligned}$$

puisque l'hypothèse $\tau_X > \tau_{X_0}$ entraîne $\bar{F}_X(\tau_{X_0}) > 0$. On conclut avec l'Inégalité (2.6).

□

**Preuve de la Proposition 7.5** Il suffit de prouver les deux inégaliés suivantes :

(i') On suppose $\tau_{X_0} < \infty$, et la densité $f_C$ bornée. Alors, pour tout $x \in (0;1)$,

$$\left|\tilde{\Phi}(x)-\Phi(x)\right| \leq \|f_C\|_{L^\infty((0;\tau_C))}\tau_{X_0}\left\|\hat{F}_{0,n_0}-F_0\right\|_{L^\infty((0;\tau_{X_0}))}.$$

(ii') On suppose $\tau_{X_0} < \tau_X$. Alors, pour tout $x \in (0;1)$,

$$\left|\hat{\Phi}(x)-\tilde{\Phi}(x)\right| \leq \frac{1}{\bar{F}_X(\tau_{X_0})}\left\|\bar{F}_X\left(\check{F}^{KM}_C-F_C\right)\right\|_{L^\infty((0;\tau_{X_0}))}.$$

En effet, si celles-ci sont démontrées, il suffit ensuite d'appliquer l'Inégalité (2.4) dans (i') (resp. l'Inégalité (2.6) dans (ii')) pour obtenir l'Inégalité (i) (resp. (ii)).

– Pour démontrer l'Inégalité (i'), on utilise d'abord l'inégalité des accroissements finis :

$$\begin{aligned}
\left|\tilde{\Phi}(x)-\Phi(x)\right| &\leq \int_0^x\left|\bar{G}\circ\hat{F}^{-1}_{0,n_0}(u)-\bar{G}\circ F_0^{-1}(u)\right|du, \\
&\leq \|G'\|_{L^\infty((0;\tau_C))}\int_0^x\left|\hat{F}^{-1}_{0,n_0}(u)-F_0^{-1}(u)\right|du, \\
&\leq \|f_C\|_{L^\infty((0;\tau_C))}\int_0^1\left|\hat{F}^{-1}_{0,n_0}(u)-F_0^{-1}(u)\right|du.
\end{aligned}$$

On utilise ensuite l'égalité suivante (voir Shorack & Wellner 1986, exercice 3, p.64), valable quelles que soient les fonctions de répartition $F$ et $G$ :

$$\int_0^1\left|F^{-1}(u)-G^{-1}(u)\right|du = \int_{\mathbb{R}}|F(x)-G(x)|\,dx.$$

On a donc,

$$\begin{aligned}
\left|\tilde{\Phi}(x)-\Phi(x)\right| &\leq \|f_C\|_{L^\infty((0;\tau_C))}\int_0^{\tau_{X_0}}\left|\hat{F}_{0,n_0}(x)-F_0(x)\right|dx, \\
&= \|f_C\|_{L^\infty((0;\tau_C))}\tau_{X_0}\left\|\hat{F}_{0,n_0}-F_0\right\|_{L^\infty((0;\tau_{X_0}))}.
\end{aligned}$$

– Pour l'Inégalité (ii'),

$$\begin{aligned}\left|\hat{\Phi}(x)-\tilde{\Phi}(x)\right| &\leq \int_0^x \left|\left(\breve{\bar{G}}_n-\bar{G}\right)\circ\hat{F}_{0,n_0}^{-1}(u)\right|du,\\ &\leq \left\|\breve{\bar{G}}_n^{(-)}-\bar{G}\right\|_{L^\infty((0;\hat{F}_{0,n_0}^{-1}(1)))},\\ &\leq \left\|\breve{\bar{G}}_n^{(-)}-\bar{G}\right\|_{L^\infty((0;\tau_{X_0}))},\end{aligned}$$

car, par définition de la répartition empirique, $\hat{F}_{0,n_0}^{-1}(1)=\max_{i_0}X_{0,i_0}\leq\tau_{X_0}$. On conclut en utilisant $\bar{F}_X(\tau_{X_0})>0$, comme dans la Proposition 7.4.

□

**Preuve de la Proposition 7.2**

La preuve est fondée sur la décomposition en trois termes de la perte de l'estimateur $g_m(.,\hat{B})$

$$\left\|\hat{g}_m(.,\hat{B})-g_m\right\|_B^2\leq 3\sum_{l=0}^{2}T_l^m,$$

avec

$$\begin{aligned}T_1^m &= \left\|\hat{g}_m(.,\hat{B})-\hat{g}_m(.,B)-\mathbb{E}\left[\hat{g}_m(.,\hat{B})-\hat{g}_m(.,B)\,|(X_-,C_-,X_0)\right]\right\|_B^2,\\ T_2^m &= \left\|\mathbb{E}\left[\hat{g}_m(.,\hat{B})-\hat{g}_m(.,B)\,|(X_-,C_-,X_0)\right]\right\|_B^2,\\ T_0^m &= \|\hat{g}_m(.,B)-g\|_B^2,\end{aligned}$$

où $\mathbb{E}[Z|\,(X_-,C_-,X_0)]$ désigne l'espérance d'une variable $Z$ conditionnellement aux échantillons $(X_{-i},C_{-i})_{i=1,\dots,n}$ et $(X_{0,i_0})_{i_0=1,\dots,n_0}$. Les lemmes ci-dessous permettent ensuite de majorer chacun de ces termes.

**Lemme 7.11.** *L'espérance du terme $T_0^m$ est majorée de la façon suivante :*

$$\mathbb{E}\left[T_0^m\right]\leq\|\Pi_{S_m}g-g\|_B^2+\frac{D_m}{n}$$

Le terme $T_0^m$ est le risque de l'estimateur quand $B$ est connu. Le Lemme 7.11 donne donc la classique décomposition biais-variance du risque dans ce cadre (pour une majoration similaire, se référer par exemple aux calculs (3.6) et (3.8) du Chapitre 3).

**Lemme 7.12.** *Sous les hypothèses de la Proposition 7.2*

$$\mathbb{E}\left[T_1^m\right] \leq 2\left\|\varphi_2'\right\|^2_{L^\infty((0;B(\tau_Z)))}\mathbb{E}\left[\delta_1\right]\left(\frac{D_m^3}{4nn_0}+\frac{\bar{c}_2}{\bar{F}_X^2(\tau_{X_0})}\frac{D_m^3}{n^2}\right).$$

La majoration de ce lemme est par exemple analogue à la majoration du Lemme 3.3 en régression (Chapitre 3), ou à la majoration du Lemme 4.5 du Chapitre 4, ou encore à la majoration du Lemme 7.4 ci-dessus. La preuve est donc omise.

**Lemme 7.13.** *Sous les hypothèses de la Proposition 7.2, il existe une constante $C$ indépendante de $n$ et $n_0$, telle que*

$$\begin{aligned}\mathbb{E}\left[T_2^m\right] &\leq 12\|g\|^2_{L^\infty((0;B(\tau_{X_0})))}\left(\frac{D_m}{n_0}+\bar{c}_2\frac{1}{\bar{F}_X^2(\tau_{X_0})}\frac{D_m}{n}\right)\\ &\quad +C\left\{\frac{D_m^7}{n^3}+\frac{D_m^7}{n_0^3}+\frac{D_m^4}{n^2}+\frac{D_m^4}{n_0^2}\right\}.\end{aligned}$$

De même la majoration de ce lemme est analogue à la majoration du Lemme 3.5 en régression (Chapitre 3), ou à la majoration du Lemme 4.5 du Chapitre 4, ou encore à la majoration du Lemme 7.5 ci-dessus. La preuve est donc également omise.

On obtient donc un majorant de la forme

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{g}_m(.,\hat{B})-g\right\|_B^2\right] &\leq 3\frac{D_m}{n}+18\|g\|^2_{L^\infty((0;B(\tau_{X_0})))}\left(\frac{D_m}{n_0}+\bar{c}_2\frac{1}{\bar{F}_X^2(\tau_{X_0})}\frac{D_m}{n}\right)\\ &\quad +c\left\{\frac{D_m^3}{nn_0}+\frac{D_m^3}{n^2}+\frac{D_m^4}{n^2}+\frac{D_m^4}{n_0^2}+\frac{D_m^7}{n^3}+\frac{D_m^7}{n_0^3}\right\},\\ &\quad +3\left\|\Pi_{S_m}g-g\right\|_B^2,\end{aligned}$$

et on conclut en utilisant les hypothèses $D_m \leq \kappa n^{1/3}$ et $D_m \leq \kappa n_0^{1/3}$.

□

**Preuve de la Proposition 7.3**

On introduit le pseudo-estimateur ci-dessous,

$$\hat{r}_m(.,\hat{B},\Phi)=\hat{g}_m(.,\hat{B})\circ\Phi,$$

et on décompose une nouvelle fois la perte de l'estimateur

$$\left\|\hat{r}_m(.,\hat{B},\hat{\Phi})-r\right\|^2_{\Phi'} \leq 3T_3^m+3T_4^m+3T_0^{m,b},$$

avec

$$\begin{aligned}T_0^{m,b} &= \left\|\hat{r}_m(.,\hat{B},\Phi)-r\right\|^2_{\Phi'},\\ T_3^m &= \left\|\hat{r}_m(.,\hat{B},\hat{\Phi})-\hat{r}_m(.,\hat{B},\Phi)-\mathbb{E}\left[\hat{r}_m(.,\hat{B},\hat{\Phi})-\hat{r}_m(.,\hat{B},\Phi)\,|(X_-,C_-,X_0)\right]\right\|^2_{\Phi'},\\ T_4^m &= \left\|\mathbb{E}\left[\hat{r}_m(.,\hat{B},\hat{\Phi})-\hat{r}_m(.,\hat{B},\Phi)\,|(X_-,C_-,X_0)\right]\right\|^2_{\Phi'}.\end{aligned}$$

Par le changement de variables $u=\Phi(x)$,

$$T_0^{m,b}=\left\|\hat{g}_m(.,\hat{B})-g\right\|^2_B,$$

et donc l'espérance de ce terme est majorée à la section précédente (voir Proposition 7.2). Il reste à contrôler $T_3^m$ et $T_4^m$. C'est l'objet des deux lemmes ci-dessous, dont on omet les démonstrations.

**Lemme 7.14.** *Sous les hypothèses de la Proposition 7.3,*

$$\mathbb{E}\left[T_3^m\right] \leq \left\|\varphi_2'\right\|^2_{L^\infty((0;B(\tau_Z)))} D_m^4 \left(\frac{\bar{c}_2}{\bar{F}_X^2(\tau_{X_0})}\frac{1}{n^2} + C_2\|f_C\|^2_{L^\infty((0;\tau_C))}\tau^2_{X_0}\frac{1}{nn_0}\right).$$

Une majoration pour un terme analogue est énoncée au Lemme 3.4 en régression (Chapitre 3), ou au Lemme 4.5 du Chapitre 4.

**Lemme 7.15.** *Sous les hypothèses de la Proposition 7.3, il existe $C > 0$ telle que*

$$\begin{aligned}\mathbb{E}\left[T_4^m\right] \leq\ & D_m^3\left(\frac{\ln(n)}{n} + \frac{\ln(n_0)}{n_0}\right)\left\{\frac{D_m}{n} + \frac{D_m}{n_0} + \frac{D_m^7}{n^3} + \frac{D_m^7}{n_0^3} + \frac{D_m^4}{n^2} + \frac{D_m^4}{n_0^2}\right\} \\ & + C\left\{\frac{D_m^3}{n^2} + \frac{D_m^3}{n_0^2} + \frac{D_m^2}{n^{3/2}} + \frac{D_m^2}{n_0^{3/2}}\right\}.\end{aligned}$$

Une majoration pour un terme analogue est énoncée au Lemme 3.6 en régression (Chapitre 3).

On obtient donc une inégalité de la forme

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{r}_m(.,\hat{B},\hat{\Phi}) - r\right\|^2_{\Phi'}\right] \\ &\leq 9\frac{D_m}{n} + 36\|g\|^2_{L^\infty((0;B(\tau_{X_0})))}\left(\frac{D_m}{n_0} + \bar{c}_2\frac{1}{\bar{F}_X^2(\tau_{X_0})}\frac{D_m}{n}\right) \\ &\quad + c\left\{\frac{D_m^3}{nn_0} + \frac{D_m^3}{n^2} + \frac{D_m^3}{n_0^2} + \frac{D_m^4}{nn_0} + \frac{D_m^4}{n^2} + \frac{D_m^4}{n_0^2} + \frac{D_m^7}{n^3} + \frac{D_m^7}{n_0^3} + \frac{D_m^2}{n^{3/2}} + \frac{D_m^2}{n_0^{3/2}}\right\} \\ &\quad + D_m^3\left(\frac{\ln(n)}{n} + \frac{\ln(n_0)}{n_0}\right)\left\{\frac{D_m}{n} + \frac{D_m}{n_0} + \frac{D_m^7}{n^3} + \frac{D_m^7}{n_0^3} + \frac{D_m^4}{n^2} + \frac{D_m^4}{n_0^2}\right\} \\ &\quad + 9\left\|\Pi_{S_m}g - g\right\|_B^2,\end{aligned}$$

et la conclusion s'obtient en utilisant les contraintes de dimension de l'énoncé de la Proposition 7.3 : $D_m \leq \kappa n_0^{1/3}/\ln^{1/3}(n_0)$ et $D_m \leq \kappa n^{1/3}/\ln^{1/3}(n)$.

□

# Bibliographie


Adams, R. A. (1975). *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London. Pure and Applied Mathematics, Vol. 65.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281. Akadémiai Kiadó, Budapest.

Akakpo, N. (2009). Estimation adaptative par sélection de partitions en rectangles dyadiques. Ph.D. thesis, Univ. Paris Sud. URL `http://tel.archives-ouvertes.fr/tel-00448753`.

Akakpo, N. & Durot, C. (2010). Histogram selection for possibly censored data. *Math. Methods Statist.* **19**, 189–218.

Akakpo, N. & Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics* **5**, 1618–1653.

Andersen, P. K. & Rønn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometrics* **51**, 323–329.

Antoniadis, A., Grégoire, G. & Vial, P. (1997). Random design wavelet curve smoothing. *Statist. Probab. Lett.* **35**, 225–232.

Audibert, J.-Y. & Catoni, O. (2011a). Robust linear least squares regression. *Ann. Statist.* **39**, 2766–2794.

Audibert, J.-Y. & Catoni, O. (2011b). Robust linear regression through pac-bayesian truncation. URL `http://fr.arxiv.org/pdf/1010.0072v2`. Technical report.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127–146 (electronic).

Baraud, Y., Giraud, C. & Huet, S. (2012). Estimator selection in the gaussian setting. *Ann. Inst. H. Poincaré Probab. Statist.* to appear (available online).

Barron, A., Birgé, L. & Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413.

Bashtannyk, D. M. & Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.* **36**, 279–298.

Baudry, J.-P., Maugis, C. & Michel, B. (2012). Slope heuristics : overview and implementation. *Stat. Comput.* **22**, 455–470.

Bell, C. B. & Doksum, K. A. (1966). "Optimal" one-sample distribution-free tests and their two-sample extensions. *Ann. Math. Statist.* **37**, 120–132.

Bertin, K. & Klutchnikoff, N. (2011). Minimax properties of beta kernel estimators. *J. Statist. Plann. Inference* **141**, 2287–2297.

Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, vol. 36 of *IMS Lecture Notes Monogr. Ser.*, pp. 113–133. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2004). Model selection for Gaussian regression with random design. *Bernoulli* **10**, 1039–1051.

Birgé, L. & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97**, 113–150.

Birgé, L. & Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pp. 55–87. Springer, New York.

Birgé, L. & Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

Birgé, L. & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**, 33–73.

Bitouzé, D., Laurent, B. & Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.* **35**, 735–763.

Bouaziz, O. (2009). Utilisation de modèles à direction révélatrice unique pour les modèles de durée. Ph.D. thesis, UPMC LSTA. URL `http://tel.archives-ouvertes.fr/tel-00569996`.

Briane, M. & Pagès, G. (2006). *Théorie de l'intégration.* Vuibert, Paris.

Brunel, E. & Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā* **67**, 441–475.

Brunel, E. & Comte, F. (2006). Adaptive nonparametric regression estimation in presence of right censoring. *Math. Methods Statist.* **15**, 233–255.

Brunel, E. & Comte, F. (2008). Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory Methods* **37**, 1284–1305.

Brunel, E. & Comte, F. (2009). Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.* **3**, 1–24.

Brunel, E., Comte, F. & Lacour, C. (2007). Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā* **69**, 734–763.

Cai, T. T. & Brown, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26**, 1783–1799.

Cao, R., Janssen, P. & Veraverbeke, N. (2000). Relative density estimation with censored data. *Canad. J. Statist.* **28**, 97–111.

Cao, R., Janssen, P. & Veraverbeke, N. (2001). Relative density estimation and local bandwidth selection for censored data. *Comput. Statist. Data Anal.* **36**, 497–510.

Cao, R., Janssen, P. & Veraverbeke, N. (2005). Relative hazard rate estimation for right censored and left truncated data. *Test* **14**, 257–280.

Carroll, R. J., Ruppert, D. & Welsh, A. H. (1994). Fitting heteroscedastic regression models. *J. Amer. Statist. Assoc.* **89**, 100–116.

Chagny, G. (2011). Régression : bases déformées et sélection de modèles par pénalisation et méthode de lepski. URL `http://hal.archives-ouvertes.fr/hal-00519556`. Technical report.

Cheng, K. F. & Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10**, 583–604.

Chesneau, C. (2007). A maxiset approach of a Gaussian noise model. *TEST* **16**, 523–546.

Chesneau, C. & Willer, T. (2012). Estimation of a cumulative distribution function under interval censoring, case 1, via warped wavelets. URL `http://hal.archives-ouvertes.fr/hal-00715260`. Submitted.

Chichignoud, M. (2010). Estimation adaptative par sélection de partitions en rectangles dyadiques. Ph.D. thesis, Univ. de Provence. URL `http://tel.archives-ouvertes.fr/tel-00540963`.

Chichignoud, M. (2012). Minimax and minimax adaptive estimation in multiplicative regression : locally Bayesian approach. *Probab. Theory Related Fields* **153**, 543–586.

Cohen, S. & Le Pennec, E. (2012). Partition-based conditional density estimation. *ESAIM Probab. Statist.* to appear (available online).

Comte, F. (2012). Estimation non-paramétrique. Polycopié de cours de Master 2.

Comte, F. & Johannes, J. (2012). Adaptive functional linear regression. *Ann. Statist.* **40**, 2765–2797.

Comte, F. & Lacour, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. H. Poincaré Probab. Statist.* **49**, 569–609.

Comte, F. & Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.* **56**, 449–473.

Ćwik, J. & Mielniczuk, J. (1993). Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Probab. Lett.* **16**, 397–405.

De Gooijer, J. G. & Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica* **57**, 159–176.

DeVore, R. A. & Lorentz, G. G. (1993). *Constructive approximation*, vol. 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369. With discussion and a reply by the authors.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539.

Doob, J. L. (1949). Heuristic approach to the kolmogorov-smirnov theorems. *The Annals of Mathematical Statistics* **20**, 393–403.

Doumic, M., Hoffmann, M., Reynaud-Bouret, P. & Rivoirard, V. (2012). Nonparametric estimation of the division rate of a size-structured population. *SIAM J. Numer. Anal.* **50**, 925–950.

Durot, C. (2008). Monotone nonparametric regression with random design. *Math. Methods Statist.* **17**, 327–341.

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27**, 642–669.

Efromovich, S. (1999). *Nonparametric curve estimation.* Springer Series in Statistics. Springer-Verlag, New York. Methods, theory, and applications.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Ann. Statist.* **35**, 2504–2535.

Efromovich, S. (2010a). Dimension reduction and adaptation in conditional density estimation. *J. Amer. Statist. Assoc.* **105**, 761–774.

Efromovich, S. (2010b). Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.* **62**, 249–275.

Fan, J. & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting : variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371–394.

Fan, J., Tong, H. & Yao, Q. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Fan, J. & Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91**, 819–834.

Fan, J.-q., Peng, L., Yao, Q.-w. & Zhang, W.-y. (2009). Approximating conditional density functions using dimension reduction. *Acta Math. Appl. Sin. Engl. Ser.* **25**, 445–456.

Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.* **100**, 2083–2099.

Ferraty, F., Laksaci, A., Tadj, A. & Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference* **140**, 335–352.

Ferraty, F., Laksaci, A. & Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes. An International Journal Devoted to Time Series Analysis and the Statistics of Continuous Time Processes and Dynamical Systems* **9**, 47–76.

Fourdrinier, D. & Pergamenshchikov, S. (2007). Improved model selection method for a regression function with dependent noise. *Ann. Inst. Statist. Math.* **59**, 435–464.

Gaïffas, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.* **11**, 344–364 (electronic).

Gastwirth, J. L. (1968). The first-median test : A two-sided version of the control median test. *J. Amer. Statist. Assoc.* **63**, 692–706.

Goldenshluger, A. & Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14**, 1150–1190.

Goldenshluger, A. & Lepski, O. (2009). Structural adaptation via $L_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143**, 41–71.

Goldenshluger, A. & Lepski, O. (2011a). Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39**, 1608–1632.

Goldenshluger, A. & Lepski, O. (2011b). Uniform bounds for norms of sums of independent random functions. *Ann. Probab.* **39**, 2318–2384.

Goldenshluger, A. & Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6**, 135–170.

Golubev, G. K. & Nussbaum, M. (1992). Adaptive spline estimates in a nonparametric regression model. *Teor. Veroyatnost. i Primenen.* **37**, 554–561.

Hall, P., Wolff, R. C. & Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**, 154–163.

Hall, P. G. & Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statist. Probab. Lett.* **64**, 181–189.

Handcock, M. S. & Janssen, P. L. (2002). Statistical inference for the relative density. *Sociol. Methods Res.* **30**, 394–424.

Handcock, M. S. & Morris, M. (1999). *Relative distribution methods in the social sciences.* Statistics for Social Science and Public Policy. Springer-Verlag, New York.

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*, vol. 129 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

Härdle, W. & Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81**, 223–242. ISSN 0304-4076.

Hirsch, F. & Lacombe, G. (1997). *Eléments d'analyse fonctionnelle. Cours et exercices.* Enseignement des Mathématiques. Masson, Paris.

Hochmuth, R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.* **12**, 179–208.

Hsieh, F. (1995). The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *J. Roy. Statist. Soc. Ser. B* **57**, 735–748.

Hsieh, F. & Turnbull, B. W. (1996a). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.* **24**, 25–40.

Hsieh, F. & Turnbull, B. W. (1996b). Nonparametric methods for evaluating diagnostic tests. *Statist. Sinica* **6**, 47–62.

Hyndman, R. J., Bashtannyk, D. M. & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.* **5**, 315–336.

Hyndman, R. J. & Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.* **14**, 259–278.

Ibragimov, R. & Sharakhmetov, S. (2001). The exact constant in the Rosenthal inequality for random variables with mean zero. *Teor. Veroyatnost. i Primenen.* **46**, 134–138.

Jewell, N. P. & van der Laan, M. (2004). Current status data : review, recent developments and open problems. In *Advances in survival analysis*, vol. 23 of *Handbook of Statist.*, pp. 625–642. Elsevier, Amsterdam.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Karunamuni, R. J. & Alberts, T. (2005). On boundary correction in kernel density estimation. *Stat. Methodol.* **2**, 191–212.

Kerkyacharian, G., Lepski, O. & Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121**, 137–170.

Kerkyacharian, G. & Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10**, 1053–1105.

Klein, T. & Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, 1060–1077.

Kohler, M. & Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Trans. Inform. Theory* **47**, 3054–3058.

Koo, J.-Y. & Lee, K.-W. (1998). $B$-spline estimation of regression functions with errors in variable. *Statist. Probab. Lett.* **40**, 57–66.

Koul, H., Susarla, V. & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9**, 1276–1288.

Kulik, R. & Raimondo, M. (2009). Wavelet regression in random design with heteroscedastic dependent errors. *Ann. Statist.* **37**, 3396–3430.

Lacour, C. (2007). Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.* **43**, 571–597.

Lacour, C. (2008). Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* **99**, 787–814.

Lepski, O. V., Mammen, E. & Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness : an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25**, 929–947.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* **36**, 645–659.

Lepskiĭ, O. V. (1992a). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.* **37**, 468–481.

Lepskiĭ, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, vol. 12 of *Adv. Soviet Math.*, pp. 87–106. Amer. Math. Soc., Providence, RI.

Lloyd, C. J. (1998). Using smoothed ROC curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.* **93**, 1356–228.

Lloyd, C. J. & Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statist. Probab. Lett.* **44**, 221–228.

Lo, S. H., Mack, Y. P. & Wang, J. L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Related Fields* **80**, 461–473.

Ma, S. & Kosorok, M. R. (2006). Adaptive penalized $M$-estimation with current status data. *Ann. Inst. Statist. Math.* **58**, 511–526.

Mallows, C. L. (1973). Comments on $C_p$. *Technometrics* **15**, 661–675.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269–1283.

Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Massart, P. & Meynet, C. (2011). The Lasso as an $\ell_1$-ball model selection procedure. *Electron. J. Stat.* **5**, 669–687.

Mehra, K. L., Ramakrishnaiah, Y. S. & Sashikala, P. (2000). Laws of iterated logarithm and related asymptotics for estimators of conditional density and mode. *Ann. Inst. Statist. Math.* **52**, 630–645.

Molanes-López, E. M. & Cao, R. (2008a). Plug-in bandwidth selector for the kernel relative density estimator. *Ann. Inst. Statist. Math.* **60**, 273–300.

Molanes-López, E. M. & Cao, R. (2008b). Relative density estimation for left truncated and right censored data. *J. Nonparametr. Stat.* **20**, 693–720.

Müller, H.-G. & Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61–76.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Application* **9**, 141–142.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, vol. 1738 of *Lecture Notes in Math.*, pp. 85–277. Springer, Berlin.

Nikol′skiĭ, S. M. (1975). *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Patil, P. N. (1993). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inference* **35**, 15–30.

Penskaya, M. (1995). Mean square consistent estimation of a ratio. *Scand. J. Statist.* **22**, 129–137.

Pham Ngoc, T. M. (2009). Regression in random design and Bayesian warped wavelets estimators. *Electron. J. Stat.* **3**, 1084–1112.

Pham Ngoc, T. M. & Rivoirard, V. (2013). The dictionary approach for spherical deconvolution. *J. Multivariate Anal.* **115**, 138–156.

Plancade, S. (2013). Estimation of conditional cumulative distribution function from current status data. *J. Statist. Plann. Inference* to appear, available online .

Polyak, B. T. & Tsybakov, A. B. (1990). Optimal orders of accuracy for search algorithms of stochastic optimization. *Problemy Peredachi Informatsii* **26**, 45–53.

Pons, O. (2007). Estimation of absolutely continuous distributions for censored variables in two-sample nonparametric and semi-parametric models. *Bernoulli* **13**, 92–113.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli* **12**, 633–661.

Reynaud-Bouret, P. & Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *Electron. J. Stat.* **4**, 172–238.

Rigollet, P. & Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16**, 260–280.

Rivoirard, V. & Stoltz, G. (2012). *Statistique mathématique en action.* Vuibert, Paris.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–837.

Rosenthal, H. P. (1970). On the subspaces of $L^p$ $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.* **8**, 273–303.

Satten, G. A. & Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist.* **55**, 207–210.

Schmisser, E. (2012). Non-parametric estimation of the diffusion coefficient from noisy data. *Stat. Inference Stoch. Process.* **15**, 193–223.

Servien, R. (2009). Estimation de la fonction de répartition : revue bibliographique. *J. SFdS* **150**, 84–104.

Shorack, G. R. & Wellner, J. A. (1986). *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *J. Roy. Statist. Soc. Ser. B* **27**, 26–33.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.* **12**, 917–926.

Stute, W. (1986a). Conditional empirical processes. *Ann. Statist.* **14**, 638–647.

Stute, W. (1986b). On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.* **14**, 891–901.

Takeuchi, I., Nomura, K. & Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput.* **21**, 533–559.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505–563.

Tanner, M. A. & Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989–993.

Triebel, H. (2006). *Theory of function spaces. III*, vol. 100 of *Monographs in Mathematics.* Birkhäuser Verlag, Basel.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14–44.

van der Vaart, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge.

Wang, Q., Yao, L. & Lai, P. (2009). Estimation of the area under ROC curve with censored data. *J. Statist. Plann. Inference* **139**, 1033–1044.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26**, 359–372.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.* **31**, 252–273.

Wellner, J. A. (2007). On an exponential bound for the Kaplan-Meier estimator. *Lifetime Data Anal.* **13**, 481–496.

Yang, S.-S. (1981). Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.* **76**, 658–662.

Youndjé, E., Sarda, P. & Vieu, P. (1994). Validation croisée pour l'estimation non-paramétrique de la densité conditionnelle. *Publ. Inst. Statist. Univ. Paris* **38**, 57–80.

Zhang, Y., Liu, W. & Wu, H. (2003). A simple nonparametric two-sample test for the distribution function of event time with interval censored data. *J. Nonparametr. Stat.* **15**, 643–652.

Zhou, Y. & Liang, H. (2005). Empirical-likelihood-based semiparametric inference for the treatment effect in the two-sample problem with censoring. *Biometrika* **92**, 271–282.

## Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie.

**Résumé :** Cette thèse présente divers problèmes d'estimation fonctionnelle adaptative par sélection d'estimateurs en projection ou à noyaux, utilisant des critères inspirés à la fois de la sélection de modèles et des méthodes de Lepski. Le point commun de nos travaux est l'utilisation de données transformées et/ou incomplètes. La première partie est consacrée à une procédure d'estimation par "déformation", dont la pertinence est illustrée pour l'estimation des fonctions suivantes : régression additive et multiplicative, densité conditionnelle, fonction de répartition dans un modèle de censure par intervalle, risque instantané pour des données censurées à droite. Le but est de reconstruire une fonction à partir d'un échantillon de couples aléatoires $(X, Y)$. Nous utilisons les données déformées $(\Phi(X), Y)$ pour proposer des estimateurs adaptatifs, où $\Phi$ est une fonction bijective que nous estimons également (par exemple la fonction de répartition de $X$). L'intérêt est double : d'un point de vue théorique, les estimateurs ont des propriétés d'optimalité au sens de l'oracle ; d'un point de vue pratique, ils sont explicites et numériquement stables. La seconde partie s'intéresse à un problème à deux échantillons : nous comparons les distributions de deux variables $X$ et $X_0$ au travers de la densité relative, définie comme la densité de la variable $F_0(X)$ ($F_0$ étant la répartition de $X_0$). Nous construisons des estimateurs adaptatifs, à partir d'un double échantillon de données, possiblement censurées. Des bornes de risque non-asymptotiques sont démontrées, et des vitesses de convergences déduites.

**Mots-Clés :** estimation adaptative, sélection de modèles, méthode de Lepski, bases et noyaux déformés, régression, données censurées, problème à deux échantillons.

## Adaptive estimation with warped or incomplete data. Application to survival analysis.

**Abstract:** This thesis presents various problems of adaptive functional estimation, using projection and kernel methods, and criterions inspired both by model selection and Lepski's methods. The common point of the studied statistical setting is to deal with transformed and/or incomplete data. The first part proposes a method of estimation with a "warping" device which permits to handle the estimation of functions such as additive and multiplicative regression, conditional density, hazard rate based on randomly right-censored data, and cumulative distribution function from current-status data. The aim is to estimate a function from a sample of random variable $(X, Y)$. We use the warped data $(\Phi(X), Y)$, to propose adaptive estimators, where $\Phi$ is a one-to-one function that we also estimate (e.g. the cumulative distribution function of $X$). The interest is twofold. From the theoretical point of view, the estimators are optimal in the oracle sense. From the practical point of view, they can be easily computed, thanks to their simple explicit expression. The second part deals with a two-sample problem: we compare the distribution of two variables $X$ and $X_0$ by studying the relative density, defined as the density of $F_0(X)$ ($F_0$ is the c.d.f. of $X_0$). We build adaptive estimators, from a double data-sample, possibly censored. Non-asymptotic risk bounds are proved, and convergence rates are also derived.

**Keywords:** adaptive estimation, model selection, Lepski's method, warped bases and kernels, regression, censored data, two-sample problem.