New insights on inverse problems : multidimensional strategies for deconvolution or regression, and ruin probability estimation

Florian Dussap



Université Paris Cité

École doctorale Sciences Mathématiques de Paris Centre

Laboratoire MAP5, CNRS UMR 8145

New Insights on Inverse Problems:
Multidimensional Strategies for Deconvolution or Regression, and Ruin Probability Estimation

Par Florian DUSSAP

Thèse de doctorat de mathématiques appliquées

Dirigée par Fabienne COMTE et Céline DUVAL

Présentée et soutenue publiquement le 24 juin 2022

Devant un jury composé de :

Fabienne COMTE	Professeure	Université de Paris	Directrice
Céline DUVAL	Professeure	Université de Lille	Directrice
Christophe GIRAUD	Professeur	Université Paris Saclay	Examinateur
Béatrice LAURENT-BONNEAU	Professeure	INSA Toulouse	Examinatrice
Oleg LEPSKI	Professeur	Aix-Marseille Université	Examinateur
Florence MERLEVÈDE	Professeure	Université Gustave Eiffel	Examinatrice
Vincent RIVOIRARD	Professeur	Université Paris Dauphine	Rapporteur
Jérôme SARACCO	Professeur	Université de Bordeaux	Rapporteur

Résumé

Dans cette thèse, on s'intéresse à plusieurs problèmes inverses de statistique non paramétrique. Nous étudions l'estimation de fonctions à plusieurs variables sur des domaines non compacts : $\mathbb{R}^d$ et $\mathbb{R}_+^d$. Nous utilisons pour cela des estimateurs par projection sur des bases orthonormées obtenues en tensorisant la base d'Hermite (cas de $\mathbb{R}^d$) et la base de Laguerre (cas de $\mathbb{R}_+^d$). Ces bases sont construites à partir de polynômes orthogonaux et ont la particularité d'être à support non compact. Cela évite la question du choix du support qui se pose avec les bases dont le support est un intervalle $[a,b]$ par exemple. Pour garantir que nos estimateurs aient de bonnes performances, la dimension de l'espace de projection nécessite d'être choisie. Nous utilisons deux procédures : la sélection de modèle par pénalisation et la méthode de Goldenshluger et Lepski. Ces procédures nous permettent de construire des estimateurs adaptatifs relativement aux espaces de régularité associés aux bases utilisées : les espaces de Sobolev–Laguerre et les espaces de Sobolev–Hermite.

Chapitre 2 On étudie le problème de l'estimation d'une densité dans le modèle additif $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$, où $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont des vecteurs aléatoires d-dimensionnels à coordonnées positives. Notre but est de retrouver la densité de $\boldsymbol{X}$ à partir d'observations indépendantes de $\boldsymbol{Z}$, en supposant que la loi de $\boldsymbol{Y}$ est connue. La densité de $\boldsymbol{Z}$ s'exprime alors comme le produit de convolution des densités de $\boldsymbol{X}$ et de $\boldsymbol{Y}$, et il faut résoudre un problème de déconvolution. Dans le cas $d = 1$, un estimateur par projection sur la base de Laguerre a déjà été étudié par Mabon (2017). On généralise cette méthode au cas multivarié : on détermine des majorations non asymptotiques sur le MISE de l'estimateur et on en déduit des vitesses de convergence sur des espaces fonctionnels anisotropes. Pour finir, on propose une procédure de sélection de modèles, inspirée des travaux de Goldenshluger & Lepski (2011) et Chagny (2013c), pour le choix du modèle qui réalise le compromis biais-variance.

Chapitre 3 On considère le modèle de Cramér–Lundbgerg, dans lequel le processus de réserve d'une compagnie d'assurance est donné par :

$$U_t = u + ct - \sum_{i=1}^{N_t} X_i,$$

où $u \geq 0$ est la réserve initiale, $c > 0$ est le taux de cotisation, $(N_t)_{t\geq 0}$ est un processus de Poisson homogène qui compte le nombre de sinistres et $(X_i)_{i\geq 1}$ sont des variables i.i.d. qui représentent les montants des sinistres. On s'intéresse au problème d'estimation de la fonction de Gerber–Shiu à partir de l'observation d'une trajectoire du processus $(U_t)_{t\in[0,T]}$. Cette fonction a été introduite par Gerber & Shiu (1998) pour étudier simultanément l'instant de ruine, le montant de la réserve avant la ruine et le déficit au moment de la ruine. Le résultat principal de Gerber & Shiu (1998) sur cette fonction est qu'elle vérifie une équation de la forme :

$$\phi = \phi * g + h,$$

avec g et h des fonctions liées à la loi du processus de Poisson composé $\sum_{i=1}^{N_t} X_i$. Du fait de l'équation de convolution satisfaite par ϕ, Zhang & Su (2018) propose un estimateur par projection sur la base de Laguerre. Cependant, ils ne donnent pas de résultat sur le MISE, mais seulement sur le comportement en O_P de l'erreur quadratique intégrée lorsque $T \to \infty$. Nous nous plaçons dans la continuité de leur travail en étudiant le MISE de l'estimateur par projection sur la base de Laguerre. On obtient ainsi des vitesses de convergence sur des boules de Sobolev–Laguerre qui dépendent de la régularité de la fonction estimé, comme c'est typiquement le cas en statistique non paramétrique. De plus, on propose une nouvelle méthode d'estimation de la fonction de Gerber–Shiu par une approche hybride Laguerre–Fourier : on utilise un estimateur par projection sur la base de Laguerre où les coefficients sont calculés grâce au théorème de Plancherel. On montre que le terme de variance de cet estimateur est majoré indépendamment de la dimension de l'espace de projection. La vitesse de convergence du MISE est alors $1/T$ (vitesse paramétrique) et ne nécessite pas de compromis biais-variance.

Chapitre 4 On étudie le problème d'estimation non paramétrique d'une fonction de régression avec un design aléatoire sur $\mathbb{R}^p$ pour $p \geq 2$. On utilise pour cela un estimateur par projection calculé avec un critère de moindres carrés. Notre contribution est de considérer des domaines d'estimation non compacts et d'étudier théoriquement le risque de l'estimateur pondéré par la loi du design. On propose une procédure de sélection de modèle dans laquelle la collection de modèles est aléatoire et prend en compte l'écart entre la norme empirique et la norme associée à la loi du design. On démontre que l'estimateur résultant optimise automatiquement le compromis biais-variance pour les deux normes.

Mots clés estimation non paramétrique, sélection de modèle, base de Laguerre, base d'Hermite, déconvolution, régression, théorie de la ruine.

Abstract

In this thesis, we are interested in several inverse problems in nonparametric statistics. We study the estimation of functions of several variables on non compact domains: $\mathbb{R}^d$ and $\mathbb{R}_+^d$. We use projection estimators on orthonormal bases constructed by tensorization of the Hermite basis (in the case of $\mathbb{R}^d$) and the Laguerre basis (in the case of $\mathbb{R}_+^d$). These two bases are formed from orthogonal polynomials and they have the property of being non-compactly supported. This avoids issues regarding the choice of the support, that one encounters with bases supported on an interval $[a,b]$ for example. In order to ensure that our estimators perform well, the dimension of the projection space must be carefully chosen. We use two procedures: the penalized model selection procedure and the Goldenshluger and Lepski's procedure. These procedures allow us to construct adaptive estimators on regularity spaces associated to the used bases: the Sobolev–Laguerre spaces and the Sobolev–Hermite spaces.

Chapitre 2 We study the density estimation problem in the additive model $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-dimensional random vectors with non-negative coordinate. Our goal is to recover the density of $\boldsymbol{X}$ from i.i.d. observations of $\boldsymbol{Z}$, under the assumption that the distribution of $\boldsymbol{Y}$ is known. The density of $\boldsymbol{Z}$ is given by the convolution product of the densities of $\boldsymbol{X}$ and $\boldsymbol{Y}$, so we have to solve a deconvolution problem. In the $d = 1$ case, a projection estimator on the Laguerre basis has already been studied by Mabon (2017). We extend this method to the multivariate case: we establish non-asymptotic bounds on the MISE of the estimator, and we provide convergence rates on anisotropic functional spaces. Finally, we propose selection model procedure inspired by the work of Goldenshluger & Lepski (2011) and Chagny (2013c) to choose the model that optimizes the bias-variance trade-off.

Chapitre 3 We consider the Cramér–Lundberg model, in which the reserve process of an insurance company is given by:

$$U_t = u + ct - \sum_{i=1}^{N_t} X_i,$$

where $u \geq 0$ is the initial reserve, $c > 0$ is the premium rate, $(N_t)_{t\geq 0}$ is an homogeneous Poisson process that counts the claims number, and $(X_i)_{i\geq 1}$ are i.i.d.

random variables that represents the claims sizes. We are interested in the problem of estimating the Gerber–Shiu function from the observation of a trajectory of the process $(U_t)_{t\in[0,T]}$. This function has been introduced by Gerber & Shiu (1998) to study simultaneously the ruin time, the reserve before the ruin, and the ruin deficit. The main result of Gerber & Shiu (1998) is that this function satisfies an equation of the form:

$$\phi = \phi * g + h,$$

where g and h are functions that depend on the distribution of the compound Poisson process $\sum_{i=1}^{N_t} X_i$. Using this equation on ϕ, Zhang & Su (2018) propose a projection estimator on the Laguerre basis. However, their results only concerns the asymptotic behavior of the estimator ISE, as $T \to \infty$. We carry on their work with a non-asymptotic study of the MISE of the projection estimator. We obtain convergence rates on Sobolev–Laguerre balls that depend on the regularity of ϕ, as it is typically the case in nonparametric statistics. Moreover, we propose a new estimation method using a hybrid approach Laguerre–Fourier: we use a projection estimator on the Laguerre basis where the coefficients are computed using Plancherel isometry. We show that the variance term of this estimator is bounded independently of the dimension of the projection space, so that the convergence rate is $1/T$ (parametric rate) without the need of optimizing the bias-variance trade-off.

Chapter 4 We study the nonparametric regression estimation problem with a random design in $\mathbb{R}^p$ with $p \geq 2$. We do so by using a projection estimator obtained by least squares minimization. Our contribution is to consider estimation domains in $\mathbb{R}^p$ that are non-compact, and to provide a theoretical study of the risk of the estimator relative to a norm weighted by the distribution of the design. We propose a model selection procedure in which the model collection is random and takes into account the discrepancy between the empirical norm and the norm associated with the distribution of design. We prove that the resulting estimator automatically optimizes the bias-variance trade-off in both norms.

Keywords nonparametric estimation, model selection, Laguerre basis, Hermite basis, deconvolution, regression, ruin theory.

Remerciements

Je commence par adresser mes premiers et plus gros remerciements à mes directrices de thèse. Fabienne, Céline, merci beaucoup pour votre implication et votre encadrement sans failles. Vous avez toujours répondu présent lorsque j'en avais besoin. Merci de m'avoir guidé dans les arcanes de la recherche, j'ai beaucoup appris à vos côtés !

Florence, je te remercie pour l'intérêt que tu as porté à mes travaux pendant ces trois années. Merci beaucoup pour ton aide avec certaines inégalités de concentration matricielles, le dernier chapitre de cette thèse en a largement bénéficié. Merci d'avoir accepté de faire partie du jury.

Merci beaucoup à Vincent Rivoirard et Jérôme Saracco d'avoir consacré leur temps et leur énergie à rapporter ce manuscrit. Je vous remercie pour votre réception si positive de mon travail. Je remercie Béatrice Laurent, Oleg Lepski et Christophe Giraud d'avoir accepté d'être membres du jury. Christophe, merci de m'avoir incité à entreprendre une thèse en statistique.

Je remercie bien sûr tous les membres du laboratoire, ainsi que le personnel administratif, pour leur accueil chaleureux et bienveillant.

Marie-Hélène, merci pour ta gentillesse, ta bonne humeur et tout le travail que tu accomplis. Anne, en tant que directrice du laboratoire, merci d'être aussi attentive à nos besoins et nos conditions de travail. Rémy, merci de t'être montré disponible pour répondre à mes diverses questions informatiques.

Je veux remercier Manon Defosseux, Antoine Marchina et Jérôme Dedecker pour les échanges constructifs que j'ai eu avec eux. Jérôme, merci d'avoir fait partie de mon comité de mi-parcours.

Je remercie Georges Koepfler, Thierry Cabanal-Duvillard et Rachid Lounes de m'avoir fait confiance pour m'occuper de leurs groupes de TD. Je tiens également à remercier Christine Graffigne pour son aide avec l'enseignement à distance lors du premier confinement.

Un grand merci à tous les « éphémères » du laboratoire. Je vous remercie pour les discussions, les parties de cartes, les sorties, les verres au *Frog*, et tous les bons moments partagés qui ont été essentiels pour moi.

Aux « anciens » et aux « anciennes » : Alessandro, Alexandre, Allan, Andrea, Arthur, Claire, Fabien, Juliana, Marta, Ousmane, Vincent, Vivien et Warith ; je vous remercie de m'avoir si bien accueilli au MAP5.

À mes camarades du bureau 750 : Adrien, Antoine M., Apolline, Chabane, Charlie, Safa et Yen ; je vous remercie pour la bonne ambiance au bureau, ç'a été un plaisir de travailler à vos côtés.

À celles et ceux que je n'ai pas encore cités : Antoine S., Anton, Ariane, Carlos, Cécile, Diala, Herb, Keanu, Laurent, Loïc, Marie, Mariem, Mehdi, Pierre-Louis, Rémi B., Rémi L., Sergio, Sinda, Sonia, Thaïs et Zoé ; merci de faire du laboratoire un endroit si agréable, vous êtes les meilleur·es !

Je remercie Alexandra Elbakyan pour son combat pour une science accessible à tous et à toutes ; son travail a grandement facilité cette thèse.

Pour finir, je remercie mes ami·es et mes proches pour tout le soutient qu'ils m'ont apporté pendant ces trois années.

Sommaire

Liste des symboles

Les quantités vectorielles, matricielles ou hypermatricielles sont notées en caractères gras.

Abréviations

p.s./a.s.	presque sûrement/almost surely
i.i.d.	indépendantes et identiquement distribuées ; independent and identically distributed
ISE	Integrated Squared Error
MISE	Mean Integrated Squared Error

Ensembles de nombres

$\mathbb{N}$	Ensemble des entiers $\{0,1,2,\ldots\}$
$\mathbb{N}_+$	Ensemble des entiers strictement positifs
$\mathbb{R}_+$	Demi-droite des nombres réels positifs ou nuls
$\widehat{\mathbb{C}}$	Ensemble des nombres complexes complété d'un point à l'infini
$\mathbb{T}$	Cercle unité ; nombres complexes de module 1
$\mathbb{D}$	Disque unité ouvert ; nombres complexes de module < 1
$\overline{\mathbb{D}}$	Disque unité fermé ; nombres complexes de module ≤ 1
$\mathscr{P}_+$	Demi-plan des nombres complexes de partie réelle positive ou nulle

Matrices et Hypermatrices

$\|\cdot\|_{\mathbb{R}^m}$	Norme euclidienne sur $\mathbb{R}^m$
$\|\mathbf{T}\|_F$	Norme de Frobenius de l'application linéaire $\mathbf{T}$
$\|\mathbf{T}\|_{\mathrm{op}}$	Norme d'opérateur de $\mathbf{T}$: $\sup_{\mathbf{v}\neq\mathbf{0}} \frac{\|\mathbf{Tv}\|}{\|\mathbf{v}\|}$
$\mathbf{0}$	Vecteur nulle ; matrice nulle ; hypermatrice nulle

$\mathbf{1}$	Vecteur $(1,\dots,1)$; longueur variable selon le contexte
$\mathbb{R}^{\boldsymbol{m}}$	Ensemble des hypermatrices de forme $m_1 \times \cdots \times m_d$
$\mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$	Ensemble des hypermatrices de forme $(m_1 \times \cdots \times m_d) \times (m_1 \times \cdots \times m_d)$
$\mathbf{I}_{\boldsymbol{m}}$	Hypermatrice identité $\boldsymbol{m} \times \boldsymbol{m}$
$\mathbf{T}(\alpha)$	Opérateur Toeplitz de symbole α
$\lambda_{\max}$, $\lambda_{\min}$	Valeur propre maximale et minimale

Fonctions, Espaces fonctionnels

$\|\cdot\|_\mu$, $\langle\cdot,\cdot\rangle_\mu$	Norme et produit scalaire intégral pondéré par la mesure μ
$\|\cdot\|_n$, $\langle\cdot,\cdot\rangle_n$	Norme et produit scalaire empirique
$\mathscr{F}u$, u^*	Transformée de Fourier de u
$\ell^p(I)$	Suites $(a_i)_{i\in I}$ telles que $\sum_{i\in I}\|a_i\|^p < +\infty$
$\mathrm{L}^p(A)$	Espace des fonctions de puissance p-ème intégrable sur A relativement à la mesure de Lebesgue.
$\mathrm{L}^p(A,\mu)$	Espace des fonctions de puissance p-ème intégrable sur A relativement à la mesure μ.
$\mathscr{L}u$	Transformée de Laplace de u
$u * v$	Produit de convolution de u et v
$\mathrm{W}^s(\mathbb{R}_+)$	Espace de Sobolev–Laguerre de régularité s
$\mathrm{W}^{\mathbf{s}}(\mathbb{R}_+^d)$	Espace de Sobolev–Laguerre multidimensionnel de régularité $\mathbf{s}$

Autres symboles

$\lceil a \rceil$	Partie entière supérieure de a
$\lfloor a \rfloor$	Partie entière inférieure de a
$\mu \ll \nu$	μ est absolument continue par rapport à ν
$A := B$	A est égal à B par définition ; A est défini par B
$a \vee b$	Maximum de a et b
$a \wedge b$	Minimum de a et b
a_+	Partie positive de a ; $\max(a,0)$
$C, C', \dots$	Constantes numériques positives
$C(\alpha,\beta,\dots)$	Constante positive qui dépend de α, β, …

Chapitre 1

Introduction

L'objectif de cette thèse est d'étudier quelques problèmes d'estimation fonctionnelle. Dans tous ces problèmes, on cherche à reconstruire une fonction à partir d'un nombre fini de réalisations de variables aléatoires dont la loi dépend de la fonction à estimer. Une particularité de ce travail est de considérer des fonctions définies sur des domaines non compact de mesure infinie, à savoir $\mathbb{R}^d$ ou $\mathbb{R}_+^d$. Du plus, à l'exception du chapitre 3, nous nous intéressons à des fonctions de plusieurs variables (i.e. $d \geq 2$).

Ce chapitre a pour but d'introduire le cadre statistique et les concepts utilisés tout au long de la thèse. Nous présentons la méthode d'estimation par projection ainsi que le problème de construction d'estimateurs adaptatifs au sens de l'oracle. Ces différentes notions sont illustrées sur l'exemple de l'estimation d'une densité à partir d'observations directes. Nous introduisons également les deux principaux outils techniques qui seront utilisés au cours de la thèse : les hypermatrices et les inégalités de concentration. Enfin, on conclut ce chapitre par une présentation des problèmes étudiés et des résultats obtenus.

Cette introduction s'appuie sur les livres de Comte (2017), Tsybakov (2009), et Massart (2007).

1.1 Généralités sur l'estimation non paramétrique

1.1.1 Cadre statistique

Ce travail se place dans le cadre de l'estimation non paramétrique ou estimation fonctionnelle. Au contraire de l'approche paramétrique, qui suppose que la loi des données peut être décrite par un vecteur fini-dimensionnel $\boldsymbol{\theta} \in \mathbb{R}^p$ sur lequel porte l'inférence, cette approche ne fait pas (ou peu) d'hypothèses sur la loi des observations. Les quantités d'intérêt sont typiquement des fonctions (de densité, de répartition, de régression, etc) qui sont supposées appartenir à des espaces fonctionnels de dimension infinie. Ceux-ci sont en général des espaces de fonctions avec une certaine régularité, par exemple l'espace des fonctions lipschitziennes, des espaces de Sobolev, ou des espaces de Besov pour en citer

quelques-uns. L'intérêt de l'approche non paramétrique est d'avoir une grande flexibilité dans la modélisation des données : on ne fait pas d'hypothèse à priori sur la forme de leur loi.

Nous adoptons également une approche *non asymptotique* : on construira et on étudiera des procédures d'estimation avec un nombre d'observations n fixé, qui ne tend pas vers l'infini. Cela n'empêche pas de se poser la question de la vitesse de convergence des estimateurs lorsque n augmente, mais cette étude passera par l'utilisation d'inégalités de concentration plutôt que par l'utilisation de théorèmes limites (théorème central limite, grandes déviations, etc) donnant une approximation du comportement des estimateurs lorsque $n \to \infty$.

Concrètement, on observe des variables aléatoires $\xi_1, \dots, \xi_n$ i.i.d. et on souhaite estimer une fonction f qui dépend de leur loi. La fonction f est supposée à valeurs réelles et définie sur un domaine A de $\mathbb{R}^d$. Une spécificité de cette thèse est de considérer des fonctions de plusieurs variables, sur des domaines qui ne sont pas nécessairement compacts. Un estimateur $\widehat{f}_n$ de f est défini comme une fonction mesurable des données :

$$\boldsymbol{x} \in A \longmapsto \widehat{f}_n(\boldsymbol{x}) = \widehat{f}_n(\boldsymbol{x}; \xi_1, \dots, \xi_n) \in \mathbb{R}.$$

La dépendance en les données, ou la taille de l'échantillon, ne sera en général pas notée explicitement dans la suite.

Exemple 1.1 (Estimation d'une densité)**.** On observe des variables i.i.d. $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ dans A, de densité commune f inconnue. On supposera dans la suite que f appartient à $\mathrm{L}^2(A)$. Cette hypothèse n'est pas nécessaire pour l'estimation d'une densité en général, mais elle l'est dans notre cas du fait de la méthode d'estimation utilisée.

1.1.2 Risque d'un estimateur

Pour comparer différents estimateurs et évaluer leur performance, on introduit une quantité appelée *risque* de l'estimateur. Pour ℓ une fonctionnelle positive, appelée *perte* ou *coût*, qui mesure « l'écart » entre deux fonctions, le risque d'un estimateur $\widehat{f}_n$ est défini par :

$$\mathscr{R}(f, \widehat{f}_n) \coloneqq \mathbb{E}\big[\ell(f, \widehat{f}_n)\big].$$

Les exemples classiques incluent bien sûr le cas où ℓ est une distance, ou une fonction croissante d'une distance, comme par exemple la perte L^p :

$$\ell(f, g) = \|f - g\|_{\mathrm{L}^p}^p \coloneqq \int_A \big|f(\boldsymbol{x}) - g(\boldsymbol{x})\big|^p \, \mathrm{d}\boldsymbol{x},$$

mais des pertes qui ne sont pas des distances sont aussi utilisées. En toute généralité, il n'y a pas vraiment de contraintes sur les propriétés que doit vérifier

une perte. Par exemple, elle n'a pas besoin d'être symétrique : en estimation de densité, la divergence de Kullback–Leibler est une perte parfois considérée :

$$\ell(f,g) = \begin{cases} \displaystyle\int_A g(\boldsymbol{x}) \log\left(\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} & \text{si } g\mathrm{d}\boldsymbol{x} \ll f\mathrm{d}\boldsymbol{x}, \\ +\infty & \text{sinon,} \end{cases}$$

qui n'est pas symétrique et peut même être infinie. Pour d'autres exemples de pertes considérées en estimation de densité, on pourra se référer au chapitre 2 de Tsybakov (2009). Dans cette thèse, on se restreindra à la perte L^2 (chapitres 2 et 3) et à la perte L^2 pondérée par une mesure μ (chapitre 4) :

$$\ell(f,g) = \|f-g\|_\mu^2 := \int_A \left|f(\boldsymbol{x}) - g(\boldsymbol{x})\right|^2 \mathrm{d}\mu(\boldsymbol{x}).$$

Ce choix est motivé par la méthode d'estimation utilisée : l'estimation par projection.

1.1.3 Vitesse de convergence

Le risque représente l'erreur moyenne de l'estimateur vis-à-vis de la fonction d'intérêt f. Si l'estimateur est bien construit, on s'attend à ce que le risque diminue et tende vers 0 lorsque la taille de l'échantillon croît vers l'infini. La question qui se pose alors est de déterminer la vitesse de convergence du risque vers 0.

Définition 1.1.1. Étant donnés un ensemble $\mathscr{F}$ de fonctions et $(\psi_n)_{n\geq 1}$ une suite qui décroit vers 0, on dit que $\widehat{f}_n$ converge à vitesse ψ_n sur $\mathscr{F}$ si :

$$\sup_{f\in\mathscr{F}} \mathscr{R}(f,\widehat{f}_n) \leq C(\mathscr{F}) \times \psi_n,$$

où $C(\mathscr{F})$ est une constante positive. De plus, on dit que la vitesse $(\psi_n)_{\geq 1}$ est minimax sur la collection $\mathscr{F}$ s'il existe $c(\mathscr{F}) > 0$ telle que :

$$\liminf_{n\to+\infty} \inf_{T_n} \sup_{f\in\mathscr{F}} \psi_n^{-1} \mathscr{R}(f,T_n) \geq c(\mathscr{F}),$$

où l'infimum porte sur tous les estimateurs.

Autrement dit, la vitesse $(\psi_n)_{n\geq 1}$ est minimax s'il n'existe pas d'estimateurs ayant une meilleure vitesse de convergence sur $\mathscr{F}$. Typiquement, les collections de fonctions $\mathscr{F}$ considérées sont des boules dans des espaces fonctionnels. On verra qu'un estimateur $\widehat{f}_n$ atteint des vitesses de convergence différentes sur des collections $\mathscr{F}$ différentes, et que cette vitesse est liée la régularité des fonctions de $\mathscr{F}$: plus les fonctions sont régulières, meilleure est la vitesse de convergence.

1.2 Estimation par projection

On suppose que f appartient à $\mathrm{L}^2(A,\nu)$ avec $A \subseteq \mathbb{R}^d$ et ν une mesure sur A. On note $\langle\cdot,\cdot\rangle$ et $\|\cdot\|$ le produit scalaire et la norme associés à cet espace :

$$\langle f,g\rangle \coloneqq \int_A f(\boldsymbol{x})g(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x}), \quad \|f\|^2 = \int_A f(\boldsymbol{x})^2\,\mathrm{d}\nu(\boldsymbol{x}).$$

L'idée de l'estimation par projection est de décomposer la fonction d'intérêt dans une base orthonormée et d'estimer ses coefficients.

1.2.1 Fonctions d'une variable

Supposons dans un premier temps que f est une fonction d'une seule variable, c'est-à-dire $d = 1$. On considère $(\varphi_k)_{k\in\mathbb{N}}$ une base orthonormée de $\mathrm{L}^2(A,\nu)$ et on décompose f dans cette base :

$$f = \sum_{k\in\mathbb{N}} a_k\varphi_k, \quad a_k = \langle f,\varphi_k\rangle,$$

avec convergence de la série au sens de la norme $\|\cdot\|$. Un estimateur par projection de f est alors de la forme :

$$\widehat{f}_m = \sum_{k=0}^{m-1} \widehat{a}_k\varphi_k,$$

où $\widehat{a}_k$ est un estimateur de a_k et m est le nombre de coefficients que l'on choisit d'estimer. Notons S_m l'espace vectoriel engendré par les m premières fonctions de la base, et notons f_m la projection de f sur S_m :

$$f_m = \sum_{k=0}^{m-1} a_k\varphi_k,$$

alors $\widehat{f}_m$ peut être vu comme un estimateur de la projection f_m, d'où le terme « estimation par projection ».

Une première façon d'estimer les coefficients est par *méthode des moments* : si a_k peut s'écrire comme l'espérance d'une fonction d'une observation, on l'estime en remplaçant l'espérance par une moyenne empirique sur les observations. Une autre façon est de procéder par *minimisation de contraste* (Birgé & Massart, 1993) : en remarquant que la projection f_m est solution du problème de minimisation :

$$f_m = \operatorname*{arg\,min}_{t\in S_m} \|f-t\|^2, \tag{1.1}$$

on l'estime en minimisant un équivalent empirique de $t \mapsto \|f-t\|^2$. Cet équivalent empirique, noté $\gamma_n(t)$, est appelé *contraste*. L'estimateur par projection est alors défini comme la solution de :

$$\widehat{f}_m \coloneqq \operatorname*{arg\,min}_{t\in S_m} \gamma_n(t).$$

1.2.2 Fonctions de plusieurs variables

Supposons à présent que f est une fonction de $d \geq 2$ variables. On suppose que le domaine A s'écrit comme un produit cartésien et que la mesure ν se décompose comme un produit de mesures :

$$A = A_1 \times \cdots \times A_d, \quad \nu = \nu_1 \otimes \cdots \otimes \nu_d,$$

où ν_i est une mesure sur A_i. On peut alors construire une base orthonormée de $\mathrm{L}^2(A,\nu)$ en tensorisant des bases orthonormées de $\mathrm{L}^2(A_i,\nu_i)$. Soit $(\varphi_k^i)_{k\in\mathbb{N}}$ une base orthonormée de $\mathrm{L}^2(A_i,\nu_i)$, on définit pour tout multi-indice $\boldsymbol{k} = (k_1,\dots,k_d)$ la fonction :

$$\forall \boldsymbol{x} \in A, \quad \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \coloneqq \big(\varphi_{k_1}^1 \otimes \cdots \otimes \varphi_{k_d}^d\big)(\boldsymbol{x}) \coloneqq \varphi_{k_1}^1(x_1) \times \cdots \times \varphi_{k_d}^d(x_d).$$

Les fonctions $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{N}^d}$ forment une base orthonormée de $\mathrm{L}^2(A,\nu)$.

L'estimation par projection de f est définie comme dans le cas $d = 1$, à ceci près que les coefficients de f sont maintenant multi-indicés :

$$f = \sum_{\boldsymbol{k}\in\mathbb{N}^d} a_{\boldsymbol{k}}\varphi_{\boldsymbol{k}}, \quad a_{\boldsymbol{k}} = \langle f, \varphi_{\boldsymbol{k}}\rangle.$$

Pour $\boldsymbol{m} = (m_1,\dots,m_d) \in \mathbb{N}_+^d$, on définit les espaces de projection :

$$S_{\boldsymbol{m}} \coloneqq \mathrm{Vect}\big(\varphi_{\boldsymbol{k}} : \forall i,\ 0 \leq k_i \leq m_i - 1\big). \tag{1.2}$$

Ceux-ci sont de dimension $D_{\boldsymbol{m}} \coloneqq m_1 \times \cdots \times m_d$. Un estimateur par projection de f est alors de la forme :

$$\widehat{f}_{\boldsymbol{m}} = \sum_{\substack{\boldsymbol{k}\in\mathbb{N}_+^d \\ \forall i,\, k_i < m_i}} \widehat{a}_{\boldsymbol{k}}\varphi_{\boldsymbol{k}},$$

où les estimateurs $\widehat{a}_{\boldsymbol{k}}$ sont obtenus par méthode des moments ou par minimisation de contraste, comme dans le cas $d = 1$.

Exemple 1.2 (Estimation d'une densité)**.** On reprend l'exemple 1.1. Les coefficients de f sont donnés par :

$$a_{\boldsymbol{k}} = \langle f, \varphi_{\boldsymbol{k}}\rangle_{\mathrm{L}^2} = \int_A f(\boldsymbol{x})\,\varphi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)\big].$$

La méthode des moments fournit l'estimateur :

$$\widehat{a}_{\boldsymbol{k}} \coloneqq \frac{1}{n}\sum_{i=1}^n \varphi_{\boldsymbol{k}}(\boldsymbol{X}_i). \tag{1.3}$$

On peut aussi procéder par minimisation de contraste ; on développe le carré de la norme dans l'équation (1.1) :

$$\operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \|f - t\|_{\mathrm{L}^2}^2 = \operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \big(\|f\|_{\mathrm{L}^2} - 2\langle f, t\rangle_{\mathrm{L}^2}^2 + \|t\|_{\mathrm{L}^2}^2\big).$$

La quantité $\|f\|^2_{\mathrm{L}^2}$ est inconnue mais ne dépend pas de t, donc elle n'intervient pas dans la minimisation du critère et peut être omise. On estime alors le produit scalaire $\langle f, t\rangle_{\mathrm{L}^2} = \mathbb{E}[t(\boldsymbol{X}_1)]$ par la moyenne empirique des $t(\boldsymbol{X}_i)$. Le contraste qui en résulte est :

$$\gamma_n(t) \coloneqq \|t\|^2_{\mathrm{L}^2} - \frac{2}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i).$$

Il est facile de vérifier que les estimateurs obtenus par méthode des moments et par minimisation de contraste coïncident dans ce cas.

1.2.3 Exemples de bases orthonormées

Citons quelques bases orthonormées classiquement utilisées en estimation fonctionnelle.

Base trigonométrique On définit sur $[0,1]$ les fonctions :

$$\varphi_0(x) \coloneqq 1, \quad \varphi_{2k}(x) \coloneqq \sqrt{2}\cos(2\pi kx), \quad \varphi_{2k+1}(x) \coloneqq \sqrt{2}\sin(2\pi kx).$$

Ces fonctions forment une base orthonormée de $\mathrm{L}^2([0,1])$ et les coefficients de f dans cette base sont ses coefficients de Fourier.

Base d'histogrammes réguliers On partitionne $[0,1]$ en D intervalles de même longueurs et on définit :

$$\forall k \in \{0,\dots,D-1\}, \quad \varphi_{D,k} = \sqrt{D}\,\mathbf{1}_{[\frac{k}{D},\frac{k+1}{D}]}.$$

Les fonctions $(\varphi_{D,k})_{0\le k\le D-1}$ sont orthonormées et forment une base de $S_D \coloneqq \mathrm{Vect}\big(\varphi_{D,k} : k \in \{0,\dots,D-1\}\big)$. Les espaces d'approximations sont donc les $(S_D)_{D\ge 1}$ et un estimateurs par projections sur S_D est une fonction en escalier. On choisit parfois de se restreindre aux D de forme 2^m (subdivision dyadique) pour que les sous-espaces soient emboités.

Base d'ondelettes Sur $\mathbb{R}$, on définit pour tous $j,k \in \mathbb{Z}$:

$$\varphi_{j,k}(x) \coloneqq 2^{j/2}\varphi\Big(2^j x - k\Big),$$

où φ est une fonction régulière sur $\mathbb{R}$. Sous certaines hypothèses sur φ, ces fonctions forment une base orthonormée de $\mathrm{L}^2(\mathbb{R})$ (cf. Meyer (1990)). Pour une utilisation de ces bases en statistique non paramétrique, on pourra se référer aux travaux de Donoho & Johnstone (1994, 1995, 1998) et Donoho *et al.* (1996).

Base de splines Des bases souvent utilisées pour l'approximation de fonctions sont les bases de splines (DeVore & Lorentz, 1993). Un estimateur par projection dans ces bases est une fonction polynomiale par morceaux avec des recollements réguliers (de classe C^k selon le degré des polynômes). Cependant, ces bases ne sont pas orthonormées. Toutefois, il est possible de les utiliser pour faire de l'estimation par projection, modulo quelques précautions techniques.

Remarque 1.2.1. À partir d'une base orthonormée de $\mathrm{L}^2([0,1])$, on construit une base orthonormée de $\mathrm{L}^2([a,b])$ via le changement de variable $x \mapsto (b-a)x+a$.

Les deux bases que nous utiliserons dans cette thèse sont construites à partir de familles de polynômes orthogonaux : les polynômes de Laguerre et les polynômes d'Hermite. Elles ont la particularité d'être à support sur $\mathbb{R}$ (Hermite) et $\mathbb{R}_+$ (Laguerre), c'est-à-dire des domaines non compacts. L'avantage d'utiliser des bases à support non compact est d'éviter le problème du choix du support. En effet, quand ce dernier est un compact, il est supposé fixé à priori dans l'analyse théorique alors qu'en pratique, il est déterminé grâce aux données.

Ces dernières années, ces bases ont été utilisées pour résoudre de nombreux problèmes statistiques en raison de leurs bonnes propriétés mathématiques, voir le chapitre 22 de Abramowitz & Stegun (1972) et l'annexe A pour ce qui concerne les fonctions de Laguerre. La base de Laguerre est utilisée par Comte *et al.* (2017), Vareschi (2015) et Mabon (2017) pour résoudre des problèmes de déconvolution sur $\mathbb{R}_+$, ainsi que par Belomestny *et al.* (2016) pour l'estimation d'une densité dans le modèle de censure multiplicative. La base d'Hermite est utilisée par Belomestny *et al.* (2019) pour l'estimation de densité et par Sacko (2020) pour la déconvolution d'une densité sur $\mathbb{R}$. Citons également les travaux de Comte *et al.* (2020) concernant l'estimation des dérivées d'une densité à l'aide de ces deux bases et ceux de Comte & Genon-Catalot (2020a) en régression non paramétrique.

Base de Laguerre On définit les fonctions de Laguerre sur $\mathbb{R}_+$ par :

$$\varphi_k(x) := \sqrt{2}\, L_k(2x)\, \mathrm{e}^{-x}, \quad L_k(x) := \sum_{j=0}^{k} \binom{k}{j} \frac{(-x)^j}{j!}. \tag{1.4}$$

Ces fonctions forment une base orthonormée de $\mathrm{L}^2(\mathbb{R}_+)$. Les polynômes $(L_k)_{k\geq 0}$ sont appelés les polynômes de Laguerre. Cette base sera utilisée aux chapitres 2 et 3. Pour une présentation détaillée de ces fonctions et de leurs propriétés, on pourra consulter l'annexe A.

Base d'Hermite On définit les fonctions d'Hermite sur $\mathbb{R}$ par :

$$\varphi_k(x) := c_k\, H_k(x)\, \mathrm{e}^{-\frac{x^2}{2}}, \quad H_k(x) := (-1)^k \mathrm{e}^{x^2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left[\mathrm{e}^{-x^2}\right], \quad c_k := \left(2^k k! \sqrt{\pi}\right)^{-1/2}.$$

Les fonctions $(\varphi_k)_{k\in\mathbb{N}}$ forment une base orthonormée de $\mathrm{L}^2(\mathbb{R})$. Les polynômes $(H_k)_{k\geq 0}$ sont appelés les polynômes d'Hermite. Nous utiliserons cette base au chapitre 4.

1.3 Calcul hypermatriciel

Pour des fonctions d'une seule variable, le choix d'une base orthonormée fournit une isométrie entre les fonctions $f \in \mathrm{L}^2(A,\nu)$ et leurs coefficients $(a_k)_{k\in\mathbb{N}} \in \ell^2(\mathbb{N})$.

Le sous-espace S_m est alors lui-même isométrique à l'espace des vecteurs de longueur m. L'intérêt de cette isométrie est qu'elle permet de remplacer certaines opérations sur les fonctions de S_m par du calcul matriciel sur des vecteurs de $\mathbb{R}^m$.

Dans le cas des fonctions de d variables, leurs coefficients dépendent de d indices et sont donc des éléments de $\ell^2(\mathbb{N}^d)$. Une fonction de $S_{\boldsymbol{m}}$ est alors représentée par un tableau à d dimensions contenant ses coefficients. Nous aurons donc besoin d'un analogue du calcul matriciel pour ces tableaux de dimension supérieure.

Les notions concernant les hypermatrices définies dans cette section sont inspirées du chapitre 15 de Hogben (2013).

1.3.1 Multi-indices

On appelle *multi-indice* un élément de $\mathbb{N}^d$ ou $\mathbb{N}_+^d$. On note $\mathbf{0}$ le multi-indice $(0,\dots,0)$ et $\mathbf{1}$ le multi-indice $(1,\dots,1)$, leur longueur dépendant du contexte. On définit un ordre partiel sur les multi-indices en les comparant coordonnée par coordonnée :

$$\forall \boldsymbol{k},\boldsymbol{\ell} \in \mathbb{N}^d, \quad \boldsymbol{k} \le \boldsymbol{\ell} \iff \forall i \in \{1,\dots,d\},\, k_i \le \ell_i.$$

Avec ces notations, l'équation (1.2) peut se réécrire :

$$S_{\boldsymbol{m}} \coloneqq \operatorname{Vect}\left(\varphi_{\boldsymbol{k}} : \mathbf{0} \le \boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}\right),$$

la soustraction $\boldsymbol{m}-\mathbf{1}$ s'effectuant terme à terme. Notons qu'on a $\boldsymbol{m} \le \boldsymbol{m}'$ si et seulement si $S_{\boldsymbol{m}} \subseteq S_{\boldsymbol{m}'}$. On note également $\wedge$ et $\vee$ respectivement le minimum et le maximum terme à terme :

$$\boldsymbol{m} \wedge \boldsymbol{m}' \coloneqq (m_1 \wedge m_1', \dots, m_d \wedge m_d'), \quad \boldsymbol{m} \vee \boldsymbol{m}' \coloneqq (m_1 \vee m_1', \dots, m_d \vee m_d').$$

Enfin, on note $|\boldsymbol{k}| \coloneqq k_1 + \cdots + k_d$ la longueur d'un multi-indice.

Concernant les vecteurs de $\mathbb{R}^d$ ou $\mathbb{C}^d$, l'addition et la multiplication par un scalaire sont définies terme à terme comme il est usuel. Nous définissons également la multiplication et la division terme à terme :

$$\boldsymbol{x} \times \boldsymbol{y} \coloneqq (x_1 \times y_1, \dots, x_d \times y_d), \quad \frac{\boldsymbol{x}}{\boldsymbol{y}} \coloneqq \left(\frac{x_1}{y_1}, \dots, \frac{x_d}{y_d}\right),$$

la division par $\boldsymbol{y}$ n'étant bien définie que si toutes les coordonnées de $\boldsymbol{y}$ sont non nulles. Le produit scalaire sera généralement noté par un crochet $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ (respectivement $\langle\cdot,\cdot\rangle_{\mathbb{C}^d}$), ou par un point médian dans certains calculs :

$$\boldsymbol{x} \cdot \boldsymbol{y} \coloneqq \begin{cases} \sum_{i=1}^d x_i\, y_i & \text{dans le cas de } \mathbb{R}^d, \\ \sum_{i=1}^d x_i\, \overline{y}_i & \text{cans le cas de } \mathbb{C}^d. \end{cases}$$

Enfin, si $k \in \mathbb{N}^d$ est un multi-indice et $\boldsymbol{x} \in \mathbb{C}^d$ est un vecteur, $\boldsymbol{x}^{\boldsymbol{k}}$ désigne le multinôme :

$$\boldsymbol{x}^{\boldsymbol{k}} \coloneqq x_1^{k_1} \cdots x_d^{k_d}.$$

Ces notations permettent d'écrire de façon compacte les calculs en dimension supérieure et seront très utilisées dans le chapitre 2. Par exemple, la formule du binôme de Newton s'étend aux vecteurs et s'écrit comme en dimension 1.

Proposition 1.3.1 (Formule multi-binomiale)**.** *Pour tout $\boldsymbol{k} \in \mathbb{N}^d$ et tous $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^d$, on a la formule :*

$$(\boldsymbol{x}+\boldsymbol{y})^{\boldsymbol{k}} = \sum_{\mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{k}} \binom{\boldsymbol{k}}{\boldsymbol{j}} \boldsymbol{x}^{\boldsymbol{j}}\, \boldsymbol{y}^{\boldsymbol{k}-\boldsymbol{j}},$$

où $\binom{\boldsymbol{k}}{\boldsymbol{j}} \coloneqq \binom{k_1}{j_1} \times \cdots \times \binom{k_d}{j_d}$.

1.3.2 Hypermatrices

On appelle *hypermatrice* un tableau de nombres multidimensionnel. Soit $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{N}_+^d$ un multi-indice, on note $\mathbb{R}^{\boldsymbol{m}} \coloneqq \mathbb{R}^{m_1 \times \cdots \times m_d}$ l'ensemble des hypermatrices ayant m_i composantes suivant la i-ème dimension. Un élément $\mathbf{a}$ de $\mathbb{R}^{\boldsymbol{m}}$ s'écrit donc :

$$\mathbf{a} = \left[a_{k_1,\ldots,k_d}\right]_{\forall i, 0 \leq k_i \leq m_i - 1} = \left[a_{\boldsymbol{k}}\right]_{\mathbf{0} \leq \boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}},$$

et on dit que $\mathbf{a}$ est une hypermatrice de forme $\boldsymbol{m}$. Un vecteur colonne et une matrice peuvent être vus comme des hypermatrices à respectivement 1 et 2 dimensions.

Remarque 1.3.2. On choisit d'indicer les vecteurs, matrices et hypermatrices en partant de 0. En effet, comme on utilise des bases construites à partir de polynômes orthogonaux, celles-ci sont naturellement ordonnées suivant le degré des polynômes, en commençant par un polynôme constant. Il est donc plus simple que les coefficients d'une fonction dans ces bases soient indicés à partir de 0.

Parfois, le multi-indice d'une hypermatrice peut se scinder en deux multi-indices qui jouent des rôles différents. On introduit pour cela une nouvelle notation : si $\boldsymbol{m} \in \mathbb{N}_+^d$ et $\boldsymbol{m}' \in \mathbb{N}_+^{d'}$ sont deux multi-indices, on note $\mathbb{R}^{\boldsymbol{m}' \times \boldsymbol{m}}$ l'espace des hypermatrices de forme $(m'_1 \times \cdots \times m'_{d'}) \times (m_1 \times \cdots \times m_d)$. Une hypermatrice $\mathbf{T} \in \mathbb{R}^{\boldsymbol{m}' \times \boldsymbol{m}}$ s'écrit :

$$\mathbf{T} = \left[T_{\boldsymbol{j},\boldsymbol{k}}\right]_{\substack{\mathbf{0} \leq \boldsymbol{j} \leq \boldsymbol{m}'-\mathbf{1} \\ \mathbf{0} \leq \boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}}}.$$

1.3.3 Opérations sur les hypermatrices

L'espace $\mathbb{R}^{\boldsymbol{m}}$ est muni d'une structure d'espace vectoriel euclidien en définissant la multiplication par un scalaire et l'addition composante par composante :

$$\left[\mathbf{a}+\mathbf{b}\right]_{\boldsymbol{k}} \coloneqq a_{\boldsymbol{k}} + b_{\boldsymbol{k}}, \quad \left[\lambda \mathbf{a}\right]_{\boldsymbol{k}} \coloneqq \lambda a_{\boldsymbol{k}},$$

et le produit scalaire et la norme par :

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{R}^{\boldsymbol{m}}} \coloneqq \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} a_{\boldsymbol{k}}\, b_{\boldsymbol{k}}, \quad \|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}^2 \coloneqq \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} a_{\boldsymbol{k}}^2.$$

Le produit matriciel est généralisé par la notion de *produit contracté*.

Definition 1.3.3 (Produit contracté)**.** Soit $p \in \mathbb{N}_+$, soient $\boldsymbol{m}, \boldsymbol{m}', \boldsymbol{m}''$ des multi-indices avec $\boldsymbol{m}' \in \mathbb{N}_+^p$ et soient $\mathbf{T} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}'}$ et $\mathbf{S} \in \mathbb{R}^{\boldsymbol{m}' \times \boldsymbol{m}''}$ deux hypermatrices. On définit leur p-produit contracté comme l'hypermatrice $\mathbf{T} \times_p \mathbf{S} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}''}$ suivante :

$$\forall \boldsymbol{j}, \forall \boldsymbol{\ell}, \quad \left[\mathbf{T} \times_p \mathbf{S}\right]_{\boldsymbol{j},\boldsymbol{\ell}} \coloneqq \sum_{\boldsymbol{k}=(k_1,\dots,k_p)} T_{\boldsymbol{j},\boldsymbol{k}} S_{\boldsymbol{k},\boldsymbol{\ell}}. \tag{1.5}$$

Le produit matriciel correspond au cas $p = 1$. Comme pour ce dernier, il est facile de voir que le produit contracté par une hypermatrice donnée est une application linéaire. Si $\mathbf{T} \in \mathbb{R}^{\boldsymbol{m}' \times \boldsymbol{m}}$ avec $\boldsymbol{m} \in \mathbb{N}_+^d$ et $\boldsymbol{m}' \in \mathbb{N}_+^{d'}$, alors $\mathbf{T}$ induit une application linéaire de $\mathbb{R}^{\boldsymbol{m}}$ dans $\mathbb{R}^{\boldsymbol{m}'}$, que l'on note encore $\mathbf{T}$:

$$\mathbf{T} \colon \mathbf{a} \in \mathbb{R}^{\boldsymbol{m}} \longmapsto \left(\mathbf{T} \times_d \mathbf{a}\right) \in \mathbb{R}^{\boldsymbol{m}'}.$$

En particulier, $\mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}}$ s'identifie avec les endomorphismes de $\mathbb{R}^{\boldsymbol{m}}$. Ainsi, si $\mathbf{T}$ est une hypermatrice $\boldsymbol{m} \times \boldsymbol{m}$, on parlera de sa trace, de ses valeurs propres, de son inverse (si elle est inversible), etc, comme on le ferait pour n'importe quel endomorphisme d'un espace vectoriel. En particulier, si on définit la norme d'opérateur et la norme de Frobenius de $\mathbf{T}$ comme :

$$\|\mathbf{T}\|_F^2 \coloneqq \sum_{\boldsymbol{j},\boldsymbol{k}} T_{\boldsymbol{j},\boldsymbol{k}}^2, \quad \|\mathbf{T}\|_{\mathrm{op}} \coloneqq \sup_{\mathbf{a} \in \mathbb{R}^{\boldsymbol{m}} \setminus \{\mathbf{0}\}} \frac{\|\mathbf{T} \times_d \mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}}{\|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}},$$

alors on a :

$$\|\mathbf{T}\|_F^2 = \mathrm{Tr}\left(\mathbf{T}^* \times_d \mathbf{T}\right), \quad \|\mathbf{T}\|_{\mathrm{op}}^2 = \lambda_{\max}\left(\mathbf{T}^* \times_d \mathbf{T}\right),$$

où $\mathbf{T}^*$ est l'adjoint de $\mathbf{T}$ pour le produit scalaire $\langle \cdot, \cdot \rangle_{\mathbb{R}^{\boldsymbol{m}}}$. Il n'est pas compliqué de voir que $\mathbf{T}^*$ est l'hypermatrice définie par :

$$\left[\mathbf{T}^*\right]_{\boldsymbol{j},\boldsymbol{k}} \coloneqq T_{\boldsymbol{k},\boldsymbol{j}}.$$

On rappelle les inégalités suivantes concernant les normes d'opérateur et de Frobenius d'un endomorphisme :

$$\frac{1}{\dim(\mathbb{R}^{\boldsymbol{m}})} \|\mathbf{T}\|_F^2 \leq \|\mathbf{T}\|_{\mathrm{op}}^2 \leq \|\mathbf{T}\|_F^2. \tag{1.6}$$

1.4 Estimation adaptative

Comme on l'a vu dans en section 1.2, les estimateurs par projection dépendent d'un hyperparamètre $\boldsymbol{m}$ qui détermine le nombre de coefficients que l'on choisit d'estimer. Le choix de $\boldsymbol{m}$ est crucial pour obtenir un estimateur performant de f et nécessite une procédure spécifique.

1.4.1 Décomposition biais-variance et oracle

Pour chaque $\boldsymbol{m}$, on commence par étudier le risque de l'estimateur $\widehat{f}_{\boldsymbol{m}}$. Pour cela, on applique le théorème de Pythagore en utilisant le fait que $f - f_{\boldsymbol{m}}$ est orthogonal à $\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}$ et on obtient la décomposition :

$$\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|^2 = \|f - f_{\boldsymbol{m}}\|^2 + \mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2, \tag{1.7}$$

$$= \inf_{t \in S_m} \|f - t\|^2 + \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \mathbb{E}\big[(\widehat{a}_{\boldsymbol{k}} - a_{\boldsymbol{k}})^2\big]. \tag{1.8}$$

La décomposition (1.7) est appelée *décomposition biais-variance* :

- Le premier terme dans (1.7) est appelé *terme de biais* (ou *erreur d'approximation*). C'est la partie déterministe du risque qui représente l'erreur que l'on commet en remplaçant f par sa projection sur $S_{\boldsymbol{m}}$. C'est aussi la distance de f au sous-espace $S_{\boldsymbol{m}}$.

- Le second terme dans (1.7) est appelé *terme de variance* (ou *erreur stochastique*). C'est la partie aléatoire du risque qui provient de l'estimation des coefficients $a_{\boldsymbol{k}}$.

Ces deux erreurs varient en sens contraire : lorsque le paramètre $\boldsymbol{m}$ croît, le terme de biais diminue tandis que le terme de variance augmente. Il en résulte qu'il doit exister une valeur optimale de $\boldsymbol{m}$ qui réalise le compromis entre biais et variance. Ce $\boldsymbol{m}$ optimal est appelé *modèle oracle* et la valeur du risque qui lui est associé est appelé *risque oracle* :

$$\boldsymbol{m}_o := \operatorname*{arg\,min}_{\boldsymbol{m}} \left(\|f - f_{\boldsymbol{m}}\|^2 + \mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2\right), \quad \mathscr{R}_o = \inf_{\boldsymbol{m}} \mathscr{R}(f, \widehat{f}_m).$$

En pratique, l'oracle n'est pas accessible car sa valeur dépend de la fonction inconnue f. La quantité $\widehat{f}_{\boldsymbol{m}_o}$ est qualifiée de *pseudo-estimateur* ou *estimateur oracle*. Son risque représente le meilleur risque qu'on puisse espérer obtenir avec un estimateur par projection $\widehat{f}_{\boldsymbol{m}}$, si un « oracle » connaissant f nous donnait le $\boldsymbol{m}$ optimal.

Exemple 1.3 (Estimation de densité)**.** On poursuit l'exemple 1.2. Analysons le terme de variance dans la décomposition (1.8). Les coefficients sont estimés par (1.3) donc on a $\mathbb{E}[\widehat{a}_{\boldsymbol{k}}] = a_{\boldsymbol{k}}$. Le terme de variance s'écrit donc :

$$\begin{aligned}
\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \mathbb{E}\big[(\widehat{a}_{\boldsymbol{k}} - a_{\boldsymbol{k}})^2\big] &= \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \operatorname{Var}(\widehat{a}_{\boldsymbol{k}}) \\
&= \frac{1}{n} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \operatorname{Var}\big(\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)\big) \\
&\leq \frac{1}{n} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)^2\big]. \qquad (1.9)
\end{aligned}$$

Si on introduit la quantité[1] :

$$L(\boldsymbol{m}) \coloneqq \sup_{\boldsymbol{x}\in A} \sum_{\boldsymbol{k}\leq \boldsymbol{m}-\mathbf{1}} \varphi_{\boldsymbol{k}}(\boldsymbol{x})^2 = \prod_{i=1}^{d} \left(\sup_{x_i \in A_i} \sum_{k_i=0}^{m_i-1} \varphi_{k_i}^i(x_i)^2 \right), \tag{1.10}$$

alors le terme de variance est majoré par $\frac{L(\boldsymbol{m})}{n}$ et le risque est majoré par :

$$\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq \|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \frac{L(\boldsymbol{m})}{n}.$$

Dans la suite de cet exemple, on se place en dimension $d = 1$. Supposons que $A = [0,1]$ et que l'on utilise la base trigonométrique, alors on a :

$$L(2p) \leq L(2p+1) = 1 + 2 \sup_{x\in[0,1]} \sum_{k=1}^{p} \left(\cos^2(2\pi k x) + \sin^2(2\pi k x) \right) = 2p+1,$$

donc on a la majoration $L(m) \leq m$. Ainsi, le terme de variance est majoré par $\frac{m}{n}$.

Concernant le terme de biais, sa décroissance est liée la régularité de f. En effet, le biais est donné par le reste de la série des coefficients de f :

$$\|f - f_m\|_{\mathrm{L}^2}^2 = \sum_{k=m}^{+\infty} a_k^2,$$

or on sait que la vitesse de décroissance des coefficients de Fourier d'une fonction traduit sa régularité. Plus précisément, en supposant que f appartienne à une boule de Sobolev :

$$\mathrm{H}^{\beta}\big([0,1], R\big) \coloneqq \left\{ f \in \mathrm{L}^2\big([0,1]\big) \,\middle|\, \sum_{j=2}^{+\infty} \left(a_{2j}^2 + a_{2j+1}^2 \right) \times (2j)^{2\beta} \leq \frac{R^2}{\pi^{2\beta}} \right\},$$

alors le terme de biais décroit à vitesse $m^{-2\beta}$. Lorsque β est un entier strictement positif, ces espaces peuvent être caractérisés en termes de la régularité de f :

$$\forall \beta \in \mathbb{N}_+, \quad f \in \mathrm{H}^{\beta}\big([0,1], R\big) \iff \begin{cases} f(0) = f(1), \\ f \text{ est } \beta - 1 \text{ fois dérivable}, \\ f^{(\beta-1)} \text{ est absolument continue}, \\ \|f^{(\beta)}\|_{\mathrm{L}^2} \leq R, \end{cases}$$

voir la proposition 1.14 de Tsybakov (2009). Ainsi, si $f \in \mathrm{H}^{\beta}\big([0,1], R\big)$, la décomposition biais-variance devient :

$$\mathbb{E}\|f - \widehat{f}_m\|_{\mathrm{L}^2}^2 \leq C(\beta, R)\, m^{-2\beta} + \frac{m}{n}. \tag{1.11}$$

[1] cette quantité mesure la différence entre les normes $\|\cdot\|_\infty$ et $\|\cdot\|_{\mathrm{L}^2}$ sur l'espace $S_{\boldsymbol{m}}$, voir le lemme 3.1 de Birgé & Massart (1998) ou notre lemme 4.3.3.

En minimisant le membre de droite, on déduit le modèle oracle et une borne sur le risque oracle :

$$m_o = \left\lfloor \left(2\beta C(\beta,R)\right)^{\frac{1}{2\beta+1}} n^{\frac{1}{2\beta+1}} \right\rfloor, \quad \mathscr{R}_o \leq C'(\beta,R)\, n^{-\frac{2\beta}{2\beta+1}}.$$

La vitesse $n^{-\frac{2\beta}{2\beta+1}}$ est donc la vitesse d'estimation de $\widehat{f}_{m_o}$ sur les boules de Sobolev de régularité β (cf. définition 1.1.1) et est connue pour être optimale au sens minimax, voir Efromovich (1986). Cette vitesse est d'autant meilleure que β est grand et s'approche de la vitesse n^{-1} lorsque $\beta \to \infty$, sans jamais l'atteindre. La vitesse n^{-1} est la vitesse que l'on trouve classiquement dans de nombreux problèmes d'estimation paramétrique. Ainsi, pour une même valeur fixée de risque, l'estimation non paramétrique requiert un nombre plus grand d'observations que l'estimation paramétrique.

De façon générale, la décroissance du terme de biais est d'autant plus rapide que f est une fonction « régulière », en un sens qui dépend de la base utilisée : à chaque base correspond un espace de régularité. Pour la base trigonométrique, ce sont les espace de Sobolev comme on l'a vu dans l'exemple précédent. Pour les bases de Laguerre et d'Hermite, ce sont les espaces de Sobolev–Laguerre et de Sobolev–Hermite, voir Bongioanni & Torrea (2006, 2009). Concernant les espaces de Sobolev–Laguerre, leur définition à partir des coefficients de Laguerre des fonctions est due à Comte & Genon-Catalot (2015).

Definition 1.4.1. Soient $s > 0$ et $L > 0$, les boules de Sobolev–Laguerre et de Sobolev–Hermite de régularité s et de rayon L sont définies par :

$$\mathrm{W}^s(A,L) := \left\{ f \in \mathrm{L}^2(A) \,\middle|\, \sum_{k=0}^{+\infty} \langle f, \varphi_k \rangle^2 \, k^s \leq L \right\},$$

où $(\varphi_k)_{k\in\mathbb{N}}$ est la base de Laguerre si $A = \mathbb{R}_+$ et la base d'Hermite si $A = \mathbb{R}$. Les espaces de Sobolev–Laguerre et de Sobolev–Hermite sont définis comme $\mathrm{W}^s(A) := \bigcup_{L>0} \mathrm{W}^s(A,L)$.

L'indice s est lié à la dérivabilité des fonctions de $\mathrm{W}^s(A)$. En effet lorsque $s \in \mathbb{N}_+$, Comte & Genon-Catalot (2015) montrent que f appartient à $\mathrm{W}^s(\mathbb{R}_+)$ si et seulement si :

1. f est $s-1$ fois dérivable et $f^{(s-1)}$ est absolument continue ;
2. pour tout $k \in \{0,\ldots,s-1\}$, $x^{\frac{k+1}{2}} \sum_{j=0}^{k+1} \binom{k+1}{j} f^{(j)} \in \mathrm{L}^2(\mathbb{R}_+)$.

Pour les espaces de Sobolev–Hermite, il découle des travaux de Bongioanni & Torrea (2006) (cf. la proposition 4 de Belomestny *et al.* (2019)) que f appartient à $\mathrm{W}^s(\mathbb{R})$ si et seulement si :

1. f est s fois dérivable,

2. $f, f', \ldots, f^{(s)} \in \mathrm{L}^2(\mathbb{R})$ et pour tout $j \in \{0, \ldots, s-1\}$ $x^{s-j} f^{(j)} \in \mathrm{L}^2(\mathbb{R})$.

Lorsque $f \in \mathrm{W}^s(A, L)$, son biais en base de Laguerre ou en base d'Hermite est majoré par $L \times m^{-s}$. À noter la différence avec les espaces de Sobolev sur $[0,1]$ et la base trigonométrique, où la vitesse de décroissance du biais est $m^{-2\beta}$ avec β la régularité de la fonction.

Remarque 1.4.2. Quand on utilise la base de Laguerre ou la base d'Hermite, la quantité $L(m)$ définie par (1.10) est majorée par $\Phi \times m$ où Φ est une constante qui dépend de la base[2] (pour la base de Laguerre, la constante optimale est $\Phi = 2$), ce qui donne un terme de variance d'ordre m/n comme pour la base trigonométrique. En réalité, on peut obtenir une meilleure borne si on évite la majoration uniforme de $\sum_{k=0}^{m-1} \varphi_k(X_1)^2$ dans (1.9). En effet, Comte & Genon-Catalot (2018) montrent que si les X_i admettent un moment d'ordre $-1/2$ (cas Laguerre) ou d'ordre $2/3$ (cas Hermite), alors :

$$\mathbb{E}\|\widehat{f}_m - f_m\|_{\mathrm{L}^2}^2 \leq \frac{1}{n} \sum_{k=0}^{m-1} \mathbb{E}\big[\varphi_k(X_1)^2\big] \leq C \frac{\sqrt{m}}{n},$$

où $C > 0$ est une constante qui dépend de $\mathbb{E}[X_1^{-1/2}]$ (cas Laguerre) ou de $\mathbb{E}[|X_1|^{2/3}]$ (cas Hermite). La majoration du risque sur les espaces de Sobolev–Laguerre ou Sobolev–Hermite est alors donnée par :

$$\sup_{f \in \mathrm{W}^s(A,L)} \mathbb{E}\|f - \widehat{f}_m\|_{\mathrm{L}^2}^2 \leq C(s, L) m^{-s} + C' \frac{\sqrt{m}}{n}.$$

Cette majoration ressemble à celle obtenue avec la base trigonométrique (1.11), en remplaçant « m » par « $\sqrt{m}$ ». Ainsi, la vitesse d'estimation sur les boules de Sobolev–Laguerre ou Sobolev–Hermite est $n^{-2s/(2s+1)}$ avec s est la régularité de f, c'est-à-dire la même que vitesse que celle obtenue sur les boules de Sobolev lorsque $A = [0,1]$.

Comme vu dans l'exemple précédent, le modèle oracle dépend de la régularité de la fonction inconnue f et n'est donc pas accessible en pratique. Notre objectif dans la suite est de choisir $\widehat{\boldsymbol{m}}$ dans une collection $\mathscr{M}_n$ uniquement à partir des observations tel que le risque de $\widehat{\boldsymbol{f}}_{\widehat{\boldsymbol{m}}}$ soit (presque) aussi bon que le risque oracle pour une large classe de fonctions.

Definition 1.4.3 (Inégalité oracle)**.** Soit $\mathscr{F}$ une collection de fonctions et soit $\mathscr{M}_n$ une collection de modèles. On dit que $\widehat{f}_{\widehat{\boldsymbol{m}}}$ est adaptatif sur $\mathscr{F}$ relativement à l'oracle sur $\mathscr{M}_n$ s'il satisfait une inégalité de la forme suivante :

$$\forall f \in \mathscr{F}, \quad \mathscr{R}(f, \widehat{f}_{\widehat{\boldsymbol{m}}}) \leq C \inf_{\boldsymbol{m} \in \mathscr{M}_n} \mathscr{R}(f, \widehat{f}_{\boldsymbol{m}}) + R_n,$$

où C est une constante absolue et R_n est un terme de reste négligeable.

[2] dans le cas de la base d'Hermite, on a en fait $L(m) \leq C\sqrt{m}$ pour $C > 0$ une constante absolue (numériquement, il semble que $C = 1/\sqrt{\pi}$ soit la constante optimale). Ce résultat est prouvé dans un travail de Comte et Lacour non publié au moment de l'écriture de cette thèse.

En général, le terme de reste est d'ordre $\frac{1}{n}$ ou $\frac{(\log n)^\delta}{n}$ avec $\delta > 0$, là où le terme principal est d'ordre $n^{-\alpha}$ avec $\alpha \in]0,1[$ qui dépend de la régularité de f. Autrement dit, le risque de l'estimateur $\widehat{f}_{\widehat{\boldsymbol{m}}}$ est le même que le risque oracle (à une constante multiplicative et un reste négligeable près) : l'estimateur adapte automatiquement son nombre de coefficients à la régularité de f.

1.4.2 Sélection de modèle par pénalisation

Puisque le modèle oracle minimise le risque parmi les modèles de $\mathscr{M}_n$, une approche naïve serait de choisir le modèle $\boldsymbol{m} \in \mathscr{M}_n$ qui minimise le risque empirique $\gamma_n(\widehat{f}_{\boldsymbol{m}})$. Cependant, on peut remarquer que si $\boldsymbol{m} \leq \boldsymbol{m}'$ (c'est-à-dire si $S_{\boldsymbol{m}} \subseteq S_{\boldsymbol{m}'}$) alors $\widehat{f}_{\boldsymbol{m}}$ est un élément de $S_{\boldsymbol{m}'}$, donc $\gamma_n(\widehat{f}_{\boldsymbol{m}'}) \leq \gamma_n(\widehat{f}_{\boldsymbol{m}})$ par définition de l'estimateur. Cette approche naïve va donc systématiquement sélectionner un modèle de grande dimension. En faisant l'hypothèse que les modèles sont emboités ou qu'il existe un modèle englobant, on se retrouve à choisir le modèle de dimension maximale. D'où l'idée de pénaliser les modèles en fonction de leur dimension.

La sélection de modèle par pénalisation consiste à choisir le modèle de $\mathscr{M}_n$ qui minimise le critère :

$$\widehat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m} \in \mathscr{M}_n} \{\gamma_n(\widehat{f}_{\boldsymbol{m}}) + \operatorname{pen}(\boldsymbol{m})\}, \tag{1.12}$$

où $\operatorname{pen}\colon \mathscr{M}_n \to \mathbb{R}_+$ est un terme de pénalité. Pour les estimateurs par projection, cette pénalité est typiquement choisie de l'ordre du terme de variance, mais peut être plus lourde si $\mathscr{M}_n$ compte de nombreux modèles de même dimension (cf. le théorème 1 de Barron *et al.* (1999) et l'exemple qui suit).

L'idée de sélectionner un modèle via l'optimisation d'un critère pénalisé par la complexité du modèle remonte aux travaux de Akaike (1973) (critère AIC) et de Mallows (1973) (C_p de Mallows). Citons également le critère BIC de Schwarz (1978).

- En estimation de densité, le critère AIC (*Akaike Information Criterion*) sélectionne le modèle qui maximise la log-vraisemblance pénalisée par la dimension du modèle, ce qui revient à (1.12) avec :

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \log t(X_i), \quad \operatorname{pen}(\boldsymbol{m}) = \frac{D_{\boldsymbol{m}}}{n}.$$

- Le critère BIC (*Bayesian Information Criterion*) est similaire au critère AIC mais avec une pénalité plus lourde faisant intervenir le log du nombre d'observations :

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \log t(X_i), \quad \operatorname{pen}(\boldsymbol{m}) = \frac{D_{\boldsymbol{m}} \log n}{2n}.$$

- En régression (linéaire), le C_p de Mallows est un critère des moindres carrés pénalisé par $2\sigma^2 \frac{D_{\boldsymbol{m}}}{n}$, c'est-à-dire (1.12) avec :

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - t(X_i)\big)^2, \quad \mathrm{pen}(\boldsymbol{m}) = 2\sigma^2\frac{D_{\boldsymbol{m}}}{n}.$$

Ces trois critère reposent sur des justifications asymptotiques (théorème de Wilks pour AIC, approximation de Laplace pour BIC) et supposent donc une collection de modèles fixée qui ne dépend pas de n. L'étude non asymptotique de la sélection de modèle par pénalisation a été développée par Birgé & Massart (1993, 1997) et Barron *et al.* (1999), et s'appuie sur l'utilisation d'inégalités de concentration. On pourra se référer au cours de Massart (2007) à l'école d'été de Saint-Flour pour une présentation approfondie de la question.

Exemple 1.4 (Estimation d'une densité)**.** On poursuit l'exemple 1.3. On considère la collection de modèles $\mathscr{M}_n := \{\boldsymbol{m} \in \mathbb{N}_+^d \,|\, D_{\boldsymbol{m}} \leq n\}$. Dans le cas de l'estimation de densité, le risque empirique s'écrit :

$$\begin{aligned}
\gamma_n(\widehat{f}_{\boldsymbol{m}}) &= \|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{\boldsymbol{m}}(X_i) \\
&= \|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\widehat{a}_{\boldsymbol{k}}\,\varphi_{\boldsymbol{k}}(X_i) \\
&= \|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - 2\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\widehat{a}_{\boldsymbol{k}}^2 \\
&= \|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - 2\|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = -\|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2.
\end{aligned}$$

Cette écriture offre une autre interprétation du critère (1.12). En remarquant que le terme de biais s'écrit $\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \|f\|_{\mathrm{L}^2}^2 - \|f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ par le théorème de Pythagore, le modèle oracle vérifie :

$$\begin{aligned}
\boldsymbol{m}_o &= \operatorname*{arg\,min}_{\boldsymbol{m}\in\mathscr{M}_n}\big\{\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2\big\} \\
&= \operatorname*{arg\,min}_{\boldsymbol{m}\in\mathscr{M}_n}\big\{-\|f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2\big\},
\end{aligned}$$

puisque le terme $\|f\|_{\mathrm{L}^2}^2$ ne dépend pas de $\boldsymbol{m}$. Le critère (1.12) s'écrit quant à lui :

$$\widehat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m}\in\mathscr{M}_n}\big\{-\|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \mathrm{pen}(\boldsymbol{m})\big\}.$$

Autrement dit, on a estimé $\|f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ par $\|\widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ et remplacé le terme de variance par le terme de pénalité. Cela suggère de choisir la pénalité du même ordre de grandeur que le terme de variance.

Le terme de variance est majoré par $L(\boldsymbol{m})/n$, où $L(\boldsymbol{m})$ est défini par (1.10). Supposons que la quantité $L(\boldsymbol{m})$ vérifie $L(\boldsymbol{m}) \leq \Phi D_{\boldsymbol{m}}/n$, avec Φ une constante positive qui dépend de la base. On a vu avec l'exemple 1.3 et la remarque 1.4.2

que c'était le cas pour la base trigonométrique, la base de Laguerre et la base d'Hermite. Le terme de variance est majoré par $\Phi D_{\boldsymbol{m}}/n$, donc on choisit une pénalité de la forme :

$$\mathrm{pen}(\boldsymbol{m}) := \kappa\Phi\frac{D_{\boldsymbol{m}}}{n},$$

avec κ une constante à ajuster plus tard. Pour établir une inégalité oracle, on commence par écrire :

$$\begin{aligned}\gamma_n(\widehat{f}_{\widehat{\boldsymbol{m}}}) + \mathrm{pen}(\widehat{\boldsymbol{m}}) &\leq \gamma_n(\widehat{f}_{\boldsymbol{m}}) + \mathrm{pen}(\boldsymbol{m}) && \text{(par définition de } \widehat{\boldsymbol{m}}\text{)}\\ &\leq \gamma_n(f_{\boldsymbol{m}}) + \mathrm{pen}(\boldsymbol{m}). && \text{(par définition de } \widehat{f}_{\boldsymbol{m}}\text{)} \quad (1.13)\end{aligned}$$

Pour tout $t \in \mathrm{L}^2(A)$, on écrit le contraste en faisant apparaître la perte L^2 :

$$\gamma_n(t) = \|t - f\|^2_{\mathrm{L}^2} - \frac{2}{n}\sum_{i=1}^{n}\left(t(\boldsymbol{X}_i) - \langle t, f\rangle_{\mathrm{L}^2}\right) - \|f\|^2_{\mathrm{L}^2}.$$

Introduisons ν_n le processus empirique défini pour tout $t \in \mathrm{L}^2(A)$ par :

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(t(\boldsymbol{X}_i) - \langle t, f\rangle_{\mathrm{L}^2}\right), \tag{1.14}$$

alors (1.13) se réécrit :

$$\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2_{\mathrm{L}^2} \leq \|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \mathrm{pen}(\boldsymbol{m}) + 2\nu_n\left(\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\right) - \mathrm{pen}(\widehat{\boldsymbol{m}}). \tag{1.15}$$

En utilisant la linéarité de $t \mapsto \nu_n(t)$ et l'inégalité $2ab \leq \frac{1}{2}a^2 + 2b^2$ valable pour tous a, b positifs, on a :

$$\begin{aligned}2\nu_n\left(\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\right) &= 2\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}\nu_n\left(\frac{\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}}{\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}}\right)\\ &= 2\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f\|_{\mathrm{L}^2}\nu_n\left(\frac{\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}}{\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}}\right) + 2\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}\nu_n\left(\frac{\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}}{\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}}\right)\\ &\leq \frac{1}{2}\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \frac{1}{2}\|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + 4\nu_n^2\left(\frac{\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}}{\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}}\right)\\ &\leq \frac{1}{2}\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \frac{1}{2}\|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + 4\sup_{t\in\mathscr{B}(\boldsymbol{m},\widehat{\boldsymbol{m}})}\nu_n^2(t),\end{aligned}$$

où $\mathscr{B}(\boldsymbol{m}, \boldsymbol{m}')$ est défini comme :

$$\mathscr{B}(\boldsymbol{m}, \boldsymbol{m}') := \left\{t \in S_{\boldsymbol{m}} + S_{\boldsymbol{m}'} \,\middle|\, \|t\|_{\mathrm{L}^2} = 1\right\}.$$

En injectant l'inégalité précédente dans (1.15), on obtient :

$$\frac{1}{2}\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2_{\mathrm{L}^2} \leq \frac{3}{2}\|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \mathrm{pen}(\boldsymbol{m}) + 4\sup_{t\in\mathscr{B}(\boldsymbol{m},\widehat{\boldsymbol{m}})}\nu_n^2(t) - \mathrm{pen}(\widehat{\boldsymbol{m}}).$$

Posons $p(\boldsymbol{m},\boldsymbol{m}') := \frac{\kappa}{4}\Phi\frac{D_{\boldsymbol{m}}+D_{\boldsymbol{m}'}}{n} = \frac{\mathrm{pen}(\boldsymbol{m})+\mathrm{pen}(\boldsymbol{m}')}{4}$, alors :

$$\begin{aligned}\frac{1}{2}\|f-\widehat{f}_{\widehat{\boldsymbol{m}}}\|_{\mathrm{L}^2}^2 &\leq \frac{3}{2}\|f-f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + 2\,\mathrm{pen}(\boldsymbol{m}) + 4\left(\sup_{t\in\mathscr{B}(\boldsymbol{m},\widehat{\boldsymbol{m}})} \nu_n^2(t) - p(\boldsymbol{m},\widehat{\boldsymbol{m}})\right)\\ &\leq \frac{3}{2}\|f-f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + 2\,\mathrm{pen}(\boldsymbol{m}) + 4\sum_{\boldsymbol{m}'\in\mathscr{M}_n}\left(\sup_{t\in\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) - p(\boldsymbol{m},\boldsymbol{m}')\right)_+.\end{aligned}$$

En prenant l'espérance, on obtient :

$$\mathbb{E}\|f-\widehat{f}_{\widehat{\boldsymbol{m}}}\|_{\mathrm{L}^2}^2 \leq 4\left(\|f-f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \mathrm{pen}(\boldsymbol{m})\right) + 8\,R_n,$$

où le terme de reste R_n est donné par :

$$R_n := \sum_{\boldsymbol{m}'\in\mathscr{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) - p(\boldsymbol{m},\boldsymbol{m}')\right)_+\right]. \tag{1.16}$$

En utilisant des inégalités de concentration sur les suprema de processus empiriques comme l'inégalité de Talagrand (théorème 1.4.8), on montre qu'il existe une constante $\kappa_0 > 0$ ($\kappa_0 = 16$ convient dans ce cas) telle que si $\kappa > \kappa_0$ alors $R_n \leq Cn^{-1}$, ce qui montre que $\widehat{f}_{\widehat{\boldsymbol{m}}}$ vérifie une inégalité oracle (voir l'exemple 1.6).

Remarque 1.4.4. Pour les bases de Laguerre et d'Hermite, on a vu dans la remarque 1.4.2 que sous des hypothèses de moment, le terme de variance est en fait d'ordre $\sqrt{D_{\boldsymbol{m}}}/n$. La pénalité utilisée dans l'exemple précédent est donc trop lourde et ne conduira pas à la vitesse optimale. Dans ce cas, on choisit une pénalité de la forme $\mathrm{pen}(\boldsymbol{m}) := \kappa\sqrt{D_{\boldsymbol{m}}}/n$ et on démontre un résultat analogue. Pour éviter ces problèmes d'évaluation de l'ordre de grandeur du terme de variance, une solution est d'utiliser une pénalité aléatoire :

$$\widehat{\mathrm{pen}}(\boldsymbol{m}) := \kappa\frac{\widehat{V}_n(\boldsymbol{m})}{n}, \quad \widehat{V}_n(\boldsymbol{m}) := \frac{1}{n}\sum_{i=1}^n\left(\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)^2\right).$$

C'est ce qui est fait par Belomestny *et al.* (2019) et Comte & Genon-Catalot (2018) en dimension 1. Ces articles établissent une inégalité oracle sur la collection de modèles $\mathscr{M}_n := \left\{m\in\mathbb{N}_+ \,\middle|\, m\leq(n/\log n)^\alpha\right\}$ avec $\alpha = 1$ dans le cas Laguerre et $\alpha = \frac{6}{5}$ dans le cas Hermite.

Dans le cas de l'estimation d'une densité, la pénalité est choisie de la forme $\mathrm{pen}(\boldsymbol{m}) = \kappa\mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2$ et l'inégalité oracle est valable si $\kappa > \kappa_0$ pour κ_0 une constante absolue. La valeur de κ_0 trouvée par la théorie est trop grande en pratique, il faut donc avoir recours à d'autres méthodes pour calibrer la pénalité.

Pour résoudre la question de la calibration de la pénalité à partir des données, Birgé & Massart (2007) ont introduit les notions de pénalités minimales et pénalités optimales. Les auteurs démontrent dans le cadre du modèle de bruit blanc gaussien que pour une pénalité connue à une constante multiplicative

près, il existe un phénomène de pénalité minimale : si $\text{pen}(m) = \kappa \, \text{pen}_{\text{min}}(m)$ avec $\kappa < 1$, la dimension du modèle sélectionné est proche de celle des modèle de plus grande dimension tandis que si $\kappa > 1$, l'estimateur vérifie une inégalité oracle. Ils démontrent également que le choix $\kappa = 2$ donne lieu à une pénalité optimale en un certain sens. Ils proposent alors une méthode heuristique, appelée *heuristique de pente*, pour déterminer la forme de pénalité minimale à partir des données, puis en déduire la pénalité optimale. Ces résultats ont ensuite été étendus par Arlot & Massart (2009) dans un modèle de régression hétéroscédastique *random design* non gaussien.

Une autre heuristique basée sur ce phénomène d'explosion de la dimension sélectionnée est l'*heuristique des sauts de dimension*. L'idée est de calculer la dimension sélectionnée en fonction de κ. La fonction obtenue est décroissante et constante par morceaux. On choisit alors κ_{min} comme la valeur de κ correspondant au plus grand saut de dimension ou au premier saut de taille supérieure à un seuil. La valeur de κ finalement retenue est $2 \times \kappa_{\text{min}}$. C'est cette heuristique que nous utiliserons.

L'heuristique de pente et l'heuristique des sauts de dimension ont été implémentées dans le package CAPUSHE développé par Baudry *et al.* (2012).

1.4.3 Méthode de Goldenshluger et Lepski

Une deuxième façon de sélectionner $\boldsymbol{m}$ est la méthode dite « de Goldenshluger et Lepski ». Cette méthode est issue des travaux de Goldenshluger & Lepski (2008, 2009, 2011) et a été développée pour faire de la sélection de fenêtre pour les estimateurs à noyau. Elle a ensuite été adaptée pour faire de la sélection de modèle par Chagny (2013b,c) dans des contextes d'estimation d'une fonction de régression et d'une densité conditionnelle. Dans ces deux articles, la fonction estimée ne dépend que d'une ou deux variables. Dans cette section, nous expliquons le principe général de cette méthode pour l'estimation par projection d'une fonction de d variables.

Les avantages de cette méthode sont ses bonnes propriétés théoriques et sa généralité. En effet, la seule hypothèse de la méthode concerne la collection de modèles : il faut que celle-ci soit stable par minimum.

Hypothèse 1.1 (min-stabilité)**.** Pour tous $\boldsymbol{m}, \boldsymbol{m}' \in \mathscr{M}_n$, on a $\boldsymbol{m} \wedge \boldsymbol{m}' \in \mathscr{M}_n$.

Remarque 1.4.5. Cette hypothèse revient à dire que $\{S_{\boldsymbol{m}} : \boldsymbol{m} \in \mathscr{M}_n\}$ est stable par intersection.

Soit $V(\boldsymbol{m})$ une fonction croissante[3] sur $\mathbb{N}_+^d$ telle que $\mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2 \leq V(\boldsymbol{m})$. La décomposition biais-variance (1.7) prend donc la forme :

$$\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|^2 \leq \|f - f_{\boldsymbol{m}}\|^2 + V(\boldsymbol{m}).$$

[3] croissante pour l'ordre partiel sur $\mathbb{N}_+^d$, c'est-à-dire coordonnée par coordonnée.

La méthode de Goldenshluger et Lepski diffère de la sélection de modèle dans la façon d'estimer le biais : avec cette méthode, le biais est estimé en comparant les estimateurs deux à deux. Une heuristique qui guide la construction de l'estimateur est la suivante. Le terme de biais s'écrit :

$$\|f - f_{\boldsymbol{m}}\|^2 = \|f - \Pi_{\boldsymbol{m}}(f)\|^2,$$

où $\Pi_{\boldsymbol{m}}$ désigne le projecteur orthogonal sur $S_{\boldsymbol{m}}$. On approche f par sa projection sur $S_{\boldsymbol{m}'}$ pour $\boldsymbol{m}'$ un modèle de $\mathscr{M}_n$:

$$\|f - \Pi_{\boldsymbol{m}}(f)\|^2 \approx \|f_{\boldsymbol{m}'} - \Pi_{\boldsymbol{m}}(f_{\boldsymbol{m}'})\|^2 = \|f_{\boldsymbol{m}'} - f_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2,$$

puisque $\Pi_{\boldsymbol{m}} \circ \Pi_{\boldsymbol{m}'}$ est le projecteur sur $S_{\boldsymbol{m}} \cap S_{\boldsymbol{m}'} = S_{\boldsymbol{m}\wedge\boldsymbol{m}'}$. Cette dernière quantité est estimée en remplaçant les projections par leurs estimateurs. Cependant, on introduit ainsi de l'aléatoire qui nécessite une correction. Cette correction s'effectue en retranchant le terme de variance de $\widehat{f}_{\boldsymbol{m}'}$ (le terme de variance de $\widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}$ étant majoré par celui de $\widehat{f}_{\boldsymbol{m}'}$) :

$$\left(\|\widehat{f}_{\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 - \kappa_1 V(\boldsymbol{m}')\right)_+, \tag{1.17}$$

la partie positive nous assurant que le résultat reste positif (le biais est positif) et où κ_1 est une constante positive à ajuster. Finalement, le choix de $\boldsymbol{m}'$ est arbitraire, donc on calcule (1.17) pour tout $\boldsymbol{m}' \in \mathscr{M}_n$ et on prend le maximum :

$$A(\boldsymbol{m}) := \max_{\boldsymbol{m}'\in\mathscr{M}_n} \left(\|\widehat{f}_{\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 - \kappa_1 V(\boldsymbol{m}')\right)_+. \tag{1.18}$$

Cette heuristique est justifié par le lemme 1.4.6 ci-dessous dans lequel on montre que :

$$A(\boldsymbol{m}) \le C\|f - f_{\boldsymbol{m}}\|^2 + R_n,$$

avec $C = 3$ et R_n un reste qui est lié aux déviations à droite de la partie « variance » du risque.

Le modèle est sélectionné en faisant un compromis entre la version empirique du biais (1.18) et le majorant de la variance :

$$\widehat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m}\in\mathscr{M}_n} \left\{A(\boldsymbol{m}) + \kappa_2 V(\boldsymbol{m})\right\}, \tag{1.19}$$

où κ_2 est une deuxième constante qu'il faut ajuster.

Lemme 1.4.6. *On suppose que $\mathscr{M}_n$ vérifie l'hypothèse 1.1. Soit $\widehat{\boldsymbol{m}}$ défini par* (1.19). *Si $\kappa_2 \ge \kappa_1 > 0$, alors pour tout $\boldsymbol{m} \in \mathscr{M}_n$, on a :*

$$\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \le C\|f - f_{\boldsymbol{m}}\|^2 + C'\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2 + (\kappa_1 + \kappa_2)V(\boldsymbol{m})\right) + C'' R_n,$$

avec :

$$R_n := \max_{\boldsymbol{m}'\in\mathscr{M}_n} \left(\|\widehat{f}_{\boldsymbol{m}'} - f_{\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6} V(\boldsymbol{m}')\right)_+. \tag{1.20}$$

Si on prend l'espérance dans l'inégalité du lemme précédent, on obtient une inégalité de forme :

$$\mathbb{E}\|f-\widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq C(\kappa_1,\kappa_2)\inf_{\boldsymbol{m}\in\mathscr{M}_n}\left(\|f-f_{\boldsymbol{m}}\|^2+V(\boldsymbol{m})\right)+C'\mathbb{E}[R_n].$$

Pour obtenir une inégalité oracle pour $\widehat{f}_{\widehat{\boldsymbol{m}}}$, il faut majorer l'espérance du terme de reste (1.20). En majorant le maximum par la somme sur tous les termes, on obtient :

$$\mathbb{E}[R_n] \leq \sum_{\boldsymbol{m}\in\mathscr{M}_n}\mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}}-f_{\boldsymbol{m}}\|^2-\frac{\kappa_1}{6}V(\boldsymbol{m})\right)_+\right].$$

Puisque $V(\boldsymbol{m})$ est un majorant du terme de variance, on voit que le problème revient à étudier les déviations de $\|\widehat{f}_{\boldsymbol{m}}-f_{\boldsymbol{m}}\|^2$ à droite de son espérance, pour tous les modèle de $\mathscr{M}_n$. En utilisant des résultats de concentration de la mesure dépendant du problème (estimation d'une densité, d'une régression, etc), on démontre qu'il existe une constant absolue $\kappa_0>0$ telle que pour tous $\kappa_2\geq\kappa_1>\kappa_0$, $\mathbb{E}[R_n]$ soit majoré par un terme d'ordre n^{-1} ou $n^{-1}(\log n)^{\delta}$.

Exemple 1.5 (Estimation d'une densité)**.** Le terme de variance est majoré par $V(\boldsymbol{m}) := \Phi\frac{D_{\boldsymbol{m}}}{n}$. Pour étudier le terme de reste (1.20), on écrit :

$$\|\widehat{f}_{\boldsymbol{m}}-f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2=\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\left(\widehat{a}_{\boldsymbol{k}}-a_{\boldsymbol{k}}\right)^2=\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\nu_n^2(\varphi_{\boldsymbol{k}}),$$

où ν_n est le processus empirique défini par (1.14). Soit $t=\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}a_{\boldsymbol{k}}\varphi_{\boldsymbol{k}}$ un élément de $S_{\boldsymbol{m}}$. Par l'inégalité de Cauchy–Schwarz, on a :

$$\begin{aligned}\nu_n(t)^2=\left(\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}a_{\boldsymbol{k}}\nu_n(\varphi_{\boldsymbol{k}})\right)^2 &\leq\left(\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}a_{\boldsymbol{k}}^2\right)\left(\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\nu_n(\varphi_{\boldsymbol{k}})^2\right)\\ &=\|t\|_{\mathrm{L}^2}^2\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\nu_n(\varphi_{\boldsymbol{k}})^2,\end{aligned}$$

avec égalité si les $a_{\boldsymbol{k}}$ sont proportionnels aux $\nu_n(\varphi_{\boldsymbol{k}})$. Ainsi, on a :

$$\|\widehat{f}_{\boldsymbol{m}}-f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2=\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}}\nu_n^2(\varphi_{\boldsymbol{k}})=\sup_{\substack{t\in S_{\boldsymbol{m}}\\ \|t\|_{\mathrm{L}^2}=1}}\nu_n^2(t).$$

Comme pour la sélection de modèle par pénalisation, la majoration du reste passe par l'étude des déviations du supremum du processus empirique ν_n.

En pratique, comme pour la sélection de modèle par pénalisation, la valeur de κ_0 obtenue par la théorie est trop grande et les constantes κ_1 et κ_2 doivent être calibrée numériquement. Le fait que deux constantes ayant des rôles différents doivent être calibrées simultanément rend cette tâche compliquée en pratique. Dans le cas de la méthode de Goldenshluger & Lepski (2011) pour la sélection de fenêtre d'un estimateur à noyau d'une densité, Lacour & Massart (2016) montre que si on prend $\kappa_1=\kappa_2=:a$ et V la variance intégrée de l'estimateur à noyau,

alors il existe une valeur critique a_0 (qui vaut 1 dans leur cas) telle que pour $a > a_0$, l'estimateur est adaptatif et pour $a < a_0$, le risque ne converge pas vers 0. Ils proposent donc d'utiliser un analogue de la méthode des sauts de dimension pour déterminer la valeur de a. Finalement, ils choisissent $\kappa_1 = a$ et $\kappa_2 = 2a$, l'utilisation de constantes différentes semblant donner de meilleurs résultats.

La méthode de Goldenshluger et Lepski sera utilisée dans le chapitre 2. Le reste de cette section est consacré à la démonstration du lemme 1.4.6.

Démonstration du lemme 1.4.6. Soit $\boldsymbol{m} \in \mathcal{M}_n$, on a par inégalité triangulaire :

$$\|\widehat{f}_{\widehat{\boldsymbol{m}}} - f\|^2 \leq 3\|\widehat{f}_{\widehat{\boldsymbol{m}}} - \widehat{f}_{\boldsymbol{m}\wedge\widehat{\boldsymbol{m}}}\|^2 + 3\|\widehat{f}_{\boldsymbol{m}\wedge\widehat{\boldsymbol{m}}} - \widehat{f}_{\boldsymbol{m}}\|^2 + 3\|\widehat{f}_{\boldsymbol{m}} - f\|^2.$$

Par définition de $A(\boldsymbol{m})$ et $\widehat{\boldsymbol{m}}$, et puisque l'on a $\kappa_1 \leq \kappa_2$, les deux premiers termes sont majorés par :

$$\begin{aligned}\|\widehat{f}_{\widehat{\boldsymbol{m}}} - \widehat{f}_{\boldsymbol{m}\wedge\widehat{\boldsymbol{m}}}\|^2 + \|\widehat{f}_{\boldsymbol{m}\wedge\widehat{\boldsymbol{m}}} - \widehat{f}_{\boldsymbol{m}}\|^2 &\leq A(\boldsymbol{m}) + \kappa_1 V(\widehat{\boldsymbol{m}}) + A(\widehat{\boldsymbol{m}}) + \kappa_1 V(\boldsymbol{m}) \\ &\leq 2A(\boldsymbol{m}) + (\kappa_1 + \kappa_2) V(\boldsymbol{m}).\end{aligned}$$

Ainsi, on obtient :

$$\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq 6\,A(\boldsymbol{m}) + 3(\kappa_1 + \kappa_2)\,V(\boldsymbol{m}) + 3\,\|f - \widehat{f}_{\boldsymbol{m}}\|^2. \tag{1.21}$$

Il reste à montrer que $A(\boldsymbol{m})$ est majoré par le terme de biais, plus un reste. Soit $\boldsymbol{m}' \in \mathcal{M}_n$, on a :

$$\|\widehat{f}_{\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 \leq 3\|\widehat{f}_{\boldsymbol{m}'} - f_{\boldsymbol{m}'}\|^2 + 3\|f_{\boldsymbol{m}'} - f_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 + 3\|f_{\boldsymbol{m}\wedge\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2.$$

Ainsi, on obtient la majoration :

$$A(\boldsymbol{m}) \leq 3\left[R_n + R'_n(\boldsymbol{m}) + B_n(\boldsymbol{m})\right], \tag{1.22}$$

où R_n, $R'_n(\boldsymbol{m})$ et $B_n(\boldsymbol{m})$ sont donnés par :

$$\begin{aligned}R_n &:= \max_{\boldsymbol{m}'\in\mathcal{M}_n}\left(\|\widehat{f}_{\boldsymbol{m}'} - f_{\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6}V(\boldsymbol{m}')\right)_+, \\ R'_n(\boldsymbol{m}) &:= \max_{\boldsymbol{m}'\in\mathcal{M}_n}\left(\|f_{\boldsymbol{m}\wedge\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6}V(\boldsymbol{m}')\right)_+, \\ B_n(\boldsymbol{m}) &:= \max_{\boldsymbol{m}'\in\mathcal{M}_n}\|f_{\boldsymbol{m}'} - f_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2.\end{aligned}$$

• Montrons que $R'_n(\boldsymbol{m})$ est borné par R_n :

$$\begin{aligned}R'_n(\boldsymbol{m}) &\leq \max_{\boldsymbol{m}'\in\mathcal{M}_n}\left(\|f_{\boldsymbol{m}\wedge\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6}V(\boldsymbol{m}\wedge\boldsymbol{m}')\right)_+ && (V \text{ croissante}) \\ &\leq \max_{\boldsymbol{m}''\in\mathcal{M}_n}\left(\|\widehat{f}_{\boldsymbol{m}''} - f_{\boldsymbol{m}''}\|^2 - \frac{\kappa_1}{6}V(\boldsymbol{m}'')\right)_+ && (\text{hypothèse 1.1}) \\ &= R_n. && (1.23)\end{aligned}$$

• Montrons que $B_n(\boldsymbol{m})$ est majoré par le terme de biais. Soit $\boldsymbol{m}' \in \mathscr{M}_n$, on introduit l'ensemble $I_{\boldsymbol{m},\boldsymbol{m}'} := \{i \,|\, m_i \leq m'_i\}$ qui sélectionne les coordonnées i pour lesquelles $[\boldsymbol{m} \wedge \boldsymbol{m}']_i = m_i$. Soit $S_{\boldsymbol{m},\boldsymbol{m}'}$ le sous-espace fermé de $\mathrm{L}^2(A,\mu)$:

$$S_{\boldsymbol{m},\boldsymbol{m}'} := \overline{\mathrm{Vect}}\{\varphi_{\boldsymbol{j}} : \boldsymbol{j} \in \mathbb{N}^p \text{ et } \forall i \in I_{\boldsymbol{m},\boldsymbol{m}'},\, j_i \leq m_i - 1\}.$$

Remarquons les relations $S_{\boldsymbol{m}'} \cap S_{\boldsymbol{m},\boldsymbol{m}'} = S_{\boldsymbol{m}\wedge\boldsymbol{m}'}$ et $S_{\boldsymbol{m}} \subseteq S_{\boldsymbol{m},\boldsymbol{m}'}$. On note $\Pi_{\boldsymbol{m}'}$ le projecteur sur $S_{\boldsymbol{m}'}$ et $\Pi_{\boldsymbol{m},\boldsymbol{m}'}$ celui sur $S_{\boldsymbol{m},\boldsymbol{m}'}$. On a :

$$\begin{aligned}
\|f_{\boldsymbol{m}'} - f_{\boldsymbol{m}\wedge\boldsymbol{m}'}\|^2 &= \|\Pi_{\boldsymbol{m}'}(f - \Pi_{\boldsymbol{m},\boldsymbol{m}'} f)\|^2 && (\Pi_{\boldsymbol{m}'} \circ \Pi_{\boldsymbol{m},\boldsymbol{m}'} = \Pi_{\boldsymbol{m}\wedge\boldsymbol{m}'}) \\
&\leq \|f - \Pi_{\boldsymbol{m},\boldsymbol{m}'} f\|^2 && (\Pi_{\boldsymbol{m}'} \text{ 1-Lipschitzienne}) \\
&\leq \|f - f_{\boldsymbol{m}}\|^2. && (S_{\boldsymbol{m}} \subseteq S_{\boldsymbol{m},\boldsymbol{m}'}) \qquad (1.24)
\end{aligned}$$

Cette majoration est valable pour tout $\boldsymbol{m}'$, donc $B_n(\boldsymbol{m}) \leq \|f - f_{\boldsymbol{m}}\|^2$.

En utilisant les majorations (1.22), (1.23) et (1.24) dans l'inégalité (1.21), on obtient :

$$\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq 18\,\|f - f_{\boldsymbol{m}}\|^2 + 36\,R_n + 3(\kappa_1 + \kappa_2)\,V(\boldsymbol{m}) + 3\,\|f - \widehat{f}_{\boldsymbol{m}}\|^2.$$

On conclut en appliquant le théorème de Pythagore à $\|f - \widehat{f}_{\boldsymbol{m}}\|^2$:

$$\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq 21\,\|f - f_{\boldsymbol{m}}\|^2 + 3\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|^2 + (\kappa_1 + \kappa_2)V(\boldsymbol{m})\right) + 36\,R_n. \qquad \square$$

1.4.4 Inégalités de concentration

Nous présentons ici quelques résultats techniques de concentration de la mesure que nous utiliserons au cours de la thèse.

Suprema de processus empiriques

Comme on l'a vu avec l'exemple de l'estimation d'une densité, que ce soit avec la méthode de contraste pénalisé ou celle de Goldenshluger et Lepski, établir une inégalité oracle nécessite de contrôler les déviations du supremum d'un processus empirique par rapport à sa moyenne. Nous utiliserons pour ce faire une inégalité de concentration de Klein & Rio (2005) basée sur les résultats de Talagrand (1996) et Ledoux (1997).

Théorème 1.4.7 (Inégalité de Talagrand, borne en probabilité)**.** *Soit $n \in \mathbb{N}_+$ et soit $\mathscr{F}$ une collection au plus dénombrable de fonctions définies sur un espace polonais E et à valeurs dans $[-1,1]^n$. Soient $\xi_1,\ldots,\xi_n$ des variables aléatoires indépendantes à valeurs dans E. On suppose que $\mathbb{E}[f_i(\xi_i)] = 0$ pour toute $\boldsymbol{f} = (f_1,\ldots,f_n)$ de $\mathscr{F}$. On considère :*

$$Z := \sup_{\boldsymbol{f}\in\mathscr{F}} \sum_{i=1}^{n} f_i(\xi_i), \quad V := \sup_{\boldsymbol{f}\in\mathscr{F}} \mathrm{Var}\left(\sum_{i=n}^{n} f_i(\xi_i)\right).$$

Posons $v := 2\mathbb{E}[Z] + V$, alors on a pour tout $x > 0$:

$$\mathbb{P}\big[Z \geq \mathbb{E}[Z] + x\big] \leq \exp\left(-\frac{x^2}{2v + x}\right).$$

En intégrant l'inégalité du théorème précédent, il est possible d'obtenir une borne sur l'espérance des déviations de Z à droite de son espérance. C'est cette version que nous utiliserons dans les chapitres suivants. La démonstration du résultat qui suit est assez technique et pourra être consultée dans le chapitre 2 de la thèse de Chagny (2013a).

Théorème 1.4.8 (Inégalité de Talagrand, borne en espérance)**.** *Soit $n \in \mathbb{N}_+$ et soit $\mathscr{F}$ une collection au plus dénombrable de fonctions mesurables sur un espace polonais E. Soient $\xi_1, \dots, \xi_n$ des variables aléatoires indépendantes à valeurs dans E. On considère ν_n, le processus empirique centré suivant :*

$$\nu_n(f) := \frac{1}{n}\sum_{i=1}^{n}\big(f(\xi_i) - \mathbb{E}\big[f(\xi_i)\big]\big), \quad f \in \mathscr{F}.$$

On suppose qu'il existe trois constantes positives M, H et v telles que :

$$\sup_{f\in\mathscr{F}} \|f\|_\infty \le M, \qquad \mathbb{E}\left[\sup_{f\in\mathscr{F}} |\nu_n(f)|\right] \le H, \qquad \sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\big(f(\xi_i)\big) \le v.$$

Alors pour tout $\delta > 0$, on a l'inégalité suivante :

$$\mathbb{E}\left[\left(\sup_{f\in\mathscr{F}} \nu_n^2(f) - 2(1+2\delta)H^2\right)_+\right] \le \frac{4}{K}\left(\frac{v}{n}\mathrm{e}^{-K\delta\frac{nH^2}{v}} + \frac{49\,M^2}{K\,C(\delta)^2\,n^2}\mathrm{e}^{-\frac{KC(\delta)\sqrt{2\delta}}{7}\frac{nH}{M}}\right),$$

avec $C(\delta) := \left(\sqrt{1+\delta} - 1\right) \wedge 1$ et $K = 1/6$.

Remarque 1.4.9. Ce théorème est énoncé pour des collections dénombrables de fonctions mais s'étend à des collections non dénombrables si celles-ci admettent une partie dénombrable dense et si ν_n est continue. Ce sera toujours le cas dans notre travail, puisque les collections considérées sont des boules d'espaces euclidiens.

Exemple 1.6 (Estimation d'une densité)**.** Que ce soit pour contrôler (1.16) pour la sélection de modèle par pénalisation ou (1.20) pour la méthode de Goldenshluger et Lepski, on a vu qu'il était nécessaire de contrôler les déviations du supremum du processus empirique ν_n (défini par (1.14)). On utilise pour cela l'inégalité du théorème 1.4.8. On traite ici l'exemple issu de la méthode par pénalisation, l'exemple pour la méthode de Goldenshluger et Lepski étant analogue. Notre but est de contrôler :

$$R_n := \sum_{\boldsymbol{m}'\in\mathscr{M}_n} \mathbb{E}\left[\left(\sup_{t\in\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) - p(\boldsymbol{m},\boldsymbol{m}')\right)_+\right],$$

où $\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')$ et $p(\boldsymbol{m},\boldsymbol{m}')$ sont définis par :

$$\mathscr{B}(\boldsymbol{m},\boldsymbol{m}') := \big\{t \in S_{\boldsymbol{m}} + S_{\boldsymbol{m}'} \,\big|\, \|t\|_{\mathrm{L}^2} = 1\big\}, \quad p(\boldsymbol{m},\boldsymbol{m}') := \frac{\kappa}{4}\Phi\frac{D_{\boldsymbol{m}} + D_{\boldsymbol{m}'}}{n},$$

et où $\mathscr{M}_n$ est la collection de modèles définie par :

$$\mathscr{M}_n := \left\{ \boldsymbol{m} \in \mathbb{N}_+^d \,\middle|\, D_{\boldsymbol{m}} \leq n \right\}.$$

Une base orthonormée de $S_{\boldsymbol{m}} + S_{\boldsymbol{m}'}$ est $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')}$ où $\mathscr{K}(\boldsymbol{m},\boldsymbol{m}')$ est défini comme :

$$\mathscr{K}(\boldsymbol{m},\boldsymbol{m}') := \left\{ \boldsymbol{k} \in \mathbb{N}_+^d \,\middle|\, \boldsymbol{k} \leq \boldsymbol{m} - \mathbf{1} \text{ ou } \boldsymbol{k} \leq \boldsymbol{m}' - \mathbf{1} \right\}.$$

Dans la suite, on note $f_{\boldsymbol{m},\boldsymbol{m}'}$ la projection de f sur $S_{\boldsymbol{m}} + S_{\boldsymbol{m}'}$, et $\widehat{f}_{\boldsymbol{m},\boldsymbol{m}'}$ l'estimateur par projection sur ce même espace. Déterminons les paramètres H, v et M du théorème 1.4.8 :

- *Valeur de H.*

$$\begin{aligned}
\mathbb{E}\left[\sup_{t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n(t) \right]^2 &\leq \mathbb{E}\left[\sup_{t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) \right] && \text{(Jensen)} \\
&= \mathbb{E}\left\| \widehat{f}_{\boldsymbol{m},\boldsymbol{m}'} - f_{\boldsymbol{m},\boldsymbol{m}'} \right\|_{\mathrm{L}^2}^2 \\
&= \sum_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')} \operatorname{Var}(\widehat{a}_{\boldsymbol{k}}) \\
&= \frac{1}{n} \sum_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')} \operatorname{Var}\left(\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)\right) \\
&\leq \frac{1}{n} \sum_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')} \mathbb{E}\left[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)^2\right] \\
&\leq \frac{1}{n} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \mathbb{E}\left[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)^2\right] + \frac{1}{n} \sum_{\boldsymbol{k} \leq \boldsymbol{m}'-\mathbf{1}} \mathbb{E}\left[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_1)^2\right] \\
&\leq \Phi \frac{D_{\boldsymbol{m}} + D_{\boldsymbol{m}'}}{n} =: H^2.
\end{aligned}$$

- *Valeur de M.* Soit $t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')$, on note $a_{\boldsymbol{k}}$ ses coefficients. Alors par l'inégalité de Cauchy–Schwarz, on a pour tout $\boldsymbol{x} \in A$:

$$\begin{aligned}
t(\boldsymbol{x})^2 &\leq \left(\sum_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')} a_{\boldsymbol{k}}^2 \right) \left(\sum_{\boldsymbol{k} \in \mathscr{K}(\boldsymbol{m},\boldsymbol{m}')} \varphi_{\boldsymbol{k}}(\boldsymbol{x})^2 \right) \\
&\leq \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \varphi_{\boldsymbol{k}}(\boldsymbol{x})^2 + \sum_{\boldsymbol{k} \leq \boldsymbol{m}'-\mathbf{1}} \varphi_{\boldsymbol{k}}(\boldsymbol{x})^2 \leq \Phi\left(D_{\boldsymbol{m}} + D_{\boldsymbol{m}'}\right).
\end{aligned}$$

  Donc pour tout $t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')$, $\|t\|_\infty^2 \leq \Phi\left(D_{\boldsymbol{m}} + D_{\boldsymbol{m}'}\right) =: M^2$.

- *Valeur de v.* On suppose que f est bornée sur A.

$$\begin{aligned}
\sup_{t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \operatorname{Var}\left(t(\boldsymbol{X}_1)\right) &\leq \sup_{t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \mathbb{E}\left[t(\boldsymbol{X}_1)^2\right] \\
&= \sup_{t \in \mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \int_A t^2(\boldsymbol{x}) f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\
&\leq \|f\|_\infty =: v.
\end{aligned}$$

Les quantités clés qui interviennent dans l'inégalité de Talagrand sont les suivantes :

$$\frac{nH^2}{v} = \frac{\Phi(D_{\boldsymbol{m}} + D_{\boldsymbol{m}'})}{\|f\|_\infty} \geq \frac{\Phi D_{\boldsymbol{m}'}}{\|f\|_{\mathrm{L}^2}}, \quad \frac{nH}{M} = \sqrt{n}, \quad \frac{M^2}{n^2} \leq \frac{2\Phi}{n}.$$

On applique l'inégalité de Talagrand avec $\delta = 1/2$ et on obtient :

$$\mathbb{E}\left[\left(\sup_{f\in\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) - 4H^2\right)_+\right] \leq \frac{C(f)}{n}\exp\left(-C'(f,\Phi)D_{\boldsymbol{m}'}\right) + \frac{C''(\Phi)}{n}\exp\left(-C'''\sqrt{n}\right).$$

En sommant sur $\boldsymbol{m}'$, on obtient la majoration :

$$\sum_{\boldsymbol{m}'\in\mathscr{M}_n} \mathbb{E}\left[\left(\sup_{f\in\mathscr{B}(\boldsymbol{m},\boldsymbol{m}')} \nu_n^2(t) - 4H^2\right)_+\right] \leq \frac{1}{n}\Sigma_n(f,\Phi),$$

avec :

$$\Sigma_n(f,\Phi) \coloneqq C(f) \sum_{\boldsymbol{m}'\in\mathscr{M}_n} \mathrm{e}^{-C'(f,\Phi)D_{\boldsymbol{m}'}} + C(\Phi)\,\mathrm{Card}(\mathscr{M}_n)\,\mathrm{e}^{-C'''\sqrt{n}}.$$

Il ne reste plus qu'à montrer que $\Sigma_n(f,\Phi)$ est borné indépendamment de n. On utilise pour cela la proposition B.2.1 et le théorème B.2.2 :

$$\begin{aligned}\Sigma_n(f,\Phi) &= C(f)\sum_{D=1}^{n} \mathrm{Card}\left\{\boldsymbol{m}'\in\mathbb{N}_+^d \,\middle|\, D_{\boldsymbol{m}'} = D\right\}\mathrm{e}^{-C(f,\Phi)D} + C(\Phi)\,\mathrm{Card}(\mathscr{M}_n)\,\mathrm{e}^{-C'''\sqrt{n}} \\ &\leq C(f)\sum_{D=1}^{n} \mathrm{o}(D)\,\mathrm{e}^{-C(f,\Phi)D} + C(\Phi)\,n(\log n)^{d-1}\,\mathrm{e}^{-C'''\sqrt{n}} \leq \Sigma_\infty(f,\Phi,d),\end{aligned}$$

avec $\Sigma_\infty(f,\Phi,d)$ une constante qui dépend de f, Φ et d. Ainsi, on a montré que si $\kappa \geq 16$, alors $R_n \leq \frac{\Sigma_\infty(f,\Phi,d)}{n}$.

Valeurs propres de matrices aléatoires

Au cours de cette thèse, nous serons amenés à étudier des matrices aléatoires symétriques définies positives. Plus précisément, nous aurons besoin de contrôler la norme d'opérateur de ces matrices et de leur inverse, avec grande probabilité. Or, si $\mathbf{G}$ est une matrice symétrique définie positive, sa norme d'opérateur n'est rien d'autre que sa plus grande valeur propre. Ainsi, on a :

$$\|\mathbf{G}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{G}), \quad \|\mathbf{G}^{-1}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{G}^{-1}) = \frac{1}{\lambda_{\min}(\mathbf{G})}.$$

Nous aurons donc besoin d'inégalités de concentration sur les valeurs propres de matrices aléatoires symétriques. On trouvera des démonstrations des inégalités suivantes dans les articles de Tropp (2012) et de Gittens & Tropp (2011).

Théorème 1.4.10 (Matrix Chernoff bound)**.** *Soient $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ des matrices aléatoires indépendantes de taille $d\times d$, symétriques semi-définies positives, et telles que $\sup_k \lambda_{\max}(\mathbf{Z}_k) \leq R$ p.s. Si on définit :*

$$\mu_{\min} \coloneqq \lambda_{\min}\left(\sum_{k=1}^{n}\mathbb{E}[\mathbf{Z}_k]\right), \quad \mu_{\max} \coloneqq \lambda_{\max}\left(\sum_{k=1}^{n}\mathbb{E}[\mathbf{Z}_k]\right)$$

alors on a les inégalités :

$$\forall \delta \in]0,1[, \quad \mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \leq (1-\delta)\mu_{\min}\right] \leq d \times \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_{\min}/R} \tag{1.25}$$

$$\forall \delta > 0, \quad \mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \geq (1+\delta)\mu_{\min}\right] \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_{\min}/R} \tag{1.26}$$

$$\forall \delta > 0, \quad \mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \geq (1+\delta)\mu_{\max}\right] \leq d \times \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_{\max}/R}$$

$$\forall \delta \in]0,1[, \quad \mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{Z}_k\right) \leq (1-\delta)\mu_{\max}\right] \leq \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_{\max}/R}.$$

Théorème 1.4.11 (Matrix Bernstein bound)**.** *Soient* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *des matrices symétriques aléatoires indépendantes de taille* $d \times d$*, telles que* $\mathbb{E}[\mathbf{Z}_k] = \mathbf{0}$ *et telles que* $\sup_k \lambda_{\max}(\mathbf{Z}_k) \leq R$ *p.s. Si* $v > 0$ *est un majorant de :*

$$\left\| \sum_{k=1}^{n} \mathbb{E}\left[\mathbf{Z}_k^2\right] \right\|_{\mathrm{op}} \leq v,$$

alors pour tout $x > 0$*, on a :*

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{Z}_i\right) \geq x\right] \leq d \times \exp\left(\frac{-x^2/2}{v + \frac{R}{3}x}\right).$$

Nous utiliserons également l'inégalité suivante tirée de l'annexe A de Chen *et al.* (2012). Cette inégalité peut être vue comme une version matricielle de l'inégalité de Rosenthal.

Théorème 1.4.12 (Matrix moment inequality)**.** *Soient* $p \geq 1$ *et* $q \geq 2$*; on fixe* $r \geq \max(q, 2\log p)$*. On considère une suite finie* $(\mathbf{Z}_i)$ *de matrices symétriques aléatoires indépendantes de taille* $p \times p$*. Alors on a l'inégalité :*

$$\left[\mathbb{E}\lambda_{\max}\left(\sum_i \mathbf{Z}_i\right)^q\right]^{1/q} \leq \sqrt{\mathrm{e}r\,\lambda_{\max}\left(\sum_i \mathbb{E}\mathbf{Z}_i^2\right)} + 2\mathrm{e}r\left[\mathbb{E}\max_i \lambda_{\max}^q(\mathbf{Z}_i)\right]^{1/q}.$$

1.5 Problèmes étudiés, résultats obtenus et perspectives

1.5.1 Introduction à la déconvolution

Dans de nombreux problèmes de statistique non paramétrique, les observations ne portent pas directement sur la fonction d'intérêt f mais sur une version bruitée de celle-ci par la convolution avec une fonction g. Ce genre de problèmes se pose par exemple en estimation de densité avec erreurs de mesure additives, en

régression non paramétrique avec erreur sur les variables, mais aussi en traitement du signal ou en traitement de l'image. Pour plus de détails sur les modèles de déconvolution étudiés en statistique non paramétrique, on pourra se référer au livre de Meister (2009).

Le problème de déconvolution consiste en la résolution de l'équation fonctionnelle :

$$h = f * g, \tag{1.27}$$

où f est la fonction inconnue du problème, g est une fonction représentant le bruit qui peut être supposée connue ou non, et h est la fonction sur laquelle portent les observations. La stratégie la plus commune pour résoudre ce problème est de passer dans le domaine fréquentiel en prenant la transformée de Fourier de (1.27). On obtient alors :

$$\mathscr{F}[h] = \mathscr{F}[f] \times \mathscr{F}[g] \iff f = \mathscr{F}^{-1}\left[\frac{\mathscr{F}[h]}{\mathscr{F}[g]}\right],$$

où $\mathscr{F}$ désigne la transformation de Fourier et $\mathscr{F}^{-1}$ son inverse. Il n'y a plus qu'à estimer h (et g quand elle est inconnue) pour obtenir un estimateur de f. Cependant, l'inversion de la transformation de Fourier peut poser problème car le quotient des transformées de Fourier n'est pas forcément intégrable une fois qu'on a estimé h (et g). Il est alors nécessaire d'avoir recours à une étape de régularisation pour pouvoir inverser la transformée de Fourier.

Récemment, Mabon (2017) et Comte *et al.* (2017) ont proposé d'utiliser la base de Laguerre, définie par (1.4), pour résoudre le problème de déconvolution (1.27) lorsque les fonctions sont à support sur $\mathbb{R}_+$. Cette méthode repose sur la relation suivante que satisfont les fonctions de Laguerre :

$$\forall k, j \in \mathbb{N}, \quad \varphi_k * \varphi_j = 2^{-\frac{1}{2}}\left(\varphi_{k+j} - \varphi_{k+j+1}\right), \tag{1.28}$$

voir la formule 22.13.14 de Abramowitz & Stegun (1972). En utilisant cette relation remarquable, on peut calculer les coefficients de Laguerre d'un produit de convolution de deux fonctions à partir de leurs coefficients respectifs.

Proposition 1.5.1. *Soient f et g dans $\mathrm{L}^2(\mathbb{R}_+)$ tels que $f * g \in \mathrm{L}^2(\mathbb{R}_+)$. Alors les coefficients de Laguerre $(c_\ell)_{\ell\in\mathbb{N}}$ de $f * g$ sont donnés par :*

$$c_\ell = \begin{cases} 2^{-\frac{1}{2}}\,(a*b)_0 & \text{si } \ell = 0, \\ 2^{-\frac{1}{2}}\left((a*b)_\ell - (a*b)_{\ell-1}\right) & \text{si } \ell \geq 1, \end{cases} \tag{1.29}$$

où $(a_k)_{k\in\mathbb{N}}$ et $(b_k)_{k\in\mathbb{N}}$ sont respectivement les coefficients de Laguerre de f et g, et où $a * b$ est le produit de convolution des suites a et b.

Démonstration. On calcule $f * g$ en décomposant f et g dans la base de Laguerre :

$$\begin{aligned}
f * g &= \sum_{k\in\mathbb{N}} \sum_{j\in\mathbb{N}} a_k\, b_j\, (\varphi_k * \varphi_j) \\
&= 2^{-\frac{1}{2}} \sum_{k\in\mathbb{N}} \sum_{j\in\mathbb{N}} a_k\, b_j\, (\varphi_{k+j} - \varphi_{k+j+1}) && \text{(d'après (1.28))} \\
&= 2^{-\frac{1}{2}} \sum_{\ell\in\mathbb{N}} \sum_{k=0}^{\ell} a_k\, b_{\ell-k}\, (\varphi_\ell - \varphi_{\ell+1}) && (\ell := k+j) \\
&= 2^{-\frac{1}{2}} \sum_{\ell\in\mathbb{N}} (a * b)_\ell\, (\varphi_\ell - \varphi_{\ell+1}) \\
&= 2^{-\frac{1}{2}}\, (a * b)_0\, \varphi_0 + 2^{-\frac{1}{2}} \sum_{\ell\in\mathbb{N}_+} \big((a * b)_\ell - (a * b)_{\ell-1}\big)\, \varphi_\ell .
\end{aligned}$$

On identifie ainsi les coefficients de $f * g$ par unicité de la décomposition. □

Si on introduit la suite $(\beta_k)_{k\in\mathbb{N}}$ formée à partir des accroissement des coefficients $(b_k)_{k\in\mathbb{N}}$:

$$\beta_k := \begin{cases} 2^{-\frac{1}{2}}\, b_0 & \text{si } k = 0, \\ 2^{-\frac{1}{2}}\, (b_k - b_{k-1}) & \text{si } k \geq 1, \end{cases}$$

alors les coefficients (1.29) s'expriment comme le produit de convolution $\beta * a$. Autrement dit, les coefficients du produit de convolution $f * g$ sont donnés par le produit de convolution (discret) des coefficients de f et des accroissements des coefficients de g. Résoudre le problème de déconvolution (1.27) est donc équivalent à inverser un produit de convolution sur $\mathbb{N}$.

Le produit de convolution par une suite donnée étant linéaire, ce problème peut-être résolu matriciellement. Soit $\mathbf{G}$ la matrice infinie triangulaire inférieure :

$$\mathbf{G} := \begin{pmatrix} \beta_0 & 0 & \cdots & \cdots & \cdots \\ \beta_1 & \beta_0 & 0 & \cdots & \cdots \\ \beta_2 & \beta_1 & \beta_0 & 0 & \cdots \\ \beta_3 & \beta_2 & \beta_1 & \beta_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \tag{1.30}$$

et soit $\mathbf{G}_m$ la sous-matrice de $\mathbf{G}$ constituée de ses m premières lignes et de ses m premières colonnes. Ces matrices ont la particularité d'avoir des diagonales constantes et sont appelées *matrices Toeplitz*. Ce type de matrices a été bien étudié dans la littérature, on pourra se référer aux livres de Böttcher & Grudsky (2000, 2005) sur la question. Les éléments de la théorie des opérateurs Toeplitz que nous utiliserons sont rappelés à la section 3.7. Si on note $\mathbf{a}_m$ et $\mathbf{c}_m$ les vecteurs des m premiers coefficients de respectivement f et h, alors on a la relation :

$$c = \beta * a \iff \forall m \in \mathbb{N}_+, \quad \mathbf{c}_m = \mathbf{G}_m\, \mathbf{a}_m .$$

Si β_0 est non nul, alors $\mathbf{G}_m$ est inversible et les coefficients de f s'expriment à partir de ceux de g et h suivant la relation :

$$\mathbf{a}_m = \mathbf{G}_m^{-1}\,\mathbf{c}_m.$$

Une fois estimé h et connaissant g (ou l'ayant estimée), le problème de déconvolution sur $\mathbb{R}_+$ revient alors à inverser une matrice triangulaire.

1.5.2 Déconvolution d'une densité

Un exemple classique de problème de déconvolution est celui de l'estimation d'une densité à partir d'observations indirectes. Dans ce modèle, on suppose qu'on observe des variables aléatoires i.i.d. $Z_1,\dots,Z_n$ dans $\mathbb{R}$ qui sont données par :

$$Z_i = X_i + Y_i,$$

où les X_i sont i.i.d. de densité f et les Y_i sont des variables de bruit i.i.d. de densité g, indépendantes des X_i. L'objectif est d'estimer la loi des X_i à partir de l'observation des Z_i. En notant h la densité des Z_i, celle-ci est donnée par le produit de convolution des densité f et g. Pour que le modèle soit identifiable, on pourra supposer que la densité du bruit est connue ou que l'on dispose d'observations du modèle « à blanc », c'est-à-dire $Y'_1,\dots,Y'_N$ i.i.d. de densité g, indépendantes des autres variables, ce qui permet de construire un estimateur de g. Dans la suite on supposera que g est connue.

Un estimateur très populaire pour résoudre ce problème est un estimateur à noyau proposé par Stefanski & Carroll (1990). Soit $K \in \mathrm{L}^1(\mathbb{R}) \cap \mathrm{L}^2(\mathbb{R})$ une fonction positive, paire et d'intégrale 1, on estime h par :

$$\widehat{h}_\delta(x) := \frac{1}{n}\sum_{i=1}^{n} K_\delta(x - Z_i), \quad K_\delta(u) := \frac{1}{\delta}K\Big(\frac{u}{\delta}\Big),$$

où $\delta > 0$ est la fenêtre (*bandwidth*) du noyau. Sous l'hypothèse que la transformée de Fourier de g ne s'annule pas, on estime f par :

$$\widehat{f}_\delta(x) := \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega x}\frac{\mathscr{F}\big[\widehat{h}_\delta\big](\omega)}{\mathscr{F}[g](\omega)}\,\mathrm{d}\omega.$$

Cet estimateur peut encore s'écrire sous la forme d'un estimateur à noyau. En effet, en utilisant les propriétés de la transformation de Fourier vis à vis des translations et des dilatations, on a :

$$\mathscr{F}\big[\widehat{h}_\delta\big](\omega) = \frac{1}{n}\sum_{i=1}^{n}\mathscr{F}\big[K_\delta(\bullet - Z_i)\big](\omega) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\mathrm{i}\omega Z_i}\right)\mathscr{F}[K](\delta\omega).$$

On réécrit alors $\widehat{f}_\delta$ comme :

$$\begin{aligned}\widehat{f}_\delta(x) &:= \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega x}\frac{\mathscr{F}\left[\widehat{h}_\delta\right](\omega)}{\mathscr{F}[g](\omega)}\,\mathrm{d}\omega \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega(x-Z_i)}\frac{\mathscr{F}[K](\delta\omega)}{\mathscr{F}[g](\omega)}\,\mathrm{d}\omega \\ &= \frac{1}{n}\sum_{i=1}^{n} K_\delta^g(x-Z_i),\end{aligned}$$

où K_δ^g est le noyau défini par :

$$K_\delta^g(u) := \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\omega u}\frac{\mathscr{F}[K](\delta\omega)}{\mathscr{F}[g](\omega)}\,\mathrm{d}\omega.$$

Avec des hypothèses de régularité sur le noyau K et un choix judicieux de fenêtre, des résultats sur la vitesse d'estimation minimax ont été établis par plusieurs auteurs. Ces résultats sont synthétisés dans la table 1.1.

Carroll & Hall (1988) montrent que dans le cas où f est k fois dérivable et les erreurs sont gaussiennes, alors la vitesse de convergence de cet estimateur est logarithmique et cette vitesse est minimax. Fan (1991) généralise ces travaux en introduisant la notion de distributions *ordinary smooth* et *supersmooth* pour le bruit. Il suppose toujours que la transformée de Fourier de g ne s'annule pas, et fait l'hypothèse suivante sur sa décroissance lorsque $|\omega| \to +\infty$:

$$C\left(1+\omega^2\right)^{-\beta}\exp\left(-2\alpha|\omega|^\rho\right) \le \left|\mathscr{F}[g](\omega)\right|^2 \le C'\left(1+\omega^2\right)^{-\beta}\exp\left(-2\alpha|\omega|^\rho\right).$$

On dit que g est *ordinary smooth* si $\alpha = 0$ ou $\rho = 0$, et *supersmooth* sinon. L'auteur montre que dans le cas d'un bruit *supersmooth* (par exemple, un bruit gaussien), la vitesse minimax sur l'espace de Hölder est logarithmique, mais que dans le cas où le bruit est *ordinary smooth* (par exemple, une loi de Laplace), cette vitesse devient polynomiale.

Plus généralement, Pensky & Vidakovic (1999) proposent de considérer des densités appartenant à la classe suivante :

$$\mathscr{A}_{a,b,r}(L) := \left\{ f \in \mathrm{L}^2(\mathbb{R}) \,\middle|\, \int_{\mathbb{R}} \left|\mathscr{F}[f](\omega)\right|^2\left(1+\omega^2\right)^b \exp\left(2a|\omega|^r\right)\mathrm{d}\omega \le 2\pi L \right\}.$$

Si $\alpha = 0$ ou $r = 0$, on reconnait une boule dans un espace de Sobolev. Les espaces de Hölder étant inclus dans les espaces de Sobolev, les vitesses minimax sur ces premiers se généralisent à ces seconds. Si $\alpha > 0$ et $r > 0$, cette classe contient des fonctions bien plus régulières (pour $r = 1$, ces fonctions admettent un prolongement analytique sur une bandes autour de l'axe des réels). Si $f \in \mathscr{A}_{a,b,r}(L)$ avec $a > 0$ et $r > 0$, on dira que f est *supersmooth*. Butucea (2004) montre que dans le cas où f est *supersmooth* et le bruit *ordinary smooth*, la vitesse minimax est quasi-paramétrique. Le cas où à la fois f et g sont *supersmooth* est plus délicat

	$\alpha = 0$ ou $\rho = 0$	$\alpha > 0$ et $\rho > 0$
$a = 0$ ou $r = 0$	$n^{-\frac{2b}{2b+2\beta+1}}$	$(\log n)^{\frac{2b}{\beta}}$
$a > 0$ et $r > 0$	$\frac{1}{n}(\log n)^{\frac{2\beta+1}{r}}$	cf. Butucea & Tsybakov (2008a,b)

TABLE 1.1 : Vitesse de convergence minimax sur $\mathscr{A}_{a,b,r}$ pour la déconvolution d'une densité sur $\mathbb{R}$, selon que g est *ordinary smooth* ou *supersmooth*.

et a été étudié par Butucea & Tsybakov (2008a,b). Tous ces résultats sont généralisés au problème de déconvolution d'une densité sur $\mathbb{R}^d$ par Comte & Lacour (2013). Rebelles (2016) étend ces résultats en construisant un estimateur à noyau adaptatif qui atteint la vitesse minimax pour le coût L^p. Récemment, Lepski & Willer (2019) étudient un modèle plus général qui mélange observations directes et indirectes, et établissent les vitesses minimax sur des classes de Nikol'skii pour le coût L^p.

D'autres estimateurs ont été proposés en dehors des estimateurs à noyaux pour résoudre le problème de déconvolution d'une densité. Pensky & Vidakovic (1999) construisent un estimateur par projection sur des bases d'ondelettes de Meyer. Leur estimateur est adaptatif et atteint la vitesse minimax sur les espaces de Sobolev, ainsi que dans le cas où f est *supersmooth* et g *ordinary smooth*. En revanche, lorsque f et g sont *supersmooth*, leur estimateur est sous optimal comme l'ont montré Butucea & Tsybakov (2008a,b). Comte *et al.* (2006) utilisent une base d'ondelette construite à partir de la fonction mère sinc et construisent un estimateur adaptatif en utilisant une procédure de sélection de modèles. Cet estimateur atteint la vitesse minimax dans les mêmes cas que celui de Pensky & Vidakovic (1999) et est meilleur dans le cas où f et g sont *supersmooth*.

Plus récemment, des estimateurs par projection sur les bases de Laguerre et d'Hermite ont été proposés. Mabon (2017) considère le problème de déconvolution d'une densité sur $\mathbb{R}_+$ lorsque le bruit est également à support sur $\mathbb{R}_+$. Elle propose d'utiliser un estimateur par projection sur la base de Laguerre et d'utiliser les propriétés de déconvolution des fonctions de Laguerre décrites en sous-section 1.5.1. Elle fait les hypothèses suivantes sur la densité g :

1. g est $r \geq 1$ fois dérivable et :
$$g^{(j)}(0) = \begin{cases} 0 & j \in \{0, \dots, r-2\}, \\ B_r \neq 0 & j = r-1. \end{cases}$$
2. $g^{(r)} \in \mathrm{L}^1(\mathbb{R}_+)$.
3. La transformée de Laplace de g ne s'annule pas sur le demi-plan des complexes de partie réelle positive.

Ces hypothèses de régularité sur g peuvent être vues comme un analogue de la régularité *ordinary smooth* en déconvolution sur $\mathbb{R}$. Elle montre que la vitesse de

convergence de l'estimateur oracle sur $\mathrm{W}^s(\mathbb{R}_+, L)$ est $n^{-\frac{s}{s+r}}$ et construit un estimateur adaptatif avec la méthode de sélection de modèle par pénalisation. Ces travaux sont généralisés au modèle avec bruit inconnu par Comte & Mabon (2017). Sacko (2020) utilise un estimateur par projection sur la base d'Hermite pour résoudre le problème de déconvolution d'une densité sur $\mathbb{R}$ en utilisant le fait que les fonctions d'Hermite sont des vecteurs propres de la transformation de Fourier. Les vitesses de convergence obtenue sur les boules de Sobolev–Hermite sont les mêmes à celles sur les espaces de Hölder et les espaces de Sobolev.

Contribution

Au chapitre 2, nous étendons les travaux de Mabon (2017) à la déconvolution de densités sur $\mathbb{R}_+^d$. Pour cela, on généralise la proposition 1.5.1 avec des coefficients dans la base de Laguerre tensorisée : on montre que les coefficients de h s'expriment linéairement en fonction des coefficients de f. Cette relation linéaire s'écrit à l'aide d'hypermatrices. Notons $a_{\boldsymbol{k}}$ (resp. $c_{\boldsymbol{k}}$) le $\boldsymbol{k}$-ème coefficient de Laguerre de f (resp. de h). Pour $\boldsymbol{m} \in \mathbb{N}_+^d$, on note $\mathbf{a}_{\boldsymbol{m}}$ l'hypermatrice des $a_{\boldsymbol{k}}$ pour $\boldsymbol{k} \leq \boldsymbol{m} - \mathbf{1}$ (de même pour $\mathbf{c}_{\boldsymbol{m}}$). On a les relations :

$$\mathbf{c}_{\boldsymbol{m}} = \mathbf{G}_{\boldsymbol{m}} \times_d \mathbf{a}_{\boldsymbol{m}} \quad \text{et} \quad \mathbf{a}_{\boldsymbol{m}} = \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \mathbf{c}_{\boldsymbol{m}} \tag{1.31}$$

où "$\times_d$" désigne le d–produit contracté d'hypermatrices et où $\mathbf{G}_{\boldsymbol{m}}$ est l'hypermatrice $\boldsymbol{m} \times \boldsymbol{m}$ suivante :

$$\big[\mathbf{G}_{\boldsymbol{m}}\big]_{\boldsymbol{\ell},\boldsymbol{k}} := 2^{-d/2} \sum_{\boldsymbol{\varepsilon} \in \{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|} b_{\boldsymbol{\ell}-\boldsymbol{k}-\boldsymbol{\varepsilon}},$$

avec la convention $b_{\boldsymbol{j}} := 0$ si $\boldsymbol{j} \notin \mathbb{N}^d$. Notons que dans la cas $d = 1$, on retrouve la matrice $\mathbf{G}_m$ définie par (1.30). Puis, on estime la densité h par projection sur la base de Laguerre tensorisée :

$$\widehat{h}_{\boldsymbol{m}} := \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \widehat{c}_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}, \quad \widehat{c}_{\boldsymbol{k}} := \frac{1}{n} \sum_{i=1}^{n} \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_i).$$

et on estime f en utilisant la relation (1.31) :

$$\widehat{f}_{\boldsymbol{m}} := \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \widehat{a}_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}, \quad \widehat{\mathbf{a}}_{\boldsymbol{m}} := \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \widehat{\mathbf{c}}_{\boldsymbol{m}}.$$

Pour étudier le biais, on introduit les espaces de Sobolev–Laguerre multivariés. Pour $\boldsymbol{s} \in]0, +\infty[^d$ et $L > 0$, on définit :

$$\mathrm{W}^{\boldsymbol{s}}\big(\mathbb{R}_+^d, L\big) := \left\{ f \in \mathrm{L}^2\big(\mathbb{R}_+^d\big) \,\middle|\, \sum_{\boldsymbol{k} \in \mathbb{N}^d} \langle f, \varphi_{\boldsymbol{k}} \rangle^2 \, \boldsymbol{k}^{\boldsymbol{s}} \leq L \right\}.$$

Lorsque f appartient à cet espace, le biais est majoré par $L \times (m_1^{-s_1} + \cdots + m_d^{-s_d})$. Sous certaines hypothèses techniques de régularité sur la transformée de Laplace

de g, on montre que le terme de variance est majoré par $\boldsymbol{m}^{2\boldsymbol{\alpha}}$ où $\boldsymbol{\alpha} \in \mathbb{N}_+^d$ est lié à la vitesse de décroissance de la transformée de Laplace de g à l'infini. C'est une généralisation des hypothèses sur g faites par Mabon (2017). On peut aussi voir ça comme un analogue de la notion de distribution *ordinary smooth*. On montre alors que la vitesse d'estimation de l'estimateur oracle sur $\mathrm{W}^{\boldsymbol{s}}\big(\mathbb{R}_+^d, L\big)$ est donnée par :

$$n^{-1/\left(1+\sum_{i=1}^d \frac{2\alpha_i}{s_i}\right)}.$$

Cette vitesse d'estimation sur les boules de Sobolev–Laguerre est similaire à celle trouvée par Comte & Lacour (2013) sur les boules Sobolev anisotropes, pour le problème de déconvolution sur $\mathbb{R}^d$ avec un estimateur à noyau, et pour un bruit *ordinary smooth*. Enfin on construit un estimateur adaptatif sur les boules de Sobolev–Laguerre en utilisant la méthode de Goldenshluger et Lepski.

Perspectives

Une extension possible de ce travail serait de considérer le problème où la densité du bruit est inconnue et doit être estimée à partir d'un échantillon indépendant de variables $\boldsymbol{Y}_1', \dots, \boldsymbol{Y}_N'$. L'hypermatrice de déconvolution $\mathbf{G}_{\boldsymbol{m}}$ serait alors remplacée par un estimateur. Dans le cas de la dimension 1, cette question a déjà été traitée par Comte & Mabon (2017). On peut penser que leurs résultats devraient se généraliser à la dimension supérieure.

Une autre question que soulève le chapitre 2 est celle de la description des espaces de Sobolev–Laguerre multidimensionnels $\mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d)$. Dans le cas de la dimension 1 et pour s entier, Comte & Genon-Catalot (2015) montrent qu'une fonction f appartient à cet espace si et seulement elles $s-1$ dérivable, $f^{(s-1)}$ est absolument continue et les dérivées de f satisfont certaines conditions d'intégrabilité sur $\mathbb{R}_+$. Dans le cas $d \geq 2$, on ne sait pour l'instant rien de ces espaces. Il serait certainement utile de savoir s'il existe une caractérisation similaire, portant sur la différentiabilité des fonctions par exemple.

1.5.3 Estimation de la fonction de Gerber–Shiu

Notre deuxième exemple de problème de déconvolution se trouve en théorie de la ruine. Cette théorie vise à modéliser l'évolution des réserves financières d'une compagnie d'assurance au cours du temps et à étudier diverses quantités liées à la ruine de la compagnie, afin de permettre une meilleure gestion des risques. Le modèle le plus simple est le modèle de Cramér–Lundberg. Dans ce modèle, le processus $(U_t)_{t\geq 0}$ des réserves de la compagnie est donné par :

$$U_t = u + ct - \sum_{i=1}^{N_t} X_i, \tag{1.32}$$

où $u \geq 0$ est la réserve initiale, $c > 0$ est le taux de cotisation (*premium rate*), $(N_t)_{t\geq 0}$ est un processus de Poisson d'intensité $\lambda > 0$ et où les montants des sinistres $(X_i)_{i\geq 1}$ sont positifs i.i.d. de moyenne μ et de densité f, et indépendants

de $(N_t)_{t\geq 0}$. Dans la suite, on note τ l'instant de ruine, c'est-à-dire le premier instant où le processus prend des valeurs négatives :

$$\tau := \inf\{t \geq 0 \,|\, U_t < 0\}.$$

Pour une présentation générale de ce modèle, et plus généralement de la théorie de la ruine, on pourra se référer au livre de Asmussen & Albrecher (2010).

Probabilité de ruine

La première quantité étudiée en théorie de la ruine est la probabilité ultime de ruine[4] :

$$\phi(u) := \mathbb{P}\big[\tau < +\infty \,|\, U_0 = u\big].$$

Dans le modèle (1.32), le processus des pertes agrégées vérifie :

$$\frac{1}{t}\sum_{i=1}^{N_t} X_i \xrightarrow[t\to+\infty]{\text{p.s.}} \lambda\mu,$$

la quantité $\lambda\mu$ représente donc les pertes moyennes par unités de temps. On définit alors $\eta := \frac{c-\lambda\mu}{\lambda\mu}$ de sorte que $c = (1+\eta)\lambda\mu$. Cette quantité est appelé *safety loading* et représente avec quelle proportion les primes excèdent les pertes moyennes par unité de temps. Comme on s'en doute, il est préférable pour la compagnie que le *safety loading* soit positif. En effet, on peut montrer que si $\eta \leq 0$, alors $\phi(u) = 1$ pour tout $u \geq 0$ et si $\eta > 0$, alors $\phi(u) < 1$ pour tout $u \geq 0$. On supposera dans la suite que $\eta > 0$ afin que la probabilité de ruine ne soit pas triviale. Cette hypothèse est appelée *safety loading condition* dans la littérature. Pour notre part, nous utiliserons plutôt la quantité $\theta := \frac{\lambda\mu}{c}$, la *safety loading condition* exprimant alors que $\theta \in [0,1[$.

Un résultat fondamental pour étudier la probabilité de ruine est la formule de Pollaczeck–Khinchine. Si on note $S(x) := \mathbb{P}[X_1 > x]$ la fonction de survie des X_i, alors ϕ vérifie la formule :

$$\phi(u) = (1-\theta)\sum_{k=1}^{+\infty} \theta^k F_k(u), \quad F_k(u) = \frac{1}{\mu^k}\int_u^{+\infty} S^{*k}(x)\,\mathrm{d}x, \tag{1.33}$$

où S^{*k} désigne la convolution de S avec elle-même k fois. Une façon équivalente d'énoncer ce résultat est le fait que ϕ vérifie une équation de renouvellement :

$$\phi = \phi * g + h, \quad g(x) := \frac{\lambda}{c}S(x), \quad h(u) := \frac{\lambda}{c}\int_u^{+\infty} S(x)\,\mathrm{d}x. \tag{1.34}$$

Ces formules permettent de calculer explicitement la fonction ϕ pour certaines distributions des X_i. C'est le cas des distributions *phase-type*, une classe de distributions qui contient par exemple les lois exponentielles, les mélanges de lois exponentielles et les lois d'Erlang.

[4] le conditionnement qui apparaît dans la définition de ϕ n'en est pas vraiment un puisque l'évènement $\{U_0 = u\}$ est certain, c'est juste une façon de faire apparaître la dépendance en u.

Dans la suite, on s'intéresse au problème de l'estimation de la probabilité de ruine dans le modèle (1.32). Les observations à partir desquels on estime ϕ varient dans la littérature : dans certains articles, on observe un n-échantillon $X_1, \dots, X_n$ de densité f (et on suppose λ connu), dans d'autres on suppose qu'on observe la trajectoire de (U_t) pour $t \in [0,T]$, enfin d'autres n'observent le processus (U_t) que sur un intervalle $[0,T]$ discrétisé à pas Δ.

La grande majorité des méthodes d'estimation de la probabilité de ruine se fondent sur la formule de Pollaczeck–Khinchine (1.33) ou l'équation de renouvellement (1.34). Les premiers travaux visant à construire un estimateur non paramétrique de la probabilité de ruine sont ceux de Frees (1986). L'auteur propose un estimateur de type Monte Carlo pour estimer ponctuellement $\phi(u)$ et démontre sa consistance. Hipp (1989) considère plusieurs scénarios selon quels paramètres parmi λ, μ et f sont connus, et propose d'estimer ϕ par *plug-in* à l'aide de la formule (1.33), en remplaçant les quantités inconnues par des versions empiriques. Il démontre que son estimateur est consistant et asymptotiquement gaussien pour l'estimation ponctuelle de $\phi(u)$, et construit un intervalle de confiance par *bootstrap*. Cependant la question du calcul de la série dans formule (1.33), ou de sa troncature, n'est pas abordée. Dans le cas où λ et μ sont connus, Croux & Veraverbeke (1990) proposent aussi d'utiliser (1.33) en estimant la fonction de répartition des X_i par la répartition empirique (ils proposent également d'utiliser un estimateur à noyau). Ils estiment alors les quantités $F_k(u)$ par des U-statistiques et construisent un estimateur ponctuel de la probabilité de ruine en tronquant la série dans la formule de Pollaczeck–Khinchine. Ils démontrent la normalité asymptotique de leur estimateur.

Les travaux précédents traitaient de l'estimation ponctuelle de $\phi(u)$. Les premiers travaux qui traitent le problème sous l'angle de l'estimation fonctionnelle sont ceux de Pitts (1994). Dans le cas où λ et μ sont connus, l'auteur utilise un estimateur similaire à ceux décrits plus haut et établit sa consistance pour la norme sup pondérée par $(1+|x|)^\beta$ avec $\beta \geq 0$:

$$\|F\|_\beta := \sup_x \left| \left(1+|x|\right)^\beta F(x) \right|,$$

sous la condition que X_1 admette un moment fini d'ordre $1+\beta'$ avec $\beta' > \beta$. Il démontre également un TCL fonctionnel pour l'estimateur. Ces résultats sont améliorés par Politis (2003) qui ne suppose plus λ et μ connus. Toutefois, ces deux derniers travaux considèrent un estimateur par *plug-in* dans la formule (1.33) sans discuter le calcul ou la troncature de la série.

Enfin, citons les travaux de Mnatsakanov *et al.* (2008). Ceux-ci se démarquent des travaux précédents en construisant leur estimateur à partir de l'équation de renouvellement et en étudiant les performances de l'estimateur pour la norme L^2. Plus précisément, les auteurs estiment la fonction $\psi(u) := 1-\phi(u)$, c'est-à-dire la probabilité de survie. Celle-ci vérifie également une équation similaire à

(1.34) qui, en prenant la transformée de Laplace, devient :

$$\mathscr{L}\psi(s) = \frac{c - \lambda\mu}{cs - \lambda\big(1 - \mathscr{L}f(s)\big)}.$$

Ils estiment alors $\mathscr{L}\psi$ par *plug-in* en estimant $\mathscr{L}f$ par la transformée de Laplace empirique et μ par la moyenne empirique (et en supposant λ connu). L'estimateur final est obtenu à l'aide une inversion régularisée de la transformation de Laplace. Malheureusement, la vitesse de convergence pour la norme L^2 sur les intervalles de forme $[0, B]$ est très lente :

$$\big\|\psi - \widehat{\psi}\big\|^2_{\mathrm{L}^2([0,B])} = \mathrm{O_P}\left(\frac{1}{\log n}\right).$$

L'utilisation de l'inversion régularisée de la transformation de Laplace, outre les difficultés numériques, semble dégrader fortement la vitesse de convergence de l'estimateur. À titre de comparaison, les travaux de Pitts (1994) et Politis (2003) cités plus haut établissent un TCL fonctionnel pour leur estimateur, ce qui suggère que la vitesse de convergence minimax devrait être $\mathrm{O_P}(n^{-1})$.

Fonction de Gerber–Shiu

Pour étudier simultanément l'instant de ruine τ, le montant de la réserve juste avant la ruine $U_{\tau-}$ et le déficit au moment de la ruine $|U_\tau|$, Gerber & Shiu (1998) introduisent la fonction suivante :

$$\phi(u) \coloneqq \mathbb{E}\Big[\mathrm{e}^{-\delta\tau}\, w\big(U_{\tau-}, |U_\tau|\big) 1_{\tau<\infty} \,\Big|\, U_0 = u\Big],$$

où $\delta \geq 0$ peut être vu comme un taux d'actualisation, et où w est une fonction positive qui représente la pénalité en cas de ruine. Cette fonction ϕ est appelée fonction de Gerber–Shiu ou *expected discounted penalty function* (EDPF). Notons que la probabilité de ruine est de cette forme avec $\delta = 0$ et $w(x, y) = 1$.

Comme la probabilité de ruine, la fonction de Gerber–Shiu vérifie une équation de renouvellement (cf. Gerber & Shiu (1998)) :

$$\phi = \phi * g + h, \tag{1.35}$$

où g et h sont maintenant données par :

$$g(x) \coloneqq \frac{\lambda}{c}\int_x^{+\infty} \mathrm{e}^{-\rho_\delta(y-x)} f(y)\,\mathrm{d}y,$$
$$h(u) \coloneqq \frac{\lambda}{c}\int_u^{+\infty} \mathrm{e}^{-\rho_\delta(x-u)} \left(\int_x^{+\infty} w(x, y-x) f(y)\,\mathrm{d}y\right)\mathrm{d}x,$$

et où ρ_δ est l'unique solution positive de l'équation de Lundberg :

$$cs - \lambda\big(1 - \mathscr{L}f(s)\big) = \delta.$$

Ce type de résultat est en fait valable dans des modèles de risque plus généraux (seules les fonctions g et h changent), comme le modèle avec bruit brownien :

$$U_t = u + ct - \sum_{i=1}^{N_t} X_i + \sigma B_t,$$

ou les modèles avec processus de Lévy :

$$U_t = u + L_t,$$

où $(L_t)_{t\geq 0}$ est un processus de Lévy vérifiant différentes hypothèses selon les modèles. Les résultats sur l'estimation de la fonction Gerber–Shiu se placent souvent dans ces modèles plus généraux, c'est pourquoi on les mentionne, mais nos résultats resteront dans le cadre du modèle (1.32).

L'étude de la question de l'estimation de la fonction de Gerber–Shiu est plus récente. Shimizu (2011, 2012) applique la même technique d'inversion de Laplace régularisé que Mnatsakanov *et al.* (2008) pour construire un estimateur de la fonction de Gerber–Shiu dans le modèle avec mouvement Brownien et le modèle avec processus de Lévy, à partir de l'observation d'une trajectoire du processus de réserve sur un intervalle de temps $[0,T]$ discrétisé. Comme Mnatsakanov *et al.* (2008), l'erreur L^2 de cet estimateur sur les intervalles de forme $[0,B]$ est un $\mathrm{O_P}(1/\log T)$, ce qui n'est pas la vitesse optimale pour ce problème comme le montrent les travaux qui suivent.

Dans le cas du modèle avec mouvement Brownien, Zhang (2017) propose un estimateur similaire à celui de Shimizu (2012), mais en utilisant la transformation de Fourier à la place de la transformation de Laplace. Il construit par *plug-in* un estimateur de $\mathscr{F}\phi$ et estime ϕ par une inversion de Fourier tronquée et à l'aide d'un développement en base sinc :

$$\begin{aligned}\phi(u) \approx \phi_m(u) &:= \frac{1}{2\pi}\int_{-m\pi}^{m\pi} \mathrm{e}^{-\mathrm{i}\omega u}\mathscr{F}\phi(\omega)\,\mathrm{d}\omega \\ &\approx \sum_{k=-K}^{K} A_{m,k}\psi_{m,k}(u), \quad A_{m,k} = \langle \phi_m, \psi_{m,k}\rangle,\end{aligned}$$

où $\psi_{m,k}(u) := \sqrt{m}\,\mathrm{sinc}(mx-k)$ et $\mathrm{sinc}(x) := \sin(\pi x)/(\pi x)$. L'estimateur obtenu est un estimateur par projection :

$$\widehat{\phi}_{m,k} := \sum_{k=-k}^{K} \widehat{A}_{m,k}\psi_{m,k}, \quad \widehat{A}_{m,k} := \frac{1}{2\pi\sqrt{m}}\int_{-\pi m}^{\pi m} \widehat{\mathscr{F}\phi}(\omega)\mathrm{e}^{-\mathrm{i}\omega\frac{k}{m}}\,\mathrm{d}\omega.$$

L'auteur montre que l'erreur L^2 de l'estimateur vérifie[5]

$$\left\|\phi - \widehat{\phi}_{m,K}\right\|_{\mathrm{L}^2}^2 = \mathrm{O}\left(\frac{1}{m}\right) + \mathrm{O}\left(\frac{m}{K}\right) + \mathrm{O_P}\left(\frac{m\log m}{T}\right).$$

[5] dans l'article de Zhang (2017), le dernier terme dans la majoration est $\mathrm{O_P}\left(\sqrt{\frac{m\log m}{T}}\right)$, c'est une erreur due à l'oubli d'un carré dans la preuve.

En choisissant $m_T = \mathrm{O}(T^{1/2})$ et $K_T = \mathrm{O}(T)$, la vitesse de convergence est finalement :

$$\left\|\phi - \widehat{\phi}_{m_T,K_T}\right\|_{\mathrm{L}^2}^2 = \mathrm{O_P}\left(\frac{\log T}{\sqrt{T}}\right),$$

ce qui constitue une nette amélioration de la vitesse obtenue par Shimizu (2012).

Récemment, Zhang & Su (2018) proposent un estimateur par projection sur la base de Laguerre en s'inspirant des travaux de Mabon (2017) et Comte *et al.* (2017). L'idée est d'estimer les fonctions g et h par projection sur la base de Laguerre puis de résoudre le problème de déconvolution (1.35) en utilisant les propriétés de cette base vues à la sous-section 1.5.1. Les coefficients des fonctions g et h sont estimés par méthode des moments et les auteurs montrent que :

$$\begin{aligned}\|g - \widehat{g}_m\|_{\mathrm{L}^2}^2 &= \|g - g_m\|_{\mathrm{L}^2}^2 + \mathrm{O_P}\left(T^{-1}\right),\\ \|h - \widehat{h}_m\|_{\mathrm{L}^2}^2 &= \|h - h_m\|_{\mathrm{L}^2}^2 + \mathrm{O_P}\left(T^{-1}\right),\end{aligned}$$

avec les $\mathrm{O_P}\left(T^{-1}\right)$ qui ne dépendent pas de m (pas de compromis biais-variance). Pour estimer ϕ, les auteurs s'appuient sur l'équation de renouvellement (1.35). En notant $(a_k)_{k\geq 0}$ et $(c_k)_{k\geq 0}$ les coefficients de Laguerre de ϕ et h, ils montrent comme dans la sous-section (1.5.1) que :

$$\mathbf{a}_m = \mathbf{A}_m^{-1}\,\mathbf{c}_m, \quad \mathbf{A}_m \coloneqq \mathbf{I}_m - \mathbf{G}_m,$$

où $\mathbf{G}_m$ est la matrice de déconvolution défini par (1.30). Ils estiment alors les coefficients de ϕ par *plug-in* :

$$\widehat{\mathbf{a}}_m \coloneqq \widehat{\mathbf{A}}_m^{-1}\,\widehat{\mathbf{c}}_m,$$

avec $\widehat{\mathbf{A}}_m$ obtenue à partir de $\mathbf{A}_m$ en remplaçant les coefficients de g par leurs estimateurs empiriques. Le fait de devoir estimer la matrice de déconvolution a d'importantes conséquences par rapport à ce qu'on a vu avec la déconvolution d'une densité sur $\mathbb{R}_+$, où la loi du bruit était supposée connue. En effet, les auteurs établissent que :

$$\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}}^2 = \mathrm{O_P}\left(mT^{-1}\right).$$

Ainsi, là où l'estimation des coefficients de g donne lieu à un terme de variance qui ne dépend pas de m, ce n'est plus le cas pour la matrice de deconvolution. Ce résultat conduit au résultat suivant sur le risque L^2 de l'estimateur :

$$\|\phi - \widehat{\phi}_m\|_{\mathrm{L}^2}^2 = \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \mathrm{O_P}\left(mT^{-1}\right),$$

avec un terme de variance qui dépend de m. Le compromis biais-variance donne alors la vitesse $\mathrm{O_P}\left(T^{-\frac{s}{s+1}}\right)$ sur les espaces $\mathrm{W}^s(\mathbb{R}_+, L)$. Cette approche a ensuite été étendue à des modèles plus généraux par Zhang & Su (2019), Su *et al.* (2019) et Su *et al.* (2020).

Contribution

Au chapitre 3, nous améliorons le travail de Zhang & Su (2018) en étudiant le risque L^2 de l'estimateur par projection sur la base de Laguerre. On suppose connu c et inconnus (λ, f, μ), et on veut estimer la fonction de Gerber–Shiu à partir des observations $\{N_T; X_1, \dots, X_{N_T}\}$, pour $T > 0$. On montre que les résultats asymptotiques en $\mathrm{O_P}$ restent vrais en espérance pour T fini. Concernant les estimateur par projection de g et h, on montre que si $\delta = 0$ alors on a :

$$\mathbb{E}\|g - \widehat{g}_m\|_{\mathrm{L}^2}^2 \leq \|g - g_m\|_{\mathrm{L}^2}^2 + \frac{\lambda}{c^2 T}\mathbb{E}[X],$$
$$\mathbb{E}\|h - \widehat{h}_m\|_{\mathrm{L}^2}^2 \leq \|h - h_m\|_{\mathrm{L}^2}^2 + \frac{\lambda}{c^2 T}\mathbb{E}[W(X)],$$

et si $\delta > 0$ alors on a :

$$\mathbb{E}\|g - \widehat{g}_m\|_{\mathrm{L}^2}^2 \leq \|g - g_m\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2 T}\left(\mathbb{E}[X] + \frac{\mathbb{E}\left[X^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),$$
$$\mathbb{E}\|h - \widehat{h}_m\|_{\mathrm{L}^2}^2 \leq \|h - h_m\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2 T}\left(\mathbb{E}[W(X)] + \frac{\mathbb{E}\left[W(X)^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),$$

où $W(X) \coloneqq \int_0^X \left(\int_u^X w(x, X-x)\,\mathrm{d}x\right)^2 \mathrm{d}u$.

Pour obtenir un résultat sur le risque de $\widehat{\phi}_m$, l'inversion de $\widehat{\mathbf{A}}_m$ pose problème. Pour contourner ce problème, on introduit une troncature. Pour $\theta_0 < 1$ fixé, on pose :

$$\widehat{\mathbf{a}}_m^{\mathrm{Lag}_1} \coloneqq \widetilde{\mathbf{A}}_{m,1}^{-1} \times \widehat{\mathbf{c}}_m, \quad \widetilde{\mathbf{A}}_{m,1}^{-1} \coloneqq \widehat{\mathbf{A}}_m^{-1}\mathbf{1}_{\Delta_{m,1}}, \quad \Delta_{m,1} \coloneqq \left\{\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} < \frac{2}{1-\theta_0}\right\}.$$

On montre que si $\theta < \theta_0$ alors on a :

$$\forall m \in \mathbb{N}^*, \quad \mathbb{E}\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + C\,\frac{m}{T},$$

où C est une constante qui dépend des paramètres du modèle.

Dans le cas $\delta = 0$, il est possible d'éviter la troncature avec un paramètre θ_0 arbitraire, mais au prix d'un facteur $\log m$ supplémentaire dans le terme de variance. Pour cela, on tronque la matrice $\widehat{\mathbf{A}}_m^{-1}$ différemment :

$$\widehat{\mathbf{a}}_m^{\mathrm{Lag}_2} \coloneqq \widetilde{\mathbf{A}}_{m,2}^{-1} \times \widehat{\mathbf{c}}_m, \quad \widetilde{\mathbf{A}}_{m,2}^{-1} \coloneqq \widehat{\mathbf{A}}_m^{-1}\mathbf{1}_{\Delta_{m,2}}, \quad \Delta_{m,2} \coloneqq \left\{\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}}^2 < \frac{cT}{m\log m}\right\}.$$

Si $m\log m \leq cT$ alors on montre que :

$$\mathbb{E}\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{cT(1-\theta)^2}\left(\frac{\mathbb{E}[W(X)]}{c} + \|\phi\|_{\mathrm{L}^2}^2(\mu + \mu^2)\,m\log m\right) + \mathrm{O}\left(\frac{1}{T^2}\right).$$

On propose également un nouvel estimateur de ϕ par une méthode hybride Laguerre–Fourier. L'idée est de calculer les coefficients de ϕ à l'aide de l'isométrie de Plancherel :

$$a_k = \langle \phi, \psi_k \rangle = \frac{1}{2\pi} \langle \mathscr{F}\phi, \mathscr{F}\psi_k \rangle = \frac{1}{2\pi} \left\langle \frac{\mathscr{F}h}{1-\mathscr{F}g}, \mathscr{F}\psi_k \right\rangle,$$

où la dernière égalité découle de (1.35). On estime alors a_k en substituant à g et h les estimateurs $\widehat{g}_{m_2}$ et $\widehat{h}_{m_3}$ dans l'égalité ci-dessus. Cependant, la division par $1-\mathscr{F}\widehat{g}$ risque de poser un problème, c'est pourquoi on a recourt à une troncature. Pour $\theta_0 < 1$ fixé, on estime a_k par :

$$\widehat{a}_{k,m_2,m_3} \coloneqq \frac{1}{2\pi} \left\langle \frac{\mathscr{F}\widehat{h}_{m_3}}{1-\widetilde{\mathscr{F}g}_{m_2}}, \mathscr{F}\psi_k \right\rangle, \qquad \widetilde{\mathscr{F}g}_{m_2} \coloneqq (\mathscr{F}\widehat{g}_{m_2})\, \mathbf{1}_{\{|\mathscr{F}\widehat{g}_{m_2}| < \theta_0\}}.$$

La fonction ϕ est estimée par $\widehat{\phi}_{m_1,m_2,m_3} \coloneqq \sum_{k=0}^{m_1-1} \widehat{a}_{k,m_2,m_3}\psi_k$. On montre que si $\theta < \theta_0$ alors on a :

$$\mathbb{E}\|\phi - \widehat{\phi}_{m_1,m_2,m_3}\|^2_{\mathrm{L}^2} \leq \|\phi - \phi_{m_1}\|^2_{\mathrm{L}^2} + \frac{C}{(1-\theta_0)^2}\left(\|g-g_{m_2}\|^2_{\mathrm{L}^2} + \|h-h_{m_3}\|^2_{\mathrm{L}^2} + \frac{1}{cT}\right),$$

où C est une constante qui dépend des paramètres du modèle. Le risque de notre estimateur est majoré par la somme des termes de biais des fonctions ϕ, g et h et un terme de variance d'ordre T^{-1} qui ne dépend pas de m. Si les fonctions appartiennent à des espaces de Sobolev–Laguerre, alors la vitesse de convergence de notre estimateur est T^{-1}.

Perspectives

On a vu que la méthode d'estimation utilisant les propriétés de déconvolution de la base de Laguerre ne donne pas la vitesse d'estimation optimale dans le modèle de Cramér–Lundberg. Or cette méthode a été utilisée par Zhang & Su (2019), Su *et al.* (2019) et Su *et al.* (2020) dans des modèles plus généraux. On peut conjecturer que les vitesses d'estimation trouvées dans ces articles ne sont pas optimales non plus. Comme on l'explique dans la remarque 3.2.2, notre méthode d'estimation hybride Laguerre–Fourier peut se généraliser à ces modèles et devrait conduire à la vitesse d'estimation paramétrique T^{-1}.

Une autre voie serait de considérer des conditions d'observation plus difficiles (et plus réalistes). Nous avons considéré qu'on observait une trajectoire complète du processus U_t pour $t \in [0,T]$. À la place, on pourrait considérer que le processus est observé sur une fenêtre de temps $[0,T]$ discrétisée à pas Δ. L'objectif serait de construire un estimateur Laguerre–Fourier dans ce contexte et d'étudier son risque sous différents régimes d'observation, par exemple $\Delta > 0$ fixé et $T \to +\infty$, ou $\Delta \to 0$ et $T \to +\infty$ (observations à haute fréquence).

1.5.4 Régression non paramétrique

Le problème de la régression non paramétrique consiste à estimer une fonction $b\colon A \to \mathbb{R}$ à partir de l'observation bruitée de cette dernière en des points $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ de A. Ces points sont appelés design et peuvent être déterministes (design fixe) ou aléatoires (design aléatoire). Plus précisément, on suppose que l'on observe des paires $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$ vérifiant :

$$Y_i = b(\boldsymbol{X}_i) + \varepsilon_i,$$

où les ε_i sont des variables de bruit. On fait l'hypothèse que les variables ε_i sont i.i.d. centrées, de variance σ^2 et indépendantes des $\boldsymbol{X}_i$ dans le cas du design aléatoire. De plus, dans le cas du design aléatoire, on suppose les variables $\boldsymbol{X}_i$ indépendantes.

Ce problème a été beaucoup étudié dans la littérature et de nombreuses approches ont été proposées (estimateurs à noyau, k plus proches voisins, etc) ; on pourra se référer aux livres de Györfi *et al.* (2002) et Tsybakov (2009) pour une présentation détaillée du sujet. Nous présentons ici l'approche par critère des moindres carrés pénalisé. Étant donnée une collection d'espaces vectoriels de dimension finie $(S_m)_{m \in \mathcal{M}_n}$, on estime b par le minimisateur du contraste des moindres carrés sur S_m :

$$\widehat{b}_m \coloneqq \operatorname*{arg\,min}_{t \in S_m} \gamma_n(t), \quad \gamma_n(t) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - t(\boldsymbol{X}_i)\big)^2.$$

Dans le cas d'un design fixe, si on note $\|\cdot\|_n$ la norme empirique associée au design, définie par :

$$\|t\|_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^{n} t(\boldsymbol{X}_i)^2,$$

alors le risque empirique de l'estimateur admet la décomposition biais-variance suivante :

$$\mathbb{E}\|b - \widehat{b}_m\|_n^2 = \inf_{t \in S_m} \|b - t\|_n^2 + \sigma^2 \frac{D_m}{n},$$

où D_m est la dimension de S_m. On choisit m avec une procédure de sélection de modèles par pénalisation :

$$\widehat{m} \coloneqq \operatorname*{arg\,min}_{m \in \mathcal{M}_n} \{\gamma_n(t) + \operatorname{pen}(m)\} = \operatorname*{arg\,min}_{m \in \mathcal{M}_n} \{-\|\widehat{b}_m\|_n^2 + \operatorname{pen}(m)\},$$

où $\operatorname{pen}\colon \mathcal{M}_n \to \mathbb{R}_+$ est la pénalité. Suivant les travaux de Birgé & Massart (1998), la pénalité est choisie de la forme $\operatorname{pen}(m) \coloneqq C_m \sigma^2 \frac{D_m}{n}$, où C_m est un terme lié à la « complexité » de la collection $\mathcal{M}_n$ et qui doit être calibré. Typiquement, les résultats qui sont obtenus disent que sous certaines hypothèses et pour un bon choix de C_m, alors :

$$\mathbb{E}\|b - \widehat{b}_{\widehat{m}}\|_n^2 \leq C \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \|b - t\|_n^2 + \operatorname{pen}(m) \right) + R_n, \tag{1.36}$$

où R_n est un reste d'ordre n^{-1}. Ainsi, toute la problématique réside dans le choix de la collection $\mathscr{M}_n$ et de la calibration C_m. D'un côté il faut choisir une collection suffisamment riche d'espaces d'approximation $(S_m)_{m\in\mathscr{M}_n}$ de façon à ce que le risque oracle :

$$\inf_{m\in\mathscr{M}_n}\left(\inf_{t\in S_m}\|b-t\|_n^2+\sigma^2\frac{D_m}{n}\right),$$

ait une bonne vitesse de convergence sur les espaces fonctionnels considérés. D'un autre côté, la collection ne doit pas être trop complexe afin que la pénalité ne soit pas trop lourde et ne dégrade pas trop l'inégalité (1.36).

Le cas d'un design aléatoire est analogue, à ceci près qu'on ne veut pas un résultat pour la norme empirique mais pour la norme pondérée par la loi du design. Notons μ_i la loi de $\boldsymbol{X}_i$ et notons $\mu \coloneqq \frac{1}{n}\sum_{i=1}^n \mu_i$ la moyenne des lois des $\boldsymbol{X}_i$. L'objectif est d'étudier le risque et de construire un estimateur adaptatif pour le risque associé à la norme de $\mathrm{L}^2(A,\mu)$. L'erreur pour cette norme peut s'interpréter comme une erreur de prédiction :

$$\forall\,\widehat{b}\text{ estimateur},\quad \|b-\widehat{b}\|_\mu^2=\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{X}_i'\sim\mu_i}\left[\left(b(\boldsymbol{X}_i')-\widehat{b}(\boldsymbol{X}_i')\right)^2\,\middle|\,\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\right],$$

à savoir l'erreur quadratique pour une nouvelle observation tirée uniformément suivant les lois μ_i.

Dans leur article, Barron *et al.* (1999) étudient de nombreux problèmes de sélection de modèle par pénalisation. Ils étudient notamment le problème de régression non paramétrique avec un design aléatoire et établissent une inégalité de type (1.36) pour la norme associée au design. Cependant, leur résultat nécessite que les variables ε_i admettent un moment exponentiel et suppose que la fonction b est bornée par une constante connue et qui est utilisée dans la pénalité.

Leur résultat est amélioré par les travaux de Baraud (2000, 2002). Dans le cas d'un design fixe, Baraud (2000) considère une pénalité de forme :

$$\mathrm{pen}(m)\coloneqq(1+\theta)\,\sigma^2\frac{D_m}{n},$$

et montre que si les ε_i admettent un moment d'ordre $q>4$, alors :

$$\mathbb{E}\|b-\widehat{b}_{\widehat{m}}\|_n^2\le C(\theta)\inf_{m\in\mathscr{M}_n}\left(\inf_{t\in S_m}\|b-t\|_n^2+\sigma^2\frac{D_m}{n}\right)+\sigma^2\frac{\Sigma_n(\theta,q)}{n},$$

où le terme de $\Sigma_n(\theta,q)$ est donné par :

$$\Sigma_n(\theta,q)\coloneqq C'(\theta,q)\frac{\mathbb{E}|\varepsilon|^q}{\sigma^q}\left(1+\sum_{m\in\mathscr{M}_n}D_m^{-\frac{q}{2}+2}\right).$$

Le terme $\Sigma_n(\theta,q)$ mesure la complexité de la collection $\mathscr{M}_n$, au sens où si $\mathscr{M}_n$ est une sous-collection de $\mathscr{M}_n'$ alors $\Sigma_n(\theta,q)\le\Sigma_n'(\theta,q)$, tout en tenant compte

de l'intégrabilité des ε_i. Remarquons que :

$$\sum_{m \in \mathscr{M}_n} D_m^{-\frac{q}{2}+2} = \sum_{D=1}^{+\infty} \mathrm{Card}\{m \in \mathscr{M}_n \,|\, D_m = D\} \times D^{-\frac{q}{2}+2},$$

donc si $q > 6$ et si le nombre de modèles de même dimension D est suffisamment petit (par exemple logarithmique en D), alors $\Sigma_n(\theta, q)$ est borné indépendamment de n.

Baraud (2002) étend ce résultat au cas d'un design aléatoire. Il fait pour cela deux hypothèses importantes. Comme la mesure μ est inconnue, il fixe une mesure de référence ν (par exemple la mesure de Lebesgue) et suppose que μ est absolument continue par rapport à ν, de densité qu'on note f. De plus, il suppose que cette densité est minorée par une constante $f_0 > 0$ et majorée par $f_1 < +\infty$. De cette façon, les espaces $\mathrm{L}^2(A,\mu)$ et $\mathrm{L}^2(A,\nu)$ sont les mêmes et les normes $\|\cdot\|_\mu$ et $\|\cdot\|_\nu$ sont équivalentes. Ainsi, il étudiera le risque de l'estimateur en norme $\|\cdot\|_\nu$ et ses hypothèses concerneront la mesure ν. Notons que l'hypothèse que la densité est minorée par une constante nécessite que $\nu(A)$ soit finie, ce qui exclut le cas où $A = \mathbb{R}^d$ et ν est la mesure de Lebesgue. La deuxième hypothèse importante concerne la collection de modèle $\mathscr{M}_n$. Il suppose que celle-ci est constituée de sous-espaces d'un même espace vectoriel de dimension finie $\mathscr{S}_n \subseteq \mathrm{L}^2(A,\nu) \cap \mathrm{L}^\infty(A,\nu)$. En notant N_n la dimension de $\mathscr{S}_n$, il suppose que $N_n < n$ et que[6] :

$$\sup_{t \in \mathscr{S}_n \setminus \{0\}} \frac{\|t\|_\infty^2}{\|t\|_\nu^2} \leq \Phi N_n, \tag{1.37}$$

pour une certaine constante $\Phi \geq 1$. Cette hypothèse se trouve déjà dans Birgé & Massart (1998) et est liée à la structure métrique de $\mathscr{S}_n$ en tant que sous-espace de $\mathrm{L}^2(A,\nu) \cap \mathrm{L}^\infty(A,\nu)$. De plus, si on note $(\varphi_\lambda)_{\lambda \in \Lambda_n}$ une base orthonormée de $\mathscr{S}_n$ pour le produit scalaire de $\mathrm{L}^2(A,\nu)$, alors on a :

$$\sup_{t \in \mathscr{S}_n \setminus \{0\}} \frac{\|t\|_\infty^2}{\|t\|_\nu^2} = \left\| \sum_{\lambda \in \Lambda_n} \varphi_\lambda^2 \right\|_\infty,$$

voir le lemme 1 de Birgé & Massart (1998) ou notre lemme 4.3.3. L'hypothèse (1.37) est donc la même que l'hypothèse $L(\boldsymbol{m}) \leq \Phi D_m$ vue dans l'exemple 1.4.

La clé pour passer d'un contrôle en norme empirique à la norme $\|\cdot\|_\nu$ est d'étudier le comportement de la quantité :

$$K_n^\nu(\mathscr{S}_n) := \sup_{t \in \mathscr{S}_n \setminus \{0\}} \frac{\|t\|_\nu^2}{\|t\|_n^2}.$$

En s'appuyant sur l'inégalité de Bernstein, Baraud (2002) montre que pour tout $\delta > f_0^{-1}$, on a :

$$\mathbb{P}\left[K_n^\nu(\mathscr{S}_n) > \delta\right] \leq N_n^2 \exp\left(-\frac{\left(f_0 - \delta^{-1}\right)^2}{4 f_1} \times \frac{n}{\Phi N_n^2}\right). \tag{1.38}$$

[6] il a également une autre hypothèse pour les bases localisées comme les bases de polynômes par morceaux ou les bases d'ondelettes, mais nous n'en parlerons pas dans cette thèse.

Cette inégalité combinée au résultat de Baraud (2000) permet d'établir une inégalité de type (1.36) pour la norme $\|\cdot\|_v$. En considérant toujours une pénalité de la forme $\text{pen}(m) := (1+\theta)\sigma^2 \frac{D_m}{n}$ et en tronquant la norme de l'estimateur $\widehat{b}_{\widehat{m}}$:

$$\widetilde{b}_n := \widehat{b}_{\widehat{m}} \mathbf{1}_{\{\|\widehat{b}_{\widehat{m}}\|_v \leq 2\exp(\log^2(n))\}},$$

l'auteur montre que si les ε_i admettent un moment d'ordre $q > 4$ et si $N_n^2 \leq \Phi^{-1} n/\log^3(n)$, alors :

$$\mathbb{E}\|b - \widetilde{b}_n\|_v^2 \leq C\left[\inf_{m\in\mathscr{M}_n}\left(\inf_{t\in S_m}\|b-t\|_v^2 + \text{pen}(m)\right) + R_n\right],$$
$$R_n := \frac{\Sigma_n(q)}{n} + \left(1+\|b\|_v^2\right)\exp\left(-2\log^2(n)\right),$$

où C est une constante qui dépend de θ, Φ, f_0, f_1, q et $\mathbb{E}|\varepsilon|^q$, et où $\Sigma_n(q)$ est défini par :

$$\Sigma_n(q) := \sum_{m\in\mathscr{M}_n} D_m^{-\frac{q}{2}+2}.$$

Cohen *et al.* (2013) étudient le problème dans le cas univarié $p = 1$. Ils supposent que $X_1, \ldots, X_n$ sont i.i.d. de loi μ connue. La motivation des auteurs n'est pas vraiment statistique, leur objectif est d'étudier l'approximation d'une fonction sur des sous-espaces S_m et de dégager un critère sur la dimension de S_m afin de connaître le niveau de régularisation requis pour assurer la stabilité de la projection des moindres carrées. D'ailleurs, ils commencent par étudier le problème sans les variables de bruit. Ils introduisent $(\varphi_k)_{1\leq k\leq m}$ une base orthonormée de S_m pour le produit scalaire de $\mathrm{L}^2(A,\mu)$, ainsi que $\widehat{\mathbf{G}}_m \in \mathbb{R}^{m\times m}$ la matrice de Gram empirique des φ_k :

$$\widehat{\mathbf{G}}_m := \left[\langle\varphi_j, \varphi_k\rangle_n\right]_{j,k}.$$

Si cette matrice est inversible, alors la solution du problème des moindres carrés est unique et peut être calculée comme :

$$\widehat{b}_m = \sum_{k=1}^m \widehat{a}_k \varphi_k, \quad \widehat{\mathbf{a}}_m := \frac{1}{n}\widehat{\mathbf{G}}_m^{-1}\widehat{\Phi}_m^*\mathbf{Y},$$

où $\widehat{\Phi}_m := \left[\varphi_k(X_i)\right]_{i,k} \in \mathbb{R}^{n\times m}$ et $\mathbf{Y}$ est le vecteur des Y_i. Ils étudient ensuite comment les normes $\|\cdot\|_n$ et $\|\cdot\|_\mu$ divergent sur S_m, ce qui revient à étudier la déviation de $\widehat{\mathbf{G}}_m$ par rapport à son espérance $\mathbf{I}_m$:

$$\|\widehat{\mathbf{G}}_m - \mathbf{I}_m\|_{\text{op}} = \sup_{t\in S_m\setminus\{0\}} \frac{\left|\|t\|_n^2 - \|t\|_\mu^2\right|}{\|t\|_\mu^2}.$$

En utilisant les inégalités de Chernov matricielles (théorème 1.4.10), ils montrent que pour tout $\delta \in]0,1[$:

$$\mathbb{P}\left[\|\widehat{\mathbf{G}}_m - \mathbf{I}_m\|_{\text{op}} > \delta\right] \leq 2m\exp\left(-c_\delta \frac{n}{L(m)}\right), \tag{1.39}$$

où $c_\delta := (1+\delta)\log(1+\delta)-\delta$ et $L(m) = \left\|\sum_{k=1}^m \varphi_k^2\right\|_\infty$. Cette inégalité est une amélioration de l'inégalité (1.38) de Baraud (2002). Pour $\delta = \frac{1}{2}$, si m vérifie la condition :

$$L(m) \leq \frac{c_{1/2}}{\alpha+1} \times \frac{n}{\log n}, \tag{1.40}$$

avec $\alpha > 0$, alors $\mathbb{P}\big[\|\widehat{\mathbf{G}}_m - \mathbf{I}_m\|_{\mathrm{op}} > \frac{1}{2}\big] \leq 2n^{-\alpha}$. La condition (1.40) est interprétée par les auteurs comme une condition de stabilité. Elle assure qu'avec grande probabilité que la matrice $\widehat{\mathbf{G}}_m$ est bien conditionnée et que la solution des moindres carrés est stable entre les normes $\|\cdot\|_n$ et $\|\cdot\|_\mu$:

$$\|\widehat{b}_m\|_\mu = \|\Pi_m^{(n)}(\mathbf{Y})\|_\mu^2 \leq C\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}, \quad C = \sqrt{6},$$

où $\Pi_m^{(n)}$ désigne le projecteur orthogonal sur S_m pour le produit scalaire empirique. Concernant la partie plus statistique de leur article, ils supposent que b est bornée par une constante L connue et définissent un estimateur tronqué par cette constante :

$$\widehat{b}_{m,L} := T_L\big(\widehat{b}_m\big), \quad T_L(t) := \mathrm{sign}(t) \times \min\big(L, |t|\big).$$

Si m vérifie la condition de stabilité (1.40), ils établissent la majoration suivante sur le risque de leur estimateur :

$$\mathbb{E}\|b - \widehat{b}_{m,L}\|_\mu^2 \leq \left(1 + \frac{8c_{1/2}}{(\alpha+1)\log n}\right) \inf_{t\in S_m} \|b-t\|_\mu^2 + 8\sigma^2\frac{m}{n} + 8L^2 n^{-\alpha}.$$

En revanche, la question de l'adaptation n'est pas abordée dans leur article. Remarquons que si $L(m) \leq \Phi m$ pour une constante Φ, la condition de stabilité peut se réécrire comme $m \leq \Phi^{-1}(\alpha+1)^{-1}c_{1/2}\, n/\log n$. En comparaison, la condition de Baraud (2002) sur la dimension des sous-espaces est d'être inférieure à $\big(\Phi^{-1} n/\log^3(n)\big)^{1/2}$.

Comte & Genon-Catalot (2020b) étudient également le problème dans le cas univarié et supposent que $X_1, \ldots, X_n$ sont i.i.d. de loi μ inconnue mais admettant une densité f par rapport à la mesure de Lebesgue. Leur but est de construire un estimateur par projection sur une base orthonormée $(\varphi_k)_{k\in\mathbb{N}}$ de $\mathrm{L}^2(A)$ et d'étudier le risque pour la norme pondérée par la loi du design. La particularité de leur travail est de considérer des domaines A non compacts, une attention particulière étant portée aux cas $A = \mathbb{R}$ (base d'Hermite) et $A = \mathbb{R}_+$ (base de Laguerre). Dans ce cas, l'hypothèse d'une minoration de la densité f par une constante ne tient plus et les normes $\|\cdot\|_\mu$ et $\|\cdot\|_{\mathrm{L}^2}$ ne sont pas équivalentes, comme c'était le cas dans l'article de Baraud (2002).

Contrairement à l'article de Cohen *et al.* (2013), la base $(\varphi_k)_{k\in\mathbb{N}}$ n'est pas orthonormée pour le produit scalaire de $\mathrm{L}^2(A,\mu)$, donc l'espérance de $\widehat{\mathbf{G}}_m$ n'est pas l'identité mais la matrice :

$$\mathbf{G}_m := \mathbb{E}\big[\widehat{\mathbf{G}}_m\big] = \big[\langle\varphi_j, \varphi_k\rangle_\mu\big]_{j,k} \in \mathbb{R}^{m\times m},$$

où $\langle\cdot,\cdot\rangle_\mu$ est le produit scalaire de $\mathrm{L}^2(A,\mu)$. Les autrices définissent l'évènement $\Omega_m(\delta)$ sur lequel les normes $\|\cdot\|_n$ et $\|\cdot\|_\mu$ sont uniformément proches sur S_m :

$$\Omega_m(\delta) \coloneqq \left\{ \sup_{t\in S_m\setminus\{0\}} \frac{\left|\|t\|_n^2 - \|t\|_\mu^2\right|}{\|t\|_\mu^2} \leq \delta \right\},$$

et montrent à l'aide de l'inégalité de Chernov matricielle une inégalité similaire à (1.39) :

$$\mathbb{P}\big[\Omega_m(\delta)^{\mathrm{c}}\big] = \mathbb{P}\left[\left\| \mathbf{G}_m^{-\frac{1}{2}} \widehat{\mathbf{G}}_m \mathbf{G}_m^{-\frac{1}{2}} - \mathbf{I}_m \right\|_{\mathrm{op}} > \delta \right] \leq 2m \exp\left(-c_\delta \frac{n}{L(m)\left(\|\mathbf{G}_m^{-1}\|_{\mathrm{op}} \vee 1\right)} \right).$$

En prenant $\delta = \frac{1}{2}$ et en supposant que m vérifie la condition :

$$L(m)\left(\|\mathbf{G}_m^{-1}\|_{\mathrm{op}} \vee 1\right) \leq \frac{\mathfrak{c}}{2} \times \frac{n}{\log n}, \quad \mathfrak{c} \coloneqq 2 \times \frac{c_{1/2}}{5}, \tag{1.41}$$

alors $\mathbb{P}\big[\Omega_m(1/2)^{\mathrm{c}}\big] \leq 2n^{-4}$, c'est-à-dire les normes $\|\cdot\|_n$ et $\|\cdot\|_\mu$ sont uniformément proches sur S_m avec grande probabilité.

Les autrices relient la difficulté à passer d'une majoration du risque en norme empirique à la majoration du risque en norme $\|\cdot\|_\mu$ au fait que le problème de régression cache un problème inverse. En effet, les Y_i admettent pour densité le produit de convolution :

$$f_Y(y) \coloneqq \int_A f_\varepsilon\big(y - b(x)\big) f(x)\,\mathrm{d}x,$$

où f_ε est la densité des ε_i, le but étant d'estimer b à partir de l'observation des couples (X_i, Y_i). Le problème inverse se voit également dans le fait que le calcul des coefficients de l'estimateur $\widehat{b}_m$ nécessite l'inversion de la matrice $\widehat{\mathbf{G}}_m$. Il n'est donc pas étonnant que la distance entre 0 et les valeurs propres de $\mathbf{G}_m$ impacte le risque de l'estimateur. Or lorsque A est compact, on peut supposer comme Baraud (2002) que f est minorée par $f_0 > 0$; on a alors la minoration $\|\mathbf{G}_m^{-1}\|_{\mathrm{op}} \leq f_0^{-1}$. Ainsi, les valeurs propres de $\mathbf{G}_m$ sont uniformément minorées et les difficultés qui découlent du problème inverse ne se posent pas vraiment. En revanche si $A = \mathbb{R}$ (resp. $\mathbb{R}_+$) et que la base utilisée est la base d'Hermite (resp. Laguerre), les autrices montrent bien que $\|\mathbf{G}_m^{-1}\|_{\mathrm{op}} \geq C\,m$ pour une certaine constante $C > 0$. Autrement dit, les valeurs propres de $\mathbf{G}_m$ ne sont pas minorées uniformément avec m et la matrice devient de plus en plus singulière à mesure que m croît. D'où l'idée d'introduire une troncature sur les valeurs propres de $\widehat{\mathbf{G}}_m$ en estimant b par :

$$\widetilde{b}_m \coloneqq \widehat{b}_m \mathbf{1}_{\left\{L(m)\left(\|\widehat{\mathbf{G}}_m^{-1}\|_{\mathrm{op}} \vee 1\right) \leq \mathfrak{c} n/\log n\right\}}.$$

Sous l'hypothèse que les ε_i admettent un moment d'ordre 4 et que $b \in \mathrm{L}^4(A,\mu)$ (et donc que les Y_i admettent un moment d'ordre 4), elles montrent que si m vérifie la condition (1.41) alors :

$$\mathbb{E}\|b - \widetilde{b}_m\|_\mu^2 \leq \left(1 + \frac{8\mathfrak{c}}{\log n}\right) \inf_{t\in S_m} \|b - t\|_\mu^2 + 8\sigma^2 \frac{m}{n} + \frac{C}{n},$$

où $C > 0$ dépend de $\mathbb{E}[\varepsilon^4]$ et $\mathbb{E}[b(X)^4]$.

Concernant l'adaptation, elles supposent que $L(m) \leq \Phi m$ pour une certaine constante $\Phi > 0$ et que la densité f est bornée. Idéalement, on voudrait sélectionner parmi les modèles vérifiant la condition $L(m)(\|\widehat{\mathbf{G}}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c} n/\log n$, qui intervient dans la définition de l'estimateur $\widetilde{b}_m$. Cependant, la preuve de leur inégalité oracle, qui utilise l'inégalité de Bernstein matricielle (théorème 1.4.11), requiert une condition plus forte. Elles considèrent donc la collection de modèle aléatoire suivante :

$$\widehat{\mathscr{M}_n} \coloneqq \left\{ m \in \mathbb{N} \,\middle|\, m\left(\|\widehat{\mathbf{G}}_m^{-1}\|_{\mathrm{op}}^2 \vee 1\right) \leq \mathfrak{d}\frac{n}{\log n} \right\}, \quad \mathfrak{d} \coloneqq \left[192\,\Phi \times \left(\left(\|f\|_\infty \vee 1\right) + \frac{1}{3} \right) \right]^{-1}, \tag{1.42}$$

ainsi que sa version théorique :

$$\mathscr{M}_n \coloneqq \left\{ m \in \mathbb{N} \,\middle|\, m\left(\|\mathbf{G}_m^{-1}\|_{\mathrm{op}}^2 \vee 1\right) \leq \frac{\mathfrak{d}}{4} \times \frac{n}{\log n} \right\},$$

et sélectionnent $\widehat{m} \in \widehat{\mathscr{M}_n}$ par critère des moindres carrés pénalisé avec une pénalité de la forme $\mathrm{pen}(m) \coloneqq \kappa\sigma^2 \frac{m}{n}$. Notons que c'est bien le carré de la norme d'opérateur qui est utilisé dans la définition de $\widehat{\mathscr{M}_n}$ et $\mathscr{M}_n$. Sous l'hypothèse que $\mathbb{E}[\varepsilon^6]$ et $\mathbb{E}[b(X)^4]$ sont finis, elles établissent une inégalité de type (1.36) sur la collection $\mathscr{M}_n$ pour la norme empirique et la norme $\|\cdot\|_\mu$, en utilisant l'inégalité de Talagrand (théorème 1.4.8) et l'inégalité de Bernstein matricielle.

Contribution

Dans le chapitre 4, on étend les résultats de Comte & Genon-Catalot (2020b) au cas multivarié. On considère un design aléatoire $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ dans lequel les variables sont indépendantes mais pas identiquement distribuées. On note μ_i la loi de $\boldsymbol{X}_i$ et on note μ la moyenne des μ_i. Comme Baraud (2002), on fixe une mesure de référence ν et on suppose que μ admet une densité bornée par rapport à ν. On suppose que le domaine A s'écrit comme un produit cartésien et que ν est une mesure produit :

$$A = A_1 \times \cdots \times A_p, \quad \nu = \nu_1 \otimes \cdots \otimes \nu_p,$$

où ν_i est une mesure sur A_i. On se donne $(\varphi_k^i)_{k\in\mathbb{N}}$ une base orthonormée de $\mathrm{L}^2(A_i, \nu_i)$ et on tensorise ces bases pour obtenir une base orthonormée $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{N}^p}$ de $\mathrm{L}^2(A, \nu)$. L'exemple qu'on a en tête est le cas où $A = \mathbb{R}^p$, ν est la mesure de Lebesgue et $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{N}^p}$ est la base d'Hermite tensorisée. On estime b par minimisation du contraste des moindres carrés sur les espaces $S_{\boldsymbol{m}} \coloneqq \mathrm{Vect}\left(\varphi_{\boldsymbol{k}} : \boldsymbol{k} \leq \boldsymbol{m} - \mathbf{1}\right)$. Notre estimateur peut être calculé à l'aide d'hypermatrices :

$$\widehat{b}_{\boldsymbol{m}} = \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \widehat{a}_{\boldsymbol{k}}^{(\boldsymbol{m})} \varphi_{\boldsymbol{k}}, \quad \widehat{\mathbf{a}}_{\boldsymbol{k}}^{(\boldsymbol{m})} = \frac{1}{n} \widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} \times_d \widehat{\Phi}_{\boldsymbol{m}}^* \times_1 \mathbf{Y},$$

où $\widehat{\Phi}^*_{\boldsymbol{m}} \coloneqq \big[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)\big]_{\boldsymbol{k},i} \in \mathbb{R}^{\boldsymbol{m}\times n}$, $\mathbf{Y}$ est le vecteur des Y_i et $\widehat{\mathbf{G}}_m$ l'hypermatrice de Gram empirique :

$$\widehat{\mathbf{G}}_m \coloneqq \big[\langle\varphi_{\boldsymbol{j}},\varphi_{\boldsymbol{k}}\rangle_n\big]_{\boldsymbol{j},\boldsymbol{k}} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}, \quad \langle t,s\rangle_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i)s(\boldsymbol{X}_i).$$

On notera $\mathbf{G}_m$ l'espérance de $\widehat{\mathbf{G}}_m$.

Nous avons également besoin de contrôler la probabilité que les normes $\|\cdot\|_\mu$ et $\|\cdot\|_n$ diffèrent l'une de l'autre sur $S_{\boldsymbol{m}}$. On introduit pour cela l'évènement :

$$\forall\delta\in\,]0,1[, \quad \Omega_{\boldsymbol{m}}(\delta) \coloneqq \left\{\forall\, t\in S_{\boldsymbol{m}},\, \|t\|_\mu^2 \le \frac{1}{1-\delta}\|t\|_n^2\right\}.$$

À la différence des travaux de Cohen *et al.* (2013) et Comte & Genon-Catalot (2020b), seule la majoration de la norme $\|\cdot\|_\mu$ par la norme empirique nous intéresse, ce qui nous permet d'avoir une inégalité un peu meilleure sur la probabilité de cet évènement. En utilisant l'inégalité de Chernov matricielle (1.25), on montre que :

$$\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}}\big] \le D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{L(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right),$$

où $h(\delta) \coloneqq \delta + (1-\delta)\log(1-\delta)$ et où $L(\boldsymbol{m}) \coloneqq \|\sum_{\forall i,k_i<m_i}\varphi_{\boldsymbol{k}}^2\|_\infty$. Ainsi, si on se restreint à la collection de modèles de la forme :

$$\mathscr{M}_{n,\alpha}^{(1)} \coloneqq \left\{\boldsymbol{m}\in\mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\big) \le \alpha\frac{n}{\log n}\right\},$$

alors on a :

$$\forall\,\boldsymbol{m}\in\mathscr{M}_{n,\alpha},\ \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}}\big] \le D_{\boldsymbol{m}}\ n^{-\frac{h(\delta)}{\alpha}} \le n^{-\frac{h(\delta)}{\alpha}+1}.$$

Ce résultat permet d'établir une majoration du risque de $\widehat{b}_{\boldsymbol{m}}$ lorsque $\boldsymbol{m}$ appartient à la collection $\mathscr{M}_{n,\alpha}^{(1)}$. Dans toute la suite, on suppose que $b\in \mathrm{L}^{2r}(\mu)$ avec $r\in\,]1,+\infty]$ et on note $r'\in[1,+\infty[$ son exposant conjugué[7]. On montre que pour tous $\alpha\in\,]0,\frac{1}{2r'+1}[$ et $\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(1)}$, on a :

$$\mathbb{E}\|b-\widehat{b}_{\boldsymbol{m}}\|_\mu^2 \le C_n(\alpha,r')\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2 + C'(\alpha,r')\,\sigma^2\frac{D_{\boldsymbol{m}}}{n} + \frac{C''\big(\|b\|_{\mathrm{L}^{2r}(\mu)},\sigma^2,\alpha\big)}{n\log n}.$$

Pour construire un estimateur adaptatif, on considère la version empirique de la collection de modèles $\mathscr{M}_{n,\alpha}^{(1)}$:

$$\widehat{\mathscr{M}}_{n,\beta}^{(1)} \coloneqq \left\{\boldsymbol{m}\in\mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m})\big(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\big) \le \beta\frac{n}{\log n}\right\},$$

avec β un paramètre positif. On sélectionne $\widehat{\boldsymbol{m}}_1$ dans cette collection par critère pénalisé avec une pénalité de la forme $\mathrm{pen}(\boldsymbol{m}) \coloneqq (1+\theta)\sigma^2\frac{D_{\boldsymbol{m}}}{n}$. En s'appuyant sur

[7] l'exposant conjugué de r est le nombre r' tel que $\frac{1}{r}+\frac{1}{r'}=1$.

le résultat de Baraud (2000) pour le cas d'un design fixe et supposant que les ε_i admettent un moment d'ordre $q > 6$, on montre que l'estimateur $\widehat{b}_{\widehat{\boldsymbol{m}}_1}$ satisfait une inégalité oracle pour la norme empirique sur les collections $\mathscr{M}_{n,\alpha}$, pour α inférieur à une quantité qui dépend de β et r. Dans le cas où A est compact, on fait l'hypothèse que la densité de μ est minorée par une constante $f_0 > 0$; la norme d'opérateur de $\mathbf{G}_m^{-1}$ est alors majorée par f_0^{-1}. En tirant à nouveau profit des inégalités de Chernov matricielles et en utilisant cette majoration, on montre que qu'il existe une constante $\beta_{f_0,r}$ telle que si $\beta < \beta_{f_0,r}$, alors l'estimateur $\hat{b}_{\widehat{\boldsymbol{m}}_1}$ vérifie une inégalité oracle en norme $\|\cdot\|_\mu$ sur les collections $\mathscr{M}_{n,\alpha}^{(1)}$, pour α inférieur à une quantité qui dépend de β et r.

Ce résultat dans le cas compact améliore les résultats de Baraud (2002) et Comte & Genon-Catalot (2020b). En effet, si on suppose que $L(\boldsymbol{m}) \leq \Phi D_{\boldsymbol{m}}$ pour constante $\Phi > 0$, et puisque $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\text{op}}$ est majoré uniformément en $\boldsymbol{m}$, la collection $\mathscr{M}_{n,\alpha}^{(1)}$ est en fait équivalente à la collection :

$$\mathscr{M}'_{n,\alpha} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, D_{\boldsymbol{m}} \leq \frac{\alpha f_0}{\Phi} \times \frac{n}{\log n} \right\}.$$

L'estimateur de Baraud (2002) est adaptatif sur une collection construite à partir des sous-espaces d'un espace vectoriel $\mathscr{S}_n$, dont la dimension doit vérifier $\dim(\mathscr{S}_n) \leq \left(\Phi^{-1} n / \log^3(n)\right)^{1/2}$. Notre procédure évite le choix d'un modèle englobant et notre condition sur la dimension des modèles est plus faible. Enfin, notre résultat améliore celui de Comte & Genon-Catalot (2020b) dans le cas compact, puisque notre collection de modèle reste définie à partir de $\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\text{op}}$ et non de son carré (cf. (1.42)).

En revanche dans le cas non compact, on fait face aux mêmes difficultés que Comte & Genon-Catalot (2020b). On suit alors une stratégie similaire à la leur, en considérant des collections plus petites :

$$\mathscr{M}_{n,\alpha}^{(2)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m}) \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\text{op}}^2 \vee 1 \right) \leq \alpha \frac{n}{\log n} \right\},$$
$$\widehat{\mathscr{M}}_{n,\beta}^{(2)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m}) \left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\text{op}}^2 \vee 1 \right) \leq \beta \frac{n}{\log n} \right\},$$

et en sélectionnant $\widehat{\boldsymbol{m}}_2$ parmi les modèles de $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$. En partant de l'inégalité oracle pour la norme empirique et en utilisant l'inégalité de Bernstein matricielle, on montre qu'il existe une constante $\beta_{f_1,r}$ qui dépend du majorant de la densité de μ et de r, telle que si $\beta < \beta_{f_1,r}$, alors $\widehat{b}_{\widehat{\boldsymbol{m}}_2}$ vérifie une inégalité oracle pour la norme $\|\cdot\|_\mu$ sur les collections $\mathscr{M}_{n,\alpha}^{(2)}$, pour α inférieur à une quantité qui dépend de β et r.

Perspectives

Une première suite de ce travail serait d'étendre nos résultats à la régression non paramétrique dans un modèle hétéroscédastique. Dans ce modèle, la variance

du bruit n'est plus constante et devient une fonction qui dépend du design :

$$Y_i = b(\boldsymbol{X}_i) + \sigma(\boldsymbol{X}_i)\varepsilon_i,$$

où les ε_i sont i.i.d. de variance 1. En dimension 1, le problème a été étudié par Comte & Genon-Catalot (2020a) qui étendent les résultats de Comte & Genon-Catalot (2020b) avec succès . Pour le moment, nos résultats dans le modèle homoscédastique s'appuient sur le travail de Baraud (2000). Pour passer au modèle hétéroscédastique, une première possibilité serait de généraliser le résultat en design fixe de Baraud (2000) au cas de variables de bruit ayant des variances différentes. On l'étendrait ensuite en design aléatoire avec la même méthode que celle utilisée au chapitre 4 pour le modèle homoscédastique. Une autre possibilité serait de suivre l'approche de Comte & Genon-Catalot (2020a), basée sur l'inégalité de Talagrand, et de voir si elle se généralise à la dimension supérieure.

Une deuxième suite serait de considérer, comme le fait Baraud (2002), des sous-espaces d'approximation S_m plus généraux, qui ne sont pas associés au choix d'une base orthonormée de $\mathrm{L}^2(A, \nu)$. Cela permettrait d'envisager des estimateurs construits à partir de bases localisées, comme des fonctions polynomiales par morceaux ou des ondelettes. Nous pensons que l'étude des déviations de la norme empirique par rapport à la norme $\|\cdot\|_\mu$ que nous avons menée est encore valable dans ce cadre plus général. Ainsi, nous pensons qu'il est possible d'étendre nos résultats d'adaptation au cas d'un domaine A non-compact et de sous-espaces S_m plus généraux, sans l'hypothèse de minoration sur la densité du design.

Chapter 2

Anisotropic Multivariate Deconvolution Using Projection on the Laguerre Basis

This chapter is a modified version of my article F. DUSSAP : Anisotropic multivariate deconvolution using projection on the Laguerre basis. *Journal of Statistical Planning and Inference*, 215:23-46, 2021.

Contents

2.1 Statistical model and motivations

In this chapter, we study the problem of recovering the distribution of a random vector $\boldsymbol{X}$ when we only observe its sum with a noise vector $\boldsymbol{Y}$ with known distribution. This is a classical problem in nonparametric statistics (see references below), but we focus on the particular case where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ have non-negative coordinates.

This assumption is quite unusual in deconvolution problems but is relevant for instance in reliability fields: we observe the failure times of several components in a system, each failure time being the sum of the lifetimes of two subcomponents. In survival analysis, for $d = 1$, X would be the time of infection of a disease and Y the incubation time. The multivariate case can then be used to study the time of infection for multiple diseases.

More precisely, we consider the following statistical model:

$$\boldsymbol{Z}_i = \boldsymbol{X}_i + \boldsymbol{Y}_i, \quad i = 1, \dots, n, \tag{2.1}$$

where the $\boldsymbol{X}_i$s, the $\boldsymbol{Y}_i$s and $\boldsymbol{Z}_i$s are random vectors in $\mathbb{R}^d$ with non-negative coordinates. We assume that the $\boldsymbol{X}_i$s are i.i.d. with unknown density f on $\mathbb{R}_+^d$, and that the $\boldsymbol{Y}_i$s are i.i.d. with known density g on $\mathbb{R}_+^d$. Moreover, we assume that the $\boldsymbol{X}_i$s and the $\boldsymbol{Y}_i$s are independent. Our goal is to provide an adaptive procedure to estimate the density f from the observations $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$.

In the univariate case, there are a lot of papers about recovering the density of a random variable when it is observed with an additive known noise. Many authors use a kernel estimator introduced by Stefanski & Carroll (1990). Fan (1991) first introduced the notion of *ordinary smooth* and *supersmooth* noise (rate of decay of the characteristic function) to study optimal rates of convergence on Hölder spaces. When f is supersmooth and the noise is ordinary smooth, Butucea (2004) showed that the kernel estimator achieves a good rate of convergence and that this rate is optimal. When both f and the noise are supersmooth, the problem is more complicated and have been studied by Butucea & Tsybakov (2008a,b) from lower bound point of view.

To provide an adaptive estimator, different procedures were proposed. When f belongs to a Sobolev space, Pensky & Vidakovic (1999) proposed a wavelet strategy that is adaptive and achieves optimal rates of convergence. For kernel estimators, Delaigle & Gijbels (2004) estimated the optimal bandwidth with a bootstrap procedure and showed its consistency. Hazelton & Turlach (2009, 2010) proposed a weighted kernel estimator with a data-driven way to choose the weights. Moreover, their weighted kernel estimator can be used in a multivariate setting. For projection with penalization strategies, Comte *et al.* (2006) used a Shannon type basis to construct an adaptive estimator that is minimax in most cases. More recently, non-compactly supported bases were used by Mabon (2017) (Laguerre basis), by Comte & Genon-Catalot (2018) (Laguerre basis and Hermite basis), and by Sacko (2020) (Hermite basis) to construct adaptive estimators on suitable functional spaces.

The Laguerre basis was also used in a regression setting to study the problem of Laplace deconvolution, see Comte *et al.* (2017) and Vareschi (2015), or more recently Benhaddou *et al.* (2019).

The multivariate deconvolution literature is more sparse. Masry (1991) generalizes the kernel estimator for stationary random processes, with a dependence structure between the variables. The noise is assumed to have i.i.d. coordinates

(isotropic noise) and no adaptive strategy is proposed, the author focuses on the problem of dependency between the variables. Youndjé & Wells (2008) propose a cross-validation strategy to estimate the optimal bandwidth of the kernel estimator in the multivariate setting, and show it is asymptotically optimal under the assumption that the noise is isotropic and ordinary smooth. Comte & Lacour (2013) use a bandwidth selection procedure inspired by Goldenshluger & Lepski (2011). Their procedure allows anisotropic noises, with both ordinary smooth and supersmooth components, and they derive rates of convergence for the pointwise risk and the L^2 risk, when f belongs to anisotropic Hölder, Sobolev, or Nikol'skii classes. For ordinary smooth noise and when f belongs to anisotropic Nikol'skii classes, Rebelles (2016) provides an adaptive kernel estimator which is minimax for the L^p loss. Recently, Lepski & Willer (2019) studied a more general model (with direct and indirect observations of $\boldsymbol{X}$) and provided an adaptive kernel estimator on anisotropic Nikol'skii classes, under the L^p loss.

Concerning our specific case of deconvolution with non-negative noise, the case $d = 1$ has already been studied by Mabon (2017) using a projection strategy on the Laguerre basis. We also use a projection strategy in the multivariate case. The main tool we use to construct our estimator is the theory of hypermatrices. Using the contraction product of hypermatrices, we show that it is possible to recover the coefficients of f from the observations. We recall the definitions of these objects in Section 1.3.

We provide rates of convergence for the MISE of our estimator on anisotropic functional spaces: Sobolev–Laguerre spaces and smooth Laguerre spaces. We propose a model selection procedure to produce an adaptive estimator, under mild assumptions on the noise density g. This procedure is inspired by the work of Goldenshluger & Lepski (2011) concerning bandwidth selection. It was introduced for model selection by Chagny (2013c) for estimation of conditional density, in a two-dimensional setting. We show this procedure can be applied to our deconvolution problem, in a d-dimensional setting, and we establish an asymptotic oracle inequality for this procedure. Moreover, the proof is written to provide general steps that can be applied to other contexts.

Outline of the paper In Section 2.2, we construct the estimator. In section 2.3, we provide non asymptotic MISE bounds and we derive convergence rates on Sobolev–Laguerre balls and smooth Laguerre balls. In Section 2.4, we give model selection procedures to construct an adaptive estimator and we establish oracle inequalities. In Section 2.5, we illustrate the procedures on simulated data. All the proofs are gathered in Section 2.6.

2.2 The estimation procedure

In the model (2.1), the $\boldsymbol{Z}_i$s are i.i.d. random vectors on $\mathbb{R}_+^d$, and they admit a density we denote by h. This density function is given by the convolution product of

f and g:

$$(f * g)(\boldsymbol{x}) := \int_{\mathbb{R}^d} f(\boldsymbol{u})\, g(\boldsymbol{x}-\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \int_{[0,x_1]\times\cdots\times[0,x_d]} \cdots \int f(\boldsymbol{u})\, g(\boldsymbol{x}-\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}.$$

We assume that f, g and h belong to $\mathrm{L}^2(\mathbb{R}_+^d)$ and we expand them in the multivariate Laguerre basis. We recall that the one-dimensional Laguerre functions $(\varphi_k)_{k\in\mathbb{N}}$ are defined by:

$$\forall x \in \mathbb{R}_+,\ \varphi_k(x) := \sqrt{2}\, L_k(2x)\, \mathrm{e}^{-x}, \text{ where } L_k(x) := \sum_{j=0}^{k} \binom{k}{j} \frac{(-x)^j}{j!},$$

and that they form an orthonormal basis of $\mathrm{L}^2(\mathbb{R}_+)$. In the multivariate case, we *tensorize* the Laguerre basis. For $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d$ a multi-index, we define the multivariate Laguerre function φ_k on $\mathbb{R}_+^d$ as the tensor product of one-dimensional Laguerre functions:

$$\varphi_{\boldsymbol{k}}(x_1, \ldots, x_d) := \big(\varphi_{k_1} \otimes \cdots \otimes \varphi_{k_d}\big)(x_1, \ldots, x_d) = \varphi_{k_1}(x_1) \times \cdots \times \varphi_{k_d}(x_d).$$

The multivariate Laguerre functions $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{N}^d}$ form a basis of $\mathrm{L}^2(\mathbb{R}_+^d)$, and we expand the functions f, g and h in this basis:

$$f = \sum_{\boldsymbol{k}\in\mathbb{N}^d} a_{\boldsymbol{k}}\, \varphi_{\boldsymbol{k}}, \qquad g = \sum_{\boldsymbol{j}\in\mathbb{N}^d} b_{\boldsymbol{j}}\, \varphi_{\boldsymbol{j}}, \qquad h = \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} c_{\boldsymbol{\ell}}\, \varphi_{\boldsymbol{\ell}}. \tag{2.2}$$

The use of the Laguerre basis is relevant in this context, as the one-dimensional Laguerre functions satisfy the relation:

$$\forall k, j \in \mathbb{N}, \quad \varphi_k * \varphi_j = 2^{-1/2}(\varphi_{k+j} - \varphi_{k+j+1}), \tag{2.3}$$

see (Abramowitz & Stegun, 1972, formula 22.13.14). Using this relation and $h = f * g$, by expanding the functions f, g and h on the Laguerre basis, we get a relation between their coefficients.

Proposition 2.2.1. *If a, b and c are the coefficients defined in* (2.2)*, then the following relation holds:*

$$\forall \boldsymbol{\ell} \in \mathbb{N}^d, \quad c_{\boldsymbol{\ell}} = 2^{-d/2} \sum_{\boldsymbol{\varepsilon}\in\{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|} (a * b)_{\boldsymbol{\ell}-\boldsymbol{\varepsilon}},$$

where $a * b$ is the discrete convolution product of a and b defined by $(a * b)_{\boldsymbol{k}} := \sum_{\boldsymbol{j}\le\boldsymbol{k}} a_{\boldsymbol{j}} b_{\boldsymbol{k}-\boldsymbol{j}}$ if $\boldsymbol{k} \in \mathbb{N}^d$ and $(a * b)_{\boldsymbol{k}} = 0$ if $\boldsymbol{k} \notin \mathbb{N}^d$.

This relation can be written as a discrete convolution product $c = \beta * a$ with $\beta \in \mathbb{N}^d$ defined by:

$$\beta_{\boldsymbol{k}} := 2^{-d/2} \sum_{\boldsymbol{\varepsilon}\in\{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|}\, b_{\boldsymbol{k}-\boldsymbol{\varepsilon}}, \tag{2.4}$$

where by convention $b_{\boldsymbol{j}} = 0$ if $\boldsymbol{j} \notin \mathbb{N}^d$. Thus, we have a linear relation between the coefficients c and a:

$$\forall \boldsymbol{\ell} \in \mathbb{N}^d,\ c_{\boldsymbol{\ell}} = \sum_{\boldsymbol{k} \le \boldsymbol{\ell}} \mathbf{G}_{\boldsymbol{\ell k}}\, a_{\boldsymbol{k}},\ \text{where } \mathbf{G}_{\boldsymbol{\ell k}} \coloneqq \begin{cases} \beta_{\boldsymbol{\ell}-\boldsymbol{k}} & \text{if } \boldsymbol{k} \le \boldsymbol{\ell}, \\ 0 & \text{else.} \end{cases} \tag{2.5}$$

If we consider $\mathbf{G}$ as an infinite hypermatrix $[\mathbf{G}_{\boldsymbol{\ell k}}]_{\ell,k \in \mathbb{N}^d} \in \mathbb{R}^{\mathbb{N}^d \times \mathbb{N}^d}$, then it is lower triangular according to the following definition.

Definition 2.2.2. A hypermatrix $\mathbf{T} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}}$ is said to be lower triangular if apart from multi-indices $\boldsymbol{\ell}, \boldsymbol{k} \le \boldsymbol{m} - \mathbf{1}$ such that $\boldsymbol{k} \le \boldsymbol{\ell}$, we have $T_{\boldsymbol{\ell k}} = 0$.
An infinite hypermatrix $\mathbf{T} \in \mathbb{R}^{\mathbb{N}^d \times \mathbb{N}^d}$ is said to be lower triangular if apart from multi-indices $\boldsymbol{\ell}, \boldsymbol{k} \in \mathbb{N}^d$ such that $\boldsymbol{k} \le \boldsymbol{\ell}$, we have $T_{\boldsymbol{\ell k}} = 0$.

In the next proposition, we show that the linear relation between a and c is invertible.

Proposition 2.2.3. *Let* $\mathbf{G}$ *be the infinite hypermatrix defined in* (2.5)*. For every* $\boldsymbol{k} \in \mathbb{N}^d$*, there exists* $[\mathbf{H}_{\boldsymbol{k\ell}}]_{\boldsymbol{\ell} \le \boldsymbol{k}}$ *such that for every* $a, c \in \mathbb{R}^{\mathbb{N}^d}$ *statisfying the relation* (2.5)*, we have:*

$$a_{\boldsymbol{\ell}} = \sum_{\boldsymbol{\ell} \le \boldsymbol{k}} \mathbf{H}_{\boldsymbol{\ell k}}\, c_{\boldsymbol{k}}. \tag{2.6}$$

We denote $\mathbf{H}_{\boldsymbol{\ell k}} =: [\mathbf{G}^{-1}]_{\boldsymbol{\ell k}}$.

We can write the linear relations between a and c using hypermatrices and contraction products. For $\boldsymbol{m} \in \mathbb{N}_+^d$, we denote by $\mathbf{a}_{\boldsymbol{m}}$ (resp. $\mathbf{c}_{\boldsymbol{m}}$) the hypermatrix $[a_{\boldsymbol{k}}]_{\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}} \in \mathbb{R}^{\boldsymbol{m}}$ (resp. $[c_{\boldsymbol{\ell}}]_{\boldsymbol{\ell} \le \boldsymbol{m}-\mathbf{1}} \in \mathbb{R}^{\boldsymbol{m}}$), and we denote by $\mathbf{G}_{\boldsymbol{m}}$ and $\mathbf{G}_{\boldsymbol{m}}^{-1}$ the hypermatrices $[\mathbf{G}_{\boldsymbol{\ell k}}]_{\boldsymbol{\ell}, \boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}}$ and $[(\mathbf{G}^{-1})_{\boldsymbol{\ell k}}]_{\boldsymbol{\ell}, \boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}}$ in $\mathbb{R}^{(\boldsymbol{m},\boldsymbol{m})}$. Then, we have:

$$\mathbf{c}_{\boldsymbol{m}} = \mathbf{G}_{\boldsymbol{m}} \times_d \mathbf{a}_{\boldsymbol{m}} \iff \mathbf{a}_{\boldsymbol{m}} = \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \mathbf{c}_{\boldsymbol{m}}, \tag{2.7}$$

where "$\times_d$" stands for the contraction product defined by (1.5).

Estimation procedure For $\boldsymbol{m} \in \mathbb{N}_+^d$, let $S_{\boldsymbol{m}}$ be the vector space spanned by the functions $\varphi_{\boldsymbol{k}}$ for $\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}$, and let $D_{\boldsymbol{m}} \coloneqq m_1 \cdots m_d$ be its dimension. We estimate f by estimating $f_{\boldsymbol{m}}$ the projection of f on $S_{\boldsymbol{m}}$. This projection is given by $f_{\boldsymbol{m}} = \sum_{\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}} a_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}$, so the problem reduces to the estimation of $\mathbf{a}_{\boldsymbol{m}}$. Because $\mathbf{a}_{\boldsymbol{m}}$ is related to $\mathbf{c}_{\boldsymbol{m}}$ by (2.7) and since $c_{\boldsymbol{\ell}} = \mathbb{E}[\varphi_{\boldsymbol{\ell}}(\boldsymbol{Z}_1)]$, we estimate f by:

$$\hat{f}_{\boldsymbol{m}} \coloneqq \sum_{\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}} \widehat{a}_{\boldsymbol{k}}\, \varphi_{\boldsymbol{k}} \text{ where } \widehat{\mathbf{a}}_{\boldsymbol{m}} \coloneqq \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \widehat{\mathbf{c}}_{\boldsymbol{m}} \text{ and } \widehat{c}_{\boldsymbol{\ell}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \varphi_{\boldsymbol{\ell}}(\boldsymbol{Z}_i).$$

2.3 Non-asymptotic error bounds

We quantify the quality of the estimator $\widehat{f}_{\boldsymbol{m}}$ by its MISE (Mean Integrated Squared Error): $\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|^2_{\mathrm{L}^2}$. In the next proposition, we decompose the MISE in a bias term and a variance term and give a bound on the MISE of $\widehat{f}_{\boldsymbol{m}}$.

Proposition 2.3.1. *If f and g are $\mathrm{L}^2(\mathbb{R}^d_+)$ functions, then we have the inequality:*

$$\forall \boldsymbol{m} \in \mathbb{N}^d_+, \quad \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} \leq \|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \frac{2^d D_{\boldsymbol{m}} \|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_{\mathrm{op}}}{n} \wedge \frac{\|h\|_\infty \|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_F}{n}.$$

Remark 2.3.2. The norm equivalence (1.6) implies that $\|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_F \leq D_{\boldsymbol{m}} \|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_{\mathrm{op}}$, so the order of magnitude of the variance term is given by $\|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_F$. The minimum is important only for the small values of $\boldsymbol{m}$ (because of the constants).

Remark 2.3.3. Using the Cauchy–Schwarz inequality, it holds $\|h\|_\infty \leq \|f\|_{\mathrm{L}^2}\|g\|_{\mathrm{L}^2}$ which is finite by assumption. Moreover, if g is bounded, we also have $\|h\|_\infty \leq \|g\|_\infty$, so that the bound on the variance term does not depend on unknown quantities. We make this assumption in the following.

Assumption 2.1. We assume that g is bounded.

Under an additional assumption, Comte & Genon-Catalot (2018) improved the variance bound in the one-dimensional case. We generalize their result to the multivariate case.

Assumption 2.2. We denote $Y_1^{(j)}$ the j-th coordinate of $\boldsymbol{Y}_1$. We assume that for every nonempty subset J of $\{1, \dots, d\}$, we have:

$$M_J(g) \coloneqq \mathbb{E}\left[\prod_{j \in J} \frac{1}{\sqrt{Y_1^{(j)}}}\right] < +\infty.$$

Proposition 2.3.4. *Under Assumption 2.2, we have:*

$$\forall \boldsymbol{m} \in \mathbb{N}^d_+, \quad \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} \leq \|f - f_{\boldsymbol{m}}\|^2_{\mathrm{L}^2} + \frac{c(g)\sqrt{D_{\boldsymbol{m}}} \|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_{\mathrm{op}}}{n} \wedge \frac{\|h\|_\infty \|\mathbf{G}^{-1}_{\boldsymbol{m}}\|^2_F}{n},$$

where $c(g)$ is a positive constant depending on $M_J(g)$ for $J \subseteq \{1, \dots, d\}$.

To study the bias term, we assume that f belongs to a Sobolev–Laguerre space. In dimension $d = 1$, these functional spaces have been introduced by Bongioanni & Torrea (2009) to study the Laguerre operator. The connection with Laguerre coefficients was established later by Comte & Genon-Catalot (2015). Following the same idea, we define Sobolev–Laguerre balls on $\mathbb{R}^d_+$.

Definition 2.3.5 (Sobolev–Laguerre ball)**.** Let $L > 0$ and $\boldsymbol{s} \in (0, +\infty)^d$, we define the Sobolev–Laguerre ball of order $\boldsymbol{s}$ and radius L by:

$$\mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L) \coloneqq \left\{ f \in \mathrm{L}^2(\mathbb{R}_+^d) \,\middle|\, \sum_{\boldsymbol{k} \in \mathbb{N}^d} a_{\boldsymbol{k}}^2(f)\, \boldsymbol{k}^{\boldsymbol{s}} \leq L \right\},$$

with $a_{\boldsymbol{k}}(f) \coloneqq \langle f, \varphi_{\boldsymbol{k}} \rangle_{\mathrm{L}^2}$ the Laguerre coefficients of f.

Assuming f to belong to $\mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L)$, the bias term decreases to 0 with polynomial rate. Indeed, for $\boldsymbol{m} \in \mathbb{N}_+^d$, we have:

$$\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \sum_{\substack{\boldsymbol{k} \in \mathbb{N}^d \\ \exists q,\, k_q \geq m_q}} a_{\boldsymbol{k}}^2(f) \leq \sum_{q=1}^{d} \sum_{\substack{\boldsymbol{k} \in \mathbb{N}^d \\ k_q \geq m_q}} a_{\boldsymbol{k}}^2(f)\, k_q^{s_q}\, k_q^{-s_q} \leq L \sum_{q=1}^{d} m_q^{-s_q}. \tag{2.8}$$

The case where the Laguerre coefficients of f decrease with exponential rate is also interesting. We define new functional spaces, "*smooth Laguerre spaces*", in the following way.

Definition 2.3.6 (Smooth Laguerre ball)**.** Let $L > 0$ and $\boldsymbol{r} \in (0, +\infty)^d$, we define the smooth Laguerre ball of order $\boldsymbol{r}$ and radius L as:

$$\mathscr{S}^{\boldsymbol{r}}(\mathbb{R}_+^d, L) \coloneqq \left\{ f \in \mathrm{L}^2(\mathbb{R}_+^d) \,\middle|\, \sum_{\boldsymbol{k} \in \mathbb{N}^d} a_{\boldsymbol{k}}^2(f)\, \mathrm{e}^{\boldsymbol{r} \cdot \boldsymbol{k}} \leq L \right\}.$$

By the same argument as previously, if f belongs to $\mathscr{S}^{\boldsymbol{r}}(\mathbb{R}_+^d, L)$, the bias term decreases to 0 with exponential rate:

$$\forall \boldsymbol{m} \in \mathbb{N}_+^d, \quad \|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq L \sum_{q=1}^{d} \mathrm{e}^{-r_q m_q}. \tag{2.9}$$

Now, we need to control the variance term in Proposition 2.3.1. In the one-dimensional case, this control is provided under assumptions on the Laplace transform G of g and under assumptions on the derivatives of g, see lemma 3.6 in (Comte *et al.*, 2017). In the next proposition, we extend this result to the multivariate case. Moreover, we make assumptions only on the behavior of the Laplace transform of g, we do not need to study its differentials.

We recall that the Laplace transform of g is the function G defined on the domain $\mathscr{P}_+^d$ by:

$$G(\boldsymbol{s}) \coloneqq \int_{\mathbb{R}_+^d} \mathrm{e}^{-\boldsymbol{s} \cdot \boldsymbol{x}} g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where $\mathscr{P}_+$ stands for the set of complex numbers with non-negative real part. In addition, we extend the set $\mathbb{C}$ of complex numbers by adding a point at infinity denoted by ∞. The control of the Frobenius norm of $\mathbf{G}_{\boldsymbol{m}}^{-1}$ relies on the behavior of G when some of its arguments take the ∞ value.

Proposition 2.3.7. *We assume that $\beta \in \ell^1(\mathbb{N}^d)$, with β defined in* (2.4)*. We assume that G is non-zero on $\mathscr{P}_+^d$ and that there exists $\boldsymbol{\alpha} \in \mathbb{N}_+^d$ such that the function:*

$$K_{\boldsymbol{\alpha}}(\boldsymbol{s}) := (\mathbf{1}+\boldsymbol{s})^{\boldsymbol{\alpha}} G(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathscr{P}_+^d,$$

can be extended as a non-zero function on $(\mathscr{P}_+ \cup \{\infty\})^d$ such that the restriction of $K_{\boldsymbol{\alpha}}$ on $(\mathrm{i}\mathbb{R} \cup \{\infty\})^d$ is continuous. Then for $\boldsymbol{m} \in \mathbb{N}^d$ satisfying $\boldsymbol{m} \geq \mathbf{4}$, there exists a constant $C > 0$ depending on β such that $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2 \leq C\boldsymbol{m}^{2\boldsymbol{\alpha}}$.

Remark 2.3.8. If $g \in \mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L)$ with $L > 0$ and $\boldsymbol{s} \in (1,+\infty)^d$, the Laguerre coefficients of g belong to $\ell^1(\mathbb{N}^d)$. Indeed, using Cauchy–Schwarz inequality,

$$\sum_{\boldsymbol{k}\in\mathbb{N}^d} |b_{\boldsymbol{k}}| = \sum_{\boldsymbol{k}\in\mathbb{N}^d} |b_{\boldsymbol{k}}|\, \boldsymbol{k}^{\frac{\boldsymbol{s}}{2}}\, \boldsymbol{k}^{-\frac{\boldsymbol{s}}{2}} \leq \left(\sum_{\boldsymbol{k}\in\mathbb{N}^d} b_{\boldsymbol{k}}^2\, \boldsymbol{k}^{\boldsymbol{s}}\right)^{\frac{1}{2}} \left(\sum_{\boldsymbol{k}\in\mathbb{N}^d} \boldsymbol{k}^{-\boldsymbol{s}}\right)^{\frac{1}{2}} < +\infty,$$

because for every $q \in \{1,\dots,d\}$, $s_q > 1$.

Remark 2.3.9. We notice that in dimension $d = 1$, the assumptions of Proposition 2.3.7 simply become:

1. The Laplace transform G does not vanish on $\mathscr{P}_+$.
2. The Fourier transform of g admits an asymptotic expansion:

$$g^*(\omega) = \omega^{-\alpha}\big(K_\alpha + \mathrm{o}(1)\big),$$

when $|\omega|$ goes to $+\infty$, for some $\alpha \in \mathbb{N}^*$ and some non-zero constant K_α.

It is easy to see that this second assumption is a consequence of the assumptions on the derivatives of g made in (Comte *et al.*, 2017, Subsection 2.5).

Remark 2.3.10. If the distribution of $\boldsymbol{Y}_i$s is a product of gamma distributions $\bigotimes_{q=1}^d \Gamma(\alpha_q, \lambda_q)$ with $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_d) \in \mathbb{N}_+^d$, then G and $K_{\boldsymbol{\alpha}}$ are given by:

$$G(\boldsymbol{s}) = \prod_{q=1}^d \left(1 + \frac{s_q}{\lambda_q}\right)^{-\alpha_q} \quad \text{and} \quad K_{\boldsymbol{\alpha}}(\boldsymbol{s}) = \boldsymbol{\lambda}^{\boldsymbol{\alpha}} \prod_{q=1}^d \left(\frac{1+s_q}{\lambda_q + s_q}\right)^{\alpha_q},$$

so the assumptions of Proposition 2.3.7 are fulfilled.

Gathering (2.8) or (2.9) with Proposition 2.3.7, we obtain convergence rates for our estimator.

Theorem 2.3.11. *Let $\boldsymbol{s}, \boldsymbol{r} \in (0,+\infty)^d$ and $L > 0$. We assume that g satisfies the assumptions of Proposition 2.3.7 with $\boldsymbol{\alpha} \in \mathbb{N}_+^d$.*

1. *For $\boldsymbol{m}_{\mathrm{opt}} \in \mathbb{N}_+^d$ given by $m_{\mathrm{opt},j} \propto n^{1/\left(s_j + s_j \sum_{i=1}^d \frac{2\alpha_i}{s_i}\right)}$, there exists a positive constant $C(s,L,g)$ depending on s, L and g such that:*

$$\sup_{f\in \mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L)} \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}_{\mathrm{opt}}}\|_{\mathrm{L}^2}^2 \leq C(s,L,g)\, n^{-1/\left(1+\sum_{i=1}^d \frac{2\alpha_i}{s_i}\right)}.$$

2. *For* $\boldsymbol{m}_{\text{opt}} \in \mathbb{N}_+^d$ *given by* $m_{\text{opt},j} \propto \frac{\log n}{r_j}$, *there exists a constant* $C(r,L,g) > 0$ *depending on* r, L *and* g *such that:*

$$\sup_{f \in \mathscr{S}^r(\mathbb{R}_+^d, L)} \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}_{\text{opt}}}\|_{\mathrm{L}^2}^2 \leq C(r,L,g)\, \frac{(\log n)^{\sum_{i=1}^d 2\alpha_i}}{n}.$$

Remark 2.3.12. Our convergence rates on Sobolev–Laguerre balls are similar to convergence rates found by Comte & Lacour (2013) on anisotropic Sobolev balls, in the context of deconvolution using a kernel estimator with an ordinary smooth noise.

Remark 2.3.13. We could have considered mixed regularities for f. For instance, if the coefficients of f satisfy:

$$\sum_{\boldsymbol{k} \in \mathbb{N}^d} a_{\boldsymbol{k}}^2\, k_1^{s_1} \cdots k_{j_0}^{s_{j_0}} \exp(r_{j_0+1} k_{j_0+1} + \cdots + r_d k_d) \leq L,$$

for some $1 \leq j_0 < d$, then we can give the following upper-bound on the bias term:

$$\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq L\left(\sum_{j=1}^{j_0} m_j^{-s_j} + \sum_{j=j_0+1}^{d} \mathrm{e}^{-r_j m_j}\right).$$

By choosing $m_j \propto \frac{\log n}{r_j}$ for $j > j_0$, the MISE of $\widehat{f}_{\boldsymbol{m}}$ is then:

$$\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|_{L^2}^2 \lesssim L \sum_{j=1}^{j_0} m_j^{-s_j} + \frac{\|h\|_\infty}{N_n} \prod_{j=1}^{j_0} m_j^{2\alpha_j} + \frac{L(d-j_0)}{n},$$

with $N_n := \frac{n}{\prod_{j>j_0} (\log(n)/r_j)^{2\alpha_j}}$. We can now choose the first j_0 components of $\boldsymbol{m}$ like in the case 1 of Theorem 2.3.11, but with a sample size N_n, and we get:

$$\mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}_{\text{opt}}}\|_{L^2}^2 \leq C\left(\frac{n}{\prod_{j>j_0} (\log(n)/r_j)^{2\alpha_j}}\right)^{-1/\left(1+\sum_{j=1}^{j_0} \frac{2\alpha_j}{s_j}\right)}.$$

It is the rate we would find on a Sobolev–Laguerre ball in dimension j_0, up to a logarithmic factor.

2.4 Adaptive estimation and oracle inequalities

In practice, we do not know the underlying regularity of f, so we can not compute the model $\boldsymbol{m}_{\text{opt}}$ of Theorem 2.3.11. We want a data-driven procedure that automatically optimizes the bias-variance compromise, without making regularity assumptions on f.

We need a restriction on the complexity of models we consider. Let $\mathscr{M}_n$ be the following model collection:

$$\mathscr{M}_n := \left\{ \boldsymbol{m} \in \mathbb{N}_+^d \,\middle|\, D_{\boldsymbol{m}} \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leq \frac{n}{\log n} \right\}. \tag{2.10}$$

We make an additional assumption on the growth of $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2$.

Assumption 2.3. For every $b > 0$, we have:

$$\sum_{\boldsymbol{m} \in \mathscr{M}_n} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \, \mathrm{e}^{-b\sqrt{D_{\boldsymbol{m}}}} \leq K(b),$$

with $K(b)$ a positive constant not depending on n.

Under Assumptions 2.1 and 2.2, we can apply Proposition 2.3.4 to control the variance term $\mathbb{E}\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ by:

$$V(\boldsymbol{m}) := \frac{c(g)\sqrt{D_{\boldsymbol{m}}}\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \wedge \frac{(\|g\|_\infty \vee 1)\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2 \log n}{n}. \tag{2.11}$$

Moreover, under Assumption 2.3, we can control the right deviation from its mean of $\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ when $\boldsymbol{m}$ belongs to $\mathscr{M}_n$.

Lemma 2.4.1. *Under Assumptions 2.1, 2.2 and 2.3, there exists a numerical constant $a_0(d) > 0$ depending on the dimension d such that for every $a > a_0(d)$:*

$$\sum_{\boldsymbol{m} \in \mathscr{M}_n} \mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - aV(\boldsymbol{m}) \right)_+ \right] \leq \frac{C(g,a)}{n},$$

with $C(g,a)$ a positive constant depending on g and a.

The one-dimensional case When $d = 1$, we consider two procedures, indexed by $i \in \{1,2\}$, the first procedure being the one studied by Mabon (2017). In both procedures, the projection space is chosen by minimizing the penalized criterion:

$$\widehat{m}_i := \operatorname*{argmin}_{m \in \mathscr{M}_n} \left[-\|\widehat{f}_m\|_{\mathrm{L}^2}^2 + \kappa_i \, \mathrm{pen}_i(m) \right], \quad i \in \{1,2\}$$

where $\mathscr{M}_n$ is defined by (2.10), κ_i is a numerical constant to be adjusted and pen_i is the penalty term. Mabon (2017) proposes the following penalty term:

$$\mathrm{pen}_1(m) := \frac{2m\|\mathbf{G}_m^{-1}\|_{\mathrm{op}}^2}{n} \wedge \frac{(\|g\|_\infty \vee 1)\|\mathbf{G}_m^{-1}\|_F^2 \log n}{n},$$

and provides an oracle inequality for the estimator $\widehat{f}_{\widehat{m}_1}$ under Assumptions 2.1 and 2.3, see Theorem 4.1 in Mabon's article. This choice is based on the bound

of the variance term of Proposition 2.3.1. However if g is bounded, then Assumption 2.2 holds automatically in dimension $d = 1$, so we can apply Proposition 2.3.4 to get a better bound on the variance term. Thus, we propose the penalty term:

$$\mathrm{pen}_2(m) := V(m) = \frac{c(g)\sqrt{m}\,\|\mathbf{G}_m^{-1}\|_{\mathrm{op}}^2}{n} \wedge \frac{(\|g\|_\infty \vee 1)\|\mathbf{G}_m^{-1}\|_F^2 \log n}{n}.$$

We show that our estimator $\widehat{f}_{\widehat{m}_2}$ also satisfies an oracle inequality.

Theorem 2.4.2. *We assume $d = 1$. Under Assumptions 2.1, 2.2 and 2.3, there exists a numerical constant $\kappa_0 > 0$ such that for every choice of $\kappa > \kappa_0$, we have the following oracle inequality:*

$$\mathbb{E}\big[\|f - \widehat{f}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2\big] \le 4 \inf_{m \in \mathscr{M}_n} \big(\|f - f_m\|_{\mathrm{L}^2}^2 + \kappa\,\mathrm{pen}_2(m)\big) + \frac{C}{n},$$

with C a positive constant depending on κ and g.

The multivariate case We use the procedure similar to Goldenshluger & Lepski (2011) for model selection introduced by Chagny (2013c) for the estimation of a conditional density function. We apply this procedure to our multivariate deconvolution problem, and we establish an oracle inequality.

We choose $\widehat{\boldsymbol{m}}$ in the model collection $\mathscr{M}_n$ defined in (2.10), minimizing:

$$\widehat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m} \in \mathscr{M}_n} \big[A(\boldsymbol{m}) + \kappa_2 V(\boldsymbol{m})\big], \tag{2.12}$$

where $V(\boldsymbol{m})$ is defined by (2.11) and $A(\boldsymbol{m})$ is a term which has the order of the bias term (see the proof):

$$A(\boldsymbol{m}) := \max_{\boldsymbol{m}' \in \mathscr{M}_n} \big(\|\widehat{f}_{\boldsymbol{m}'} - \widehat{f}_{\boldsymbol{m} \wedge \boldsymbol{m}'}\|_{\mathrm{L}^2}^2 - \kappa_1 V(\boldsymbol{m}')\big)_+,$$

and where κ_1, κ_2 are two numerical constants to be adjusted.

Theorem 2.4.3 (Oracle inequality)**.** *Under Assumptions 2.1, 2.2 and 2.3, there exists a numerical constant $\kappa_0(d) > 0$ depending on the dimension d such that for every choice of κ_1, κ_2 satisfying $\kappa_0(d) < \kappa_1 \le \kappa_2$, we have the following oracle inequality:*

$$\mathbb{E}\big[\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|_{\mathrm{L}^2}^2\big] \le C \inf_{\boldsymbol{m} \in \mathscr{M}_n} \big(\|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + V(\boldsymbol{m})\big) + \frac{C'}{n},$$

with C a positive constant depending on κ_1 and κ_2, and C' a positive constants depending on g, d and κ_1.

2.5 Numerical illustrations

2.5.1 Estimators comparison in one-dimensional case

We want to compare the two estimators $\widehat{f}_{\widehat{m}_1}$ and $\widehat{f}_{\widehat{m}_2}$ defined in Section 2.4, with different distributions for X and Y, and with different sample size n. We compare their MISE:

$$\mathscr{R}_i = \mathscr{R}_i(f,g,n) \coloneqq \mathbb{E}\|f - \widehat{f}_{\widehat{m}_i}\|_{\mathrm{L}^2}^2, \quad i \in \{1,2\}.$$

We also compute the oracle risk:

$$\mathscr{R}_o = \mathscr{R}_o(f,g,n) \coloneqq \min_{1\le m\le m^*} \mathbb{E}\|f - \widehat{f}_m\|_{\mathrm{L}^2}^2,$$

where m^* is the maximal element of $\mathscr{M}_n$. We compute an approximation of $\|\cdot\|_{\mathrm{L}^2}^2$ with Simpson's rule on a bounded interval I of $\mathbb{R}_+$ with 1000 points. We compute the expectations with an empirical mean over 500 samples (we use the same samples for $\mathscr{R}_1$, $\mathscr{R}_2$ and $\mathscr{R}_o$).

Remark 2.5.1. Since f is non-negative, $\widehat{f}_m$ is replaced by $(\widehat{f}_m)_+ \coloneqq \max(\widehat{f}_m, 0)$ in the following.

Distributions for X We consider several distributions for X. In each case, we normalize the distribution for the variance to be 1. We use the same examples as Mabon (2017).

- Exponential $\mathscr{E}(1)$, $I = [0,5]$.
- Gamma $\Gamma(20, \sqrt{20})$, $I = [0,10]$.
- Rayleigh, $f(x) = \frac{x}{\sigma^2}\exp(-\frac{x^2}{2\sigma^2})$ with $\sigma^2 = \frac{2}{4-\pi}$, $I = [0,6]$.
- Weibull, $f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-(x/\lambda)^k}$, with $k = \frac{3}{2}$ and $\lambda = 1/\sqrt{\Gamma(1+\frac{4}{3}) - \Gamma(1+\frac{2}{3})^2}$, $I = [0,5]$.
- Beta B(4,5) with normalisation $\frac{9}{\sqrt{2}}$, $I = [0,8]$.
- Gamma mixture $0.4\,\Gamma(2,2) + 0.6\,\Gamma(16,4)$ with normalisation $\frac{1}{\sqrt{2.96}}$, $I = [0,5]$.

Distributions for Y We choose gamma distributions for Y. We recall the density of the gamma distribution with parameters $\alpha \ge 1$ and $\lambda > 0$:

$$g(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\lambda x}, \quad x \ge 0.$$

with Γ the gamma function. These distributions satisfy Assumptions 2.1, 2.2 and 2.3, so we can use the procedure described in Section 2.4. We can compute the Laguerre coefficients of g exactly:

$$\forall k \in \mathbb{N}, \quad b_k = \sqrt{2}\left(\frac{\lambda}{\lambda+1}\right)^\alpha \sum_{j=0}^{k}\binom{k}{j}\binom{\alpha-1+j}{j}\left(\frac{-2}{1+\lambda}\right)^j,$$

where $\binom{\alpha-1+j}{j} := \frac{(\alpha-1+j)\times(\alpha-2+j)\times\cdots\times\alpha}{j!}$. Denoting $\mathbf{b}_m$ the vectors $[b_k]_{k\leq m-1}$, we can compute:

$$\mathbf{b}_m = \sqrt{2}\left(\frac{\lambda}{\lambda+1}\right)^{\alpha}\mathbf{P}_m\mathbf{v}_m,$$

where $\mathbf{P}_m$ is the $m\times m$ matrix with components $P_{ij} := \binom{i}{j}$ for $0\leq i,j\leq m-1$, and $\mathbf{v}_m$ is the vector of size m:

$$\mathbf{v}_m := \left[\binom{\alpha-1+j}{j}\left(\frac{-2}{1+\lambda}\right)^{j}\right]_{0\leq j\leq m-1}.$$

Therefore, we can compute easily and efficiently the matrix $\mathbf{G}_m$ and its inverse. We choose two distributions for Y : the $\Gamma(2,\sqrt{20})$ distribution which has variance $\frac{1}{10}$ and the $\Gamma(2,\sqrt{8})$ distribution which has variance $\frac{1}{4}$.

The constant $c(g)$ This constant in the penalty pen_2 is not known, so we have to evaluate it numerically. Following the proof of the Proposition 2.3.4, we see this constant appears in the upper bound:

$$\frac{\|\mathbf{G}_m^{-1}\|_{\text{op}}^2}{n}\sum_{k=0}^{m-1}\mathbb{E}\big[\varphi_k(Z_1)^2\big] \leq \frac{c(g)\sqrt{m}\,\|\mathbf{G}_m^{-1}\|_{\text{op}}^2}{n},$$

which is true for every density of X and for every m. Therefore,

$$c(g) = \sup_{X_1\sim f}\sup_{m\in\mathbb{N}^*}\frac{1}{\sqrt{m}}\sum_{k=0}^{m-1}\mathbb{E}\big[\varphi_k(X_1+Y_1)^2\big].$$

For every distribution of X we considered earlier and for every m from 1 to 50, we computed:

$$\frac{1}{\sqrt{m}}\sum_{k=0}^{m-1}\mathbb{E}\big[\varphi_k(X+Y)^2\big],$$

using an empirical mean over 1000 realizations of (X,Y) to compute the expectation, then we took the maximum. For the two distributions we considered for Y, the worst constant was about 0.4 and was reached for $X\sim\mathscr{E}(1)$. For the other distributions of X, we found constants between 0.2 and 0.3. To be safe, we chose a bigger constant $c(g) = 0.5$ for both $\Gamma(2,\sqrt{20})$ and $\Gamma(2,\sqrt{8})$ distributions.

The constants κ We use the same constant $\kappa_1 = 0.03$ as Mabon (2017). Concerning the constant κ_2, we made several simulations to calibrate it and we chose $\kappa_2 = 0.04$. During the calibration of κ_2, we realized that the collection $\mathscr{M}_n$ did not have enough models. The selected dimensions were too small and lead to bad performances. Therefore, we considered the following collection:

$$\mathscr{M}_n' := \left\{1\leq m\leq m^* \;\middle|\; m\left(\|\mathbf{G}_m^{-1}\|_{\text{op}}^2\vee 1\right)\leq 10\frac{n}{\log n}\right\},$$

where m^* is a fixed maximal model.

Procedure We take the maximal dimension to be $m^* = 20$. We use two sample sizes: $n = 200$ and $n = 2000$. Given a distribution for X, a distribution for Y, a sample size n, and a subdivision $\Sigma(I)$ of I, do:

1. Compute the matrix $\mathbf{G}_{20}$ and its inverse.
2. Compute the collection $\mathscr{M}_n'$.
3. Compute $f(x)$ for $x \in \Sigma(I)$.
4. Repeat 500 times:
   a) Generate a sample $Z_i = X_i + Y_i$, $i = 1, \dots, n$.
   b) Compute the Laguerre coefficients $\widehat{\mathbf{c}}_{20}$, and compute $\widehat{\mathbf{a}}_{20} = \mathbf{G}_{20}^{-1}\widehat{\mathbf{c}}_{20}$.
   c) Compute $\widehat{m}_1$ and $\widehat{m}_2$ minimizing:
   $$\widehat{m}_i = \underset{m \in \mathscr{M}_n'}{\arg\min} \left(-\sum_{k=0}^{m-1} \widehat{a}_k^2 + \kappa_i \operatorname{pen}_i(m) \right), \quad i \in \{1,2\},$$
   and compute $\left(\widehat{f}_{\widehat{m}_1}(x)\right)_+$ and $\left(\widehat{f}_{\widehat{m}_2}(x)\right)_+$ for $x \in \Sigma(I)$.
   d) Compute $J_1 = \|f - (\widehat{f}_{\widehat{m}_1})_+\|_{\mathrm{L}^2}^2$ and $J_2 = \|f - (\widehat{f}_{\widehat{m}_2})_+\|_{\mathrm{L}^2}^2$ using Simpson's rule.
   e) For m from 1 to 20, compute $\left(\widehat{f}_m(x)\right)_+$ for $x \in \Sigma(I)$; then compute $J(m) = \|f - (\widehat{f}_m)_+\|_{\mathrm{L}^2}^2$ using Simpson's rule.
5. Compute $\mathscr{R}_1$ (resp. $\mathscr{R}_2$) as the mean of J_1 (resp. J_2) over the 500 samples.
6. For each m, compute the mean of $J(m)$ over the 500 samples, then compute $\mathscr{R}_o$ as the minimum of these quantities.

Results We show our results in Table 2.1. We note that in the cases $X \sim \mathscr{E}(1)$ and $X \sim \Gamma(20, \sqrt{20})$, both estimators performed badly compared to the oracle. We see that our estimator $\widehat{f}_{\widehat{m}_2}$ is better when X has gamma, Rayleigh, beta and mixture gamma distribution. The estimator $\widehat{f}_{\widehat{m}_1}$ is better when X has Weibull and exponential distribution.

For illustration, Figure 2.1 shows several estimations when the distribution of X is a mixture gamma. It's a bimodal distribution, so it's interesting to see if the estimators are able to recover the two "peaks" and the "hollow" of the true density. For small samples ($n = 200$), the estimator $\widehat{f}_{\widehat{m}_1}$ seems to fail more often than the estimator $\widehat{f}_{\widehat{m}_2}$. For large samples ($n = 2000$), both estimators locate well the two pikes (by overestimating the first one and underestimating the second one), but the estimator $\widehat{f}_{\widehat{m}_1}$ locates less well the hollow and underestimates more the second pike.

Distribution of X	MISE	$\sigma_Y^2 = 1/10$		$\sigma_Y^2 = 1/4$	
		$n = 200$	$n = 2000$	$n = 200$	$n = 2000$
$\mathscr{E}(1)$	$\mathscr{R}_1$	139	8.57	113	9.95
	$\mathscr{R}_2$	204	30.4	203	37.2
	$\mathscr{R}_o$	10.6	0.93	14.7	1.50
$\Gamma(20, \sqrt{20})$	$\mathscr{R}_1$	127	32.1	407	109
	$\mathscr{R}_2$	80.4	16.4	350	58.0
	$\mathscr{R}_o$	26.8	3.84	34.7	5.24
Rayleigh	$\mathscr{R}_1$	68.1	7.13	88.7	8.23
	$\mathscr{R}_2$	42.1	6.44	50.3	9.13
	$\mathscr{R}_o$	29.8	4.76	35.0	7.33
Weibull	$\mathscr{R}_1$	55.2	4.89	79.0	6.80
	$\mathscr{R}_2$	64.3	8.56	70.3	10.5
	$\mathscr{R}_o$	30.8	4.84	42.0	6.59
B(4,5)	$\mathscr{R}_1$	122	16.8	184	29.8
	$\mathscr{R}_2$	39.6	6.50	53.0	16.5
	$\mathscr{R}_o$	30.9	4.52	35.5	5.59
Γ mixture	$\mathscr{R}_1$	286	63.8	717	82.9
	$\mathscr{R}_2$	162	47.1	595	71.2
	$\mathscr{R}_o$	151	34.0	228	49.4

Table 2.1: MISE computation over 500 samples. MISE are multiplied by 10^4. The column "$\sigma_Y^2 = 1/10$" corresponds to the case $Y_i \sim \Gamma(2, \sqrt{20})$ and the column "$\sigma_Y^2 = 1/4$" to the case $Y_i \sim \Gamma(2, \sqrt{8})$.

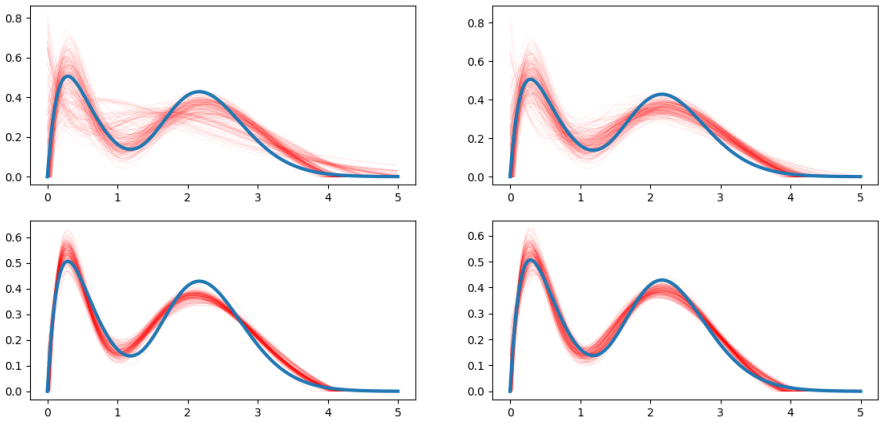

Figure 2.1: Density estimation when X is a mixture of gamma distributions and when $Y \sim \Gamma(2, \sqrt{20})$. The thick blue curve represents the true density and each thin red curve represents the estimation obtained with one sample (total: 200 samples). First line: $n = 200$, second line: $n = 2000$. Left: $\widehat{f}_{\widehat{m}_1}$, right: $\widehat{f}_{\widehat{m}_2}$.

2.5.2 Model selection in two-dimensional case

In this subsection, our goal is to illustrate the procedure similar to Goldenshluger & Lepski (2011) on two examples, in the case $d = 2$.

Distributions for $\boldsymbol{X}$ To generate a random vector $\boldsymbol{X} = (X^{(1)}, X^{(2)})$, we do the following: we generate a random variable $W^{(1)}$ with some distribution and we generate $W^{(2)}$ independently with some other distribution, and we compute:

$$\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0.2 & 1 \end{bmatrix} \begin{bmatrix} W^{(1)} \\ W^{(2)} \end{bmatrix},$$

so that the coordinates of $\boldsymbol{X}$ are not independent. We consider the following distributions for $W^{(1)}$ and $W^{(2)}$:

- Gamma $\Gamma(3,1)$
- Beta $\mathrm{B}(4,5)$ with normalisation $\frac{9}{\sqrt{2}}$
- Rayleigh, $f(x) = \frac{x}{\sigma^2}\exp(-\frac{x^2}{2\sigma^2})$ with $\sigma^2 = \frac{2}{4-\pi}$
- Weibull, $f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-(x/\lambda)^k}$, with $k = \frac{3}{2}$ and $\lambda = 1/\sqrt{\Gamma(1+\frac{4}{3}) - \Gamma(1+\frac{2}{3})^2}$

Distribution for $\boldsymbol{Y}$ We choose distributions of the form $\Gamma(\alpha_1,\lambda_1)\otimes\Gamma(\alpha_2,\lambda_2)$. These distributions satisfy the Assumptions 2.1, 2.2 and 2.3 of the Section 2.4. Moreover, the Laguerre coefficients can be computed easily: if γ_i is the density of the distribution $\Gamma(\alpha_i,\lambda_i)$, then we have $g = \gamma_1\otimes\gamma_2$, so:

$$\forall \boldsymbol{k}\in\mathbb{N}^2, \quad b_{\boldsymbol{k}} = \langle\gamma_1,\varphi_{k_1}\rangle_{\mathrm{L}^2}\times\langle\gamma_2,\varphi_{k_2}\rangle_{\mathrm{L}^2}.$$

Refer to Subsection 2.5.1 for the computation of $\langle\gamma_i,\varphi_{k_i}\rangle_{\mathrm{L}^2}$. We chose the distribution $\Gamma(2,\sqrt{20})\otimes\Gamma(2,\sqrt{20})$ for $\boldsymbol{Y}$.

The κ constants After several simulations, we chose $\kappa_1 = \kappa_2 = 10^{-5}$. However, this calibration is rough. Like in the one-dimensional case, the model collection $\mathscr{M}_n$ defined by (2.10) was too small, so we considered the following collection:

$$\mathscr{M}_n' := \left\{ \boldsymbol{m}\le\boldsymbol{m}^* \;\middle|\; D_{\boldsymbol{m}}\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\vee 1\right)\le 10^4\,\frac{n}{\log n}\right\},$$

where $\boldsymbol{m}^*$ is a fixed maximal model.

W distribution		Gamma	Beta	Rayleigh	Weibull
Gamma	$n = 500$	7.78 (4.6, 4.8)	12.99 (3.6, 7.0)	11.73 (4.7, 5.9)	20.30 (5.1, 6.4)
	$n = 5000$	1.36 (4.5, 5.6)	3.40 (3.8, 8.9)	1.96 (4.9, 6.5)	6.30 (4.6, 8.5)
Beta	$n = 500$	11.75 (6.5, 3.9)	22.04 (6.2, 6.3)	18.73 (6.9, 5.0)	28.15 (6.7, 5.4)
	$n = 5000$	2.78 (8.3, 4.6)	6.66 (7.1, 7.8)	3.89 (8.9, 5.8)	12.19 (7.6, 7.0)
Rayleigh	$n = 500$	11.82 (5.8, 4.7)	20.17 (4.6, 6.9)	18.08 (5.7, 5.8)	32.78 (6.2, 6.8)
	$n = 5000$	1.86 (6.3, 5.2)	4.21 (5.5, 8.5)	3.07 (6.6, 6.3)	8.39 (6.5, 7.7)
Weibull	$n = 500$	17.91 (6.6, 5.2)	24.01 (5.4, 6.8)	40.37 (7.1, 6.3)	64.66 (8.0, 7.4)
	$n = 5000$	4.29 (7.9, 4.7)	8.54 (6.8, 7.7)	5.89 (7.7, 6.5)	11.06 (7.2, 8.0)

Table 2.2: MISE computation over 100 samples of size n, for $n = 500$ and $n = 5000$. In rows: distribution of $W^{(1)}$. In columns: distribution of $W^{(2)}$. In each cell, we show the MISE (multiplied by 10^4) and the mean selected model.

Results We took the maximal model to be $\boldsymbol{m}^* = (12, 12)$. In each case, we simulated a sample of size $n = 500$ and $n = 5000$, then we computed the model $\widehat{\boldsymbol{m}}$ defined by (2.12). We computed $\mathbb{E}\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|_{\mathrm{L}^2}^2$ by approximating the double integral by a Riemann sum:

$$\iint_{\mathbb{R}_+^2} \left(f(x,y) - \widehat{f}_{\widehat{m}}(x,y)\right)^2 \mathrm{d}x\,\mathrm{d}y \approx \sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} \left(f(x_i,y_j) - \widehat{f}_{\widehat{m}}(x_i,y_j)\right)^2 \Delta x_i\,\Delta y_j$$

on a grid with step size 0.01, and by computing the expectation with an empirical mean over 100 samples.

We show our results in Table 2.2. In each case, we compute the MISE and the mean selected model. We emphasize that in the majority of the samples, the selected model is anisotropic: the components of $\widehat{\boldsymbol{m}}$ are not equal. It is a property we expected from a procedure similar to Goldenshluger & Lepski (2011). The estimator adapts to the regularity of f.

For illustration, Figure 2.2 shows the result of the estimation when $W^{(1)}$ has a Gamma distribution and $W^{(2)}$ has a Weibull distribution, for a sample of size 5000. We show both the adaptive estimator $\widehat{f}_{\widehat{\boldsymbol{m}}}$ and the estimator with no selection procedure $\widehat{f}_{(12,12)}$, where we simply choose the maximum model. We see that the procedure selected the model $\widehat{\boldsymbol{m}} = (5, 8)$, which has a smaller dimension than the maximum model.

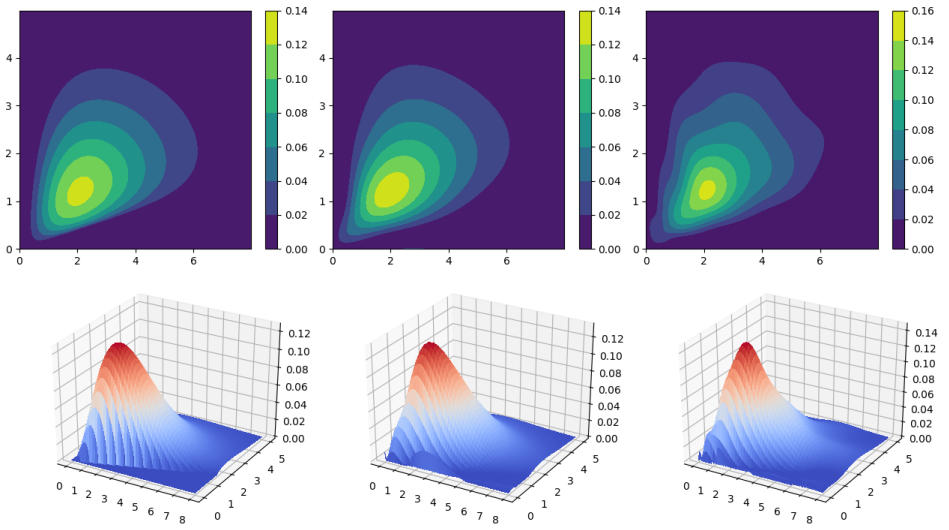

Figure 2.2: Density estimation when $W^{(1)}$ has a Gamma distribution and $W^{(2)}$ has a Weibull distribution, for a sample of size $n = 5000$. First column: true density, second column: adaptive estimator $\widehat{f}_{\widehat{\boldsymbol{m}}}$, third column: max model estimator $\widehat{f}_{(12,12)}$. The selected model is $\widehat{\boldsymbol{m}} = (5,8)$.

2.6 Proofs

2.6.1 Proofs of Sections 2.2 and 2.3

Proof of Proposition 2.2.1. Using the relation (2.3), we have:

$$f * g = \sum_{\boldsymbol{k}\in\mathbb{N}^d} \sum_{\boldsymbol{j}\in\mathbb{N}^d} a_{\boldsymbol{k}} b_{\boldsymbol{j}} (\varphi_{\boldsymbol{k}} * \varphi_{\boldsymbol{j}}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} \sum_{\boldsymbol{j}\in\mathbb{N}^d} a_{\boldsymbol{k}} b_{\boldsymbol{j}} \bigotimes_{q=1}^{d} 2^{-1/2} (\varphi_{k_q+j_q} - \varphi_{k_q+j_q+1}).$$

By setting $\ell_q = k_q + j_q$, we get:

$$f * g = 2^{-d/2} \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} \sum_{\substack{\boldsymbol{k}\in\mathbb{N}^d \\ \boldsymbol{k}\le\boldsymbol{\ell}}} a_{\boldsymbol{k}} b_{\boldsymbol{\ell}-\boldsymbol{k}} \bigotimes_{q=1}^{d} (\varphi_{\ell_q} - \varphi_{\ell_q+1}) = 2^{-d/2} \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} (a*b)_{\boldsymbol{\ell}} \bigotimes_{q=1}^{d} (\varphi_{\ell_q} - \varphi_{\ell_q+1}).$$

Using the tensor product multi-linearity:

$$\begin{aligned} \bigotimes_{q=1}^{d} (\varphi_{\ell_q} - \varphi_{\ell_q+1}) = \bigotimes_{q=1}^{d} \sum_{\varepsilon_q=0}^{1} (-1)^{\varepsilon_q} \varphi_{\ell_q+\varepsilon_q} &= \sum_{\varepsilon_1=0}^{1} \cdots \sum_{\varepsilon_d=0}^{1} \bigotimes_{q=1}^{d} (-1)^{\varepsilon_q} \varphi_{\ell_q+\varepsilon_q} \\ &= \sum_{\boldsymbol{\varepsilon}\in\{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|} \varphi_{\boldsymbol{\ell}+\boldsymbol{\varepsilon}}. \end{aligned}$$

Thus, we get:

$$f * g = \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} (a*b)_{\boldsymbol{\ell}} \left[2^{-d/2} \sum_{\boldsymbol{\varepsilon}\in\{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|} \varphi_{\boldsymbol{\ell}+\boldsymbol{\varepsilon}} \right]$$
$$= \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} \left[2^{-d/2} \sum_{\boldsymbol{\varepsilon}\in\{0,1\}^d} (-1)^{|\boldsymbol{\varepsilon}|} (a*b)_{\boldsymbol{\ell}-\boldsymbol{\varepsilon}} \right] \varphi_{\boldsymbol{\ell}}.$$

Since $h = f * g$, by uniqueness of the Laguerre coefficients of h, we obtain the desired relation. □

Proof of Proposition 2.2.3. First, we notice that $\forall \boldsymbol{\ell},\ \mathbf{G}_{\boldsymbol{\ell}\boldsymbol{\ell}} = \mathbb{E}[\mathrm{e}^{-(Y_1^{(1)}+\cdots+Y_1^{(d)})}] > 0$. We proceed by induction on $|\boldsymbol{k}|$.

- If $|\boldsymbol{k}| = 0$, then $\boldsymbol{k} = \mathbf{0}$ and $a_{\mathbf{0}} = (\mathbf{G}_{\mathbf{0},\mathbf{0}})^{-1} c_{\mathbf{0}}$.
- Let $r \in \mathbb{N}$, we suppose (2.6) is true for every $\boldsymbol{k} \in \mathbb{N}^d$ such that $|\boldsymbol{k}| \leq r$. Let $\boldsymbol{k} \in \mathbb{N}^d$ with $|\boldsymbol{k}| = r+1$. From (2.5),

$$c_{\boldsymbol{k}} = \sum_{\substack{\boldsymbol{\ell}\leq\boldsymbol{k} \\ \boldsymbol{\ell}\neq\boldsymbol{k}}} \mathbf{G}_{\boldsymbol{k}\boldsymbol{\ell}}\, a_{\boldsymbol{\ell}} + \mathbf{G}_{\boldsymbol{k}\boldsymbol{k}}\, a_{\boldsymbol{k}}.$$

  If $\boldsymbol{\ell} \leq \boldsymbol{k}$ with $\boldsymbol{\ell} \neq \boldsymbol{k}$ then $|\boldsymbol{\ell}| < |\boldsymbol{k}|$, so we can use the induction assumption:

$$a_{\boldsymbol{k}} = (\mathbf{G}_{\boldsymbol{k}\boldsymbol{k}})^{-1} \left(c_{\boldsymbol{k}} - \sum_{\substack{\boldsymbol{\ell}\leq\boldsymbol{k} \\ \boldsymbol{\ell}\neq\boldsymbol{k}}} \mathbf{G}_{\boldsymbol{k}\boldsymbol{\ell}} \sum_{\boldsymbol{j}\leq\boldsymbol{\ell}} \mathbf{H}_{\boldsymbol{\ell}\boldsymbol{j}}\, c_{\boldsymbol{j}} \right).$$

  Thus, by setting $\mathbf{H}_{\boldsymbol{k}\boldsymbol{k}} := (\mathbf{G}_{\boldsymbol{k}\boldsymbol{k}})^{-1}$ and $\mathbf{H}_{\boldsymbol{k}\boldsymbol{j}} := (\mathbf{G}_{\boldsymbol{k}\boldsymbol{k}})^{-1} \sum_{\boldsymbol{j}\leq\boldsymbol{\ell}\leq\boldsymbol{k},\boldsymbol{\ell}\neq\boldsymbol{k}} \mathbf{G}_{\boldsymbol{k}\boldsymbol{\ell}} \mathbf{H}_{\boldsymbol{\ell}\boldsymbol{j}}$ for every $\boldsymbol{j} \leq \boldsymbol{k},\ \boldsymbol{j} \neq \boldsymbol{k}$, we've just proved (2.6) for all $\boldsymbol{k}$ such that $|\boldsymbol{k}| = r+1$. □

Proof of Proposition 2.3.1. By Pythagoras theorem, $\|f - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \|f - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 + \|f_{\boldsymbol{m}} - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$. We decompose the second term on the Laguerre basis:

$$\|f_{\boldsymbol{m}} - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}} (a_{\boldsymbol{k}} - \widehat{a}_{\boldsymbol{k}})^2 = \|\mathbf{a}_{\boldsymbol{m}} - \widehat{\mathbf{a}}_{\boldsymbol{m}}\|_{\mathbb{R}^m}^2.$$

We now give an upper bound on the last quantity in two different ways.

1. The first way is a bound using the spectral norm.

$$\begin{aligned}
\mathbb{E}\|f_{\boldsymbol{m}} - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 &= \mathbb{E}\|\mathbf{G}_{\boldsymbol{m}}^{-1}(\mathbf{c}_{\boldsymbol{m}} - \widehat{\mathbf{c}}_{\boldsymbol{m}})\|_{\mathbb{R}^m}^2 \\
&\leq \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\, \mathbb{E}\|\mathbf{c}_{\boldsymbol{m}} - \widehat{\mathbf{c}}_{\boldsymbol{m}}\|_{\mathbb{R}^m}^2 \\
&= \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\, \mathbb{E}\left[\sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}} \left(\frac{1}{n} \sum_{i=1}^n \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_i) - \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)\big] \right)^2 \right] \\
&= \frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}} \mathrm{Var}\big(\varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)\big) \leq \frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \sum_{\boldsymbol{k}\leq\boldsymbol{m}-\mathbf{1}} \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)^2\big].
\end{aligned}$$

Yet for every $\boldsymbol{x} \in \mathbb{R}_+^d$, $\varphi_{\boldsymbol{k}}(\boldsymbol{x})^2 \leq 2^d$ and $\mathrm{Card}\{\boldsymbol{k} \in \mathbb{N}^d \,|\, \boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}\} = \prod_{q=1}^d m_q$, so we have:

$$\mathbb{E}\|f_{\boldsymbol{m}} - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq \frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)^2\big] \leq \frac{2^d \left(\prod_{q=1}^d m_q\right) \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n}. \quad (2.13)$$

2. For the second way, we develop the square.

$$\begin{aligned}
\mathbb{E}\|\mathbf{G}_{\boldsymbol{m}}^{-1}(\mathbf{c}_{\boldsymbol{m}} - \widehat{\mathbf{c}}_{\boldsymbol{m}})\|_{\mathbb{R}^m}^2 &= \mathbb{E}\left[\sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} (c_{\boldsymbol{k}} - \widehat{c}_{\boldsymbol{k}})\right)^2\right] \\
&= \mathbb{E}\left[\sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \sum_{\boldsymbol{k}' \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} (c_{\boldsymbol{k}} - \widehat{c}_{\boldsymbol{k}}) [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}'} (c_{\boldsymbol{k}'} - \widehat{c}_{\boldsymbol{k}'})\right] \\
&= \sum_{\boldsymbol{\ell} \leq m-\mathbf{1}} \sum_{\boldsymbol{k} \leq m-\mathbf{1}} \sum_{\boldsymbol{k}' \leq m-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}'} \operatorname{Cov}(\widehat{c}_{\boldsymbol{k}}, \widehat{c}_{\boldsymbol{k}'}) \\
&= \sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \sum_{\boldsymbol{k}' \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}'} \frac{\operatorname{Cov}\big(\varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1), \varphi_{\boldsymbol{k}'}(\boldsymbol{Z}_1)\big)}{n} \\
&= \frac{1}{n} \sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \operatorname{Cov}\left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1), \sum_{\boldsymbol{k}' \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}'} \varphi_{\boldsymbol{k}'}(\boldsymbol{Z}_1)\right) \\
&= \frac{1}{n} \sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \operatorname{Var}\left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)\right).
\end{aligned}$$

We control the variance by the expectation of the square:

$$\begin{aligned}
\operatorname{Var}\left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)\right) &\leq \mathbb{E}\left[\left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{Z}_1)\right)^2\right] \\
&= \int_{\mathbb{R}_+^d} \left(\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{x})\right)^2 h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&\leq \|h\|_\infty \left\| \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}} \varphi_{\boldsymbol{k}} \right\|_{\mathrm{L}^2}^2 \\
&= \|h\|_\infty \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}}^2.
\end{aligned}$$

In the end, we obtain:

$$\mathbb{E}\|f_{\boldsymbol{m}} - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq \frac{\|h\|_\infty}{n} \sum_{\boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}} \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{\ell}\boldsymbol{k}}^2 = \frac{\|h\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2}{n}. \qquad \square$$

To prove Proposition 2.3.4, we will use the following lemma about the order of magnitude of the functions φ_k for $k \in \mathbb{N}$.

Lemma 2.6.1 (Askey & Wainger 1965)**.** *There exists $k_0 \in \mathbb{N}$ and $C > 0$ such that for every $k \geq k_0$,*

$$\left|\varphi_k\Big(\frac{x}{2}\Big)\right| \leq C \times \begin{cases} 1 & \text{if } 0 \leq x \leq \frac{1}{k}, \\ k^{-\frac{1}{4}}\, x^{-\frac{1}{4}} & \text{if } \frac{1}{k} \leq x \leq \delta k, \\ k^{-\frac{1}{4}}\, (\nu - x)^{-\frac{1}{4}} & \text{if } \delta k \leq x \leq \nu - \nu^{\frac{1}{3}}, \\ k^{-\frac{1}{3}} & \text{if } \nu - \nu^{\frac{1}{3}} \leq x \leq \nu + \nu^{\frac{1}{3}}, \\ k^{-\frac{1}{4}}\, (x - \nu)^{-\frac{1}{4}} \exp\big(-\eta (x-\nu)^{\frac{3}{2}} \nu^{-\frac{1}{2}}\big) & \text{if } \nu + \nu^{\frac{1}{3}} \leq x \leq (1+\lambda)\nu, \\ \exp(-\xi x) & \text{if } (1+\lambda)\nu \leq x, \end{cases}$$

where $\nu := 4k + 2$, where δ et λ are small enough positive constants, and where η and ξ are fixed positive constants.

Proof of Proposition 2.3.4. Following the proof of Proposition 2.3.1, we see we need to improve the upper bound in (2.13). Using Lemma 2.6.1, we have for all $k \in \mathbb{N}$,

$$\big|\varphi_k(x)\big|^2 \leq C \begin{cases} 1 & \text{if } 0 \leq 2x \leq \frac{1}{k}, \\ 1/\sqrt{kx} & \text{if } \frac{1}{k} \leq 2x \leq \delta k, \\ R_k & \text{if } 2x \geq \delta k. \end{cases}$$

where $R_k = \mathrm{o}(1/\sqrt{k})$ does not depend on x. Since "$0 \leq x \leq \frac{1}{k}$" is equivalent to "$1 \leq \frac{1}{\sqrt{kx}}$", we get $|\varphi_k(x)|^2 \leq C(\frac{1}{\sqrt{kx}} + R_k)$. Thus for $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d$, we get:

$$\begin{aligned} \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{Z})^2\big] = \mathbb{E}\left[\prod_{j=1}^{d} \varphi_{k_j}(Z^{(j)})^2\right] &\leq \mathbb{E}\left[\prod_{j=1}^{d} C\left(\frac{1}{\sqrt{k_j Z^{(j)}}} + R_{k_j}\right)\right] \\ &\leq \mathbb{E}\left[\prod_{j=1}^{d} C\left(\frac{1}{\sqrt{k_j Y^{(j)}}} + R_{k_j}\right)\right] \\ &= \frac{C^d}{\sqrt{k_1 \cdots k_d}} \mathbb{E}\left[\frac{1}{\sqrt{Y^{(1)} \ldots Y^{(d)}}}\right] + \tilde{C}\, \mathrm{o}\left(\frac{1}{\sqrt{k_1 \cdots k_d}}\right), \end{aligned}$$

where $\tilde{C}$ depends on M_J for $J \subseteq \{1, \ldots, d\}$. Therefore:

$$\sum_{\boldsymbol{k} \leq \boldsymbol{m} - \boldsymbol{1}} \mathbb{E}\big[\varphi_{\boldsymbol{k}}(\boldsymbol{Z})^2\big] \leq c(g) \sqrt{D_{\boldsymbol{m}}},$$

with $c(g)$ a positive constant depending on M_J for $J \subseteq \{1, \ldots, d\}$. □

Proof of Theorem 2.3.11. Using Proposition 2.3.1, Equation (2.8), and Proposition 2.3.7, if f belongs to $\mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L)$, we have:

$$\forall\, \boldsymbol{m} \in \mathbb{N}_+^d, \quad \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq L \sum_{i=1}^{d} m_i^{-s_i} + \frac{\|h\|_\infty}{n} \prod_{i=1}^{d} m_i^{2\alpha_i}.$$

By Remark 2.3.3, $\|h\|_\infty$ is controlled by $\|f\|_{\mathrm{L}^2}\|g\|_{\mathrm{L}^2}$. Because $f \in \mathrm{W}^{\boldsymbol{s}}(\mathbb{R}_+^d, L)$, we have $\|f\|_{\mathrm{L}^2}^2 \leq L$, so $\|h\|_\infty \leq \sqrt{L}\|g\|_{L^2}$. Thus,

$$\forall \boldsymbol{m} \in \mathbb{N}_+^d, \quad \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq L\sum_{i=1}^d m_i^{-s_i} + \frac{\sqrt{L}\|g\|_{\mathrm{L}^2}}{n}\prod_{i=1}^d m_i^{2\alpha_i} =: \psi_n(\boldsymbol{m}).$$

Let $\boldsymbol{m}_{\mathrm{opt}}$ minimizing ψ_n, then the gradient of ψ_n vanishes on $\boldsymbol{m}_{\mathrm{opt}}$, so we have:

$$\forall j \in \{1,\dots,d\}, \quad \frac{L s_j}{2\alpha_j} m_{\mathrm{opt},j}^{-s_j} = \frac{\sqrt{L}\|g\|_{\mathrm{L}^2}}{n}\prod_{i=1}^d m_{\mathrm{opt},i}^{2\alpha_i}. \tag{2.14}$$

Therefore, the coordinates of $\boldsymbol{m}_{\mathrm{opt}}$ must satisfy $\frac{s_i}{2\alpha_i} m_{\mathrm{opt},i}^{-s_i} = \frac{s_j}{2\alpha_j} m_{\mathrm{opt},j}^{-s_j}$ for every i and j in $\{1,\dots,d\}$. Using these relations in (2.14), we obtain:

$$\forall j \in \{1,\dots,d\}, \quad m_{\mathrm{opt},j} = C_j(s,L,g)\, n^{1/\left(s_j + s_j\sum_{i=1}^d \frac{2\alpha_i}{s_i}\right)},$$

where $C_j(s,L,g)$ are constants depending on s, L and g. The minimum value of ψ_n is then:

$$\psi_n(\boldsymbol{m}_{\mathrm{opt}}) = C(s,L,g)\, n^{-1/\left(1+\sum_{i=1}^d \frac{2\alpha_i}{s_i}\right)},$$

where $C(s,L,g)$ is a constant depending in s, L and g.

Now, if f belongs to $\mathscr{S}^{\boldsymbol{r}}(\mathbb{R}_+^d, L)$, we use the bound (2.9) on the bias term:

$$\forall \boldsymbol{m} \in \mathbb{N}_+^d, \quad \mathbb{E}\|f - \widehat{f}_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \leq L\sum_{i=1}^d \mathrm{e}^{-r_i m_i} + \frac{\sqrt{L}\|g\|_\infty}{n}\prod_{i=1}^d m_i^{2\alpha_i} =: \phi_n(\boldsymbol{m}).$$

We minimize the function $\phi_n(\boldsymbol{m})$ as we did above, and we find the following relations:

$$\forall j \in \{1,\dots,d\}, \quad r_j\, m_{\mathrm{opt},j}\, \mathrm{e}^{-r_j m_{\mathrm{opt},j}} = C_j(r,L,g)\,\frac{(r_j\, m_{\mathrm{opt},j})^{\sum_{i=1}^d 2\alpha_j}}{n},$$

where $C_j(r,L,g)$ are constants depending on r, L and g. Taking the log, we find:

$$r_j\, m_{\mathrm{opt},j} + \left(\sum_{i=1}^d 2\alpha_i - 1\right)\log(r_j\, m_{\mathrm{opt},j}) = \log n - \log C_j(r,L,g).$$

Thus, when n goes to $+\infty$, we have $r_j\, m_{\mathrm{opt},j} \sim \log n$. Taking $\boldsymbol{m}_{\mathrm{opt}} \in \mathbb{N}_+^d$ such that $m_{\mathrm{opt},j} \propto \frac{\log n}{r_j}$, we find:

$$\phi_n(\boldsymbol{m}_{\mathrm{opt}}) \leq C(r,L,g)\,\frac{(\log n)^{\sum_{i=1}^d 2\alpha_i}}{n},$$

where $C(r,L,g)$ is a constant depending on r, L and g. □

2.6.2 Proposition 2.3.7

Preliminary results

To prove this proposition, we first need to extend the theory of Toeplitz matrices to hypermatrices. For more details about Toeplitz matrices, see Böttcher & Grudsky (2005). We say that $\mathbf{T}$ is an infinite lower triangular Toeplitz hypermatrix if it is lower triangular according to Definition 2.2.2 and if the value of $T_{\boldsymbol{\ell k}}$ depends only on the difference $\boldsymbol{\ell} - \boldsymbol{k}$:

$$\mathbf{T} = \left[T_{\boldsymbol{\ell k}}\right]_{\boldsymbol{\ell},\boldsymbol{k}\in\mathbb{N}^d}, \qquad T_{\boldsymbol{\ell k}} = a_{\boldsymbol{\ell}-\boldsymbol{k}},$$

with $a = [a_{\boldsymbol{k}}]_{\boldsymbol{k}\in\mathbb{Z}^d}$ and $a_{\boldsymbol{k}} = 0$ if $\boldsymbol{k} \notin \mathbb{N}^d$. So there is a bijection that takes $a \in \mathbb{R}^{\mathbb{N}^d}$ and returns the corresponding Toeplitz hypermatrix $\mathbf{T}(a)$. We can see $\mathbf{T}(a)$ as a linear map on $\mathbb{R}^{\mathbb{N}^d}$:

$$x = \left[x_k\right]_{k\in\mathbb{N}^d} \longmapsto \left[\sum_{\boldsymbol{k}\in\mathbb{N}^d} T_{\boldsymbol{\ell k}}\, x_{\boldsymbol{k}}\right]_{\boldsymbol{\ell}\in\mathbb{N}^d} = \left[\sum_{\boldsymbol{k}\leq\boldsymbol{\ell}} a_{\boldsymbol{\ell}-\boldsymbol{k}}\, x_{\boldsymbol{k}}\right]_{\boldsymbol{\ell}\in\mathbb{N}^d} = a * x.$$

The associativity of the convolution product gives that for every a and b, we have $\mathbf{T}(a) \times \mathbf{T}(b) = \mathbf{T}(a * b)$.

Notation If $\boldsymbol{m} \in \mathbb{N}_+^d$, we denote by $\mathbf{T}_{\boldsymbol{m}}(a) \in \mathbb{R}^{(\boldsymbol{m},\boldsymbol{m})}$ the sub-hypermatrix of $\mathbf{T}(a)$ constructed by taking only the coefficients $[\mathbf{T}(a)]_{\boldsymbol{\ell},\boldsymbol{k}}$ for $\boldsymbol{\ell},\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}$. Note that as a linear map on $\mathbb{R}^{\boldsymbol{m}}$, we have $\mathbf{T}_{\boldsymbol{m}}(a) = \mathbf{T}(a)_{|\mathbb{R}^{\boldsymbol{m}}}$ because of the triangular structure.

Operator on $\ell^p(\mathbb{N}^d)$ If $a \in \ell^1(\mathbb{N}^d)$ and if x is in $\ell^p(\mathbb{N}^d)$ with $p \in [1,+\infty]$, then $a * x$ belongs to $\ell^p(\mathbb{N}^d)$ we have $\|a * x\|_{\ell^p} \leq \|a\|_{\ell^1}\|x\|_{\ell^p}$. In other words, $\mathbf{T}(a)$ is an operator on $\ell^p(\mathbb{N}^d)$, and its operator norm is bounded by $\|a\|_{\ell^1}$.

The ℓ^2 case We will need to study $\mathbf{T}(a)$ as an operator on $\ell^2(\mathbb{N}^d)$ and give a bound on its norm with milder assumptions than $a \in \ell^1(\mathbb{N}^d)$. We define a subspace of $\mathrm{L}^2(\mathbb{T}^d)$ that plays an important role in this matter.

Definition 2.6.2 (Hardy space)**.** We define the Hardy space as the following subset of $\mathrm{L}^2(\mathbb{T}^d)$:

$$\mathrm{H}^2(\mathbb{T}^d) := \left\{u \in \mathrm{L}^2(\mathbb{T}^d) \;\middle|\; \forall \boldsymbol{k} \notin \mathbb{N}^d,\, c_{\boldsymbol{k}}(u) = 0\right\},$$

where $c_{\boldsymbol{k}}(u)$ denotes the k-th Fourier coefficient of u and $\mathbb{T}$ is the set of complex numbers with unitary module.

The map that takes a function and returns its Fourier coefficient is then an isometric bijection between $\mathrm{H}^2(\mathbb{T}^d)$ and $\ell^2(\mathbb{N}^d)$.

$$\begin{array}{rccc} \mathscr{F}\colon & \mathrm{H}^2(\mathbb{T}^d) & \longrightarrow & \ell^2(\mathbb{N}^d) \\ & f & \longmapsto & \left[c_{\boldsymbol{k}}(f)\right]_{\boldsymbol{k}\in\mathbb{N}^d} \end{array}.$$

We see that if u and v are $\mathrm{H}^2(\mathbb{T}^d)$ functions, the identity $\mathscr{F}[u] * \mathscr{F}[v] = \mathscr{F}[u \times v]$ translates for Toeplitz hypermatrices into $\mathbf{T}\big(\mathscr{F}[u]\big) \times \mathbf{T}\big(\mathscr{F}[v]\big) = \mathbf{T}\big(\mathscr{F}[u \times v]\big)$.

Under the additional assumption that $\mathscr{F}^{-1}[a]$ belongs to $\mathrm{L}^\infty(\mathbb{T}^d)$, we show in the next proposition that $\mathbf{T}(a)$ defines an operator on $\ell^2(\mathbb{N}^d)$.

Proposition 2.6.3. *Let $a \in \ell^2(\mathbb{N}^d)$ such that $u \coloneqq \mathscr{F}^{-1}[a] \in \mathrm{L}^\infty(\mathbb{T}^d)$. Then $\mathbf{T}(a)$ defines an operator on $\ell^2(\mathbb{N}^d)$, and its operator norm is bounded by $\|u\|_\infty$.*

Proof. Let $x \in \ell^2(\mathbb{N}^d)$ and let $v \coloneqq \mathscr{F}^{-1}[x]$. Then:

$$\|\mathbf{T}(a)x\|_{\ell^2} = \|\mathscr{F}[u] * \mathscr{F}[v]\|_{\ell^2} = \|\mathscr{F}[u \times v]\|_{\ell^2} = \|u \times v\|_{\mathrm{L}^2} \leq \|u\|_\infty \|v\|_{\mathrm{L}^2} = \|u\|_\infty \|x\|_{\ell^2}.$$

□

Before we can prove Proposition 2.3.7, some facts need to be established. We denote by $\widehat{\mathbb{C}} \coloneqq \mathbb{C} \cup \{\infty\}$ the Riemann sphere. We will use the following functions:

- We make the assumption that β belongs to $\ell^1(\mathbb{N}^d)$ so that the power series $B(\boldsymbol{z}) \coloneqq \sum_{\boldsymbol{k} \in \mathbb{N}^d} \beta_{\boldsymbol{k}} \boldsymbol{z}^{\boldsymbol{k}}$ is normally convergent on $\overline{\mathbb{D}}^d$ and defines a function which is continuous on $\overline{\mathbb{D}}^d$ and holomorphic on $\mathbb{D}^d$.
- We denote by G the Laplace transform of g. This function is defined on $\mathscr{P}_+^d$, continuous on $\mathscr{P}_+^d$, and holomorphic on $\big\{s \in \mathbb{C}^d \,\big|\, \forall q,\ \mathfrak{Re}(s_q) > 0\big\}$.
- If u is a $\mathrm{L}^1(\mathbb{R}_+^d)$ or $\mathrm{L}^2(\mathbb{R}_+^d)$ function, we denote by:

$$u^*(\boldsymbol{\omega}) \coloneqq \int_{\mathbb{R}_+^d} \mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{x}}\, u(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\omega} \in \mathbb{R}^d,$$

  its Fourier transform, where $\boldsymbol{\omega} \cdot \boldsymbol{x} \coloneqq \sum_i \omega_i x_i$. The Fourier transform of the Laguerre functions $\varphi_{\boldsymbol{k}}$, $\boldsymbol{k} \in \mathbb{N}^d$ can be computed from the case $d = 1$:

$$\varphi_{\boldsymbol{k}}^*(\boldsymbol{\omega}) = \prod_{q=1}^d \varphi_{k_q}^*(\omega_q) = \prod_{q=1}^d (-1)^{k_q} \sqrt{2} \frac{(1 + \mathrm{i}\omega_q)^{k_q}}{(1 - \mathrm{i}\omega_q)^{k_q+1}}.$$

We will need to understand the behaviour of the map $z \mapsto \frac{1+z}{1-z}$.

Lemma 2.6.4. *Let $\eta \colon \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ be the homographic function $\eta(s) \coloneqq \frac{s-1}{s+1}$ (with $\eta(\infty) = 1$ and $\eta(-1) = \infty$).*

1. *The function η is invertible and its inverse is $\eta^{-1}(z) = \frac{1+z}{1-z}$.*
2. *The image of $\{s \in \mathbb{C} \,|\, \mathfrak{Re}(s) > 0\}$ by η is $\mathbb{D}$.*
3. *The image of $\{s \in \mathbb{C} \,|\, \mathfrak{Re}(s) = 0\}$ by η is $\mathbb{T} \setminus \{1\}$.*

In the next proposition that generalizes lemma C.1 from Comte *et al.* (2017), we show that the functions G and B are linked through the transformation η.

Proposition 2.6.5. *If $\boldsymbol{\theta} \in \mathbb{R}^d$, we denote $\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}} := (\mathrm{e}^{\mathrm{i}\theta_1}, \dots, \mathrm{e}^{\mathrm{i}\theta_d})$. If $\beta \in \ell^1(\mathbb{N}^d)$, then $[\beta_{\boldsymbol{k}}]_{\boldsymbol{k}\in\mathbb{N}^d}$ are the Fourier coefficients of $\boldsymbol{\theta} \mapsto G(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}})$ and we have:*

$$\forall \boldsymbol{\theta} \in \mathbb{R}^d, \quad G\left(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}\right) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} \beta_{\boldsymbol{k}}\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}} = B\left(\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}\right),$$

with normal convergence of the series.

Remark 2.6.6. If $g \in \mathrm{W}^{\boldsymbol{s}}(\mathbb{R}^d_+, L)$ with $\boldsymbol{s} \in (1, +\infty)^d$, then β belongs to $\ell^1(\mathbb{N}^d)$, see Remark 2.3.8.

Proof of Proposition 2.6.5. We start from the expression of $\beta_{\boldsymbol{k}}$ and we use the Plancherel isometry:

$$\begin{aligned}\beta_{\boldsymbol{k}} &= 2^{-d/2} \sum_{\boldsymbol{k}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|}\, b_{\boldsymbol{k}-\boldsymbol{\varepsilon}} = 2^{-d/2} \sum_{\boldsymbol{k}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \langle g, \varphi_{\boldsymbol{k}-\boldsymbol{\varepsilon}}\rangle_{\mathrm{L}^2} \\ &= 2^{-d/2} \sum_{\boldsymbol{k}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \frac{1}{(2\pi)^d} \langle g^*, \varphi^*_{\boldsymbol{k}-\boldsymbol{\varepsilon}}\rangle_{\mathrm{L}^2} = \left\langle g^*, \frac{2^{-d/2}}{(2\pi)^d} \sum_{\boldsymbol{\varepsilon}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \varphi^*_{\boldsymbol{k}-\boldsymbol{\varepsilon}} \right\rangle_{\mathrm{L}^2}.\end{aligned}$$

Let us compute the second factor in the scalar product. The reader is referred to Section 1.3 for the notations governing vectors and multi-indices computation.

$$\begin{aligned}\frac{2^{-d/2}}{(2\pi)^d} \sum_{\boldsymbol{\varepsilon}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \varphi^*_{\boldsymbol{k}-\boldsymbol{\varepsilon}}(\boldsymbol{\omega}) &= \frac{2^{-d/2}}{(2\pi)^d} \sum_{\boldsymbol{\varepsilon}\le\mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \prod_{q=1}^{d} (-1)^{k_q-\varepsilon_q} \sqrt{2} \frac{(1+\mathrm{i}\omega_q)^{k_q-\varepsilon_q}}{(1-\mathrm{i}\omega_q)^{k_q-\varepsilon_q+1}} \\ &= \frac{(-1)^{|\boldsymbol{k}|}}{(2\pi)^d} \sum_{\boldsymbol{\varepsilon}\le\mathbf{1}} \left(\frac{\mathbf{1}+\mathrm{i}\boldsymbol{\omega}}{\mathbf{1}-\mathrm{i}\boldsymbol{\omega}}\right)^{\boldsymbol{k}-\boldsymbol{\varepsilon}} \left(\frac{\mathbf{1}}{\mathbf{1}-\mathrm{i}\boldsymbol{\omega}}\right)^{\mathbf{1}}\end{aligned}$$

Using the multibinomial theorem (Proposition 1.3.1),

$$\begin{aligned}&= \left(\frac{\mathrm{i}\boldsymbol{\omega}+\mathbf{1}}{\mathrm{i}\boldsymbol{\omega}-\mathbf{1}}\right)^{\boldsymbol{k}} \frac{1}{(2\pi)^d} \left(\frac{\mathbf{1}-\mathrm{i}\boldsymbol{\omega}}{\mathbf{1}+\mathrm{i}\boldsymbol{\omega}} + \mathbf{1}\right)^{\mathbf{1}} \left(\frac{\mathbf{1}}{\mathbf{1}-\mathrm{i}\boldsymbol{\omega}}\right)^{\mathbf{1}} \\ &= \left(\frac{\mathrm{i}\boldsymbol{\omega}+\mathbf{1}}{\mathrm{i}\boldsymbol{\omega}-\mathbf{1}}\right)^{\boldsymbol{k}} \frac{1}{\pi^d} \prod_{q=1}^{d} \frac{1}{1+\omega_q^2}.\end{aligned}$$

Hence, we reach the following expression:

$$\beta_{\boldsymbol{k}} = \frac{1}{\pi^d} \int_{\mathbb{R}^d} g^*(\boldsymbol{\omega}) \left(\frac{\mathrm{i}\boldsymbol{\omega}-\mathbf{1}}{\mathrm{i}\boldsymbol{\omega}+\mathbf{1}}\right)^{\boldsymbol{k}} \frac{\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d}{(1+\omega_1^2)\cdots(1+\omega_d^2)}.$$

The change of variable $\mathrm{e}^{-\mathrm{i}\theta_q} = \frac{\mathrm{i}\omega_q - 1}{\mathrm{i}\omega_q + 1} = \eta(\mathrm{i}\omega_q)$ yields:

$$\begin{aligned}\beta_{\boldsymbol{k}} &= \frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} g^*\left(\frac{\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}+\mathbf{1}}{\mathrm{i}(\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}-\mathbf{1})}\right) \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}\, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} G\left(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}\right) \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}\, \mathrm{d}\boldsymbol{\theta}.\end{aligned}$$

Therefore, if $\boldsymbol{k} \in \mathbb{N}^d$, the $\boldsymbol{k}$-th Fourier coefficient of the function $\boldsymbol{\theta} \mapsto G(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}})$ is $\beta_{\boldsymbol{k}}$. On the other hand, if $\boldsymbol{k} \notin \mathbb{N}^d$, let us show that the Fourier coefficients vanish. Without loss of generality, we assume that $k_1 < 0$, and we compute the $\boldsymbol{k}$-th Fourier coefficient:

$$\begin{aligned}
&\frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} G\left(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}\right) \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}} \,\mathrm{d}\boldsymbol{\theta} \\
&\quad = \frac{1}{\pi^d} \int_{\mathbb{R}^d} g^*(\boldsymbol{\omega}) \left(\frac{\mathrm{i}\omega_1 - 1}{\mathrm{i}\omega_1 + 1}\right)^{k_1} \prod_{q=2}^{d} \left(\frac{\mathrm{i}\omega_q - 1}{\mathrm{i}\omega_q + 1}\right)^{k_q} \frac{\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d}{(1+\omega_1^2)\cdots(1+\omega_d^2)} \\
&\quad = \frac{1}{\pi^d} \int_{\mathbb{R}^d} g^*(\boldsymbol{\omega}) \left(\frac{\mathrm{i}(-\omega_1) - 1}{\mathrm{i}(-\omega_1) + 1}\right)^{-k_1} \prod_{q=2}^{d} \left(\frac{\mathrm{i}\omega_q - 1}{\mathrm{i}\omega_q + 1}\right)^{k_q} \frac{\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d}{(1+\omega_1^2)\cdots(1+\omega_d^2)} \\
&\quad = \frac{-1}{\pi^d} \int_{\mathbb{R}^d} g^*(-\omega_1, \omega_2, \ldots, \omega_d) \left(\frac{\mathrm{i}\boldsymbol{\omega} - \mathbf{1}}{\mathrm{i}\boldsymbol{\omega} + \mathbf{1}}\right)^{\boldsymbol{k}'} \frac{\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d}{(1+\omega_1^2)\cdots(1+\omega_d^2)},
\end{aligned}$$

where $\boldsymbol{k}' \coloneqq (-k_1, k_2, \ldots, k_d)$. Let δ be the map on $\mathbb{R}^d$ defined by:

$$\delta(\boldsymbol{\omega}) \coloneqq (-\omega_1, \omega_2, \ldots, \omega_d),$$

then we have:

$$\begin{aligned}
&\frac{-1}{\pi^d} \int_{\mathbb{R}^d} g^*\big(\delta(\boldsymbol{\omega})\big) \left(\frac{\mathrm{i}\boldsymbol{\omega} - \mathbf{1}}{\mathrm{i}\boldsymbol{\omega} + \mathbf{1}}\right)^{\boldsymbol{k}'} \frac{\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d}{(1+\omega_1^2)\cdots(1+\omega_d^2)} \\
&\qquad = \left\langle -g^* \circ \delta, \frac{2^{-d/2}}{(2\pi)^d} \sum_{\boldsymbol{\varepsilon} \le \mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \varphi^*_{\boldsymbol{k}'-\boldsymbol{\varepsilon}} \right\rangle_{\mathrm{L}^2} \\
&\qquad = 2^{-d/2} \sum_{\boldsymbol{k} \le \mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \frac{1}{(2\pi)^d} \langle -g^* \circ \delta, \varphi^*_{\boldsymbol{k}'-\boldsymbol{\varepsilon}} \rangle_{\mathrm{L}^2} \\
&\qquad = 2^{-d/2} \sum_{\boldsymbol{k} \le \mathbf{1}} (-1)^{|\boldsymbol{\varepsilon}|} \langle g \circ \delta, \varphi_{\boldsymbol{k}'-\boldsymbol{\varepsilon}} \rangle_{\mathrm{L}^2},
\end{aligned}$$

because $(g \circ \delta)^* = -g^* \circ \delta$. This last expression is always zero:

$$\langle g \circ \delta, \varphi_{\boldsymbol{k}'-\boldsymbol{\varepsilon}} \rangle_{\mathrm{L}^2} = \int_{\mathbb{R}^d} g(-x_1, x_2, \ldots, x_d)\, \varphi_{\boldsymbol{k}'-\boldsymbol{\varepsilon}}(x) \,\mathrm{d}\boldsymbol{x} = 0,$$

because the function g is zero of $x_1 > 0$, and $\varphi_{\boldsymbol{k}'-\boldsymbol{\varepsilon}}$ is zero if $x_1 < 0$. Thus, the Fourier coefficients of $\boldsymbol{\theta} \mapsto G(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}})$ are $[\beta_{\boldsymbol{k}}]_{\boldsymbol{k} \in \mathbb{N}^d}$.

The function G is continuous on $(\mathrm{i}\mathbb{R} \cup \{\infty\})^d$, so the function $G \circ (\eta^{-1})^{\otimes d}$ is continuous on $\mathbb{T}^d$ by Lemma 2.6.4. Therefore, since the Fourier series $\boldsymbol{\theta} \mapsto G(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}})$ is normally convergent, this function is equal to its Fourier series at each point:

$$\forall \boldsymbol{\theta} \in \mathbb{R}^d, \quad G\left(\frac{\mathbf{1}+\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}{\mathbf{1}-\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}}}\right) = \sum_{\boldsymbol{k} \in \mathbb{N}^d} \beta_{\boldsymbol{k}}\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}. \qquad \square$$

We need a last technical lemma before we start the proof of Proposition 2.3.7.

Lemma 2.6.7. *Let $\boldsymbol{\alpha} \in \mathbb{N}_+^d$, then $(\mathbf{1}-\boldsymbol{z})^{-\boldsymbol{\alpha}}$ admits a power series expansion on $\mathbb{D}^d$ given by:*

$$\forall \boldsymbol{z} \in \mathbb{D}^d, \quad (\mathbf{1}-\boldsymbol{z})^{-\boldsymbol{\alpha}} = \sum_{\boldsymbol{j}\in\mathbb{N}^d} \binom{\boldsymbol{\alpha}-\mathbf{1}+\boldsymbol{j}}{\boldsymbol{j}} \boldsymbol{z}^{\boldsymbol{j}}.$$

Moreover, denoting ζ_j the j-th coefficient in the power series above, for $\boldsymbol{m} \geq \mathbf{4}$ we have $\|\mathbf{T}_{\boldsymbol{m}}(\zeta)\|_F^2 \leq \boldsymbol{m}^{2\boldsymbol{\alpha}}$ where $\mathbf{T}(\zeta)$ is the Toeplitz hypermatrix contructed from the coefficients ζ_j.

Proof of Lemma 2.6.7. We recall the following identity: for $z \in \mathbb{D}$ and $r \in \mathbb{N}^*$, we have $(1-z)^{-r} = \sum_{j=0}^{\infty} \binom{r-1+j}{j} z^j$. Thus, for $\boldsymbol{z} \in \mathbb{D}^d$ and $\boldsymbol{\alpha} \in \mathbb{N}_+^d$, we have:

$$(\mathbf{1}-\boldsymbol{z})^{-\boldsymbol{\alpha}} = \prod_{q=1}^{d} (1-z_q)^{-\alpha_q} = \prod_{q=1}^{d} \sum_{j_q=0}^{\infty} \binom{\alpha_q - 1 + j_q}{j_q} z_q^{j_q} = \sum_{\boldsymbol{j}\in\mathbb{N}^d} \binom{\boldsymbol{\alpha}-\mathbf{1}+\boldsymbol{j}}{\boldsymbol{j}} \boldsymbol{z}^{\boldsymbol{j}}.$$

Therefore:

$$\begin{aligned}\|\mathbf{T}_{\boldsymbol{m}}(\zeta)\|_F^2 &= \sum_{\boldsymbol{j}\leq\boldsymbol{m}-\mathbf{1}} \binom{\boldsymbol{\alpha}-\mathbf{1}+\boldsymbol{j}}{\boldsymbol{j}}^2 \operatorname{Card}\Big\{(\boldsymbol{k},\boldsymbol{\ell}) \in \mathbb{N}^d \times \mathbb{N}^d \;\Big|\; \boldsymbol{k} \leq \boldsymbol{\ell} \leq \boldsymbol{m}-\mathbf{1}, \boldsymbol{\ell}-\boldsymbol{k} = \boldsymbol{j}\Big\} \\ &= \prod_{q=1}^{d} \sum_{j_q=0}^{m_q-1} \binom{\alpha-1+j_q}{j_q}^2 (m_q - j_q),\end{aligned}$$

and it reduces to the case $d = 1$ which was already solved in (Comte *et al.*, 2017, appendix C). So if $m_q \geq 4$ for every q, then $\|\mathbf{T}_{\boldsymbol{m}}(\zeta)\|_F^2 \leq \prod_{q=1}^{d} m_q^{2\alpha_q} = \boldsymbol{m}^{2\boldsymbol{\alpha}}$. □

Proof of Proposition 2.3.7

From Proposition 2.6.5, we get $G = B \circ \eta^{\otimes d}$. We define a function w on $(\overline{\mathbb{D}} \setminus \{1\})^d$ by:

$$\forall \boldsymbol{z} \in \left(\overline{\mathbb{D}} \setminus \{1\}\right)^d, \quad w(\boldsymbol{z}) := (\mathbf{1}-\boldsymbol{z})^{-\boldsymbol{\alpha}} B(\boldsymbol{z}).$$

This function is related to $K_{\boldsymbol{\alpha}}$ by the identity:

$$\forall \boldsymbol{z} \in \left(\overline{\mathbb{D}} \setminus \{1\}\right)^d, \quad w(\boldsymbol{z}) = \mathbf{2}^{-\boldsymbol{\alpha}} K_{\boldsymbol{\alpha}}\big((\eta^{-1})^{\otimes d}(\boldsymbol{z})\big).$$

Thus, the function w can be extended as a function on $\overline{\mathbb{D}}^d$, still denoted w, and according to our assumptions on $K_{\boldsymbol{\alpha}}$ and Lemma 2.6.4, w satisfies:

- $w_{|\mathbb{T}^d}$ is continuous;
- w is continuous on $\left(\overline{\mathbb{D}} \setminus \{1\}\right)^d$;
- w is holomorphic on $\mathbb{D}^d$;
- w doesn't vanish on $\overline{\mathbb{D}}^d$.

Thus, the function $w^{-1} = 1/w$ is well defined on $\overline{\mathbb{D}}^d$ and has the same properties. In particular, since it is holomorphic on $\mathbb{D}^d$, it admits a power series expansion:

$$\forall \boldsymbol{z} \in \mathbb{D}^d, \quad w^{-1}(\boldsymbol{z}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} d_{\boldsymbol{k}}\, \boldsymbol{z}^{\boldsymbol{k}}.$$

Let us have look on what is happening on $\mathbb{T}^d$. For $r \in [0,1[$, we denote w_r^{-1} the function on $\mathbb{T}^d$ defined by:

$$\forall \boldsymbol{t} \in \mathbb{T}^d, \quad w_r^{-1} = w^{-1}(r\,\boldsymbol{t}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} d_{\boldsymbol{k}}\, r^{|\boldsymbol{k}|}\, \boldsymbol{t}^{\boldsymbol{k}}.$$

On the one hand, the Fourier coefficients of w_r^{-1} are $d_{\boldsymbol{k}}\, r^{|\boldsymbol{k}|}$ (we set $d_{\boldsymbol{k}} = 0$ if $\boldsymbol{k} \notin \mathbb{N}^d$). On the other hand, we can compute the Fourier coefficients and we get:

$$\forall \boldsymbol{k} \in \mathbb{Z}^d, \quad d_{\boldsymbol{k}}\, r^{|\boldsymbol{k}|} = \frac{1}{(2\pi)^d} \int\limits_{[0,2\pi]^d} w^{-1}(r\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}\, \mathrm{d}\boldsymbol{\theta}.$$

Since w^{-1} is continuous on $\left(\overline{\mathbb{D}} \setminus \{1\}\right)^d$, we have $w_r^{-1} \to w_{|\mathbb{T}^d}^{-1}$ a.e. on $\mathbb{T}^d$. By dominated convergence, we obtain:

$$\forall \boldsymbol{k} \in \mathbb{Z}^d, \quad d_{\boldsymbol{k}} = \frac{1}{(2\pi)^d} \int\limits_{[0,2\pi]^d} w^{-1}(\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}\, \mathrm{d}\boldsymbol{\theta}.$$

Therefore, $(d_{\boldsymbol{k}})$ are the Fourier coefficients of $w_{|\mathbb{T}^d}^{-1}$. Thus, we have shown that $w_{|\mathbb{T}^d}^{-1} \in \mathrm{H}^2(\mathbb{T}^d)$ and:

$$w^{-1}(\boldsymbol{t}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} d_{\boldsymbol{k}}\, \boldsymbol{t}^{\boldsymbol{k}},$$

with $\mathrm{L}^2(\mathbb{T}^d)$–convergence of the series.
By Lemma 2.6.7, the function $(\mathbf{1} - \boldsymbol{z})^{-\boldsymbol{\alpha}}$ admits a power series expansion on $\mathbb{D}^d$ given by $(\mathbf{1} - \boldsymbol{z})^{-\boldsymbol{\alpha}} = \sum_{\boldsymbol{k}\in\mathbb{N}^d} \zeta_{\boldsymbol{k}} \boldsymbol{z}^{\boldsymbol{k}}$. Thus, the power series equality:

$$B(\boldsymbol{z}) \times (\mathbf{1} - \boldsymbol{z})^{-\boldsymbol{\alpha}} \times w^{-1}(\boldsymbol{z}) = 1,$$

on the domain $\mathbb{D}^d$ translates to their coefficients into the relation $\beta * \zeta * d = \delta_{\mathbf{0}}$, where $\delta_{\mathbf{0}}$ is the element of $\mathbb{R}^{\mathbb{N}^d}$ defined by:

$$\left[\delta_{\mathbf{0}}\right]_{\boldsymbol{k}} = \begin{cases} 1 & \text{if } \boldsymbol{k} = \mathbf{0}, \\ 0 & \text{else.} \end{cases}$$

Taking the corresponding Toeplitz hypermatrices, we get $\mathbf{G} \times \mathbf{T}(\zeta) \times \mathbf{T}(d) = \mathbf{I}_{\mathbb{N}^d}$, where $\mathbf{I}_{\mathbb{N}^d}$ is the infinite hypermatrix given by:

$$\left[\mathbf{I}_{\mathbb{N}^d}\right]_{\boldsymbol{\ell}\boldsymbol{k}} = \begin{cases} 1 & \text{if } \boldsymbol{\ell} = \boldsymbol{k}, \\ 0 & \text{else.} \end{cases}$$

Thus, for $\boldsymbol{m} \in \mathbb{N}_+^d$, we get $\mathbf{G}_{\boldsymbol{m}}^{-1} = \mathbf{T}_{\boldsymbol{m}}(\zeta) \times \mathbf{T}_{\boldsymbol{m}}(d)$. Taking the Frobenius norm, we obtain the following inequality:

$$\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2 \leq \|\mathbf{T}_{\boldsymbol{m}}(\zeta)\|_F^2 \times \|\mathbf{T}_{\boldsymbol{m}}(d)\|_{\mathrm{op}}^2.$$

From Proposition 2.6.3, we have:

$$\|\mathbf{T}_{\boldsymbol{m}}(d)\|_{\mathrm{op}}^2 = \sup_{\mathbf{a} \in \mathbb{R}^{\boldsymbol{m}} \setminus \{\mathbf{0}\}} \frac{\|\mathbf{T}_{\boldsymbol{m}}(d) \times_d \mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}^2}{\|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}^2} \leq \sup_{x \in \ell^2(\mathbb{N}^d) \setminus \{0\}} \frac{\|\mathbf{T}(d)\, a\|_{\ell^2}^2}{\|a\|_{\ell^2}^2} = \|w_{|\mathbb{T}^d}^{-1}\|_\infty^2,$$

and by Lemma 2.6.7, we have $\|\mathbf{T}_{\boldsymbol{m}}(\zeta)\|_F^2 \leq \boldsymbol{m}^{2\boldsymbol{\alpha}}$ if $\boldsymbol{m} \geq \mathbf{4}$. □

2.6.3 Proofs of Section 2.4

Proof of Theorem 2.4.2. The proof is identical to the one in Mabon (2017), but we use our Lemma 2.4.1 instead of Mabon's proposition 7.1. □

Proof of Lemma 2.4.1. Let $\boldsymbol{m} \in \mathbb{N}_+^d$, we prove that $\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2$ can be written as the supremum of an empirical process, then we use Talagrand's inequality.

First, the coefficient $\widehat{a}_{\boldsymbol{k}}$ is given by:

$$\widehat{a}_{\boldsymbol{k}} = \left[\mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \widehat{\mathbf{c}}_{\boldsymbol{m}}\right]_{\boldsymbol{k}} = \sum_{\boldsymbol{j} \leq \boldsymbol{k}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{k}\boldsymbol{j}}\, \widehat{c}_{\boldsymbol{j}} = \frac{1}{n} \sum_{i=1}^n \left(\sum_{\boldsymbol{j} \leq \boldsymbol{k}} [\mathbf{G}_{\boldsymbol{m}}^{-1}]_{\boldsymbol{k}\boldsymbol{j}}\, \varphi_{\boldsymbol{j}}(\boldsymbol{Z}_i) \right).$$

We rewrite this expression with hypermatrix calculus: for $\boldsymbol{x} \in \mathbb{R}_+^d$, let $\Phi_{\boldsymbol{m}}(\boldsymbol{x}) \in \mathbb{R}^{\boldsymbol{m}}$ be the hypermatrix $[\varphi_{\boldsymbol{j}}(\boldsymbol{x})]_{\boldsymbol{j} \leq \boldsymbol{m}-\mathbf{1}}$, and let $\mathbf{E}_{\boldsymbol{k}} \in \mathbb{R}^{\boldsymbol{m}}$ be the elementary hypermatrix with 1 in position $\boldsymbol{k}$ and zeros elsewhere; the coefficient $\widehat{a}_{\boldsymbol{k}}$ can be written as:

$$\widehat{a}_{\boldsymbol{k}} = \left\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \widehat{\mathbf{c}}_{\boldsymbol{m}}, \mathbf{E}_{\boldsymbol{k}} \right\rangle_{\mathbb{R}^{\boldsymbol{m}}} = \frac{1}{n} \sum_{i=1}^n \left\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_i), \mathbf{E}_{\boldsymbol{k}} \right\rangle_{\mathbb{R}^{\boldsymbol{m}}}.$$

For $t \in S_{\boldsymbol{m}}$, let us introduce the empirical process:

$$\nu_n(t) \coloneqq \frac{1}{n} \sum_{i=1}^n \left(\left\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_i), \mathbf{t}_{\boldsymbol{m}} \right\rangle_{\mathbb{R}^{\boldsymbol{m}}} - \mathbb{E}\left[\left\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_i), \mathbf{t}_{\boldsymbol{m}} \right\rangle_{\mathbb{R}^{\boldsymbol{m}}} \right] \right),$$

where $\mathbf{t}_{\boldsymbol{m}} \in \mathbb{R}^{\boldsymbol{m}}$ is the hypermatrix of the Laguerre coefficients of t. Since $\mathbf{E}_{\boldsymbol{k}}$ is the hypermatrix of the coefficients of $\varphi_{\boldsymbol{k}}$, we have:

$$\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} (\widehat{a}_{\boldsymbol{k}} - a_{\boldsymbol{k}})^2 = \sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \nu_n^2(\varphi_{\boldsymbol{k}}).$$

The map $t \mapsto \nu_n(t)$ being linear on $S_{\boldsymbol{m}}$ and $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}}$ being a basis of $S_{\boldsymbol{m}}$, the Cauchy–Schwarz inequality implies that:

$$\sum_{\boldsymbol{k} \leq \boldsymbol{m}-\mathbf{1}} \nu_n^2(\varphi_{\boldsymbol{k}}) = \sup_{t \in B_{\boldsymbol{m}}} \nu_n^2(t), \quad B_{\boldsymbol{m}} \coloneqq \left\{ t \in S_{\boldsymbol{m}} \,|\, \|t\|_{\mathrm{L}^2} = 1 \right\}.$$

Thus, the following holds:

$$\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 = \sup_{t \in B_{\boldsymbol{m}}} \nu_n^2(t), \quad B_{\boldsymbol{m}} \coloneqq \big\{ t \in S_{\boldsymbol{m}} \,|\, \|t\|_{\mathrm{L}^2} = 1 \big\}.$$

Secondly, we apply[1] Talagrand's inequality (Theorem 1.4.8). We need the constants M, H and v that appear in this inequality.

- *Computation of M.* By the Cauchy–Schwarz inequality, we have:

$$\begin{aligned} \big| \big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x}), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m} \big|^2 &\le \|\mathbf{t}_{\boldsymbol{m}}\|_{\mathbb{R}^m}^2 \, \|\mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{\mathbb{R}^m}^2 \\ &\le \|t\|_{\mathrm{L}^2}^2 \, \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \sum_{\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}} \varphi_{\boldsymbol{k}}(\boldsymbol{x})^2. \end{aligned}$$

  Hence, we get:

$$\sup_{t \in B_{\boldsymbol{m}}} \sup_{\boldsymbol{x} \in \mathbb{R}_+^d} \big| \big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x}), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m} \big|^2 \le 2^d \, D_{\boldsymbol{m}} \, \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 =: M^2.$$

- *Computation of H.*

$$\mathbb{E}\left[\sup_{t \in B_{\boldsymbol{m}}} |\nu_n(t)| \right]^2 = \mathbb{E}\big[\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2} \big]^2 \le \mathbb{E}\big[\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 \big].$$

  We recognize the variance term, that is bounded using Proposition 2.3.4:

$$\mathbb{E}\left[\sup_{t \in B_{\boldsymbol{m}}} |\nu_n(t)| \right]^2 \le \frac{c(g)\sqrt{D_{\boldsymbol{m}}} \, \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \wedge \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2}{n} =: H^2.$$

- *Computation of v.*

$$\begin{aligned} \mathrm{Var}\big(\big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_1), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m} \big) &\le \mathbb{E}\Big[\big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_1), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m}^2 \Big] \\ &= \int_{\mathbb{R}_+^d} \big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x}), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m}^2 \, h(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &\le \|h\|_\infty \int_{\mathbb{R}_+^d} \big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x}), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m}^2 \, \mathrm{d}\boldsymbol{x}. \end{aligned}$$

  The norm of h is bounded by $\|g\|_\infty$. We compute the integral:

$$\begin{aligned} \int_{\mathbb{R}_+^d} \big\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{x}), \mathbf{t}_{\boldsymbol{m}} \big\rangle_{\mathbb{R}^m}^2 \, \mathrm{d}\boldsymbol{x} &= \int_{\mathbb{R}_+^d} \left(\sum_{\boldsymbol{k}, \boldsymbol{j} \le \boldsymbol{m}-\mathbf{1}} [\mathbf{t}_{\boldsymbol{m}}]_{\boldsymbol{k}} \, \mathbf{G}_{\boldsymbol{k}\boldsymbol{j}}^{-1} \, \varphi_{\boldsymbol{j}}(\boldsymbol{x}) \right)^2 \mathrm{d}\boldsymbol{x} \\ &= \left\| \sum_{\boldsymbol{k}, \boldsymbol{j} \le \boldsymbol{m}-\mathbf{1}} [\mathbf{t}_{\boldsymbol{m}}]_{\boldsymbol{k}} \, \mathbf{G}_{\boldsymbol{k}\boldsymbol{j}}^{-1} \, \varphi_{\boldsymbol{j}} \right\|_{\mathrm{L}^2}^2 \\ &= \sum_{\boldsymbol{j} \le \boldsymbol{m}-\mathbf{1}} \left(\sum_{\boldsymbol{k} \le \boldsymbol{m}-\mathbf{1}} [\mathbf{t}_{\boldsymbol{m}}]_{\boldsymbol{k}} \mathbf{G}_{\boldsymbol{k}\boldsymbol{j}}^{-1} \right)^2 \\ &= \big\| \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \mathbf{t}_{\boldsymbol{m}} \big\|_{\mathbb{R}^m}^2 \le \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \|\mathbf{t}_{\boldsymbol{m}}\|_{\mathbb{R}^m}^2. \end{aligned}$$

[1] this inequality concerns countable families of functions, but it's not a problem here since ν_n is continuous on $S_{\boldsymbol{m}}$, and $B_{\boldsymbol{m}}$ is separable.

Hence, we have:

$$\sup_{t\in B_{\boldsymbol{m}}} \mathrm{Var}\left(\left\langle \mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \Phi_{\boldsymbol{m}}(\boldsymbol{Z}_1), \mathbf{t}_{\boldsymbol{m}} \right\rangle_{\mathbb{R}^m}\right) \leq \|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 =: v.$$

We consider two cases.

1. If $c(g)\sqrt{D_{\boldsymbol{m}}}\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq (\|g\|_\infty \vee 1)\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2$, then we have:

$$\frac{nH^2}{v} = \frac{c(g)}{\|g\|_\infty}\sqrt{D_{\boldsymbol{m}}}, \qquad \frac{nH}{M} = \frac{\sqrt{c(g)}}{2^{d/2}}\sqrt{n}D_{\boldsymbol{m}}^{-1/4} \geq \frac{\sqrt{c(g)}}{2^{d/2}}n^{1/4},$$

since $D_{\boldsymbol{m}} \leq n$. For $\delta > 0$, Talagrand's inequality yields:

$$\mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - 2(1+2\delta)H^2\right)_+\right] \lesssim \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n}\exp\left(-K\delta\frac{c(g)}{\|g\|_\infty}\sqrt{D_{\boldsymbol{m}}}\right) + \frac{D_{\boldsymbol{m}}\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{C(\delta)^2 n^2}\exp\left(-\frac{KC(\delta)\sqrt{2\delta}}{7}\frac{\sqrt{c(g)}}{2^{d/2}}n^{1/4}\right).$$

We control the first term using Assumption 2.3:

$$\sum_{\boldsymbol{m}\in\mathcal{M}_n} \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n}\exp\left(-K\delta\frac{c(g)}{\|g\|_\infty}\sqrt{D_{\boldsymbol{m}}}\right) \leq \frac{\|g\|_\infty K(\delta)}{n},$$

where $K(\delta)$ is a constant not depending on n. To control the second term, we use that $D_{\boldsymbol{m}}\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq n$ when $\boldsymbol{m} \in \mathcal{M}_n$:

$$\sum_{\boldsymbol{m}\in\mathcal{M}_n} \frac{D_{\boldsymbol{m}}\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{C(\delta)^2 n^2}\exp\left(-\frac{KC(\delta)\sqrt{2\delta}}{7}\frac{\sqrt{c(g)}}{2^{d/2}}n^{1/4}\right) \lesssim \exp\left(-\tilde{K}(\delta)\, n^{1/4}\right),$$

so we have our result.

2. If $c(g)\sqrt{D_{\boldsymbol{m}}}\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \geq (\|g\|_\infty \vee 1)\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2$, then we have:

$$\frac{nH^2}{v} = \frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F^2}{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2} \geq 1, \qquad \frac{nH}{M} = \frac{\sqrt{n}\,(\|g\|_\infty \vee 1)^{1/2}\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_F}{2^{d/2}\sqrt{D_{\boldsymbol{m}}}\,\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \geq \frac{\sqrt{n}}{2^{d/2}}D_{\boldsymbol{m}}^{-1/2}.$$

Let us show that $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \geq 1$. We notice that $\frac{2^{d/2}}{b_{\mathbf{0}}}$ is an eigenvalue of $\mathbf{G}_{\boldsymbol{m}}^{-1}$: indeed, if $\mathbf{E_0}$ is the elementary hypermatrix with 1 in position $(0,\ldots,0)$ and zeros elsewhere, then:

$$\mathbf{G}_{\boldsymbol{m}}^{-1} \times_d \mathbf{E_0} = \frac{2^{d/2}}{b_{\mathbf{0}}}\mathbf{E_0}.$$

Since the operator norm is bigger than the spectral radius, we have:

$$\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \geq \frac{2^{d/2}}{b_{\mathbf{0}}} = \frac{1}{\mathbb{E}\left[\mathrm{e}^{-(Y_1^{(1)}+\cdots+Y_1^{(d)})}\right]} \geq 1.$$

Thus, if $\boldsymbol{m} \in \mathscr{M}_n$ we have $D_{\boldsymbol{m}} \leq D_{\boldsymbol{m}} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq \frac{n}{\log n}$. Therefore, $\frac{nH}{M} \geq \frac{\sqrt{\log n}}{2^{d/2}}$. We apply Talagrand's inequality with $\delta = a \log n$ and $a > 0$ to be chosen later:

$$\mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - 2(1+2a\log n)H^2\right)_+\right] \lesssim \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n} \frac{1}{n^{Ka}} + \frac{D_{\boldsymbol{m}} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{C(a\log n)^2 n^2} \exp\left(-\frac{KC(a\log n)\sqrt{2a}}{7} \frac{\log n}{2^{d/2}}\right).$$

We assume that[2] $n \geq \exp(3/a)$, such that $C(a\log n) = 1$. Hence, we get:

$$\mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}} - f_{\boldsymbol{m}}\|_{\mathrm{L}^2}^2 - 2(1+2a\log n)H^2\right)_+\right] \lesssim \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n^{1+Ka}} + \frac{D_{\boldsymbol{m}} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n^2} \exp\left(-\frac{K\sqrt{2a}}{7\times 2^{d/2}} \log n\right).$$

We control the first term using that if $\boldsymbol{m} \in \mathscr{M}_n$ then $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq n$:

$$\sum_{\boldsymbol{m}\in\mathscr{M}_n} \frac{\|g\|_\infty \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n^{1+Ka}} \leq \frac{\|g\|_\infty}{n^{Ka-1}} \leq \frac{\|g\|_\infty}{n},$$

if $a \geq 2/K = 12$. To control the second term, we use that if $\boldsymbol{m} \in \mathscr{M}_n$ then $D_{\boldsymbol{m}} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \leq n$, so that we obtain:

$$\sum_{\boldsymbol{m}\in\mathscr{M}_n} \frac{D_{\boldsymbol{m}} \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2}{n^2} \exp\left(-\frac{K\sqrt{2a}}{7\times 2^{d/2}} \log n\right) \leq n^{-\frac{K\sqrt{2a}}{7\times 2^{d/2}}} \leq \frac{1}{n},$$

if $a \geq \frac{2^d \times 7^2}{2K^2}$. □

Proof of Theorem 2.4.3. Using Lemma 1.4.6, we have for all $\boldsymbol{m} \in \mathscr{M}_n$:

$$\mathbb{E}\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq C\|f - f_{\boldsymbol{m}}\|^2 + C'(1+\kappa_1+\kappa_2)V(\boldsymbol{m}) + C''\mathbb{E}[R_n],$$

with:

$$R_n := \max_{\boldsymbol{m}'\in\mathscr{M}_n} \left(\|\widehat{f}_{\boldsymbol{m}'} - f_{\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6} V(\boldsymbol{m}')\right)_+.$$

Taking the infimum on $\boldsymbol{m}$, we obtain:

$$\mathbb{E}\|f - \widehat{f}_{\widehat{\boldsymbol{m}}}\|^2 \leq \max\left(C, C'(1+\kappa_1+\kappa_2)\right) \inf_{\boldsymbol{m}\in\mathscr{M}_n} \left(\|f - f_{\boldsymbol{m}}\|^2 + V(\boldsymbol{m})\right) + C''\mathbb{E}[R_n].$$

We control the rest with a sum on all the models:

$$\mathbb{E}[R_n] \leq \sum_{\boldsymbol{m}'\in\mathscr{M}_n} \mathbb{E}\left[\left(\|\widehat{f}_{\boldsymbol{m}'} - f_{\boldsymbol{m}'}\|^2 - \frac{\kappa_1}{6} V(\boldsymbol{m}')\right)_+\right].$$

We conclude by applying Lemma 2.4.1. □

[2] actually, we assume later that $a \geq 12$, so there is no assumptions on n after all.

Chapter 3

Nonparametric Estimation of the Expected Discounted Penalty Function in the Compound Poisson Model

This chapter is a modified version of my article F. DUSSAP : Nonparametric estimation of the expected discounted penalty function in the compound Poisson model. *Electronic Journal of Statistics*, 16(1):2124-2174, 2022.

Contents

3.1 Introduction

3.1.1 The statistical problem

We consider the classical risk model (compound Poisson model) for the risk reserve process $(U_t)_{t\geq 0}$ of an insurance company:

$$U_t = u + ct - \sum_{i=1}^{N_t} X_i, \quad t \geq 0 \tag{3.1}$$

where $u \geq 0$ is the initial capital; $c > 0$ is the premium rate; the claim number process (N_t) is a homogeneous Poisson process with intensity λ; the claim sizes (X_i) are positive and i.i.d. with density f and mean μ, independent of (N_t). We denote by $\tau(u)$ the ruin time:

$$\tau(u) := \inf\left\{ t \geq 0 \,\middle|\, \sum_{i=1}^{N_t} X_i - ct > u \right\} \in \mathbb{R}_+ \cup \{+\infty\}$$

and we make the following assumption to ensure that $\tau(u)$ is not almost surely finite.

Assumption 3.1 (safety loading condition)**.** Let $\theta := \frac{\lambda\mu}{c}$, we assume that $\theta < 1$.

To study simultaneously the ruin time, the deficit at ruin, and the surplus level before the ruin, Gerber & Shiu (1998) introduced the function:

$$\phi(u) := \mathbb{E}\Big[\mathrm{e}^{-\delta\tau(u)} w\big(U_{\tau(u)-}, |U_{\tau(u)}|\big) \mathbf{1}_{\tau(u)<+\infty} \Big], \tag{3.2}$$

where $\delta \geq 0$, and w is a non-negative function of the surplus before the ruin and the deficit at ruin. This function is called the *expected discounted penalty function*, but it will also be referred to as the *Gerber–Shiu function* in the following. For more information concerning the compound Poisson model and the Gerber–Shiu function, see Asmussen & Albrecher (2010).

Example 3.1.1. Several quantities of interest can be put in the form (3.2),

1. if $\delta = 0$ and $w(x,y) = 1$, then $\phi(u)$ is the probability of ruin;
2. if $\delta > 0$ and $w(x,y) = 1$, then $\phi(u)$ is the Laplace transform of $\tau(u)$;
3. if $\delta = 0$ and $w(x,y) = x + y$, then $\phi(u)$ is the expected jump size causing the ruin.

Observations and goal In all this chapter, we suppose that the premium rate c is known, but the parameters of the aggregate claims process are not, that is (λ, μ, f) is unknown. We suppose we have observed the process $(U_t)_{t\geq 0}$ during the interval $[0, T]$, with $T > 0$ fixed, so we have access to the number of claims

and their size. Our goal is to recover the Gerber–Shiu function from the observations $(N_T; X_1, \ldots, X_{N_T})$.

Several authors have considered the problem of estimating the Gerber–Shiu function using nonparametric methods. The first articles had an asymptotic approach: Frees (1986), Croux & Veraverbeke (1990), Pitts (1994), Politis (2003), and Masiello (2014) constructed nonparametric estimators of the ruin probability, and established their consistency and their asymptotic normality.

Concerning non-asymptotic approaches, Mnatsakanov, Ruymgaart, & Ruymgaart (2008) introduced a regularized Laplace inversion method to estimate the ruin probability in the compound Poisson model. Shimizu (2011, 2012) then extended this method to the estimation of the Gerber–Shiu function in more general risk models. However, this method suffers from poor rates of convergence, and numerical difficulties to compute the estimator.

In their paper, Zhang & Su (2018) introduced a projection estimator on the Laguerre basis to overcome these drawbacks. The choice of this basis is motivated by the work of Comte, Cuenod, Pensky, & Rozenholc (2017), where the properties of the Laguerre functions relative to the convolution product are used to solve a Laplace deconvolution problem. The same method was then used in more general risk models: Zhang & Su (2019) estimate the Gerber–Shiu function in a Lévy risk model, where the aggregate claims is a pure-jump Lévy process; Su, Yong, & Zhang (2019) estimate the Gerber–Shiu function in the compound Poisson model perturbed by a Brownian motion; and Su, Shi, & Wang (2020) study the model where both the income and the aggregate claims are compound Poisson processes. Recently, Su & Yu (2020) showed the Laguerre projection estimator of the Gerber–Shiu function in the compound Poisson model is pointwise asymptotically normal in the case $\delta = 0$.

In this paper, we construct an estimator of the Gerber–Shiu function (3.2) in the compound Poisson model (3.1). As Zhang & Su (2018), our estimator is a projection estimator on the Laguerre basis, but we compute the coefficients using Plancherel theorem instead of using a Laguerre deconvolution method. We emphasize that our estimator achieves parametric rates of convergence on Sobolev–Laguerre spaces regardless of the regularity of the Gerber–Shiu function, and without needing to find a compromise between the bias and the variance.

We also improve the previous results concerning the Laguerre deconvolution method. Previous rates were given in $\mathrm{O_P}$, and we propose a non-asymptotic bound on the MISE (Mean Integrated Squared Error) of the estimator. To achieve this goal, we introduce two modified versions of the Laguerre deconvolution estimator: the first one depends on a truncation parameter, whereas the second one does not, but it is only defined in the case $\delta = 0$.

To control the MISE of the second version of the Laguerre deconvolution estimator, we had to prove that the primitives of the Laguerre functions were uniformly bounded (see Theorem A.3.1). This result is interesting in itself, the proof

relies on the study of the properties of the ODE's satisfied by Laguerre polynomials. The interested reader can find all the details in Appendix A.3.

Outline of the paper In the remaining part of this section, we introduce the notations and we give preliminary results on the Gerber–Shiu function. In Section 3.2, we construct our estimator and we study its MISE. In Section 3.3, we introduce two modified versions of the Laguerre deconvolution estimator and we study their MISE. In Section 3.4, we compute convergence rates of the different estimators considered on Sobolev–Laguerre spaces and also in the case where the claim sizes are exponentially distributed. In Section 3.5, we compare numerically the estimators on simulated data. We gathered all the proofs in Section 3.7.

3.1.2 Preliminaries on the Gerber–Shiu function

The key result to estimate the Gerber–Shiu function is the following theorem.

Theorem 3.1.2 (Gerber & Shiu (1998))**.** *Under Assumption 3.1, the Gerber–Shiu function satisfies the equation:*

$$\phi = \phi * g + h. \tag{3.3}$$

where g and h are given by:

$$\begin{aligned} g(x) &:= \frac{\lambda}{c}\int_x^{+\infty} \mathrm{e}^{-\rho_\delta(y-x)} f(y)\,\mathrm{d}y, \\ h(u) &:= \frac{\lambda}{c}\int_u^{+\infty} \mathrm{e}^{-\rho_\delta(x-u)} \left(\int_x^{+\infty} w(x,y-x) f(y)\,\mathrm{d}y\right)\mathrm{d}x, \end{aligned} \tag{3.4}$$

and where ρ_δ is the unique non-negative solution of the so-called Lundberg equation:

$$cs - \lambda\big(1 - \mathscr{L}f(s)\big) = \delta. \tag{3.5}$$

Remark 3.1.3. Since $\mathscr{L}f(s) \in [0,1]$, it is easy to see that $\rho_\delta \in \left[\frac{\delta}{c}, \frac{\delta+\lambda}{c}\right]$. Moreover, we know that $\rho_\delta = 0$ when $\delta = 0$.

We need to ensure that ϕ, g and h belong to $\mathrm{L}^2(\mathbb{R}_+)$ in order to use a projection estimator. We see that $\sup_x g(x) \le \sup_x \frac{\lambda}{c}\mathbb{P}[X > x] \le \frac{\lambda}{c}$ and $\int_0^\infty g(x)\,\mathrm{d}x \le \frac{\lambda}{c}\int_0^\infty \mathbb{P}[X > x]\,\mathrm{d}x = \theta$, hence $g \in \mathrm{L}^1(\mathbb{R}_+) \cap \mathrm{L}^\infty(\mathbb{R}_+)$, therefore $g \in \mathrm{L}^2(\mathbb{R}_+)$. To ensure that $h \in \mathrm{L}^2(\mathbb{R}_+)$ we make the following assumption.

Assumption 3.2. We assume that $\int_0^\infty (1+x)\left(\int_x^\infty w(x,y-x)f(y)\,\mathrm{d}y\right)\mathrm{d}x$ is finite.

Under this assumption, we have:

$$\begin{aligned} \sup_{u>0} h(u) &\le \frac{\lambda}{c}\int_0^\infty \left(\int_x^\infty w(x,y-x)f(y)\,\mathrm{d}y\right)\mathrm{d}x < +\infty, \\ \int_0^\infty h(u)\,\mathrm{d}u &\le \frac{\lambda}{c}\int_0^\infty x\left(\int_x^\infty w(x,y-x)f(y)\,\mathrm{d}y\right)\mathrm{d}x < +\infty. \end{aligned}$$

Hence, h belongs to $\mathrm{L}^1(\mathbb{R}_+) \cap \mathrm{L}^\infty(\mathbb{R}_+)$, so $h \in \mathrm{L}^2(\mathbb{R}_+)$. Integrating Equation (3.3) yields:

$$\|\phi\|_{\mathrm{L}^1} = \int_0^\infty \phi(u)\,\mathrm{d}u = \frac{\int_0^\infty h(u)\,\mathrm{d}u}{1 - \int_0^\infty g(x)\,\mathrm{d}x},$$

which is finite under Assumption 3.1 since $\int_0^\infty g(x)\,\mathrm{d}x \leq \theta < 1$. Since ϕ belongs to $\mathrm{L}^1(\mathbb{R}_+)$ and g belongs to $\mathrm{L}^2(\mathbb{R}_+)$, their convolution product belongs to $\mathrm{L}^2(\mathbb{R}_+)$, hence $\phi = \phi * g + h$ belongs to $\mathrm{L}^2(\mathbb{R}_+)$ as well.

Remark 3.1.4. Assumption 3.2 has already been considered by Shimizu & Zhang (2017), and Zhang & Su (2018). Actually, the quantity:

$$\omega(x) \coloneqq \int_x^{+\infty} w(x, y-x) f(y)\,\mathrm{d}y = \mathbb{E}\big[w(x, X-x)\mathbf{1}_{X \geq x}\big],$$

can be found on several occasion in the study of the Gerber–Shiu function. The assumption that $\int_0^\infty \omega(x)\,\mathrm{d}x$ is finite ensures that $\phi(u)$ is finite for all u (Asmussen & Albrecher, 2010, Chapter X, Section 1). The additional requirement that:

$$\int_0^\infty x\,\omega(x)\,\mathrm{d}x < +\infty,$$

serves to prove that ϕ belongs to $\mathrm{L}^1(\mathbb{R}_+)$, so that its Fourier transform is well defined. As we have seen, it also ensures that ϕ belongs to $\mathrm{L}^2(\mathbb{R}_+)$.

3.2 The Laguerre–Fourier estimator

We use the Laguerre functions $(\psi_k)_{k\in\mathbb{N}}$ as an orthonormal basis of $\mathrm{L}^2(\mathbb{R}_+)$:

$$\forall x \in \mathbb{R}_+, \psi_k(x) \coloneqq \sqrt{2}\,L_k(2x)\,\mathrm{e}^{-x}, \qquad L_k(x) \coloneqq \sum_{j=0}^{k} \binom{k}{j} \frac{(-x)^j}{j!}. \tag{3.6}$$

We choose this basis for several reasons. First, the support of the Laguerre functions is $\mathbb{R}_+$, which is well suited since the functions we want to estimate are defined on $\mathbb{R}_+$. Moreover, exponential functions (and more broadly mixtures of gamma functions, see the proof of Lemma 3.9 in Mabon (2017)) have an exponentially small bias in this basis, which is interesting because when the claim sizes distribution is exponential and w is a polynomial, then g and h will be given by products of polynomials with exponentials. Finally, the Fourier transform of the Laguerre function is known explicitly:

$$\forall \omega \in \mathbb{R}, \quad \mathscr{F}\psi_k(\omega) = (-1)^k \sqrt{2} \frac{(1+\mathrm{i}\omega)^k}{(1-\mathrm{i}\omega)^{k+1}}, \tag{3.7}$$

which is helpful for the computation of the estimated coefficients (3.8).

We denote by $(a_k)_{k\geq 0}$ the Laguerre coefficients of ϕ. If $m \in \mathbb{N}^*$, we denote by ϕ_m the projection of ϕ on the subspace of $\mathrm{L}^2(\mathbb{R}_+)$ spanned by the first m Laguerre functions $\psi_0, \dots, \psi_{m-1}$, that is:

$$\phi_m := \sum_{k=0}^{m-1} a_k \psi_k, \text{ with } a_k = \langle \phi, \psi_k \rangle.$$

The Laguerre coefficients of ϕ can be computed using Plancherel theorem:

$$a_k = \langle \phi, \psi_k \rangle = \frac{1}{2\pi} \langle \mathscr{F}\phi, \mathscr{F}\psi_k \rangle.$$

Taking the Fourier transform in equation (3.3), we see that $\mathscr{F}\phi = \frac{\mathscr{F}h}{1-\mathscr{F}g}$. Let $\widehat{g}, \widehat{h} \in \mathrm{L}^2(\mathbb{R}_+)$ be some estimators of g and h (we provide these estimators later in equation (3.14)), we estimate the coefficients of ϕ by:

$$\widehat{a}_k := \frac{1}{2\pi} \left\langle \frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}}, \mathscr{F}\psi_k \right\rangle, \tag{3.8}$$

where $\widehat{\mathscr{F}g} := (\mathscr{F}\widehat{g})\mathbf{1}_{|\mathscr{F}\widehat{g}| \leq \theta_0}$ for some truncation parameter $\theta_0 < 1$. The estimator of ϕ is then:

$$\widehat{\phi}_{m_1} := \sum_{k=0}^{m_1-1} \widehat{a}_k \psi_k,$$

where m_1 is the dimension of the projection space.

Proposition 3.2.1. *Under Assumptions 3.1 and 3.2, if $\theta < \theta_0$, we have:*

$$\begin{aligned} \|\phi - \widehat{\phi}_{m_1}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_{m_1}\|_{\mathrm{L}^2}^2 + \frac{2}{(1-\theta_0)^2} \|h - \widehat{h}\|_{\mathrm{L}^2}^2 \\ + \frac{2\|h\|_{\mathrm{L}^1}^2}{(1-\theta_0)^2(1-\theta)^2} \left(1 + \frac{\|g\|_{\mathrm{L}^1}^2}{(\theta_0-\theta)^2}\right) \|g - \widehat{g}\|_{\mathrm{L}^2}^2. \end{aligned}$$

Remark 3.2.2. We emphasize the fact that this result is proven using only two properties: the function ϕ satisfies the equation (3.3) and $\theta_0 > \theta > \|g\|_{\mathrm{L}^1}$. Hence, it can be applied to other problems where the target function satisfies an equation of the form (3.3). For example, it is the case in Zhang & Su (2019), Su *et al.* (2019) and Su *et al.* (2020).

We now need to provide good estimators of g and h. We choose to estimate them by projection on the Laguerre basis too. Let $(b_k)_{k\geq 0}$ and $(c_k)_{k\geq 0}$ be the coefficients of g and h, that is:

$$g = \sum_{k=0}^{+\infty} b_k \psi_k, \quad \text{with} \quad b_k := \langle g, \psi_k \rangle, \tag{3.9}$$

$$h = \sum_{k=0}^{+\infty} c_k \psi_k, \quad \text{with} \quad c_k := \langle h, \psi_k \rangle. \tag{3.10}$$

By Fubini's theorem and using equation (3.4):

$$\begin{aligned} b_k &= \int_0^{+\infty} g(x)\,\psi_k(x)\,\mathrm{d}x \\ &= \frac{\lambda}{c}\int_0^{+\infty}\left(\int_x^{+\infty} \mathrm{e}^{-\rho_\delta(y-x)} f(y)\,\mathrm{d}y\right)\psi_k(x)\,\mathrm{d}x \\ &= \frac{\lambda}{c}\int_0^{+\infty}\left(\int_0^{y} \mathrm{e}^{-\rho_\delta(y-x)}\psi_k(x)\,\mathrm{d}x\right) f(y)\,\mathrm{d}y \\ &= \frac{\lambda}{c}\,\mathbb{E}\left[\int_0^{X} \mathrm{e}^{-\rho_\delta(X-x)}\psi_k(x)\,\mathrm{d}x\right]. \end{aligned}$$

The same calculation for c_k yields:

$$c_k = \frac{\lambda}{c}\,\mathbb{E}\left[\int_0^{X}\left(\int_u^{X} \mathrm{e}^{-\rho_\delta(x-u)}\, w(x, X-x)\,\mathrm{d}x\right)\psi_k(u)\,\mathrm{d}u\right].$$

We estimate these coefficients by empirical means. However, we first need to estimate ρ_δ. Since ρ_δ is the non-negative solution of the Lundberg equation (3.5), we estimate it by $\widehat{\rho}_\delta$ the non-negative solution of the empirical Lundberg equation:

$$cs - \widehat{\lambda}(1 - \widehat{\mathscr{L}f}(s)) = \delta, \tag{3.11}$$

where $\widehat{\lambda} := \frac{N_T}{T}$ and $\widehat{\mathscr{L}f}(s) := \frac{1}{N_T}\sum_{i=1}^{N_T} \mathrm{e}^{-sX_i}$. When $\delta = 0$ we know that $\rho_\delta = 0$ so we do not need to estimate it, thus we set $\widehat{\rho}_0 = 0$. The estimated coefficients of g and h are:

$$\widehat{b}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{X_i} \mathrm{e}^{-\widehat{\rho}_\delta(X_i-x)}\psi_k(x)\,\mathrm{d}x, \tag{3.12}$$

$$\widehat{c}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{X_i}\left(\int_u^{X_i} \mathrm{e}^{-\widehat{\rho}_\delta(x-u)}\, w(x, X_i-x)\,\mathrm{d}x\right)\psi_k(u)\,\mathrm{d}u, \tag{3.13}$$

and the estimators of g and h are:

$$\widehat{g}_{m_2} := \sum_{k=0}^{m_2-1}\widehat{b}_k\psi_k, \quad \widehat{h}_{m_3} := \sum_{k=0}^{m_3-1}\widehat{c}_k\psi_k, \tag{3.14}$$

where m_2 and m_3 are the dimensions of the projection spaces. We denote by g_{m_2} and h_{m_3} the projections of g and h on the subspaces $\mathrm{Span}(\psi_0,\dots,\psi_{m_2-1})$ and $\mathrm{Span}(\psi_0,\dots,\psi_{m_3-1})$.

Remark 3.2.3. The dimensions m_1, m_2, m_3 do not have to be the same for the estimation of ϕ, g and h. In practice, we will choose different dimensions.

In order to give a bound on the mean integrated squared error of our estimators $\widehat{g}_{m_2}$ and $\widehat{h}_{m_3}$, we need to make an additional assumption.

Assumption 3.3. Let $W(X) := \int_0^X \left(\int_u^X w(x, X-x)\,\mathrm{d}x\right)^2 \mathrm{d}u$. If $\delta = 0$, we assume that $\mathbb{E}[W(X)]$ is finite, and if $\delta > 0$, we assume that $\mathbb{E}[W(X)^2]$ is finite.

Remark 3.2.4 (Applicability of Assumptions 3.2 and 3.3)**.** Assumptions 3.2 and 3.3 can be thought as moment conditions on the claim sizes distribution, with respect to w. In the special case where w is given by $w(x,y) = x^k(x+y)^\ell$ for $k,\ell \geq 0$, we have:

$$\int_0^{+\infty}(1+x)\left(\int_x^{+\infty} w(x,y-x)f(y)\,\mathrm{d}y\right)\mathrm{d}x = \mathbb{E}\left[\frac{X^{k+\ell+1}}{k+1}\right] + \mathbb{E}\left[\frac{X^{k+\ell+2}}{k+2}\right],$$

$$W(X) = \frac{X^{2k+2\ell+3}}{(k+1)^2(2k+3)},$$

so Assumptions 3.2 and 3.3 reduce to the moment condition $\mathbb{E}\left[X^{2k+2\ell+3}\right] < +\infty$ (if $\delta = 0$). Notice that the functions of Example 3.1.1 correspond to the cases $(k,\ell) = (0,0)$ or $(0,1)$, so that the corresponding moment condition is $\mathbb{E}[X^3] < \infty$ or $\mathbb{E}[X^5] < \infty$. Hence, heavy-tailed distributions can fit into these assumptions, provided they admit sufficiently large moments. On the other hand, if w grows with an exponential rate, for example if $w(x,y-x) \coloneqq \exp(\gamma(x+y))$, then we also need an exponential moment for X, so that we are restricted to light tailed distributions.

Theorem 3.2.5. *Under Assumptions 3.1, 3.2 and 3.3, if $\delta = 0$ then it holds:*

$$\mathbb{E}\|g - \widehat{g}_{m_2}\|_{\mathrm{L}^2}^2 \leq \|g - g_{m_2}\|_{\mathrm{L}^2}^2 + \frac{\lambda}{c^2 T}\mathbb{E}[X],$$
$$\mathbb{E}\|h - \widehat{h}_{m_3}\|_{\mathrm{L}^2}^2 \leq \|h - h_{m_3}\|_{\mathrm{L}^2}^2 + \frac{\lambda}{c^2 T}\mathbb{E}[W(X)],$$

and if $\delta > 0$ then it holds:

$$\mathbb{E}\|g - \widehat{g}_{m_2}\|_{\mathrm{L}^2}^2 \leq \|g - g_{m_2}\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2 T}\left(\mathbb{E}[X] + \frac{\mathbb{E}\left[X^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),$$
$$\mathbb{E}\|h - \widehat{h}_{m_3}\|_{\mathrm{L}^2}^2 \leq \|h - h_{m_3}\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2 T}\left(\mathbb{E}[W(X)] + \frac{\mathbb{E}\left[W(X)^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),$$

where $C(\lambda)$ is a $\mathrm{O}(\lambda^2)$.

Remark 3.2.6. The variance terms do not depend on m_2 nor m_3, so no compromise between the bias and the variance is needed: we just have to take m_2 and m_3 as large as possible such that the bias is smaller than $1/T$. See Section 3.4 for a discussion concerning the choice of m_2 and m_3 when the functions g and h belong to a Sobolev–Laguerre space.

Let $m_1, m_2, m_3 \in \mathbb{N}^*$, we estimate g by $\widehat{g}_{m_2}$ and h by $\widehat{h}_{m_3}$. We plug these estimators in (3.8) and we estimate ϕ by:

$$\widehat{\phi}_{m_1,m_2,m_3} \coloneqq \sum_{k=0}^{m_1-1}\left\langle \frac{\mathscr{F}\widehat{h}_{m_3}}{1-\widehat{\mathscr{F}g_{m_2}}}, \mathscr{F}\psi_k\right\rangle \psi_k,$$

with $\widehat{\mathscr{F}g_{m_2}} := \mathscr{F}\widehat{g}_{m_2}\mathbf{1}_{|\mathscr{F}\widehat{g}_{m_2}|\leq\theta_0}$. Combining Proposition 3.2.1 with Theorem 3.2.5, we obtain:

Corollary 3.2.7. *Under Assumptions 3.1, 3.2 and 3.3, if $\theta < \theta_0$ then it holds:*

$$\mathbb{E}\|\phi - \widehat{\phi}_{m_1,m_2,m_3}\|^2_{\mathrm{L}^2} \leq \|\phi - \phi_{m_1}\|^2_{\mathrm{L}^2} + \frac{C}{(1-\theta_0)^2}\left(\|g - g_{m_2}\|^2_{\mathrm{L}^2} + \|h - h_{m_3}\|^2_{\mathrm{L}^2} + \frac{1}{cT}\right),$$

where C is a constant depending on λ, c, θ, $\|g\|_{\mathrm{L}^1}$, $\|h\|_{\mathrm{L}^1}$, $\mathbb{E}[X]$, $\mathbb{E}[W(X)]$ and $\theta_0 - \theta$; and also δ, $\mathbb{E}[X^2]$, $\mathbb{E}[W(X)^2]$ if $\delta > 0$.

We want to compare our estimator with the Laguerre deconvolution method. However, there is no result on the MISE of this method for estimating the Gerber–Shiu function, so we study it in the next section.

3.3 The Laguerre deconvolution estimator

For the Laguerre deconvolution method, we need an additional assumption on the coefficients of g.

Assumption 3.4. The coefficients $(b_k)_{k\geq 0}$, defined by (3.9), satisfy $(b_{k+1} - b_k)_{k\geq 0} \in \ell^1(\mathbb{N})$.

Remark 3.3.1. If g belongs to a Sobolev–Laguerre space $\mathrm{W}^s(\mathbb{R}_+)$ with regularity $s > 1$, then Assumption 3.4 holds automatically. The spaces $\mathrm{W}^s(\mathbb{R}_+)$ are regularity spaces associated with the Laguerre basis, see Definition 3.4.1 below. Indeed, by the Cauchy–Schwarz inequality, we have:

$$\sum_{k=0}^{+\infty} |b_k| = \sum_{k=0}^{+\infty} |b_k|(1+k)^{\frac{s}{2}}(1+k)^{-\frac{s}{2}} \leq \left(\sum_{k=0}^{+\infty} |b_k|^2(1+k)^s\right)^{\frac{1}{2}} \left(\sum_{k=0}^{+\infty}(1+k)^{-s}\right)^{\frac{1}{2}},$$

which is finite if $g \in \mathrm{W}^s(\mathbb{R}_+)$ and $s > 1$. Hence, $(b_k)_{k\geq 0}$ is in $\ell^1(\mathbb{N})$ and so is $(b_{k+1} - b_k)_{k\geq 0}$.

The reason why the Laguerre basis is well suited for deconvolution on $\mathbb{R}_+$ is the following relation satisfied by the Laguerre functions:

$$\forall k, j \in \mathbb{N}, \quad \psi_k * \psi_j = \frac{1}{\sqrt{2}}(\psi_{k+j} - \psi_{k+j+1}),$$

see formula 22.13.14 in Abramowitz & Stegun (1972). The reader interested in the use of the Laguerre basis for deconvolution problems is referred to Mabon (2017). Expanding the renewal equation (3.3) on the Laguerre basis, one easily obtains the following relation between the coefficients of ϕ, g and h:

$$\forall k \in \mathbb{N}, \quad a_k = (\beta * a)_k + c_k,$$

where the sequence $(\beta_k)_{k\geq 0}$ is defined by $\beta_0 := \frac{b_0}{\sqrt{2}}$ and $\beta_k := \frac{b_k - b_{k-1}}{\sqrt{2}}$ for $k \geq 1$. This relation can be written in a matrix form: if $\mathbf{a}_m := (a_0, \dots, a_{m-1})^T$ and $\mathbf{c}_m := (c_0, \dots, c_{m-1})^T$ are the vectors of the m first coefficients of ϕ and h, then it holds:

$$\mathbf{A}_m \times \mathbf{a}_m = \mathbf{c}_m \iff \mathbf{a}_m = \mathbf{A}_m^{-1} \times \mathbf{c}_m, \tag{3.15}$$

where $\mathbf{A}_m$ is the lower triangular Toeplitz matrix:

$$\forall i, j \in \{0, \dots, m-1\}, \quad (\mathbf{A}_m)_{i,j} := \begin{cases} 1 - \frac{1}{\sqrt{2}} b_0 & \text{if } i = j, \\ \frac{1}{\sqrt{2}} (b_{i-j-1} - b_{i-j}) & \text{if } i > j, \\ 0 & \text{else.} \end{cases} \tag{3.16}$$

This matrix is invertible if and only if $1 - \frac{b_0}{\sqrt{2}} \neq 0$, which is the case because $\frac{b_0}{\sqrt{2}} \leq \theta < 1$ under Assumption 3.1.

Lemma 3.3.2. *Under Assumption 3.4, we have* $\|\mathbf{A}_m^{-1}\|_{\mathrm{op}} \leq \frac{2}{1-\theta}$ *for all* $m \in \mathbb{N}^*$.

This lemma is borrowed from Zhang & Su (2018) (Lemma 4.3 in their article). There were missing elements in their proof, so we give a new proof of this lemma, for the sake of completeness.

The naive Laguerre deconvolution estimator consists in estimating the matrix $\mathbf{A}_m$ and the coefficients $\mathbf{c}_m$ in (3.15), to obtain an estimation of the coefficients of ϕ. More precisely, the matrix $\mathbf{A}_m$ is estimated by plugging $\widehat{b}_k$, defined by (3.12), in (3.16):

$$\forall i, j \in \{0, \dots, m-1\}, \quad (\widehat{\mathbf{A}}_m)_{i,j} := \begin{cases} 1 - \frac{1}{\sqrt{2}} \widehat{b}_0 & \text{if } i = j, \\ \frac{1}{\sqrt{2}} (\widehat{b}_{i-j-1} - \widehat{b}_{i-j}) & \text{if } i > j, \\ 0 & \text{else.} \end{cases} \tag{3.17}$$

This matrix is invertible if and only if $\frac{\widehat{b}_0}{\sqrt{2}} \neq 1$, which is almost surely the case since $\frac{\widehat{b}_0}{\sqrt{2}} = \frac{1}{cT} \sum_{i=1}^{N_T} (1 - \mathrm{e}^{-X_i})$ is a continuous random variable. The coefficients of ϕ are estimated by:

$$\widehat{\mathbf{a}}_m^{\mathrm{Lag}_0} := \widehat{\mathbf{A}}_m^{-1} \times \widehat{\mathbf{c}}_m, \tag{3.18}$$

where $\widehat{\mathbf{c}}_m := (\widehat{c}_0, \dots, \widehat{c}_{m-1})^T$. Under Assumptions 3.1, 3.2, 3.3, and 3.4, Zhang & Su (2018) show that if $\mathbb{E}[X^2]$ is finite and if $m = o(T)$, then $\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_0}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \mathrm{O_P}\left(\frac{m}{T}\right)$.

In the following, we propose two ways inspired by Comte & Mabon (2017) to estimate the Gerber–Shiu function, using the Laguerre deconvolution method. To obtain a non asymptotic result on the MISE of the estimator, a cutoff is required when inverting the matrix $\widehat{\mathbf{A}}_m$. Let $\theta_0 < 1$ be a truncation parameter, we estimate $\mathbf{A}_m^{-1}$ by:

$$\widetilde{\mathbf{A}}_{m,1}^{-1} := \widehat{\mathbf{A}}_m^{-1} \mathbf{1}_{\Delta_m^1} \quad \text{where} \quad \Delta_m^1 := \left\{ \|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} \leq \frac{2}{1-\theta_0} \right\},$$

and we estimate the coefficients $\mathbf{a}_m$ by $\widehat{\mathbf{a}}_m^{\mathrm{Lag}_1} := \widetilde{\mathbf{A}}_{m,1}^{-1} \times \widehat{\mathbf{c}}_m$.

Theorem 3.3.3. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, if $\theta < \theta_0$ then it holds:*

$$\forall m \in \mathbb{N}^*, \quad \mathbb{E}\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + C\frac{m}{T},$$

where C is a constant depending on λ, c, θ, $\mathbb{E}[X]$, $\mathbb{E}[W(X)]$ and $\theta_0 - \theta$; and also δ, $\mathbb{E}[X^2]$, $\mathbb{E}[W(X)^2]$ if $\delta > 0$.

We propose a second way to estimate ϕ using the Laguerre deconvolution method, in the case $\delta = 0$. It avoids the use of a truncation parameter θ_0, but at the expense of an extra $\log(m)$ factor in the upper bound of the MISE, and it uses an additional independence assumption. We estimate the Laguerre coefficients of g by (3.12), that is in this case:

$$\widehat{b}_k := \frac{1}{cT}\sum_{i=1}^{N_T} \Psi_k(X_i),$$

where $\Psi_k(x) := \int_0^x \psi_k(t)\,\mathrm{d}t$. The matrix $\mathbf{A}_m$ is still estimated by (3.17).

Proposition 3.3.4. *If $\delta = 0$, $p \geq 1$ and $\log m \geq p$, then it holds:*

$$\mathbb{E}\left[\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}^{2p}\right] \leq C(p,\lambda)\mu^p\left(\frac{m\log m}{cT}\right)^p + C(p)\left(\frac{m\log m}{cT}\right)^{2p},$$

where $C(p,\lambda)$ is a $\mathrm{O}(\lambda^p)$, and $C(p)$ is a constant depending on p.

This time, we estimate the inverse of the matrix $\mathbf{A}_m$ by:

$$\widetilde{\mathbf{A}}_{m,2}^{-1} := \widehat{\mathbf{A}}_m^{-1}\mathbf{1}_{\Delta_m^2}, \quad \text{where} \quad \Delta_m^2 := \left\{\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}}^2 \leq \frac{cT}{m\log m}\right\},$$

we estimate the coefficients of ϕ by $\widehat{\mathbf{a}}_m^{\mathrm{Lag}_2} := \widetilde{\mathbf{A}}_{m,2}^{-1} \times \widehat{\mathbf{c}}_m$, and we estimate ϕ by:

$$\widehat{\phi}_m^{\mathrm{Lag}_2} := \sum_{k=0}^{m-1} \widehat{a}_k^{\mathrm{Lag}_2}\psi_k.$$

To provide an upper bound on the MISE of $\widehat{\phi}_m^{\mathrm{Lag}_2}$, we need $\widetilde{\mathbf{A}}_{m,2}^{-1}$ and $\widehat{\mathbf{c}}_m$ to be independent. For this reason, we assume that we have a second observation set $\{N_T'; X_1', \ldots, X_{N_T'}'\}$ identical in law but independent from the main one[1]. We use this second set to estimate $\widetilde{\mathbf{A}}_{m,2}^{-1}$.

[1] alternatively, we could split the data $\{X_1, \ldots, X_{N_T}\}$ in two parts: we use half of the data to estimate $\widetilde{\mathbf{A}}_{m,2}^{-1}$, and the other half to estimate $\widehat{\mathbf{c}}_m$.

Theorem 3.3.5. *We assume that $\delta = 0$. Under Assumptions 3.1, 3.2, 3.3 and 3.4, if $m \log m \leq cT$ then it holds:*

$$\mathbb{E}\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2 \leq \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{cT(1-\theta)^2}\left(\frac{\mathbb{E}[W(X)]}{c} + \|\phi\|_{\mathrm{L}^2}^2(\mu+\mu^2)m\log m\right) + \mathrm{O}\left(\frac{1}{T^2}\right),$$

with $C(\lambda) = \mathrm{O}(\lambda \vee \lambda^2)$.

Remark 3.3.6. Contrary to the Laguerre–Fourier method, there is only one bias term with the Laguerre deconvolution method. However, the variance term is more complicated and a bias-variance compromise is needed. It leads to non-parametric rates of convergence, which are slower than the parametric rate $\frac{1}{T}$.

3.4 Convergence rates of the Laguerre estimators

3.4.1 Sobolev–Laguerre spaces

To study the bias of a function in the Laguerre basis, we consider the Sobolev–Laguerre spaces. These functional spaces have been introduced by Bongioanni & Torrea (2009) to study the Laguerre operator. The connection with the Laguerre coefficients was established later by Comte & Genon-Catalot (2015).

Definition 3.4.1. For $s > 0$, we define the Sobolev–Laguerre ball of radius $L > 0$ and regularity s as:

$$\mathrm{W}^s(\mathbb{R}_+, L) := \left\{ v \in \mathrm{L}^2(\mathbb{R}_+) \;\middle|\; \sum_{k=0}^{+\infty} \langle v, \psi_k \rangle^2 k^s \leq L \right\},$$

and we define the Sobolev–Laguerre space as $\mathrm{W}^s(\mathbb{R}_+) := \bigcup_{L>0} \mathrm{W}^s(\mathbb{R}_+, L)$.

By Proposition 7.2 in Comte & Genon-Catalot (2015), when s is a natural number, $v \in \mathrm{W}^s(\mathbb{R}_+)$ if and only if v is $(s-1)$ times differentiable, $v^{(s-1)}$ is absolutely continuous, and for all $0 \leq k \leq s-1$ we have $x^{\frac{k+1}{2}} \sum_{j=0}^{k+1} \binom{k+1}{j} v^{(j)} \in \mathrm{L}^2(\mathbb{R}_+)$. In particular, a function v belongs to $\mathrm{W}^1(\mathbb{R}_+)$ if and only if it is absolutely continuous and $\sqrt{x}(v + v') \in \mathrm{L}^2(\mathbb{R}_+)$.

We are interested in the Sobolev–Laguerre spaces because of the following observation. If v belongs to a Sobolev–Laguerre ball $\mathrm{W}^s(\mathbb{R}_+, L)$, then its bias is controlled by:

$$\|v - v_m\|_{\mathrm{L}^2}^2 = \sum_{k=m}^{+\infty} \langle v, \psi_k \rangle^2 = \sum_{k=m}^{+\infty} \langle v, \psi_k \rangle^2 k^s k^{-s} \leq L m^{-s}.$$

Combining this upper bound on the bias term with Corollary 3.2.7, and Theorems 3.3.3 and 3.3.5, we obtain convergence rates for the Laguerre–Fourier estimator and the Laguerre deconvolution estimators, on Sobolev–Laguerre spaces.

Theorem 3.4.2. *Under Assumptions 3.1, 3.2 and 3.3, if $\theta < \theta_0$ and if $\phi \in \mathrm{W}^{s_1}(\mathbb{R}_+)$, $g \in \mathrm{W}^{s_2}(\mathbb{R}_+)$ and $h \in \mathrm{W}^{s_3}(\mathbb{R}_+)$, then choosing $m_i > (cT)^{\frac{1}{s_i}}$ for all $i \in \{1,2,3\}$ yields:*

$$\mathbb{E}\|\phi - \widehat{\phi}_{m_1,m_2,m_3}\|^2_{\mathrm{L}^2} = \mathrm{O}\left(\frac{1}{cT}\right).$$

Remark 3.4.3. If ϕ, g and h belong to some Sobolev–Laguerre spaces with a regularity index greater than 1, we can just choose $m_1 = m_2 = m_3 = \lceil cT \rceil$ and obtain the parametric rate $\mathrm{O}(\frac{1}{cT})$ for the Laguerre–Fourier estimator.

Theorem 3.4.4. *We make Assumptions 3.1, 3.2, 3.3 and 3.4, and we assume that $\phi \in \mathrm{W}^s(\mathbb{R}_+)$.*

1. *If $\theta < \theta_0$, then choosing $m_{\mathrm{opt}} \propto (cT)^{\frac{1}{1+s}}$ yields:*

$$\mathbb{E}\|\phi - \widehat{\phi}^{\mathrm{Lag}_1}_{m_{\mathrm{opt}}}\|^2_{\mathrm{L}^2} = \mathrm{O}\left((cT)^{-\frac{s}{1+s}}\right).$$

2. *If $\delta = 0$, then choosing $m_{\mathrm{opt}} \propto (cT)^{\frac{1}{1+s}}$ yields:*

$$\mathbb{E}\|\phi - \widehat{\phi}^{\mathrm{Lag}_2}_{m_{\mathrm{opt}}}\|^2_{\mathrm{L}^2} = \mathrm{O}\left((cT)^{-\frac{s}{1+s}} \log(cT)\right).$$

Remark 3.4.5. The Fourier–Laguerre estimator and the Laguerre deconvolution estimator $\widehat{\phi}^{\mathrm{Lag}_1}_m$ both depend on a truncation parameter θ_0 that needs to be chosen such that $\theta < \theta_0$. We see two ways to ensure that.

1. We can assume that we know some $\theta_0 < 1$ such that $\theta < \theta_0$. Then our convergence rates are those of Theorems 3.4.2 and 3.4.4.
2. We can choose $\theta_0 = 1 - (\log T)^{1/2}$. Then for T large enough (more precisely $T > \mathrm{e}^{(1-\theta)^2}$), the convergence rates of the Laguerre–Fourier estimator and $\widehat{\phi}^{\mathrm{Lag}_1}_m$ are those of Theorems 3.4.2 and 3.4.4 multiplied by $\log(T)$.

In our simulations, we chose the first way.

3.4.2 The exponential case

In this section, we want to compute the convergence rate of the estimators, in the exponential case: $X \sim \mathrm{Exp}(1/\mu)$. This distribution is often considered in risk theory and closed forms of the Gerber–Shiu function are available in this case. Indeed, the Gerber–Shiu functions of Example 3.1.1 are given by:

$$\phi(u) = \begin{cases} \theta \exp\left(-\dfrac{1-\theta}{\mu} u\right) & \text{(ruin probability)}, \\ \dfrac{\theta}{1+\mu\rho_\delta} \exp\left(-\left[\dfrac{1-\theta}{\mu} + \rho_\delta - \dfrac{\delta}{c}\right] u\right) & \text{(Laplace transform of } \tau\text{)}, \\ \mu(1+2\theta)\exp\left(-\dfrac{1-\theta}{\mu} u\right) - \mu \exp\left(-\dfrac{u}{\mu}\right) & \text{(jump size causing the ruin)}. \end{cases} \tag{3.19}$$

These formulas are obtained by Laplace inversion, see Asmussen & Albrecher (2010), chapter XII. We use the following lemma to compute the bias terms of the functions ϕ, g and h.

Lemma 3.4.6. *Let C, γ be positive numbers and let $F(x) = C\exp(-\gamma x)\mathbf{1}_{\mathbb{R}_+}(x)$. The Laguerre coefficients of F are given by:*

$$\langle F, \psi_k \rangle = \frac{C\sqrt{2}}{\gamma+1}\left(\frac{\gamma-1}{\gamma+1}\right)^k.$$

Hence if $m \geq 0$ we have:

$$\sum_{k=m}^{+\infty} \langle F, \psi_k \rangle^2 = \frac{C^2}{2\gamma}\left(\frac{\gamma-1}{\gamma+1}\right)^{2m}.$$

Proposition 3.4.7. *If the density of X is $f(x) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}}$, then the bias term of ϕ is given by:*

$$\|\phi - \phi_m\|^2_{\mathrm{L}^2} \leq L\mathrm{e}^{-rm}$$

with L and r given by:

1. *Ruin probability: $L = \frac{\theta^2}{2\gamma}$, $r = 2\log\left|\frac{\gamma+1}{\gamma-1}\right|$ and $\gamma = \frac{1-\theta}{\mu}$.*
2. *Laplace transform of the ruin time: $L = \frac{\theta^2}{2\gamma(1+\mu\rho_\delta)^2}$, $r = 2\log\left|\frac{\gamma+1}{\gamma-1}\right|$ and $\gamma = \frac{1-\theta}{\mu} + \rho_\delta - \frac{\delta}{c}$.*
3. *Jump size causing the ruin: $L = \mu^3\frac{(1+2\theta)^2}{1-\theta}$, $r = 2\log\left(\left|\frac{1-\theta+\mu}{1-\theta-\mu}\right| \wedge \left|\frac{1+\mu}{1-\mu}\right|\right)$.*

Combining this Proposition with Theorems 3.3.3 and 3.3.5, we easily obtain convergence rates for the Gerber–Shiu functions we are interested in.

Theorem 3.4.8. *We assume that the density of X is $f(x) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}}$, we make Assumptions 3.1, 3.2, 3.3 and 3.4, and we assume that the bias term of ϕ decreases as:*

$$\|\phi - \phi_m\|^2_{\mathrm{L}^2} \leq L\mathrm{e}^{-rm}.$$

1. *If $\theta < \theta_0$, then choosing $m_{\mathrm{opt}} = \lceil \frac{1}{r}\log(cT)\rceil$ yields:*

$$\mathbb{E}\|\phi - \widehat{\phi}^{\mathrm{Lag}_1}_{m_{\mathrm{opt}}}\|^2 = \mathrm{O}\left(\frac{\log(cT)}{cT}\right).$$

2. *If $\delta = 0$, then choosing $m_{\mathrm{opt}} = \lceil \frac{1}{r}\log(cT)\rceil$ yields:*

$$\mathbb{E}\|\phi - \widehat{\phi}^{\mathrm{Lag}_2}_{m_{\mathrm{opt}}}\|^2_{\mathrm{L}^2} = \mathrm{O}\left(\frac{\log(cT)\log\log(cT)}{cT}\right).$$

For the Laguerre–Fourier estimator, we also need to know the decreasing rate of the bias term of g and h. For the ruin probability, the Laplace transform of τ, and the jump size causing the ruin, direct calculations show that g and h are given by a positive multiple of $\mathrm{e}^{-x/\mu}$. Thus, Lemma 3.4.6 yields that their bias term is less than $\exp(-r'm)$, with $r' := 2\log|\frac{1+\mu}{1-\mu}|$. Together with Corollary 3.2.7, we obtain the convergence rates of the Laguerre–Fourier estimator.

Theorem 3.4.9. *If the density of X is $f(x) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}}$, under Assumptions 3.1, 3.2 and 3.3, if $\theta < \theta_0$ and if the bias term of ϕ decreases as:*

$$\|\phi - \phi_m\|^2_{\mathrm{L}^2} \leq L\mathrm{e}^{-rm},$$

then choosing $m_1 > \lceil \frac{1}{r}\log(cT)\rceil$ and $m_2, m_3 > \lceil \frac{1}{r'}\log(cT)\rceil$ with $r' := 2\log|\frac{1+\mu}{1-\mu}|$ yields:

$$\mathbb{E}\|\phi - \widehat{\phi}_{m_1,m_2,m_3}\|^2_{\mathrm{L}^2} = \mathrm{O}\left(\frac{1}{cT}\right).$$

3.5 Numerical illustrations

3.5.1 Risk comparison

In this subsection, we compare the performance of the Laguerre–Fourier estimator (see Section 3.2) and the Laguerre deconvolution estimators (see Section 3.3) on simulated data. We consider the three Gerber–Shiu functions of Example 3.1.1:

1. $\phi^{(1)}(u) = \mathbb{P}[\tau(u) < \infty]$ the ruin probability;
2. $\phi^{(2)}(u) = \mathbb{E}[(U_{\tau(u)-} + |U_{\tau(u)}|)\mathbf{1}_{\tau(u)<\infty}]$ the expected claim size causing the ruin;
3. $\phi^{(3)}(u) = \mathbb{E}[\mathrm{e}^{-\delta\tau(u)}\mathbf{1}_{\tau(u)<\infty}]$ the Laplace transform of the ruin time, for $\delta = 0.1$.

We also consider three sets of parameters:

1. X follows an exponential distribution, $\lambda = 1$, $\mu = 1$, $c = 1.5$. In this setting, $\theta \approx 0.67$.
2. X follows an exponential distribution, $\lambda = 1.25$, $\mu = 2$, $c = 3$. In this setting, $\theta \approx 0.83$
3. X follows a $\Gamma(2, \frac{\mu}{2})$ distribution, $\lambda = 1.25$, $\mu = 2$, $c = 3$. In this setting, $\theta \approx 0.83$.

Using Laplace inversion techniques (Asmussen & Albrecher, 2010, Chapter XII), we have access to explicit formulas for these Gerber–Shiu functions. In all cases, they are given by a sum of products of polynomials and exponentials, hence they belong to $\mathrm{W}^s(\mathbb{R}_+)$ for all $s > 0$.

Computation of the estimators The Laguerre functions and their primitives are computed using recursive relations, see Appendix A. The expression of $\widehat{b}_k$ and $\widehat{c}_k$ depends on the value of δ and the form of w:

1. *Ruin probability.* The estimators of the coefficients b_k and c_k are in this case:
$$\widehat{b}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\Psi_k(X_i), \qquad \widehat{c}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{X_i}\Psi_k(x)\,\mathrm{d}x.$$
We compute the integrals in $\widehat{c}_k$ using Romberg's method with $2^{10}+1$ points.

2. *Expected claim size causing the ruin.* The estimators of the coefficients b_k and c_k are in this case:
$$\widehat{b}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\Psi_k(X_i), \qquad \widehat{c}_k = \frac{1}{cT}\sum_{i=1}^{N_T}X_i\int_0^{X_i}\Psi_k(x)\,\mathrm{d}x.$$
We compute the integrals in $\widehat{c}_k$ using Romberg's method with $2^{10}+1$ points.

3. *Laplace transform.* The estimators of the coefficients b_k and c_k are in this case:
$$\widehat{b}_k = \frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{X_i}\mathrm{e}^{-\widehat{\rho}_\delta(X_i-x)}\psi_k(x)\,\mathrm{d}x, \qquad \widehat{c}_k = \frac{1}{\widehat{\rho}_\delta}\left(\frac{1}{cT}\sum_{i=1}^{N_T}\Psi_k(X_i)-\widehat{b}_k\right),$$
where we used integration by parts to obtain this expression of $\widehat{c}_k$. We compute the integrals in $\widehat{b}_k$ using Romberg's method with $2^{10}+1$ points. We compute $\widehat{\rho}_\delta$, the solution of Equation (3.11), with Newton's method using the initial condition $\frac{\delta+\widehat{\lambda}/2}{c}$.

For the Laguerre–Fourier estimator, once $(\widehat{b}_k)_{0\le k<m_2}$ and $(\widehat{c}_k)_{0\le k<m_3}$ have been computed, we can compute $\widehat{a}_k$ defined by (3.8):

$$\widehat{a}_k = \frac{1}{2\pi}\int_{\mathbb{R}}\frac{\mathscr{F}\widehat{h}(\omega)}{1-\widehat{\mathscr{F}g}(\omega)}\overline{\mathscr{F}\psi_k(\omega)}\,\mathrm{d}\omega,$$
$$\widehat{\mathscr{F}g}(\omega) = \begin{cases}\mathscr{F}\widehat{g}(\omega) & \text{if } |\mathscr{F}\widehat{g}(\omega)|\le\theta_0\\ 0 & \text{if } |\mathscr{F}\widehat{g}(\omega)|>\theta_0\end{cases}, \quad \mathscr{F}\widehat{g} = \sum_{k=0}^{m_2-1}\widehat{b}_k\mathscr{F}\psi_k, \quad \mathscr{F}\widehat{h} = \sum_{k=0}^{m_3-1}\widehat{c}_k\mathscr{F}\psi_k,$$

where $\mathscr{F}\psi_k$ is given by (3.7), and where the integral in $\widehat{a}_k$ is computed with Romberg's method on a discretization of $[-10^3,10^3]$ with $2^{15}+1$ points.

For the Laguerre deconvolution estimators, once we have computed the coefficients $(\widehat{b}_k)_{0\le k<m}$ and $(\widehat{c}_k)_{0\le k<m}$, we can compute the matrix $\widehat{\mathbf{A}}_m$ defined by (3.17) and then compute the coefficients $\widehat{a}_m^{\mathrm{Lag}_i}$ as described in Section 3.3.

Remark 3.5.1. While the Gerber–Shiu function is always positive, this is not necessarily the case of the estimators. However, we can always take their positive part, since it does not increase their risk:

$$\mathbb{E}\|\phi - (\widehat{\phi})_+\|^2 \leq \mathbb{E}\|\phi - \widehat{\phi}\|^2.$$

In Figures 3.1 and 3.2, we observe that the estimators stay positive where ϕ is positive, and that they can take small negative values when ϕ becomes small (as u tends to $+\infty$). Hence, it is reasonable to use the estimators without taking their positive part. We choose to do so, in the simulations.

Model selection Each estimator we consider depends on one or several parameters that need to be chosen. The Laguerre–Fourier estimator and the Laguerre deconvolution estimator depend on a truncation parameter θ_0, which needs to be chosen such that $\theta < \theta_0$. We choose $\theta_0 = 0.95$ in our simulations.

- The Laguerre–Fourier estimator depends on four parameters: m_1, m_2 and m_3, the dimensions of the projection spaces for the functions ϕ, g and h, and θ_0 the truncation parameter in the estimation of $\widehat{\mathscr{F}g}$. As said in Remark 3.4.3, we can choose $m_1 = m_2 = m_3 = \lceil cT \rceil$, no selection procedure is required. Still, we propose a *model reduction procedure* for the choice of m_2 and m_3, that we describe in Subsection 3.5.2.

- The naive Laguerre deconvolution estimator $\widehat{\phi}_m^{\mathrm{Lag}_0}$, defined by (3.18), depends on one parameter: m, the dimension of the projection space for ϕ. However, there is no model selection procedure for m. In their numerical section, Zhang & Su (2018) only consider (as we do) Gerber–Shiu functions with exponential decay; hence the bias term also decays with exponential rate. Using this fact, they chose $m = \lfloor 5T^{1/10} \rfloor$. We make the same choice in our simulations and we write $\widehat{\phi}^{\mathrm{ZS}}$ this estimator.

- The Laguerre deconvolution estimators $\widehat{\phi}_m^{\mathrm{Lag}_1}$ and $\widehat{\phi}_m^{\mathrm{Lag}_2}$ also depend on m. For $i \in \{1,2\}$, we choose $\widehat{m}^{\mathrm{Lag}_i}$ as the minimizer of a penalized criterion:

$$\widehat{m}^{\mathrm{Lag}_i} \in \underset{m \in \mathscr{M}_i}{\arg\min} \left\{ -\|\widehat{\phi}_m^{\mathrm{Lag}_i}\|_{\mathrm{L}^2}^2 + \kappa_i \,\mathrm{pen}_i(m) \right\} \tag{3.20}$$

  where the model collections are:

$$\mathscr{M}_1 := \left\{ 1 \leq m \leq M \,\middle|\, \|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} \leq \frac{1}{1-\theta_0} \right\}$$
$$\mathscr{M}_2 := \left\{ 1 \leq m \leq M \,\middle|\, \|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}}^2 \leq \frac{cT}{m\log(m)} \right\}$$

  with $M = \lceil cT \rceil \wedge 500$ (we do not compute more than 500 coefficients, because of computation time).

In the following, if $F(X)$ is a function of X, we write $\overline{F(X)} := \frac{1}{N_T}\sum_{i=1}^{N_T} F(X_i)$ its empirical mean from the sample $\{X_1,\dots,X_{N_T}\}$. For the penalty terms, we choose empirical versions of the variance terms in Theorems 3.3.3 and 3.3.5:

$$\mathrm{pen}_1(m) := \frac{1}{(1-\theta_0)^2}\left(\|\widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2 m\widehat{V}_g + \widehat{V}_h\right)$$
$$\mathrm{pen}_2(m) := (\widehat{\lambda}\vee\widehat{\lambda}^2)\left(\frac{\overline{W(X)}}{c} + \|\widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2\left(\overline{X}\vee\overline{X}^2\right)m\log(m)\right),$$

with:

$$\widehat{V}_g := \begin{cases} \frac{\widehat{\lambda}}{c^2T}\overline{X} & \text{if } \delta = 0, \\ \frac{\widehat{\lambda}^2}{c^2T}\left(\overline{X} + \frac{\left(\overline{X^2}\right)^{1/2}}{\delta(1-\theta_0)^2}\right) & \text{if } \delta > 0, \end{cases}$$
$$\widehat{V}_h := \begin{cases} \frac{\widehat{\lambda}}{c^2T}\overline{W(X)} & \text{if } \delta = 0, \\ \frac{\widehat{\lambda}^2}{c^2T}\left(\overline{W(X)} + \frac{\left(\overline{W(X)^2}\right)^{1/2}}{\delta(1-\theta_0)^2}\right) & \text{if } \delta > 0. \end{cases}$$

The constants κ_1 and κ_2 are calibrated following the "minimum penalty heuristic" (Arlot & Massart, 2009). On several preliminary simulations, we compute the selected dimension $\widehat{m}$ as a function of κ, and we find $\kappa_{\min}$ such that for $\kappa < \kappa_{\min}$ the dimension is too high and for $\kappa > \kappa_{\min}$ it is acceptable. Then, the selected constant is $2\kappa_{\min}$. In our cases, we choose:

- $\kappa_1 = 0.01$, $\kappa_2 = 0.01$ for the ruin probability;
- $\kappa_1 = 0.1$, $\kappa_2 = 1$ for the expected claim size causing the ruin;
- $\kappa_1 = 10^{-8}$ for the Laplace transform of the ruin time, $\delta = 0.1$.

There is no constant κ_2 in the last case because the Laguerre deconvolution estimator $\widehat{\phi}_m^{\mathrm{Lag}_2}$ is defined only if $\delta = 0$.

We write $\widehat{\phi}^{\mathrm{Lag}_1} := \widehat{\phi}_{\widehat{m}^{\mathrm{Lag}_1}}^{\mathrm{Lag}_1}$ and $\widehat{\phi}^{\mathrm{Lag}_2} := \widehat{\phi}_{\widehat{m}^{\mathrm{Lag}_2}}^{\mathrm{Lag}_2}$ in the following.

MISE calculation We compare the estimators by looking at their MISE: $\mathbb{E}\|\phi - \widehat{\phi}\|_{\mathrm{L}^2}^2$. We compute the norm $\|\cdot\|_{\mathrm{L}^2}$ with Romberg's method using a discretization of $[0, u_{\max}]$ with $2^{11}+1$ points. The value of $u_{\max}$ varies from 12 to 50, depending on the parameters set. We compute the expectation by an empirical mean over $n = 200$ paths of the process $(U_t)_{t\in[0,T]}$. We also compute a 95% confidence interval for the MISE, using the asymptotic confidence interval for a mean (CLT approximation):

$$\mathrm{CI} = \left[\overline{\mathrm{ISE}}_n \pm q_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right], \quad \alpha = 5\%,$$

where $\overline{\mathrm{ISE}}_n$ is the empirical mean of the ISEs, $q_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the normal distribution, and S_n^2 is the empirical variance of the ISEs. We have two goals in this section:

1. To compare the performance of our Laguerre–Fourier estimator with the Laguerre deconvolution estimators.

2. To see if the model selection procedures (3.20) for the Laguerre deconvolution estimators lead to the same performance than the naive choice $m = \lfloor 5T^{1/10} \rfloor$.

The code that performed the simulations can be obtained on request.

Results We display our results in Tables 3.1, 3.2 and 3.3. Concerning the estimation of the ruin probability (Table 3.1), we see that all the estimators perform well with the first set of parameter (exponential distribution, $\theta = 0.67$). However, with the two other sets of parameters (exponential distribution and Gamma(2) distribution, $\theta = 0.83$), the difference is clear: the Laguerre–Fourier estimator has the smallest risk, followed by the estimator of Zhang & Su (2018), and the Laguerre deconvolution estimators come last. We notice that $\widehat{\phi}^{\mathrm{Lag}_2}$ seems to be better than $\widehat{\phi}^{\mathrm{Lag}_1}$ in this case.

Concerning the estimation of the expected jump size causing the ruin (Table 3.2), the difference is even clearer. With the first set of parameters, we see that the Laguerre–Fourier is better for small sample size ($\mathbb{E}[N_T] = 100$), but equivalent to the other estimators for larger sample sizes. We also notice that the estimator $\widehat{\phi}^{\mathrm{ZS}}$ and $\widehat{\phi}^{\mathrm{Lag}_2}$ have the same risk. With the two other sets of parameters, we find again that the Laguerre–Fourier estimator is better than the estimator $\widehat{\phi}^{\mathrm{ZS}}$, which is better than the Laguerre deconvolution estimators. This time, we see that $\widehat{\phi}^{\mathrm{Lag}_1}$ has better performances than $\widehat{\phi}^{\mathrm{Lag}_2}$.

Concerning the estimation of Laplace transform of the ruin time (Table 3.3), we see no difference between the MISE of the Laguerre–Fourier estimator and the Laguerre deconvolution estimators.

For illustration purposes, on Figures 3.1 and 3.2, we show the estimations of the ruin probability and the expected claim size causing the ruin, on 50 independent samples, with the second set of parameters (exponential distribution, $\theta = 0.83$). Qualitatively, we see that the Laguerre–Fourier estimator is better than the others. In contrast, the non data-driven choice of m for estimator of Zhang & Su (2018) seems not appropriate in this setting.

To conclude, we can say that our Laguerre–Fourier estimator has better performances than the Laguerre deconvolution estimators on simulated data, even in the exponential case where they have theoretically the same MISE (up to a log factor). Furthermore, the Laguerre deconvolution estimators with the model selection procedure (3.20) fail to match the performance of the estimator of Zhang & Su (2018), for which we choose the parameter m knowing the bias decay rate of ϕ, in most cases.

Parameters	Estimator	$\mathbb{E}[N_T] = 100$	$\mathbb{E}[N_T] = 200$	$\mathbb{E}[N_T] = 400$
$X \sim \text{Exp}(1)$ $\lambda = 1$ $c = 1.5$ $\theta = 0.67$	LagFou	0.14 [0.07, 0.21] $m_1 = 150$	0.053 [0.039, 0.067] $m_1 = 300$	0.022 [0.017, 0.027] $m_1 = 500$
	ZS	0.23 [0.02, 0.44] $m = 8$	0.053 [0.039, 0.068] $m = 9$	0.022 [0.017, 0.028] $m = 10$
	LagDec1	0.25 [0.01, 0.48] $\overline{\widehat{m}} = 3.3$	0.055 [0.042, 0.069] $\overline{\widehat{m}} = 3.8$	0.024 [0.019, 0.029] $\overline{\widehat{m}} = 4.2$
	LagDec2	0.23 [0.02, 0.45] $\overline{\widehat{m}} = 6.0$	0.053 [0.039, 0.068] $\overline{\widehat{m}} = 6.5$	0.023 [0.017, 0.028] $\overline{\widehat{m}} = 7.0$
$X \sim \text{Exp}(1/2)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	0.95 [0.80, 1.09] $m_1 = 240$	0.67 [0.53, 0.80] $m_1 = 480$	0.43 [0.31, 0.55] $m_1 = 500$
	ZS	1.57 [1.26, 1.89] $m = 8$	1.02 [0.74, 1.30] $m = 9$	0.54 [0.46, 0.61] $m = 9$
	LagDec1	8.51 [5.07, 11.95] $\overline{\widehat{m}} = 11.0$	3.82 [1.57, 6.07] $\overline{\widehat{m}} = 12.8$	0.72 [0.36, 1.07] $\overline{\widehat{m}} = 14.4$
	LagDec2	2.96 [2.11, 3.82] $\overline{\widehat{m}} = 14.2$	1.94 [1.08, 2.80] $\overline{\widehat{m}} = 18.1$	0.64 [0.36, 0.92] $\overline{\widehat{m}} = 21.8$
$X \sim \Gamma(2, 1/4)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	0.64 [0.52, 0.77] $m_1 = 240$	0.46 [0.37, 0.56] $m_1 = 480$	0.30 [0.22, 0.38] $m_1 = 500$
	ZS	1.77 [1.07, 2.47] $m = 8$	0.62 [0.45, 0.78] $m = 9$	0.30 [0.23, 0.36] $m = 9$
	LagDec1	7.22 [4.25, 10.19] $\overline{\widehat{m}} = 9.3$	1.71 [0.87, 2.56] $\overline{\widehat{m}} = 10.6$	0.45 [0.21, 0.70] $\overline{\widehat{m}} = 11.6$
	LagDec2	2.71 [1.80, 3.62] $\overline{\widehat{m}} = 12.1$	1.07 [0.66, 1.47] $\overline{\widehat{m}} = 15.7$	0.41 [0.22, 0.60] $\overline{\widehat{m}} = 18.0$

Table 3.1: **Ruin Probability**. We compare the Laguerre–Fourier estimator (LagFou), the estimator of Zhang & Su (2018) (ZS), and the Laguerre deconvolution estimators (LagDec1 and LagDec2). In each case, we display the estimation of the MISE over 200 samples with a 95% confidence interval and the model used ($\overline{\widehat{m}}$ is the mean selected model in the case of the Laguerre deconvolution estimators).

Parameters	Estimator	$\mathbb{E}[N_T] = 100$	$\mathbb{E}[N_T] = 200$	$\mathbb{E}[N_T] = 400$
$X \sim \text{Exp}(1)$ $\lambda = 1$ $c = 1.5$ $\theta = 0.67$	LagFou	1.71 [1.09, 2.32] $m_1 = 150$	0.60 [0.46, 0.73] $m_1 = 300$	0.34 [0.27, 0.40] $m_1 = 500$
	ZS	1.80 [1.07, 2.53] $m = 8$	0.62 [0.47, 0.77] $m = 9$	0.34 [0.27, 0.41] $m = 10$
	LagDec1	1.41 [1.19, 1.64] $\overline{\widehat{m}} = 1.9$	0.84 [0.74, 0.93] $\overline{\widehat{m}} = 2.2$	0.44 [0.39, 0.50] $\overline{\widehat{m}} = 2.8$
	LagDec2	1.86 [1.11, 2.61] $\overline{\widehat{m}} = 3.7$	0.64 [0.49, 0.78] $\overline{\widehat{m}} = 4.1$	0.35 [0.28, 0.42] $\overline{\widehat{m}} = 4.6$
$X \sim \text{Exp}(1/2)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	46.2 [30.0, 62.3] $m_1 = 240$	28.1 [21.7, 34.5] $m_1 = 480$	20.5 [15.1, 25.8] $m_1 = 500$
	ZS	96.3 [62.4, 130.2] $m = 8$	48.0 [31.3, 64.6] $m = 9$	27.9 [23.0, 32.7] $m = 9$
	LagDec1	77.5 [71.6, 83.5] $\overline{\widehat{m}} = 3.0$	56.3 [45.1, 67.4] $\overline{\widehat{m}} = 4.9$	38.9 [29.7, 48.1] $\overline{\widehat{m}} = 7.1$
	LagDec2	197.7 [115.1, 280.3] $\overline{\widehat{m}} = 9.7$	96.7 [47.4, 146.0] $\overline{\widehat{m}} = 12.5$	48.5 [27.5, 69.6] $\overline{\widehat{m}} = 14.6$
$X \sim \Gamma(2, 1/4)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	11.7 [9.4, 14.0] $m_1 = 240$	9.2 [7.5, 10.9] $m_1 = 480$	6.2 [4.1, 8.3] $m_1 = 500$
	ZS	18.5 [12.0, 25.0] $m = 8$	13.6 [10.1, 17.1] $m = 9$	5.9 [4.6, 7.2] $m = 9$
	LagDec1	19.2 [18.2, 20.2] $\overline{\widehat{m}} = 2.8$	15.0 [13.1, 17.0] $\overline{\widehat{m}} = 4.3$	8.4 [7.1, 9.7] $\overline{\widehat{m}} = 5.9$
	LagDec2	28.2 [19.4, 37.1] $\overline{\widehat{m}} = 7.8$	24.6 [16.5, 32.7] $\overline{\widehat{m}} = 10.0$	8.3 [5.6, 11.1] $\overline{\widehat{m}} = 11.3$

Table 3.2: **Expected claim size causing the ruin**. We compare the Laguerre–Fourier estimator (LagFou), the estimator of Zhang & Su (2018) (ZS), and the Laguerre deconvolution estimators (LagDec1 and LagDec2). In each case, we display the estimation of the MISE over 200 samples with a 95% confidence interval and the model used ($\overline{\widehat{m}}$ is the mean selected model in the case of the Laguerre deconvolution estimators).

Parameters	Estimator	$\mathbb{E}[N_T] = 100$	$\mathbb{E}[N_T] = 200$	$\mathbb{E}[N_T] = 400$
$X \sim \text{Exp}(1)$ $\lambda = 1$ $c = 1.5$ $\theta = 0.67$	LagFou	2.50 [1.91, 3.09] $m_1 = 150$	1.09 [0.87, 1.31] $m_1 = 300$	0.64 [0.52, 0.77] $m_1 = 500$
	ZS	2.50 [1.93, 3.07] $m = 8$	1.10 [0.88, 1.33] $m = 9$	0.66 [0.53, 0.79] $m = 10$
	LagDec1	2.52 [1.95, 3.08] $\overline{\widehat{m}} = 4.2$	1.11 [0.89, 1.34] $\overline{\widehat{m}} = 4.6$	0.67 [0.54, 0.80] $\overline{\widehat{m}} = 4.9$
$X \sim \text{Exp}(1/2)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	11.81 [9.26, 14.36] $m_1 = 240$	5.60 [4.49, 6.72] $m_1 = 480$	2.51 [2.04, 2.98] $m_1 = 500$
	ZS	12.47 [10.53, 14.41] $m = 8$	6.13 [5.07, 7.19] $m = 9$	3.30 [2.86, 3.75] $m = 9$
	LagDec1	13.22 [10.61, 15.84] $\overline{\widehat{m}} = 10.2$	5.82 [4.57, 7.06] $\overline{\widehat{m}} = 11.2$	2.65 [2.15, 3.14] $\overline{\widehat{m}} = 12.4$
$X \sim \Gamma(2, 1/4)$ $\lambda = 1.25$ $c = 3$ $\theta = 0.83$	LagFou	10.26 [8.08, 12.45] $m_1 = 240$	4.09 [3.28, 4.91] $m_1 = 480$	2.01 [1.57, 2.46] $m_1 = 500$
	ZS	9.76 [8.07, 11.45] $m = 8$	4.16 [3.36, 4.96] $m = 9$	2.20 [1.77, 2.63] $m = 9$
	LagDec1	10.39 [8.39, 12.38] $\overline{\widehat{m}} = 8.8$	4.15 [3.29, 5.00] $\overline{\widehat{m}} = 9.5$	2.05 [1.59, 2.52] $\overline{\widehat{m}} = 10.3$

Table 3.3: **Laplace transform**, $\delta = 0.1$. We compare the estimators of the Laplace transform of the ruin time: the Laguerre–Fourier estimator (LagFou), the estimator of Zhang & Su (2018) (ZS) and the Laguerre deconvolution estimator (LagDec1). In each case, we display the estimation of the MISE (multiplied by 10^2 in this table) over 200 samples with a 95% confidence interval and the model used ($\overline{\widehat{m}}$ is the mean selected model in the case of the Laguerre deconvolution estimator).

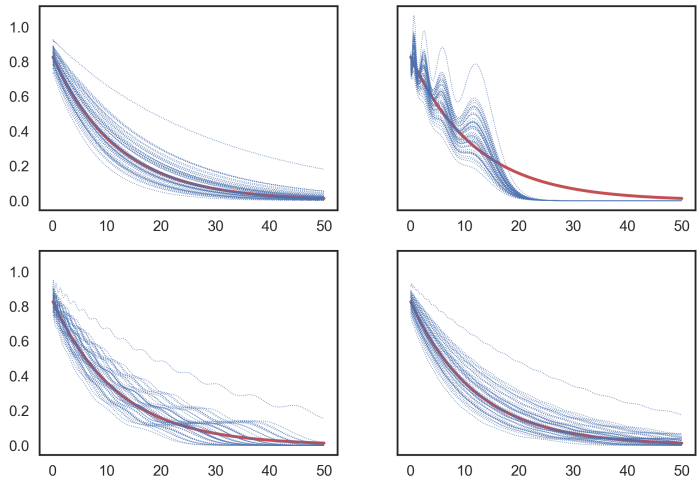

Figure 3.1: Estimation of the ruin probability when the parameters of the model are $\lambda = 1.25$, $\mu = 2$, $c = 3$, $X \sim \text{Exp}(1/\mu)$ and $T = 800$ (so that $\mathbb{E}[N_T] = 1000$). For each estimation procedure, we plot the estimation of ϕ from 50 independent samples. The true function is in bold red and the estimated functions are in dotted blue. **Top left:** Laguerre–Fourier. **Top right:** estimator of Zhang & Su (2018). **Bottom left:** Laguerre deconvolution 1. **Bottom right:** Laguerre deconvolution 2.

Remark 3.5.2. In Tables 3.1, 3.2 and 3.3, the MISEs of the estimators are not normalized by $\|\phi\|^2_{\mathrm{L}^2}$, the size of the estimated function. Hence, it is normal that the order of magnitude of the results varies from one function to another. For example, in Table 3.2, $\|\phi\|^2_{\mathrm{L}^2}$ equals respectively 5, 100 and 50, for each set of parameters.

3.5.2 Model reduction procedure

We propose a *model reduction procedure* to choose the dimensions m_2 and m_3, defined by (3.14). We explain the method for the choice of m_2 in the case $\delta = 0$. Let us assume we have estimated the M first coefficients of g, for a large M. By Remark 3.2.6, we know that the best estimator is $\widehat{g}_M$. Our goal is to choose $\widehat{m}_2$ smaller than M that achieves a similar MISE. This provides a parsimonious version of the estimator without degrading its MISE. By Theorem 3.2.5, the MISE of $\widehat{g}_m$ is given by:

$$\mathbb{E}\|g - \widehat{g}_m\|^2_{\mathrm{L}^2} \leq \|g - g_m\|^2_{\mathrm{L}^2} + \frac{\lambda}{c^2 T}\mathbb{E}[X].$$

Ideally, we would like to choose the first m such that the bias term $\|g - g_m\|^2_{\mathrm{L}^2}$ is smaller than the variance term $\frac{\lambda}{c^2 T}\mathbb{E}[X]$. Since these terms are unknown, we

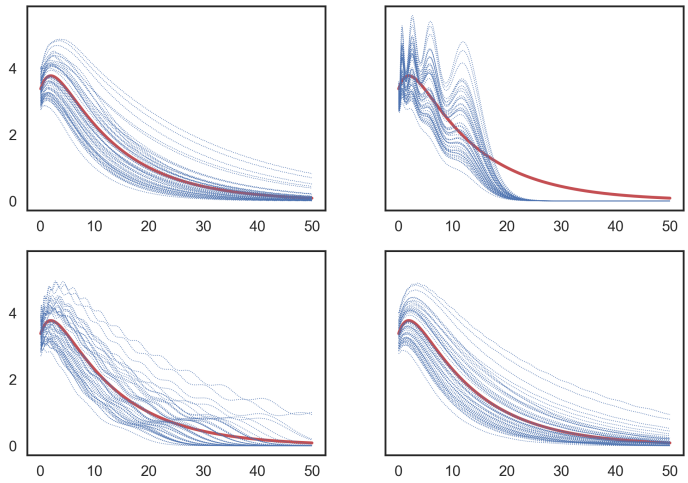

Figure 3.2: Estimation of the expected claim size causing the ruin when the parameters of the model are $\lambda = 1.25$, $\mu = 2$, $c = 3$, $X \sim \text{Exp}(1/\mu)$ and $T = 800$ (so that $\mathbb{E}[N_T] = 1000$). For each estimation procedure, we plot the estimation of ϕ from 50 independent samples. The true function is in bold red and the estimated functions are in dotted blue. **Top left:** Laguerre–Fourier. **Top right:** estimator of Zhang & Su (2018). **Bottom left:** Laguerre deconvolution 1. **Bottom right:** Laguerre deconvolution 2.

estimate them by $\sum_{k=m}^{M-1} \widehat{b}_k^2$ and $\frac{1}{(cT)^2}\sum_{i=1}^{N_T} X_i$ respectively. We choose $\widehat{m}_2$ as:

$$\widehat{m}_2 = \min\left\{1 \le m \le M-1 \,\middle|\, \sum_{k=m}^{M-1} \widehat{b}_k^2 \le \frac{\kappa_2}{(cT)^2}\sum_{i=1}^{N_T} X_i\right\}, \tag{3.21}$$

with κ_2 an adjustment constant. The next proposition shows that the MISE of $\widehat{g}_{\widehat{m}_2}$ does not exceed the MISE of $\widehat{g}_M$ by more than $\kappa_2 \times$ (variance term).

Proposition 3.5.3. *Let $\kappa_2 > 0$, if $\widehat{m}_2$ is chosen as* (3.21) *then the MISE of $\widehat{g}_{\widehat{m}_2}$ is:*

$$\mathbb{E}\|g - \widehat{g}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2 \le \|g - g_M\|_{\mathrm{L}^2}^2 + (1+\kappa_2)\frac{\lambda}{c^2 T}\mathbb{E}[X].$$

Proof. By Pythagoras Theorem:

$$\begin{aligned}
\|g - \widehat{g}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2 &= \|g - \widehat{g}_M\|_{\mathrm{L}^2}^2 + \|\widehat{g}_M - \widehat{g}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2 \\
&= \|g - \widehat{g}_M\|_{\mathrm{L}^2}^2 + \sum_{k=\widehat{m}_2}^{M-1} \widehat{b}_k^2 \\
&\le \|g - \widehat{g}_M\|_{\mathrm{L}^2}^2 + \frac{\kappa_2}{(cT)^2}\sum_{i=1}^{N_T} X_i.
\end{aligned}$$

We take the expectation, and we apply Theorem 3.2.5:

$$\mathbb{E}\|g-\widehat{g}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2 \leq \|g-g_M\|_{\mathrm{L}^2}^2 + (1+\kappa_2)\frac{\lambda}{c^2T}\mathbb{E}[X]. \qquad \square$$

The same goes for $\widehat{m}_3$: we estimate the bias term by $\sum_{k=m}^{M-1}\widehat{c}_k^2$ and the variance term by $\frac{1}{(cT)^2}\sum_{i=1}^{N_T}W(X_i)$; we choose $\widehat{m}_3$ as:

$$\widehat{m}_3 = \min\left\{1\leq m\leq M-1 \,\middle|\, \sum_{k=m}^{M-1}\widehat{c}_k^2 \leq \frac{\kappa_3}{(cT)^2}\sum_{i=1}^{N_T}W(X_i)\right\}.$$

By the same arguments, the MISE of $\widehat{h}_{m_3}$ is given by:

$$\mathbb{E}\|h-\widehat{h}_{\widehat{m}_3}\|_{\mathrm{L}^2}^2 \leq \|h-h_M\|_{\mathrm{L}^2}^2 + (1+\kappa_3)\frac{\lambda}{c^2T}\mathbb{E}[W(X)].$$

In the case $\delta>0$, we choose the same $\widehat{m}_2$ and $\widehat{m}_3$ as in the case $\delta=0$. By the same arguments, we obtain:

$$\begin{aligned}
\mathbb{E}\|g-\widehat{g}_{\widehat{m}_2}\|_{\mathrm{L}^2}^2 &\leq \|g-g_M\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2T}\left(\mathbb{E}[X] + \frac{\mathbb{E}\left[X^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right) + \kappa_2\frac{\lambda}{c^2T}\mathbb{E}[X] \\
&\leq \|g-g_M\|_{\mathrm{L}^2}^2 + (1+\kappa_2)\frac{C(\lambda)}{c^2T}\left(\mathbb{E}[X] + \frac{\mathbb{E}\left[X^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right). \\
\mathbb{E}\|h-\widehat{h}_{\widehat{m}_3}\|_{\mathrm{L}^2}^2 &\leq \|h-h_M\|_{\mathrm{L}^2}^2 + \frac{C(\lambda)}{c^2T}\left(\mathbb{E}[W(X)] + \frac{\mathbb{E}\left[W(X)^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right) + \kappa_3\frac{\lambda}{c^2T}\mathbb{E}[W(X)] \\
&\leq \|h-h_M\|_{\mathrm{L}^2}^2 + (1+\kappa_3)\frac{C(\lambda)}{c^2T}\left(\mathbb{E}[W(X)] + \frac{\mathbb{E}\left[W(X)^2\right]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right).
\end{aligned}$$

Numerically, we compared the MISE's of the Laguerre–Fourier estimator with and without the model reduction procedure for $\widehat{m}_2$ and $\widehat{m}_3$, with the choice $\kappa_2=\kappa_3=0.3$. We show the results in Table 3.4. We see that the model reduction procedure does not affect the MISE of the estimator and we emphasize that the selected dimensions are far lower than the maximum dimension ($\widehat{m}$'s are less than 10 whereas the maximum dimension is 100).

3.6 Conclusion

Using a projection estimator on the Laguerre basis, and computing the coefficients with Fourier transforms, we constructed an estimator of the Gerber–Shiu function that achieves parametric rates of convergence, without needing a model selection procedure. It is worth noticing that our results are non-asymptotic and concern the MISE of the estimator. In comparison, the Laguerre deconvolution

	Model Reduction	No Model Reduction
Ruin Probability	1.06 [0.86, 1.26] $\overline{\widehat{m}_2} = 4.2$, $\overline{\widehat{m}_3} = 4.1$	1.06 [0.86, 1.25] $m_2 = m_3 = N_T$
Jump size	40.1 [33.7, 46.9] $\overline{\widehat{m}_2} = 4.1$, $\overline{\widehat{m}_3} = 4.3$	41.6 [34.7, 48.4] $m_2 = m_3 = N_T$
Laplace transform, $\delta = 0.1$	0.092 [0.075, 0.110] $\overline{\widehat{m}_2} = 3.9$, $\overline{\widehat{m}_3} = 4.0$	0.098 [0.079, 0.117] $m_2 = m_3 = N_T$

Table 3.4: Comparison between the MISE of the Laguerre–Fourier estimator with and without model reduction. In each case, we chose the following parameters: $X \sim \text{Exp}(1/2)$, $\lambda = 1.25$, $c = 3$, $T = 80$. With this set of parameters, $\mathbb{E}[N_T] = 100$. Each cell displays an estimation of the MISE over 200 samples with a 95% confidence interval, and the mean selected models $\widehat{m}_2$ and $\widehat{m}_3$. In every case, m_1 is equal to N_T.

estimators have slower rates of convergence and necessitate a model selection procedure in practice. The better performances of our procedure are confirmed by a numerical study, on simulated data.

Knowing that the Laguerre deconvolution method does not achieve the best rate of convergence in the compound Poisson model is important. Indeed, this method is used to estimate the Gerber–Shiu function in more general models, see Zhang & Su (2019), Su *et al.* (2019) and Su *et al.* (2020). These papers have one thing in common: they all want to estimate a function ϕ that satisfies an equation of the form $\phi = \phi * g + h$, with g and h functions that depend on the specificity of each problem. If we applied the procedure described in the beginning of Section 3.2, we could obtain an estimator that would achieve the same rate of convergence as the estimators of g and h (see Remark 3.2.2). Hence the Laguerre deconvolution method used in these papers is not optimal since a factor m appears in the variance term in the construction step of $\widehat{\phi}_m$ from $\widehat{g}_m$ and $\widehat{h}_m$.

3.7 Proofs

Proof of Proposition 3.2.1. We start with the decomposition bias-variance of the risk of $\widehat{\phi}_m$:

$$\|\phi - \widehat{\phi}_m\|_{\mathrm{L}^2}^2 = \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \|\phi_m - \widehat{\phi}_m\|_{\mathrm{L}^2}^2.$$

Let Π_m be the projector on $\mathrm{Span}(\mathscr{F}\psi_0,\dots,\mathscr{F}\psi_{m-1})$. Since $\|\mathscr{F}\psi_k\|^2 = 2\pi$, we get:

$$\begin{aligned}\|\phi_m - \widehat{\phi}_m\|_{\mathrm{L}^2}^2 &= \sum_{k=0}^{m-1} (\widehat{a}_k - a_k)^2 \\ &= \frac{1}{(2\pi)^2}\sum_{k=0}^{m-1}\left\langle \frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\mathscr{F}g}, \mathscr{F}\psi_k\right\rangle^2 \\ &= \frac{1}{2\pi}\left\|\Pi_m\left(\frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\mathscr{F}g}\right)\right\|_{\mathrm{L}^2}^2 \\ &\le \frac{1}{2\pi}\left\|\frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\mathscr{F}g}\right\|_{\mathrm{L}^2}^2. \end{aligned} \tag{3.22}$$

Then since $|\widehat{\mathscr{F}g}| \le \theta_0$ by definition, and $|\mathscr{F}g| \le \|g\|_{\mathrm{L}^1} \le \theta$, we obtain:

$$\begin{aligned}\left\|\frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\mathscr{F}g}\right\|_{\mathrm{L}^2}^2 &\le 2\left\|\frac{\mathscr{F}\widehat{h}}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\widehat{\mathscr{F}g}}\right\|_{\mathrm{L}^2}^2 + 2\left\|\frac{\mathscr{F}h}{1-\widehat{\mathscr{F}g}} - \frac{\mathscr{F}h}{1-\mathscr{F}g}\right\|_{\mathrm{L}^2}^2 \\ &\le \frac{2}{(1-\theta_0)^2}\|\mathscr{F}\widehat{h} - \mathscr{F}h\|_{\mathrm{L}^2}^2 + \frac{2}{(1-\theta_0)^2}\frac{\|h\|_{\mathrm{L}^1}^2}{(1-\theta)^2}\|\widehat{\mathscr{F}g} - \mathscr{F}g\|_{\mathrm{L}^2}^2. \end{aligned} \tag{3.23}$$

To control the last term, we decompose according to the set $\{|\mathscr{F}\widehat{g}| \le \theta_0\}$ and its complement:

$$\|\widehat{\mathscr{F}g} - \mathscr{F}g\|_{\mathrm{L}^2}^2 \le \|\mathscr{F}\widehat{g} - \mathscr{F}g\|_{\mathrm{L}^2}^2 + \|g\|_{\mathrm{L}^1}^2 \,\mathrm{Leb}(\{|\mathscr{F}\widehat{g}| > \theta_0\}).$$

Thus if $\theta < \theta_0$, then $\{|\mathscr{F}\widehat{g}| > \theta_0\} \subseteq \{|\mathscr{F}\widehat{g} - \mathscr{F}g| \ge \theta_0 - \theta\}$, therefore Markov inequality yields:

$$\|\widehat{\mathscr{F}g} - \mathscr{F}g\|_{\mathrm{L}^2}^2 \le \|\mathscr{F}\widehat{g} - \mathscr{F}g\|_{\mathrm{L}^2}^2 + \frac{\|g\|_{\mathrm{L}^1}^2}{(\theta_0-\theta)^2}\|\mathscr{F}\widehat{g} - \mathscr{F}g\|_{\mathrm{L}^2}^2. \tag{3.24}$$

Finally, gathering (3.22), (3.23) and (3.24), and using Plancherel theorem yield the desired result. □

3.7.1 Proof of Theorem 3.2.5

We start with some preliminary lemmas.

Lemma 3.7.1. *Let $Y_1,\dots,Y_n$ be i.i.d non-negative random variables. We denote by $\mathscr{L}(s) := \mathbb{E}[\mathrm{e}^{-sY_1}]$ their Laplace transform and we denote by $\widehat{\mathscr{L}}(s) := \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{-sY_i}$ the empirical Laplace transform. Then for $p \ge 1$, we have:*

$$\mathbb{E}\left[\sup_{s>0}\left|\widehat{\mathscr{L}}(s) - \mathscr{L}(s)\right|^{2p}\right] \le \frac{p!}{2^{p-1}n^p}.$$

Proof. Let $\widehat{F}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \le x\}}$ be the empirical c.d.f. of the Y_i's, and let $F(x) := \mathbb{P}[Y \le x]$ be their c.d.f. We notice that for $s > 0$:

$$\int_0^{+\infty} s\mathrm{e}^{-sx}\widehat{F}(x)\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int_0^{+\infty} s\mathrm{e}^{-sx}\mathbf{1}_{\{Y_i\le x\}}\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{-sX_i} =: \widehat{\mathscr{L}}(s),$$

and by the same argument, $\mathscr{L}(s) = \int_0^{+\infty} s\mathrm{e}^{-sx}F(x)\,\mathrm{d}x$. Thus:

$$\sup_{s>0}\left|\widehat{\mathscr{L}}(s) - \mathscr{L}(s)\right| \le \sup_{s>0}\int_0^{+\infty} s\mathrm{e}^{-sx}\left|\widehat{F}(x) - F(x)\right|\mathrm{d}x \le \|\widehat{F} - F\|_\infty.$$

We take the expectation and we get:

$$\mathbb{E}\left[\sup_{s>0}\left|\widehat{\mathscr{L}}(s) - \mathscr{L}(s)\right|^{2p}\right] \le \mathbb{E}\left[\|\widehat{F} - F\|_\infty^{2p}\right] = 2p\int_0^{+\infty} t^{2p-1}\,\mathbb{P}\left[\|\widehat{F} - F\|_\infty \ge t\right]\mathrm{d}t. \tag{3.25}$$

By Massart (1990), $\mathbb{P}\left[\sqrt{n}\|\widehat{F} - F\|_\infty \ge x\right] \le 2\mathrm{e}^{-2x^2}$, so by setting $t = x/\sqrt{n}$ in (3.25), we obtain:

$$\begin{aligned}\mathbb{E}\left[\sup_{s>0}\left|\widehat{\mathscr{L}}(s) - \mathscr{L}(s)\right|^{2p}\right] &\le \frac{2p}{n^p}\int_0^{+\infty} x^{2p-1}2\mathrm{e}^{-2x^2}\,\mathrm{d}x \\ &= \frac{2p}{n^p}\int_0^{+\infty} u^{p-1}\mathrm{e}^{2u}\,\mathrm{d}u \\ &= \frac{p!}{n^p\,2^{p-1}}.\end{aligned}$$

□

Lemma 3.7.2. *Let $Z \sim \mathscr{P}(\lambda)$ and $m_j(\lambda) := \mathbb{E}[(Z-\lambda)^j]$ be the j-th central moment of Z. Then, for all $r \ge 2$ we have:*

$$m_r(\lambda) = \lambda\sum_{j=0}^{r-2}\binom{r-1}{j}m_j(\lambda).$$

Proof. Let $\mathscr{L}(\lambda, t) := e^{\lambda(\mathrm{e}^t - t - 1)} = \mathbb{E}[\mathrm{e}^{t(Z-\lambda)}]$ and $\varphi(t) := \mathrm{e}^t - t - 1$. Then $m_r(\lambda) = \frac{\partial^r \mathscr{L}}{\partial t^r}(\lambda, 0)$. By Leibniz's rule:

$$\frac{\partial^r \mathscr{L}}{\partial t^r}(\lambda, t) = \frac{\partial^{r-1}}{\partial t^{r-1}}\Big(\lambda\varphi'(t)\,\mathscr{L}(\lambda, t)\Big) = \lambda\sum_{j=0}^{r-1}\binom{r-1}{j}\frac{\partial^j \mathscr{L}}{\partial t^j}(\lambda, t)\,\varphi^{(r-j)}(t).$$

Taking $t = 0$ gives the result since $\varphi'(0) = 0$ and $\varphi^{(k)}(0) = 1$ if $k \ge 2$. □

Corollary 3.7.3. *The central moments $m_{2r}(\lambda)$ and $m_{2r+1}(\lambda)$ are polynomials of degree r in λ.*

The next proposition provides an upper bound on the L^p-risk of $\widehat{\rho}_\delta$.

Proposition 3.7.4. *Under Assumption 3.1, for $p \ge 1$, we have:*

$$\mathbb{E}\left[(\widehat{\rho}_\delta - \rho_\delta)^{2p}\right] \le \frac{C(p, \lambda)}{c^{2p}(1-\theta)^{2p}T^p},$$

where $C(p, \lambda)$ is a $\mathrm{O}(\lambda^p)$.

Proof. By definition, ρ_δ is a solution of the Lundberg equation, so it is a zero of the function:

$$\ell_\delta(s) := cs - (\lambda + \delta) + \lambda \mathscr{L} f(s).$$

The estimator $\widehat{\rho}_\delta$ is then a zero of the function:

$$\widehat{\ell}_\delta(s) := cs - (\widehat{\lambda} + \delta) + \widehat{\lambda}\, \widehat{\mathscr{L} f}(s).$$

We use a Taylor–Lagrange expansion:

$$\widehat{\ell}_\delta(\widehat{\rho}_\delta) = 0 = \ell_\delta(\rho_\delta) = \ell_\delta(\widehat{\rho}_\delta) + \ell'_\delta(z)(\rho_\delta - \widehat{\rho}_\delta),$$

where z is between ρ_δ and $\widehat{\rho}_\delta$.

$$|\ell'_\delta(z)| = \left| c - \lambda \int_0^{+\infty} x \mathrm{e}^{-zx} f(x)\,\mathrm{d}x \right| \geq c - \lambda \int_0^{+\infty} x f(x)\,\mathrm{d}x = c - \lambda\mu > 0,$$

under the safety loading condition. Thus:

$$\begin{aligned}
|\rho_\delta - \widehat{\rho}_\delta| &\leq \frac{1}{c - \lambda\mu} \left| \widehat{\ell}_\delta(\widehat{\rho}_\delta) - \ell_\delta(\widehat{\rho}_\delta) \right| \\
&= \frac{1}{c(1-\theta)} \left| \widehat{\lambda} \left(\widehat{\mathscr{L} f}(\widehat{\rho}_\delta) - \mathscr{L} f(\widehat{\rho}_\delta) \right) + \left(\widehat{\lambda} - \lambda \right) \left(1 - \mathscr{L} f(\widehat{\rho}_\delta) \right) \right| \\
&\leq \frac{1}{c(1-\theta)} \left(|\widehat{\lambda}| \| \widehat{\mathscr{L} f} - \mathscr{L} f \|_\infty + 2 |\widehat{\lambda} - \lambda| \right) \\
\mathbb{E}\left[(\widehat{\rho}_\delta - \rho_\delta)^{2p} \right] &\leq \frac{1}{c^{2p}(1-\theta)^{2p}} \left(2^{2p-1} \mathbb{E}\left[\widehat{\lambda}^{2p} \| \widehat{\mathscr{L} f} - \mathscr{L} f \|_\infty^{2p} \right] + 2^{4p-1} \mathbb{E}\left[|\widehat{\lambda} - \lambda|^{2p} \right] \right).
\end{aligned}$$

For the second term, we use Corollary 3.7.3: $\mathbb{E}|\widehat{\lambda} - \lambda|^{2p} = \frac{\mathbb{E}|N_T - \lambda T|^{2p}}{T^{2p}} = \frac{\mathrm{O}(\lambda^p)}{T^p}$. For the first term, we apply Lemma 3.7.1 conditional to N_T:

$$\mathbb{E}\left[\widehat{\lambda}^{2p} \| \widehat{\mathscr{L} f} - \mathscr{L} f \|_\infty^{2p} \right] \leq \mathbb{E}\left[C(p) \frac{\widehat{\lambda}^{2p}}{N_T^p} \right] = \frac{C(p)}{T^p} \mathbb{E}\left[\widehat{\lambda}^p \right] = \frac{\mathrm{O}(\lambda^p)}{T^p}.$$

Finally:

$$\mathbb{E}\left[(\widehat{\rho}_\delta - \rho_\delta)^{2p} \right] \leq \frac{C(p,\lambda)}{c^{2p}(1-\theta)^{2p} T^p},$$

with $C(p,\lambda) = \mathrm{O}(\lambda^p)$. □

Now, we can prove Theorem 3.2.5.

Proof of Theorem 3.2.5. By Pythagoras theorem:

$$\begin{aligned}
\mathbb{E}\| g - \widehat{g}_{m_2} \|_{\mathrm{L}^2}^2 &= \| g - g_{m_2} \|_{\mathrm{L}^2}^2 + \mathbb{E}\| g_{m_2} - \widehat{g}_{m_2} \|_{\mathrm{L}^2}^2 = \| g - g_{m_2} \|_{\mathrm{L}^2}^2 + \sum_{k=0}^{m_2 - 1} \mathbb{E}\left[(\widehat{b}_k - b_k)^2 \right], \\
\mathbb{E}\| h - \widehat{h}_{m_3} \|_{\mathrm{L}^2}^2 &= \| h - h_{m_3} \|_{\mathrm{L}^2}^2 + \mathbb{E}\| h_{m_3} - \widehat{h}_{m_3} \|_{\mathrm{L}^2}^2 = \| h - h_{m_3} \|_{\mathrm{L}^2}^2 + \sum_{k=0}^{m_3 - 1} \mathbb{E}\left[(\widehat{c}_k - c_k)^2 \right],
\end{aligned}$$

hence we need to control the variance terms $\sum_{k=0}^{m_2-1}\mathbb{E}[(\widehat{b}_k-b_k)^2]$ and $\sum_{k=0}^{m_3-1}\mathbb{E}[(\widehat{c}_k-c_k)^2]$.

Using equations (4.17) to (4.21) and (4.10) to (4.14) in Zhang & Su (2018), we can obtain equations (3.31) and (3.32) below. Still, we give the proofs of these equations for the sake of completeness.

We notice that $\widehat{b}_k$ and $\widehat{c}_k$ (defined by (3.12) and (3.13)) can be written as:

$$\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\widehat{\rho}_\delta)\psi_k(u)\,\mathrm{d}u,$$

and that the coefficients b_k and c_k (defined by (3.9) and (3.10)) can be written as:

$$\mathbb{E}\left[\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\rho_\delta)\psi_k(u)\,\mathrm{d}u\right],$$

where F is given by:

$$F(u,X,\rho) \coloneqq \begin{cases} \mathrm{e}^{-\rho(X-u)}\mathbf{1}_{X>u} & \text{for the coefficients of } g, \\ \int_u^X \mathrm{e}^{-\rho(X-x)}w(x,X-x)\,\mathrm{d}x\,\mathbf{1}_{X>u} & \text{for the coefficients of } h. \end{cases} \tag{3.26}$$

Thus, we need to give an upper bound on quantities of the form:

$$V_m \coloneqq \sum_{k=0}^{m-1}\mathbb{E}[I_k], \tag{3.27}$$

$$I_k \coloneqq \left(\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\widehat{\rho}_\delta)\psi_k(u)\,\mathrm{d}u - \mathbb{E}\left[\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\rho_\delta)\psi_k(u)\,\mathrm{d}u\right]\right)^2.$$

The bound on V_m is based on the following decomposition:

$$\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\widehat{\rho}_\delta)\psi_k(u)\,\mathrm{d}u - \mathbb{E}\left[\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\rho_\delta)\psi_k(u)\,\mathrm{d}u\right] = Z_k + \Delta_k \tag{3.28}$$

where:

$$Z_k \coloneqq \frac{1}{cT}\left(\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\widehat{\rho}_\delta)\psi_k(u)\,\mathrm{d}u - \mathbb{E}\left[\sum_{i=1}^{N_T}\int_0^{+\infty} F(u,X_i,\widehat{\rho}_\delta)\psi_k(u)\,\mathrm{d}u\right]\right)$$

$$\Delta_k \coloneqq \frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty}\left[F(u,X_i,\widehat{\rho}_\delta) - F(u,X_i,\rho_\delta)\right]\psi_k(u)\,\mathrm{d}u.$$

Let us notice that if $\delta = 0$, then $\widehat{\rho}_\delta = \rho_\delta = 0$, so $\Delta_k = 0$ and the decomposition reduces to Z_k.

- Bound on $\sum_{k=0}^{m-1}\mathbb{E}\big[Z_k^2\big]$. This bound is obtained by a projection argument:

$$\begin{aligned}\sum_{k=0}^{m-1}\mathbb{E}\big[Z_k^2\big] &= \sum_{k=0}^{m-1}\mathrm{Var}\left(\frac{1}{cT}\sum_{i=1}^{N_T}\int_0^{+\infty}F(u,X_i,\rho_\delta)\psi_k(u)\,\mathrm{d}u\right)\\ &= \sum_{k=0}^{m-1}\frac{\lambda}{c^2T}\mathbb{E}\left[\left(\int_0^{+\infty}F(u,X,\rho_\delta)\psi_k(u)\,\mathrm{d}u\right)^2\right]\\ &\le \frac{\lambda}{c^2T}\mathbb{E}\left[\int_0^{+\infty}F(u,X,\rho_\delta)^2\,\mathrm{d}u\right],\end{aligned}$$

  where the last inequality comes from the fact that $(\psi_k)_{k\ge0}$ is an orthonormal basis of $\mathrm{L}^2(\mathbb{R}_+)$. From (3.26), we see that:

$$\frac{\lambda}{c^2T}\mathbb{E}\left[\int_0^{+\infty}F(u,X,\rho_\delta)^2\,\mathrm{d}u\right]\le\begin{cases}\frac{\lambda}{c^2T}\mathbb{E}[X] & \text{for the coefficients of } g,\\ \frac{\lambda}{c^2T}\mathbb{E}[W(X)] & \text{for the coefficients of } h.\end{cases}\tag{3.29}$$

  where $W(X)$ is defined in Assumption 3.3. In the $\delta=0$ case, this gives the desired results.

- Bound on $\sum_{k=0}^{m-1}\Delta_k^2$. We use a projection argument again:

$$\begin{aligned}\sum_{k=0}^{m-1}\Delta_k^2 &\le \sum_{k=0}^{m-1}\frac{N_T}{c^2T^2}\sum_{i=1}^{N_T}\left(\int_0^{+\infty}\big[F(u,X_i,\widehat{\rho}_\delta)-F(u,X_i,\rho_\delta)\big]\psi_k(u)\,\mathrm{d}u\right)^2\\ &\le \frac{\widehat{\lambda}}{c^2T}\sum_{i=1}^{N_T}\int_0^{+\infty}|F(u,X_i,\widehat{\rho}_\delta)-F(u,X_i,\rho_\delta)|^2\,\mathrm{d}u,\end{aligned}$$

  where $\widehat{\lambda}:=\frac{N_T}{T}$. By Remark 3.1.3, we know that $\rho_\delta\in[\frac{\delta}{c},\frac{\delta+\lambda}{c}]$ and that $\widehat{\rho}_\delta\in[\frac{\delta}{c},\frac{\delta+\widehat{\lambda}}{c}]$, so by the mean value theorem:

$$|F(u,X_i,\widehat{\rho}_\delta)-F(u,X_i,\rho_\delta)|\le|\widehat{\rho}_\delta-\rho_\delta|\sup_{\rho\ge\frac{\delta}{c}}\left|\frac{\partial F}{\partial\rho}(u,X_i,\rho)\right|.$$

  Since the function $te^{-\rho t}\mathbf{1}_{t>0}$ achieves its maximum at $t=\frac{1}{\rho}$, we see that:

$$\sup_{\rho\ge\frac{\delta}{c}}\left|\frac{\partial F}{\partial\rho}(u,X_i,\rho)\right|\le\begin{cases}\frac{c}{\mathrm{e}\delta}\mathbf{1}_{X_i>u} & \text{for the coefficients of } g,\\ \frac{c}{\mathrm{e}\delta}\int_u^{X_i}w(x,X_i-x)\,\mathrm{d}x\mathbf{1}_{X_i>u} & \text{for the coefficients of } h.\end{cases}$$

  Thus,

$$\sum_{k=0}^{m-1}\Delta_k^2\le\frac{\widehat{\lambda}\big|\widehat{\rho}_\delta-\rho_\delta\big|^2}{\mathrm{e}^2\delta^2}\times\begin{cases}\frac{1}{T}\sum_{i=1}^{N_T}X_i & \text{for the coefficients of } g,\\ \frac{1}{T}\sum_{i=1}^{N_T}W(X_i) & \text{for the coefficients of } h.\end{cases}\tag{3.30}$$

Using the decomposition (3.28) in (3.27), we get:

$$V_m\le 2\sum_{k=0}^{m-1}\mathbb{E}[Z_k^2]+2\sum_{k=0}^{m-1}\mathbb{E}[\Delta_k^2].$$

Combining (3.29) and (3.30) yields:

$$\mathbb{E}\|\widehat{g}_{m_2} - g_{m_2}\|^2_{\mathrm{L}^2} \leq 2\frac{\lambda}{c^2 T}\mathbb{E}[X] + 2\mathbb{E}\left[\frac{\widehat{\lambda}|\widehat{\rho}_\delta - \rho_\delta|^2}{\mathrm{e}^2\delta^2}\frac{1}{T}\sum_{i=1}^{N_T} X_i\right], \tag{3.31}$$

$$\mathbb{E}\|\widehat{h}_{m_3} - h_{m_3}\|^2_{\mathrm{L}^2} \leq 2\frac{\lambda}{c^2 T}\mathbb{E}[W(X)] + 2\mathbb{E}\left[\frac{\widehat{\lambda}|\widehat{\rho}_\delta - \rho_\delta|^2}{\mathrm{e}^2\delta^2}\frac{1}{T}\sum_{i=1}^{N_T} W(X_i)\right]. \tag{3.32}$$

We apply Hölder's inequality on the second term in (3.31) and we use Proposition 3.7.4:

$$\begin{aligned}\mathbb{E}\left[\widehat{\lambda}|\widehat{\rho}_\delta - \rho_\delta|^2\frac{1}{T}\sum_{i=1}^{N_T} X_i\right] &\leq \mathbb{E}[\widehat{\lambda}^4]^{1/4}\mathbb{E}\big[|\widehat{\rho}_\delta - \rho_\delta|^8\big]^{1/4}\mathbb{E}\left[\left(\frac{1}{T}\sum_{i=1}^{N_T} X_i\right)^2\right]^{1/2} \\ &\leq \frac{C(\lambda)}{c^2(1-\theta)^2 T}\mathbb{E}\left[\left(\frac{1}{T}\sum_{i=1}^{N_T} X_i)\right)^2\right]^{1/2},\end{aligned}$$

with $C(\lambda) = \mathrm{O}(\lambda^2)$. We need to evaluate this last expectation:

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{i=1}^{N_T} X_i\right)^2\right] \leq \mathbb{E}\left[\frac{N_T}{T^2}\sum_{i=1}^{N_T} X_i^2\right] = \mathbb{E}\left[\frac{N_T^2}{T^2}\right]\mathbb{E}[X^2] = \left(\frac{\lambda}{T} + \lambda^2\right)\mathbb{E}[X^2].$$

Thus, we obtain:

$$\mathbb{E}\|\widehat{g}_m - g_m\|^2_{\mathrm{L}^2} \leq 2\frac{\lambda}{c^2 T}\mathbb{E}[X] + 2\frac{C(\lambda)}{c^2 T(1-\theta)^2\delta^2}\mathbb{E}[X^2]^{1/2}$$

with $C(\lambda) = \mathrm{O}(\lambda^2)$. We make the same reasoning for h, replacing X_i by $W(X_i)$. □

3.7.2 Proofs of Section 3.3

Let us recall some facts about Toeplitz matrices; the interested reader can find more details in the book of Böttcher & Grudsky (2000). Given $(\alpha_n)_{n\in\mathbb{Z}}$ a sequence of complex numbers, a Toeplitz matrix is an infinite matrix of the form:

$$\begin{pmatrix} \alpha_0 & \alpha_{-1} & \alpha_{-2} & \cdots \\ \alpha_1 & \alpha_0 & \alpha_{-1} & \cdots \\ \alpha_2 & \alpha_1 & \alpha_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{3.33}$$

The classical result from O. Toeplitz says that this matrix induces a bounded operator on $\ell^2(\mathbb{N})$ if and only if $(\alpha_n)_{n\in\mathbb{Z}}$ are the Fourier coefficients of some function $\alpha \in \mathrm{L}^\infty(\mathbb{T})$, where $\mathbb{T}$ denotes the complex unit circle. We denote both the matrix (3.33) and its induced operator on $\ell^2(\mathbb{N})$ by $\mathbf{T}(\alpha)$, the function α being called *the symbol* of the Toeplitz matrix. Finally, if $m \in \mathbb{N}^*$ and if $\mathbf{T}(\alpha)$ is a Toeplitz matrix, we denote by $\mathbf{T}_m(\alpha)$ the $m \times m$ matrix:

$$\mathbf{T}_m(\alpha) \coloneqq \begin{pmatrix} \alpha_0 & \cdots & \alpha_{-(m-1)} \\ \vdots & \ddots & \vdots \\ \alpha_{m-1} & \cdots & \alpha_0 \end{pmatrix}. \tag{3.34}$$

The operator norm of $\mathbf{T}(\alpha)$ depends on the properties of its symbol. In the case where $\alpha_k = 0$ for all $k < 0$, we have the following lemma.

Lemma 3.7.5. *Let $(\alpha_k)_{k\geq 0} \in \ell^1(\mathbb{N})$ be a sequence of complex numbers. Then the Toeplitz matrix $\mathbf{T}(\alpha)$ is lower triangular and we have:*

$$\forall x \in \ell^2(\mathbb{N}), \quad \mathbf{T}(\alpha) \times x = \alpha * x.$$

In particular, $\|\mathbf{T}(\alpha)\|_{\mathrm{op}} \leq \|\alpha\|_{\ell^1}$.

Proof. The fact that $\mathbf{T}(\alpha)$ is lower triangular and that $\mathbf{T}(\alpha) \times x = \alpha * x$ is clear from the definition of a Toeplitz matrix. Then, Young's inequality for convolution yields $\|\alpha * x\|_{\ell^2} \leq \|\alpha\|_{\ell^1} \|x\|_{\ell^2}$. □

Concerning the inverse of a Toeplitz matrix, its norm depends on the position of zero relatively to the range of the symbol. More precisely, we use the following result.

Lemma 3.7.6 (Lemma 3.8 in Böttcher & Grudsky (2000))**.** *Let $\alpha \in \mathrm{L}^\infty(\mathbb{T})$ and let $E(\alpha)$ be the convex hull of the essential range of α. If $d := \mathrm{dist}(0, E(\alpha)) > 0$, then $\mathbf{T}_m(\alpha)$ is invertible for all $m \geq 1$, and we have $\|\mathbf{T}_m^{-1}(\alpha)\|_{\mathrm{op}} < \frac{2}{d}$.*

The matrix $\mathbf{A}_m$ defined by (3.16) is a Toeplitz matrix and its symbol is given by:

$$\alpha(t) := \sum_{k=0}^{+\infty} \alpha_k \, t^k, \quad \text{with} \quad \alpha_k := \begin{cases} 1 - \frac{b_0}{\sqrt{2}} & \text{if } k = 0, \\ \frac{b_{k-1} - b_k}{\sqrt{2}} & \text{if } k \geq 1. \end{cases}$$

Let us notice that under Assumption 3.4, we have $(\alpha_k)_{k\geq 0} \in \ell^1(\mathbb{N})$ so the symbol α is continuous on $\mathbb{T}$, and thus $\alpha \in \mathrm{L}^\infty(\mathbb{T})$.

Proof of Lemma 3.3.2. We apply[2] Lemma C.1 in Comte *et al.* (2017) to the coefficients of g: the sequence $(\beta_k)_{k\geq 0}$, defined by $\beta_0 := \frac{b_0}{\sqrt{2}}$ and $\beta_k := \frac{b_k - b_{k-1}}{\sqrt{2}}$ for $k \geq 1$, are the Fourier coefficients of the function $t \in \mathbb{T} \mapsto \mathscr{L}g(\frac{1+t}{1-t}) \in \mathbb{C}$. Thus, we have:

$$\forall t \in \mathbb{T}, \quad \mathscr{L}g\left(\frac{1+t}{1-t}\right) = \sum_{k=0}^{+\infty} \beta_k \, t^k,$$

with the convention $\mathscr{L}g(\infty) = 0$. Since $\alpha(t) = 1 - \sum_{k\geq 0} \beta_k t^k$, we get:

$$\forall t \in \mathbb{T}, \quad \alpha(t) = 1 - \mathscr{L}g\left(\frac{1+t}{1-t}\right),$$

[2] this lemma is stated for the generalized Laguerre basis, which depends on a parameter a in their article. This parameter is equal to 1 in our case.

We notice that if $t \in \mathbb{T} \setminus \{1\}$, then there exists $\omega \in \mathbb{R}$ such that $\frac{1+t}{1-t} = \mathrm{i}\omega$. Thus:

$$\begin{aligned}
\forall\, t \in \mathbb{T} \setminus \{1\}, \quad \mathfrak{Re}\, \alpha(t) &= 1 - \mathfrak{Re}\left[\mathscr{L} g\left(\frac{1+t}{1-t}\right)\right] \\
&= 1 - \mathfrak{Re}\big[\mathscr{L} g(\mathrm{i}\omega)\big] \\
&= 1 - \mathfrak{Re}\left[\int_0^{+\infty} \mathrm{e}^{-\mathrm{i}\omega x} g(x)\,\mathrm{d}x\right] \\
&= 1 - \int_0^{+\infty} \cos(\omega x)\, g(x)\,\mathrm{d}x \\
&\geq 1 - \int_0^{+\infty} g(x)\,\mathrm{d}x \geq 1 - \theta.
\end{aligned}$$

This inequality holds for $t = 1$ as well, hence $\alpha(\mathbb{T})$ is included in the half-plane $\{z \in \mathbb{C} \,|\, \mathfrak{Re}(z) \geq 1-\theta\}$, and so is its convex hull. By Lemma 3.7.6:

$$\|\mathbf{A}_m^{-1}\|_{\mathrm{op}} \leq \frac{2}{1-\theta}. \qquad \square$$

Remark 3.7.7. In their article, Zhang & Su (2018) show that $\inf_{|z|=1} |\alpha(z)| \geq 1-\theta > 0$, that is $\mathrm{dist}(0, \alpha(\mathbb{T})) > 0$, which is not sufficient to apply Lemma 3.7.6.

Proof of Theorem 3.3.3

Proposition 3.7.8. *Under Assumption 3.4, if $\theta < \theta_0$ then it holds:*

$$\forall\, m \in \mathbb{N}^*, \quad \mathbb{E}\|(\widetilde{\mathbf{A}}_{m,1}^{-1} - \mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2 \leq \frac{C(\theta,\theta_0)}{(1-\theta_0)^2}\, \|\phi_m\|_{\mathrm{L}^2}^2\, \mathbb{E}\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}^2,$$

where $C(\theta,\theta_0)$ is a constant satisfying $C(\theta,\theta_0) \lesssim \frac{(1-\theta_0)^2}{(\theta_0-\theta)^2}$.

Proof. We decompose the expectation according to the event $\Delta_m := \{\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} \leq \frac{2}{1-\theta_0}\}$:

$$\begin{aligned}
\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,1}^{-1} - \mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2 &= \|\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \mathbb{P}\big[\Delta_m^c\big] + \mathbb{E}\big[\|\widetilde{\mathbf{A}}_{m,1}^{-1}(\mathbf{A}_m - \widehat{\mathbf{A}}_m)\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \mathbf{1}_{\Delta_m}\big] \\
&\leq \|\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \left(\mathbb{P}\big[\Delta_m^c\big] + \frac{4}{(1-\theta_0)^2}\mathbb{E}\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}^2\right) \\
&= \|\mathbf{a}_m\|_{\ell^2}^2 \left(\mathbb{P}\big[\Delta_m^c\big] + \frac{4}{(1-\theta_0)^2}\mathbb{E}\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}^2\right).
\end{aligned}$$

Since $\theta < \theta_0$ and $\|\mathbf{A}_m^{-1}\|_{\mathrm{op}} \leq \frac{2}{1-\theta}$ (see Lemma 3.3.2), we get:

$$\begin{aligned}
\mathbb{P}\big[\Delta_m^c\big] &\leq \mathbb{P}\left[\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \frac{2}{1-\theta_0} - \frac{2}{1-\theta}\right] \\
&= \mathbb{P}\left[\left\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \frac{2}{1-\theta_0} - \frac{2}{1-\theta}\right\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\quad + \mathbb{P}\left[\left\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \frac{2}{1-\theta_0} - \frac{2}{1-\theta}\right\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} \geq \frac{1}{2}\right\}\right].
\end{aligned}$$

- First term. We apply Theorem B.1.2 and we conclude using Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\left[\left\{\|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\mathrm{op}}>\frac{2}{1-\theta_0}-\frac{2}{1-\theta}\right\}\cap\left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m-\mathbf{A}_m)\|_{\mathrm{op}}<\frac{1}{2}\right\}\right]\\
&\leq\mathbb{P}\left[\left\{\frac{\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}}{1-\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m-\mathbf{A}_m)\|_{\mathrm{op}}}>\frac{2}{1-\theta_0}-\frac{2}{1-\theta}\right\}\cap\left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m-\mathbf{A}_m)\|_{\mathrm{op}}<\frac{1}{2}\right\}\right]\\
&\leq\mathbb{P}\left[\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}>\frac{1}{1-\theta_0}-\frac{1}{1-\theta}\right]\\
&\leq\frac{(1-\theta)^2(1-\theta_0)^2}{(\theta_0-\theta)^2}\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^4\,\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}^2\leq\frac{16(1-\theta_0)^2}{(\theta_0-\theta)^2(1-\theta)^2}\,\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}^2.
\end{aligned}$$

- Second term. We use Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\left[\left\{\|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\mathrm{op}}>\frac{2}{1-\theta_0}-\frac{2}{1-\theta}\right\}\cap\left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m-\mathbf{A}_m)\|_{\mathrm{op}}\geq\frac{1}{2}\right\}\right]\\
&\leq\mathbb{P}\left[\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m-\mathbf{A}_m)\|_{\mathrm{op}}\geq\frac{1}{2}\right]\\
&\leq 4\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\,\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}^2\leq\frac{16}{(1-\theta)^2}\,\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}^2.
\end{aligned}$$

Thus, we obtain:

$$\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,1}^{-1}-\mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2\leq\|\phi_m\|_{\mathrm{L}^2}^2\left(\frac{16}{(1-\theta)^2}\left(1+\frac{(1-\theta_0)^2}{(\theta_0-\theta)^2}\right)+\frac{4}{(1-\theta_0)^2}\right)\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{op}}^2.$$

□

We can now prove Theorem 3.3.3.

Proof of Theorem 3.3.3. We start with the decomposition bias-variance of the risk of $\widehat{\phi}_m^{\mathrm{Lag}_1}$:

$$\|\phi-\widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2=\|\phi-\phi_m\|_{\mathrm{L}^2}^2+\|\phi_m-\widehat{\phi}_m^{\mathrm{Lag}}\|_{\mathrm{L}^2}^2.$$

We decompose the variance term as:

$$\begin{aligned}
&\mathbb{E}\|\phi_m-\widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2=\mathbb{E}\|\widehat{\mathbf{a}}_m^{\mathrm{Lag}_1}-\mathbf{a}_m\|_{\ell^2}^2=\mathbb{E}\|\widetilde{\mathbf{A}}_{m,1}^{-1}\widehat{\mathbf{c}}_m-\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2\\
&\leq 3\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,1}^{-1}-\mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2+3\mathbb{E}\|(\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,1}^{-1})(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2+3\mathbb{E}\|\mathbf{A}_m^{-1}(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2.
\end{aligned}$$

- First term. We apply Proposition 3.7.8 with Lemma 3.7.5:

$$\begin{aligned}
\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,1}^{-1}-\mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2&\leq C(\theta,\theta_0)\,\|\phi_m\|_{\mathrm{L}^2}^2\,\mathbb{E}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_{\mathrm{L}^2}^2\\
&\leq C(\theta,\theta_0)\,\|\phi_m\|_{\mathrm{L}^2}^2\,\mathbb{E}\|\widehat{\mathbf{b}}_m-\mathbf{b}_m\|_{\ell^1}^2\\
&\leq C(\theta,\theta_0)\,\|\phi_m\|_{\mathrm{L}^2}^2\,m\,\mathbb{E}\|\widehat{\mathbf{b}}_m-\mathbf{b}_m\|_{\ell^2}^2\\
&=C(\theta,\theta_0)\,\|\phi\|_{\mathrm{L}^2}^2\,m\,\mathbb{E}\|\widehat{g}_m-g_m\|_{\mathrm{L}^2}^2.
\end{aligned}$$

- Second term.

$$\begin{aligned}\mathbb{E}\|(\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,1}^{-1})(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2 &\leq \mathbb{E}\Big[\|\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,1}^{-1}\|_{\mathrm{op}}^2\|\mathbf{c}_m-\widehat{\mathbf{c}}_m\|_{\ell^2}^2\Big]\\ &\leq \left(\frac{8}{(1-\theta)^2}+\frac{8}{(1-\theta_0)^2}\right)\mathbb{E}\|\widehat{h}_m-h_m\|_{\mathrm{L}^2}^2.\end{aligned}$$

- Third term.

$$\begin{aligned}\mathbb{E}\|\mathbf{A}_m^{-1}(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2 &\leq \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\,\mathbb{E}\|\mathbf{c}_m-\widehat{\mathbf{c}}_m\|_{\ell^2}^2\\ &\leq \frac{4}{(1-\theta)^2}\mathbb{E}\|\widehat{h}_m-h_m\|_{\mathrm{L}^2}^2.\end{aligned}$$

$$\mathbb{E}\|\phi_m-\widehat{\phi}_m^{\mathrm{Lag}_1}\|_{\mathrm{L}^2}^2 \leq 3\times\frac{C(\theta,\theta_0)}{(1-\theta_0)^2}\,\|\phi\|_{\mathrm{L}^2}^2\,m\mathbb{E}\|\widehat{g}_m-g_m\|_{\mathrm{L}^2}^2+\frac{60}{(1-\theta_0)^2}\mathbb{E}\|\widehat{h}_m-h_m\|_{\mathrm{L}^2}^2,$$

with $C(\theta,\theta_0)\lesssim\frac{(1-\theta_0)^2}{(1-\theta)^2}$. To conclude, we use the upper bounds established in the proof of Theorem 3.2.5. If $\delta=0$, we have:

$$\mathbb{E}\|\widehat{g}_m-g_m\|_{\mathrm{L}^2}^2\leq\frac{\lambda}{c^2T}\mathbb{E}[X],\qquad\mathbb{E}\|\widehat{h}_m-h_m\|_{\mathrm{L}^2}^2\leq\frac{\lambda}{c^2T}\mathbb{E}[W(X)],$$

and if $\delta>0$, we have:

$$\begin{aligned}\mathbb{E}\|\widehat{g}_m-g_m\|_{\mathrm{L}^2}^2&\leq\frac{C(\lambda)}{c^2T}\left(\mathbb{E}[X]+\frac{\mathbb{E}\big[X^2\big]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),\\ \mathbb{E}\|\widehat{h}_m-h_m\|_{\mathrm{L}^2}^2&\leq\frac{C(\lambda)}{c^2T}\left(\mathbb{E}[W(X)]+\frac{\mathbb{E}\big[W(X)^2\big]^{\frac{1}{2}}}{(1-\theta)^2\delta^2}\right),\end{aligned}$$

with $C(\lambda)=\mathrm{O}(\lambda^2)$. □

Proof of Proposition 3.3.4

Let us introduce the sequence of functions $(D_k)_{k\geq0}$ as:

$$D_k(x):=\begin{cases}\frac{\Psi_0(x)}{\sqrt{2}} & \text{if } k=0,\\ \frac{\Psi_k(x)-\Psi_{k-1}(x)}{\sqrt{2}} & \text{if } k\geq1.\end{cases}$$

so we can rewrite $\mathbf{A}_m=\mathbf{I}_m-\frac{\lambda}{c}\mathbf{T}_m(\mathbb{E}[D(X)])$ and $\widehat{\mathbf{A}}_m=\mathbf{I}_m-\frac{1}{cT}\sum_{i=1}^{N_T}\mathbf{T}_m(D(X_i))$, with $\mathbf{T}_m(\bullet)$ defined by (3.34). Now, the difference between $\widehat{\mathbf{A}}_m$ and $\mathbf{A}_m$ can be decomposed as:

$$\widehat{\mathbf{A}}_m-\mathbf{A}_m=\frac{1}{cT}\sum_{i=1}^{N_T}\big\{\mathbf{T}_m(D(X_i))-\mathbb{E}\big[\mathbf{T}_m(D(X_i))\big]\big\}+\frac{N_T}{cT}\mathbf{T}_m\big(\mathbb{E}[D(X)]\big)-\frac{\lambda}{c}\mathbf{T}_m(\mathbb{E}[D(X)]).\tag{3.35}$$

The next lemma gives a control on the first term in the decomposition (3.35).

Lemma 3.7.9. *Let* $\mathbf{S}_n \coloneqq \sum_{i=1}^n \mathbf{Z}_i$, *with* $\mathbf{Z}_i \coloneqq \mathbf{T}_m(D(X_i)) - \mathbb{E}\big[\mathbf{T}_m(D(X_i))\big]$. *Then for* $p \geq 1$ *and* $\log m \geq p$, *we have:*

$$\mathbb{E}\|\mathbf{S}_n\|_{\mathrm{op}}^{2p} \leq C(p)\Big[(n\mu m \log m)^p + (m\log m)^{2p}\Big],$$

with $C(p)$ *a constant depending on* p.

Proof. We want to apply Theorem 1.4.12. First, we need upper bounds on $\|\mathbf{Z}_i\|_{\mathrm{op}}$ and $\lambda_{\max}(\mathbb{E}[\mathbf{S}_n^\top \mathbf{S}_n])$.

- Bound on $\|\mathbf{Z}_i\|_{\mathrm{op}}$:

$$\begin{aligned}
\|\mathbf{Z}_i\|_{\mathrm{op}} &= \sup_{\|x\|_{\ell^2}\leq 1} \big\|\big(\mathbf{T}_m(D(X_i)) - \mathbb{E}[\mathbf{T}_m(D(X_i))]\big)x\big\|_{\ell^2} \\
&= \sup_{\|x\|_{\ell^2}\leq 1} \|(D(X_i) - \mathbb{E}[D(X_i)]) * x\|_{\ell^2} \\
&\leq \|D(X_i) - \mathbb{E}[D(X_i)]\|_{\ell^1} \\
&\leq \sqrt{2}\sum_{k=0}^{m-1} |\Psi_k(X_i) - \mathbb{E}[\Psi_k(X_i)]| \\
&\leq 2\sqrt{2}\sum_{k=0}^{m-1} \|\Psi_k\|_\infty
\end{aligned}$$

  By Theorem A.3.1, there exists an absolute constant $C > 0$ such that for all k, we have $\|\Psi_k\|_\infty \leq C$, hence $\|\mathbf{Z}_i\|_{\mathrm{op}} \leq C2\sqrt{2}m$.

- Bound on $\lambda_{\max}\big(\mathbb{E}[\mathbf{S}_n^\top \mathbf{S}_n]\big)$:

$$\begin{aligned}
\lambda_{\max}\big(\mathbb{E}[\mathbf{S}_n^\top \mathbf{S}_n]\big) &= \sup_{\|x\|_{\ell^2}=1} x^\top \mathbb{E}[\mathbf{S}_n^\top \mathbf{S}_n] x \\
&= n \sup_{\|x\|_{\ell^2}=1} x^\top \mathbb{E}[\mathbf{Z}_1^\top \mathbf{Z}_1] x \\
&= n \sup_{\|x\|_{\ell^2}=1} \mathbb{E}\big[\|\mathbf{Z}_1 x\|_{\ell^2}^2\big] \\
&= n \sup_{\|x\|_{\ell^2}=1} \mathbb{E}\big[\|(D(X_1) - \mathbb{E}[D(X_1)]) * x\|_{\ell^2}^2\big].
\end{aligned}$$

  If $x \in \mathbb{R}^m$, we have:

$$\begin{aligned}
\mathbb{E}\big[\|(D(X_1) - \mathbb{E}[D(X_1)]) * x\|_{\ell^2}^2\big] &= \sum_{j=0}^{m-1} \mathbb{E}\Big[(\{D(X_1) - \mathbb{E}[D(X_1)]\} * x)_j^2\Big] \\
&= \sum_{j=0}^{m-1} \mathrm{Var}\big[(D(X_1) * x)_j\big] \leq \sum_{j=0}^{m-1} \mathbb{E}\Big[(D(X_1) * x)_j^2\Big],
\end{aligned}$$

and by Cauchy–Schwarz inequality:

$$(D(X_1) * x)_j^2 \leq \left(\sum_{k=0}^{j} D_k(X_1)^2\right)\left(\sum_{k=0}^{j} x_k^2\right)$$
$$\leq \|x\|_{\ell^2}^2 \sum_{k=0}^{j} \Psi_k(X_1)^2 \leq \|x\|_{\ell^2}^2 \|\mathbf{1}_{X_1>\bullet}\|_{\mathrm{L}^2}^2 = \|x\|_{\ell^2}^2 X_1,$$

because $\Psi_k(X_1) = \langle \mathbf{1}_{X_1>\bullet}, \psi_k \rangle$ and (ψ_k) is an orthonormal basis of $\mathrm{L}^2(\mathbb{R}_+)$. Hence, we obtain $\lambda_{\max}\big(\mathbb{E}[\mathbf{S}^\top \mathbf{S}]\big) \leq nm\mu$.

We want apply Theorem 1.4.12 to our matrix $\mathbf{S}_n$, which is not Hermitian. We use the following trick, called the Paulsen dialtation. For $\mathbf{M}$ a rectangular matrix, we define:

$$\mathbf{M} \mapsto \mathscr{H}(\mathbf{M}) = \begin{pmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\dagger & 0 \end{pmatrix},$$

where $\mathbf{M}^\dagger$ denotes the conjugate transpose of $\mathbf{M}$. Now, $\mathscr{H}(\mathbf{M})$ is an Hermitian matrix, and:

$$\mathscr{H}(\mathbf{M})^2 = \begin{pmatrix} \mathbf{M}\mathbf{M}^\dagger & 0 \\ 0 & \mathbf{M}^\dagger\mathbf{M} \end{pmatrix},$$

hence $\lambda_{\max}\big(\mathscr{H}(\mathbf{M})^2\big) = \|\mathbf{M}\|_{\mathrm{op}}^2$ and $\lambda_{\max}\big(\mathscr{H}(\mathbf{M})\big) = \|\mathbf{M}\|_{\mathrm{op}}$. We can now apply Theorem 1.4.12: for $\mathbf{M} = \mathbf{S}_n$, we have:

$$\mathscr{H}(\mathbf{S}_n) = \begin{pmatrix} 0 & \sum_i \mathbf{Z}_i \\ \sum_i \mathbf{Z}_i^\top & 0 \end{pmatrix} = \sum_i \begin{pmatrix} 0 & \mathbf{Z}_i \\ \mathbf{Z}_i^\top & 0 \end{pmatrix} = \sum_i \mathscr{H}(\mathbf{Z}_i),$$

thus for $p \geq 1$ and $r \geq \max(2p, 2\log m)$, we get that:

$$\begin{aligned}
\left(\mathbb{E}\|\mathbf{S}_n\|_{\mathrm{op}}^{2p}\right)^{1/2p} &= \left(\mathbb{E}\lambda_{\max}\left(\mathscr{H}\left(\sum_i \mathbf{Z}_i\right)\right)^{2p}\right)^{1/2p} \\
&\leq \sqrt{\mathrm{e}r}\,\lambda_{\max}^{1/2}\left(\sum_i \mathbb{E}\mathscr{H}(\mathbf{Z}_i)^2\right) + 2\mathrm{e}r\left(\mathbb{E}\max_i \lambda_{\max}\big(\mathscr{H}(\mathbf{Z}_i)\big)^{2p}\right)^{1/2p} \\
&\leq \sqrt{\mathrm{e}r\,\lambda_{\max}\big(\mathbb{E}\mathbf{S}_n^\top \mathbf{S}_n\big)} + 2\mathrm{e}r\left(\mathbb{E}\max_i \|\mathbf{Z}_i\|_{\mathrm{op}}^{2p}\right)^{1/2p} \\
&\leq \sqrt{\mathrm{e}r\,nm\mu} + C4\sqrt{2}\mathrm{e}r\,m.
\end{aligned}$$

If $\log m \geq p$, then $r = 2\log m$ and we get:

$$\mathbb{E}\|\mathbf{S}_n\|_{\mathrm{op}}^{2p} \lesssim 2^{2p-1}(n\mu\, m\log m)^p + 2^{6p-1}(m\log m)^{2p}. \qquad \square$$

Now we can prove Proposition 3.3.4.

Proof of Proposition 3.3.4. From the decomposition (3.35), we get:

$$\mathbb{E}\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}^{2p} \leq 2^{2p-1}\frac{1}{(cT)^{2p}}\mathbb{E}\|\mathbf{S}_{N_T}\|_{\mathrm{op}}^{2p} + 2^{2p-1}\frac{\mathbb{E}|N_T - \lambda T|^{2p}}{(cT)^{2p}}\|\mathbf{T}_m(\mathbb{E}[D(X)])\|_{\mathrm{op}}^{2p}.$$

For the first term, we apply Lemma 3.7.9 conditional on N_T:

$$\begin{aligned}\frac{1}{(cT)^{2p}}\mathbb{E}\|\mathbf{S}_{N_T}\|_{\mathrm{op}}^{2p} &\leq C(p)\left[\frac{\mathbb{E}[N_T^p]\mu^p(m\log m)^p}{(cT)^{2p}} + \left(\frac{m\log m}{cT}\right)^{2p}\right] \\ &= C(p)\left[\mu^p\mathbb{E}\left[\left(\frac{N_T}{cT}\right)^p\right]\left(\frac{m\log m}{cT}\right)^p + \left(\frac{m\log m}{cT}\right)^{2p}\right],\end{aligned}$$

with $\mathbb{E}\left[\left(\frac{N_T}{cT}\right)^p\right] = \mathrm{O}(\lambda^p)$. For the second term, we know from Corollary 3.7.3 that $\mathbb{E}[(N_T - \lambda T)^{2p}] = \mathrm{O}(\lambda^p T^p)$, and:

$$\begin{aligned}\|\mathbf{T}_m(\mathbb{E}[D(X)])\|_{\mathrm{op}} \leq \sum_{k=0}^{m-1}|\mathbb{E}[D_k(X)]| &\leq \sqrt{2}\sum_{k=0}^{m-1}|\mathbb{E}[\Psi_k(X)]| \\ &\leq \sqrt{2m}\left(\sum_{k=0}^{m-1}\mathbb{E}\left[\Psi_k(X)^2\right]\right)^{1/2} \\ &= \sqrt{2m}\left(\mathbb{E}\left[\sum_{k=0}^{m-1}\langle \mathbf{1}_{X>\bullet}, \psi_k\rangle^2\right]\right)^{1/2} \leq \sqrt{2m\mu},\end{aligned}$$

thus:

$$\frac{\mathbb{E}|N_T - \lambda T|^{2p}}{(cT)^{2p}}\|\mathbf{T}_m(\mathbb{E}[D(X)])\|_{\mathrm{op}}^{2p} \leq \frac{\mathrm{O}(\lambda^p)}{T^p}\mu^p m^p. \qquad \square$$

Proof of Theorem 3.3.5

The following results are based on the proofs of Lemma 3.1 and Corollary 3.2 in Comte & Mabon (2017).

Proposition 3.7.10. *If $m\log m \leq cT$, then it holds:*

$$\mathbb{E}\|\widetilde{\mathbf{A}}_{m,2}^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p} \leq C(p,\lambda)\left(\mu^p\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}\right) \wedge \left((\mu^p + \mu^{2p})\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{4p}\left(\frac{m\log m}{cT}\right)^p\right),$$

with $C(p,\lambda) = \mathrm{O}(\lambda^p \vee \lambda^{2p})$.

Proof. We decompose the expectation according to the event $\Delta_m := \{\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}}^1 \leq \frac{cT}{m\log m}\}$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\mathrm{op}}^{2p}\right] &= \mathbb{E}\left[\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}\mathbf{1}_{\Delta_m^c} + \|\widehat{\mathbf{A}}_m^{-1}(\mathbf{A}_m - \widehat{\mathbf{A}}_m)\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}\mathbf{1}_{\Delta_m}\right] \\ &= \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}\mathbb{P}\left[\Delta_m^c\right] + \mathbb{E}\left[\|\widehat{\mathbf{A}}_m^{-1}(\mathbf{A}_m - \widehat{\mathbf{A}}_m)\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}\mathbf{1}_{\Delta_m}\right]. \quad (3.36)\end{aligned}$$

We now give two bounds on (3.36), depending on the value of $\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}$.

• **First case:** $\|\mathbf{A}_m^{-1}\|_{\text{op}} > \frac{1}{2}\sqrt{\frac{cT}{m\log m}}$.
Starting from Equation (3.36) and using the set Δ_m^2, we have that:

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\text{op}}^{2p}\Big] &\leq \|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}+\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{E}\Big[\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p}\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]\\ &\leq \|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}+\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\left(\frac{cT}{m\log m}\right)^p\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\Big].\end{aligned}$$

We apply Proposition 3.3.4 and get:

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\text{op}}^{2p}\Big] &\leq \|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}+\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\left(\frac{cT}{m\log m}\right)^p C(p,\lambda)\mu^p\left(\frac{m\log m}{cT}\right)^p\\ &\leq (1+C(p,\lambda)\mu^p)\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p},\end{aligned}$$

with $C(p,\lambda)=\mathrm{O}(\lambda^p)$.
• **Second case:** $\|\mathbf{A}_m^{-1}\|_{\text{op}} < \frac{1}{2}\sqrt{\frac{cT}{m\log m}}$.
Starting from (3.36) again, we get:

$$\mathbb{E}\Big[\|\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\text{op}}^{2p}\Big] \leq \|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{P}\big[\Delta_m^c\big]+\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big].$$

Let us give an upper bound on $\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]$. First we notice that:

$$\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p} \leq 2^{2p-1}\|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}+2^{2p-1}\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}.$$

Applying Proposition 3.3.4, we get:

$$\begin{aligned}&\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]\\ &\leq 2^{2p-1}\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]+2^{2p-1}\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]\\ &\leq 2^{2p-1}\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]+2^{2p-1}\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\mathbb{E}\Big[\|\mathbf{A}_m-\widehat{\mathbf{A}}_m\|_{\text{op}}^{4p}\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}^{2p}\mathbf{1}_{\Delta_m}\Big]\\ &\leq C(p,\lambda)\mu^p\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\left(\frac{m\log m}{cT}\right)^p+\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\left(\frac{cT}{m\log m}\right)^p C(2p,\lambda)\mu^{2p}\left(\frac{m\log m}{cT}\right)^{2p}\\ &\leq C'(p,\lambda)(\mu^p+\mu^{2p})\|\mathbf{A}_m^{-1}\|_{\text{op}}^{2p}\left(\frac{m\log m}{cT}\right)^p, \qquad (3.37)\end{aligned}$$

with $C'(p,\lambda)=\mathrm{O}(\lambda^p\vee\lambda^{2p})$.
Now let us give an upper bound on $\mathbb{P}[\Delta_m^{\mathrm{c}}]=\mathbb{P}\Big[\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}>\sqrt{\frac{cT}{m\log m}}\Big]$. From the triangular inequality, we get:

$$\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}} \leq \|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\text{op}}+\|\mathbf{A}_m^{-1}\|_{\text{op}}$$

we obtain:

$$\mathbb{P}\left[\|\widehat{\mathbf{A}}_m^{-1}\|_{\text{op}}>\sqrt{\frac{cT}{m\log m}}\right] \leq \mathbb{P}\left[\|\widehat{\mathbf{A}}_m^{-1}-\mathbf{A}_m^{-1}\|_{\text{op}}>\sqrt{\frac{cT}{m\log m}}-\|\mathbf{A}_m^{-1}\|_{\text{op}}\right].$$

Moreover we have assumed that $\|\mathbf{A}_m^{-1}\|_{\mathrm{op}} < \frac{1}{2}\sqrt{\frac{cT}{m\log m}}$, so:

$$\mathbb{P}\left[\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} > \sqrt{\frac{cT}{m\log m}}\right] \leq \mathbb{P}\big[\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big].$$

Now let us rewrite this probability, as:

$$\begin{aligned}
&\mathbb{P}\big[\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big] \\
&\quad = \mathbb{P}\left[\big\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\qquad + \mathbb{P}\left[\big\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} \geq \frac{1}{2}\right\}\right] \\
&\quad \leq \mathbb{P}\left[\big\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\qquad + \mathbb{P}\left[\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} \geq \frac{1}{2}\right]. \qquad (3.38)
\end{aligned}$$

To control the second term, we apply Markov inequality and Proposition 3.3.4:

$$\begin{aligned}
\mathbb{P}\left[\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} \geq \frac{1}{2}\right] &\leq \mathbb{P}\left[\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}} \geq \frac{1}{2}\right] \\
&\leq C(p,\lambda)\mu^p\left(\frac{m\log m}{cT}\right)^p \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}. \qquad (3.39)
\end{aligned}$$

Next, to control the first term on the right hand side of Equation (3.38), we apply Theorem B.1.2:

$$\begin{aligned}
&\mathbb{P}\left[\big\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\quad \leq \mathbb{P}\left[\left\{\frac{\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}}\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2}{1 - \|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\right\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\quad \leq \mathbb{P}\left[\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\|_{\mathrm{op}} > \frac{1}{2}\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{-1}\right]. \qquad (3.40)
\end{aligned}$$

We apply Markov inequality again, along with Proposition 3.3.4:

$$\begin{aligned}
&\mathbb{P}\left[\big\{\|\widehat{\mathbf{A}}_m^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}} > \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}\big\} \cap \left\{\|\mathbf{A}_m^{-1}(\widehat{\mathbf{A}}_m - \mathbf{A}_m)\|_{\mathrm{op}} < \frac{1}{2}\right\}\right] \\
&\quad \leq C(p,\lambda)\mu^p\left(\frac{m\log m}{cT}\right)^p \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}.
\end{aligned}$$

So starting from Equation (3.38) and gathering Equations (3.39) and (3.40) gives:

$$\mathbb{P}\left[\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}} > \sqrt{\frac{cT}{m\log m}}\right] \leq C(p,\lambda)\mu^p\left(\frac{m\log m}{cT}\right)^p \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^{2p}, \qquad (3.41)$$

with $C(p,\lambda) = \mathrm{O}(\lambda^p)$.
Finally gathering Equations (3.37) with (3.41), we get that

$$\mathbb{E}\Big[\|\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\mathrm{op}}^{2p}\Big] \leq C(p,\lambda)(\mu^p + \mu^{2p})\left(\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^4 \frac{m\log m}{cT}\right)^p.$$

with $C(p,\lambda) = \mathrm{O}(\lambda^p \vee \lambda^{2p})$. □

The next proposition is a variant of the last one. It gives a better bound than applying directly Proposition 3.7.10 to $\mathbb{E}\Big[\|\widetilde{\mathbf{A}}_{m,2}^{-1} - \mathbf{A}_m^{-1}\|_{\mathrm{op}}^2 \|\mathbf{c}_m\|_{\ell^2}^2\Big]$.

Proposition 3.7.11. *If $m\log m \leq cT$, then it holds:*

$$\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,2}^{-1} - \mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2 \leq C(\lambda)\|\phi_m\|_{\mathrm{L}^2}^2\left(\mu \wedge \left\{(\mu+\mu^2)\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2 \frac{m\log m}{cT}\right\}\right),$$

with $C(\lambda) = \mathrm{O}(\lambda \vee \lambda^2)$.

Proof. The proof follows the lines of the proof of of Proposition 3.7.10, but starting from the following decomposition:

$$\begin{aligned}\mathbb{E}\|(\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1})\mathbf{c}_m\|_{\ell^2}^2 &= \|\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \mathbb{P}\big[\Delta_m^c\big] + \mathbb{E}\big[\|\widehat{\mathbf{A}}_m^{-1}(\mathbf{A}_m - \widehat{\mathbf{A}}_m)\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \mathbf{1}_{\Delta_m}\big] \\ &= \|\mathbf{a}_m\|_{\ell^2}^2 \mathbb{P}\big[\Delta_m^c\big] + \mathbb{E}\big[\|\widehat{\mathbf{A}}_m^{-1}(\mathbf{A}_m - \widehat{\mathbf{A}}_m)\mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \mathbf{1}_{\Delta_m}\big],\end{aligned}$$

It yields the following upper bound:

$$\mathbb{E}\big[\|(\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1})\mathbf{c}_m\|_{\ell^2}^2\big] \leq \|\mathbf{a}_m\|_{\ell^2}^2 \mathbb{P}\big[\Delta_m^c\big] + \|\mathbf{a}_m\|_{\ell^2}^2 \mathbb{E}\Big[\|\widehat{\mathbf{A}}_m^{-1}\|_{\mathrm{op}}^2 \|\mathbf{A}_m - \widehat{\mathbf{A}}_m\|_{\mathrm{op}}^2 \mathbf{1}_{\Delta_m}\Big].$$

Following the proof of Proposition 3.7.10, we get:

$$\mathbb{E}\big[\|(\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1})\mathbf{c}_m\|_{\ell^2}^2\big] \leq C(\lambda)\|\mathbf{a}_m\|_{\mathrm{L}^2}^2\left(\mu \wedge \left\{(\mu+\mu^2)\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2 \frac{m\log m}{cT}\right\}\right),$$

with $C(\lambda) = \mathrm{O}(\lambda \vee \lambda^2)$. □

Now we can prove Theorem 3.3.5.

Proof of Theorem 3.3.5. The decomposition bias-variance of the risk of $\widehat{\phi}_m^{\mathrm{Lag}_2}$ is:

$$\|\phi - \widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2 = \|\phi - \phi_m\|_{\mathrm{L}^2}^2 + \|\phi_m - \widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2.$$

In the proof of Theorem 3.2.5, we saw that:

$$\mathbb{E}\|\widehat{\mathbf{c}}_m - \mathbf{c}_m\|_{\ell^2}^2 = \mathbb{E}\|\widehat{h}_m - h_m\|_{\mathrm{L}^2}^2 \leq \frac{\lambda}{c^2 T}\mathbb{E}[W(X)].$$

We decompose the variance term in three terms:

$$\begin{aligned}&\mathbb{E}\|\phi_m - \widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2 = \mathbb{E}\|\widehat{\mathbf{a}}_m^{\mathrm{Lag}_2} - \mathbf{a}_m\|_{\ell^2}^2 = \mathbb{E}\|\widetilde{\mathbf{A}}_{m,2}^{-1}\widehat{\mathbf{c}}_m - \mathbf{A}_m^{-1}\mathbf{c}_m\|_{\ell^2}^2 \\ &\leq 3\,\mathbb{E}\|(\widetilde{\mathbf{A}}_{m,2}^{-1} - \mathbf{A}_m^{-1})\mathbf{c}_m\|_{\ell^2}^2 + 3\,\mathbb{E}\|(\mathbf{A}_m^{-1} - \widetilde{\mathbf{A}}_{m,2}^{-1})(\mathbf{c}_m - \widehat{\mathbf{c}}_m)\|_{\ell^2}^2 + 3\,\mathbb{E}\|\mathbf{A}_m^{-1}(\mathbf{c}_m - \widehat{\mathbf{c}}_m)\|_{\ell^2}^2.\end{aligned}$$

For the first term, we apply Proposition 3.7.10. For the second term, we use the fact that $\widetilde{\mathbf{A}}_{m,2}^{-1}$ and $\widehat{\mathbf{c}}_m$ are independent, and we apply Proposition 3.7.11:

$$\mathbb{E}\|(\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,2}^{-1})(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2 \le \mathbb{E}\|\mathbf{A}_m^{-1}-\widetilde{\mathbf{A}}_{m,2}^{-1}\|_{\mathrm{op}}^2 \times \mathbb{E}\|\mathbf{c}_m-\widehat{\mathbf{c}}_m\|_{\ell^2}^2 = \mathrm{O}\left(\frac{1}{T^2}\right).$$

For the third term:

$$\mathbb{E}\|\mathbf{A}_m^{-1}(\mathbf{c}_m-\widehat{\mathbf{c}}_m)\|_{\ell^2}^2 \le \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\,\mathbb{E}\|\widehat{\mathbf{c}}_m-\mathbf{c}_m\|_{\ell^2}^2 \le \|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\frac{\lambda}{c^2T}\mathbb{E}[W(X)].$$

We apply Lemma 3.3.2 and we obtain the following bound, with $C(\lambda)=\mathrm{O}(\lambda\vee\lambda^2)$:

$$\begin{aligned}\mathbb{E}\|\phi_m-\widehat{\phi}_m^{\mathrm{Lag}_2}\|_{\mathrm{L}^2}^2 &\le 3\,\|\mathbf{A}_m^{-1}\|_{\mathrm{op}}^2\,\frac{C(\lambda)}{cT}\left(\|\phi_m\|_{\mathrm{L}^2}^2(\mu+\mu^2)m\log(m)+\frac{\mathbb{E}[W(X)]}{c}\right)+\mathrm{O}\left(\frac{1}{T^2}\right)\\ &\le 12\,\frac{C(\lambda)}{cT(1-\theta)^2}\left(\|\phi_m\|_{\mathrm{L}^2}^2(\mu+\mu^2)m\log(m)+\frac{\mathbb{E}[W(X)]}{c}\right)+\mathrm{O}\left(\frac{1}{T^2}\right).\ \square\end{aligned}$$

3.7.3 Proofs of Section 3.4

Proof of Lemma 3.4.6. It follows from direct calculation using (3.6):

$$\begin{aligned}\langle F,\psi_k\rangle &= \int_0^{+\infty} C\exp(-\gamma x)\psi_k(x)\,\mathrm{d}x\\ &= C\sqrt{2}\sum_{j=0}^{k}\binom{k}{j}\frac{(-2)^j}{j!}\int_0^{+\infty}x^j\mathrm{e}^{-(1+\gamma)x}\,\mathrm{d}x\\ &= C\sqrt{2}\sum_{j=0}^{k}\binom{k}{j}\frac{(-2)^j}{(1+\gamma)^{j+1}}\\ &= \frac{C\sqrt{2}}{\gamma+1}\left(1-\frac{2}{\gamma+1}\right)^k = \frac{C\sqrt{2}}{\gamma+1}\left(\frac{\gamma-1}{\gamma+1}\right)^k.\end{aligned}$$

Since γ is positive, we have $\left|\frac{\gamma-1}{\gamma+1}\right|<1$ and we can compute the geometric series:

$$\sum_{k=m}^{+\infty}\langle F,\psi_k\rangle^2 = \frac{2C}{(\gamma+1)^2}\,\frac{\left(\frac{\gamma-1}{\gamma+1}\right)^{2m}}{1-\left(\frac{\gamma-1}{\gamma+1}\right)^2} = \frac{C^2}{2\gamma}\left(\frac{\gamma-1}{\gamma+1}\right)^{2m}.\qquad\square$$

Proof of Proposition 3.4.7. For the ruin probability and the Laplace transform of the ruin time, we start from (3.19) and we apply Lemma 3.4.6. For the expected jump size causing the ruin, we also start from (3.19) and we write $\phi=F_1+F_2$ with:

$$F_1(u) := \mu(1+2\theta)\mathrm{e}^{-\frac{1-\theta}{\mu}u}\mathbf{1}_{u>0},\quad F_2(u) := \mu\mathrm{e}^{-u/\mu}\mathbf{1}_{u>0}.$$

Hence, $\|\phi - \phi_m\|^2_{\mathrm{L}^2} \leq 2\sum_{k=m}^{+\infty}\langle F_1, \psi_k\rangle^2 + 2\sum_{k=m}^{+\infty}\langle F_2, \psi_k\rangle^2$. We apply Lemma 3.4.6:

$$\begin{aligned}
\|\phi - \phi_{m_1}\|^2_{\mathrm{L}^2} &\leq \frac{\mu^2(1+2\theta)^2}{\frac{1-\theta}{\mu}}\left(\frac{\frac{1-\theta}{\mu}-1}{\frac{1-\theta}{\mu}+1}\right)^{2m} + \mu^3\left(\frac{\frac{1}{\mu}-1}{\frac{1}{\mu}+1}\right)^{2m} \\
&= \frac{\mu^3(1+2\theta)^2}{1-\theta}\left(\frac{1-\theta-\mu}{1-\theta+\mu}\right)^{2m} + \mu^3\left(\frac{1-\mu}{1+\mu}\right)^{2m} \\
&\leq \frac{\mu^3(1+2\theta)^2}{1-\theta}\left(\left|\frac{1-\theta-\mu}{1-\theta+\mu}\right| \vee \left|\frac{1-\mu}{1+\mu}\right|\right)^{2m}. \qquad \square
\end{aligned}$$

Acknowledgments

I want to thank fedja for their help with Theorem A.3.1 regarding the primitives of the Laguerre functions, and Emmanuel Rio for his hints for Lemma 3.7.1.

Chapter 4

Nonparametric Multiple Regression on Non-compact Domains

This chapter is a modified version of a paper submitted in 2022 to *Annals of the Institute of Statistical Mathematics* that is under review at the moment I am writing this thesis.

Contents

4.1 Introduction

We consider the following random design regression model:

$$Y_i = b(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the variables $\boldsymbol{X}_i \in \mathbb{R}^p$ are independent but not necessarily identically distributed, the noise variables $\varepsilon_i \in \mathbb{R}$ are i.i.d. centered with finite variance σ^2 and

independent from the $\boldsymbol{X}_i$s, and $b\colon \mathbb{R}^p \to \mathbb{R}$ is a regression function. We seek to recover the function b on a domain $A \subseteq \mathbb{R}^p$ from the observations $(\boldsymbol{X}_i, Y_i)_{i=1,\dots,n}$.

More precisely, we consider the following framework. We assume that the variance of the noise σ^2 is known. We assume that the variables $\boldsymbol{X}_i$ are independent but not identically distributed, we call μ_i the distribution of $\boldsymbol{X}_i$, but we do not assume that μ_i is known. However, we fix ν a reference measure on A and we assume that $\mu := \frac{1}{n}\sum_{i=1}^n \mu_i$ admits a bounded density with respect to ν, so that we have $\mathrm{L}^2(A,\mu) \subseteq \mathrm{L}^2(A,\nu)$. In particular, this assumption implies that $\mathrm{supp}(\mu) \subseteq A$. Finally, we consider domains $A \subseteq \mathbb{R}^p$ of the form $A_1 \times \cdots \times A_p$ where $A_k \subseteq \mathbb{R}$ and we consider a measure ν on A that is of the form $\nu_1 \otimes \cdots \otimes \nu_p$ with ν_k supported on A_k. Our goal is to estimate the regression function b on the domain A and to control the expected error with respect to the norm $\|\cdot\|_\mu$ associated with the distribution of the $\boldsymbol{X}_i$s:

$$\forall\, t \in \mathrm{L}^2(A,\mu), \quad \|t\|_\mu^2 := \int_A t(\boldsymbol{x})^2 \,\mathrm{d}\mu(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \int_A t(\boldsymbol{x})^2 \,\mathrm{d}\mu_i(\boldsymbol{x}).$$

We can interpret the error with respect to this norm as a prediction risk:

$$\forall\, \hat{b} \text{ estimator}, \quad \|b-\hat{b}\|_\mu^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{X}\sim\mu_i}\Big[\big(b(\boldsymbol{X})-\hat{b}(\boldsymbol{X})\big)^2 \,\Big|\, \boldsymbol{X}_1,\dots,\boldsymbol{X}_n\Big],$$

which is the mean quadratic error of a new observation drawn uniformly from one of the distributions μ_i.

Nonparametric regression problems have a long history, and a large number of methods have been proposed. In this introduction, we focus on two main families of methods: kernel estimators and projection estimators. For reference books on the subject, see Efromovich (1999) regarding the projection method and Györfi *et al.* (2002) for the kernel method.

The classical estimator of Nadaraya (1964) and Watson (1964) consists of a quotient of estimators $\widehat{bf}/\hat{f}$, where $\widehat{bf}$ and $\hat{f}$ are kernel estimators of the functions bf and f (the function f being the common density of the $\boldsymbol{X}_i$s in the i.i.d case). This estimator can also be interpreted as locally fitting a constant by averaging the Y_is, the locality being determined by the kernel, see the book of Györfi *et al.* (2002) or Tsybakov (2009). This method can then be generalized by replacing the local constant by a local polynomial, leading to the so-called *local polynomial estimator*.

The main drawback of the Nadaraya–Watson estimator is that it relies on an estimator of the density of the $\boldsymbol{X}_i$s. As such, the rate of convergence depends on the regularity of f, and two smoothing parameters have to be chosen. A popular solution is to choose the same bandwidth for both estimators using leave-one-out cross validation. This method works well in practice and has been proven consistent by Hardle & Marron (1985) (see also Chapter 8 in Györfi *et al.* (2002)). Recently, Comte & Marie (2021) have proposed to use the Penalized Comparison to Overfitting method (PCO), a bandwidth selection method developed by

Lacour *et al.* (2017) for kernel density estimation, to select separately the bandwidths of the numerator and the denominator of the Nadaraya–Watson estimator. Their estimator matches the performances of the single bandwidth CV estimator when the noise is high, but the latter is better when the noise is small. Other bandwidth selection methods exist such as plug-in or bootstrap; see Köhler *et al.* (2014) for an extensive survey and comparison of the different bandwidth selection methods for the local linear estimator.

Another approach is to use a projection estimator. The idea is to minimize a least squares contrast over finite-dimensional spaces of functions $\{S_{\boldsymbol{m}} : \boldsymbol{m} \in \mathscr{M}_n\}$ called *models*:

$$\hat{b}_{\boldsymbol{m}} := \operatorname*{arg\,min}_{t \in S_{\boldsymbol{m}}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - t(\boldsymbol{X}_i)\big)^2,$$

the model collection $\mathscr{M}_n$ being allowed to depend on the number of observations. This method overcomes the problems of the Nadaraya–Watson estimator: it does not need to estimate the density of the $\boldsymbol{X}_i$s, and only one model selection procedure is required. Moreover, it can provide a sparse representation of the estimator. This approach was developed in a fixed design setting by Birgé & Massart (1998), Barron *et al.* (1999) and Baraud (2000). In particular, the papers of Baraud (2000, 2002) provide a model selection procedure that optimizes the bias-variance compromise under weak assumptions on the moments of the noise distribution. They obtain an estimator that is adaptive both in the fixed and random design setting when the domain A is compact.

The non-compact case have been studied recently in the simple regression setting ($p = 1$) by Comte & Genon-Catalot (2020b,a). They use non-compactly supported bases, specifically the Hermite basis (supported on $\mathbb{R}$) and the Laguerre basis (supported on $\mathbb{R}_+$), to construct their estimator. Significant attention has been paid to these bases in the past years since they exhibit nice mathematical properties that are useful for solving inverse problems (Mabon, 2017; Comte & Genon-Catalot, 2018; Sacko, 2020). Non-compactly supported bases also avoid issues concerning the choice of support. When A is compact, the theory assumes it is fixed *a priori*. In practice, however, the support is generally determined using the data, although this dependency between data and support is not taken into account in the theoretical development. Working with a non-compact domain, for example $\mathbb{R}$ or $\mathbb{R}_+$, allows us to bypass this issue.

Concerning the regression problem, difficulties arise when we go from the compact case to the non-compact case. When A is compact, it is usual to assume that the density of the $\boldsymbol{X}_i$s is bounded from below by some positive constant f_0. In the non-compact case, this assumption fails. Instead, the study of the minimum eigenvalue of some random matrix must be done. This question has been studied in the simple regression case ($p = 1$) by Cohen *et al.* (2013) by using the matrix concentration inequalities of Tropp (2012). However, their results are obtained under the assumption that the regression function is bounded by a known quantity and they do not provide a model selection procedure.

We make the following contributions in our paper. We extend the results of Comte & Genon-Catalot (2020b) to the multiple regression case ($p \geq 2$) with more general assumptions on the design, and we improve their result on the oracle inequality under the empirical norm (see Theorem 4.4.1). Our work generalizes the results of Baraud (2002) to the non-compact case and improves their results in the compact case (see Theorem 4.4.4). We do so by combining the fixed design results of Baraud (2000) with a more refined study of the discrepancy between the empirical norm and the μ-norm. This discrepancy is expressed in terms of the deviation of the minimum eigenvalue of a random matrix, of which we control the probability with the concentration inequalities of Tropp (2012) and Gittens & Tropp (2011). Finally, our estimator is constructed as a projection estimator on a tensorized basis whose coefficients are computed using hypermatrix calculus and can be implemented in practice. This feasibility is illustrated in Section 4.5 which also shows that the procedure works well.

Outline of the paper In Section 4.2 we define the projection estimator. In Section 4.3 we study the probability that the empirical norm and the μ-norm depart from each other and we derive an upper bound on the μ-risk of our estimator. In Section 4.4 we propose a model selection procedure and we prove that it satisfies an oracle inequality both in empirical norm and in μ-norm. Finally, in Section 4.5 we study numerically the performance of our estimator. All the proofs are gathered in Section 4.6.

Notations

- $\mathbb{E}_{\boldsymbol{X}} \coloneqq \mathbb{E}[\cdot \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n]$, $\mathbb{P}_{\boldsymbol{X}} \coloneqq \mathbb{P}[\cdot \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n]$, $\mathrm{Var}_{\boldsymbol{X}} \coloneqq \mathrm{Var}(\cdot \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$.
- If π is a measure on A, we write $\|\cdot\|_\pi$ and $\langle \cdot, \cdot \rangle_\pi$ the norm and the inner product weighted by the measure π.
- We denote by $\langle \cdot, \cdot \rangle_n$ and $\|\cdot\|_n$ the empirical inner product and the empirical norm[1], defined as $\langle t, s \rangle_n \coloneqq \frac{1}{n} \sum_{i=1}^n t(\boldsymbol{X}_i) s(\boldsymbol{X}_i)$ and $\|t\|_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n t(\boldsymbol{X}_i)^2$. If $\mathbf{u} \in \mathbb{R}^n$ is a vector, we also write $\|\mathbf{u}\|_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n u_i^2$.

4.2 Projection estimator

In our setting, the domain is a Cartesian product $A = A_1 \times \cdots \times A_p$ and $\nu = \nu_1 \otimes \cdots \otimes \nu_p$ where ν_k is supported on A_k. For each $i \in \{1, \dots, p\}$, we consider $(\varphi_j^i)_{j \in \mathbb{N}}$ an orthonormal basis of $\mathrm{L}^2(A_i, \mathrm{d}\nu_i)$ and we form an orthonormal basis of $\mathrm{L}^2(A, \mathrm{d}\nu)$ by tensorization:

$$\forall \boldsymbol{j} \in \mathbb{N}^p, \quad \forall \boldsymbol{x} \in A, \quad \varphi_{\boldsymbol{j}}(\boldsymbol{x}) \coloneqq (\varphi_{j_1}^1 \otimes \cdots \otimes \varphi_{j_p}^p)(\boldsymbol{x}) \coloneqq \varphi_{j_1}^1(x_1) \times \cdots \times \varphi_{j_p}^p(x_p).$$

[1] in general it is a semi-norm but we will only consider subspaces on which it is a norm.

For $\boldsymbol{m} \in \mathbb{N}_+^p$, we set $S_{\boldsymbol{m}} \coloneqq \mathrm{Span}(\varphi_{\boldsymbol{j}} : \boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1})$ and we write $D_{\boldsymbol{m}} \coloneqq m_1 \cdots m_p$ its dimension. We estimate b by minimizing a least squares contrast on $S_{\boldsymbol{m}}$:

$$\hat{b}_{\boldsymbol{m}} \coloneqq \operatorname*{arg\,min}_{t \in S_{\boldsymbol{m}}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - t(\boldsymbol{X}_i)\big)^2.$$

If we expand $\hat{b}_{\boldsymbol{m}}$ on the basis $(\varphi_{\boldsymbol{j}})_{\boldsymbol{j} \in \mathbb{N}^p}$, this problem can be written as:

$$\hat{b}_{\boldsymbol{m}} = \sum_{\boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}} \hat{a}_{\boldsymbol{j}}^{(\boldsymbol{m})} \varphi_{\boldsymbol{j}}, \qquad \hat{\mathbf{a}}^{(\boldsymbol{m})} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^{\mathbf{m}}} \|\mathbf{Y} - \widehat{\Phi}_{\boldsymbol{m}} \times_p \mathbf{a}\|_{\mathbb{R}^n}^2, \tag{4.1}$$

where $\mathbf{Y} \coloneqq (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ and $\widehat{\Phi}_{\boldsymbol{m}} \in \mathbb{R}^{n \times \boldsymbol{m}}$ is defined as:

$$\forall\, i \in \{1, \ldots, n\}, \quad \forall\, \boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}, \quad \big[\widehat{\Phi}_{\boldsymbol{m}}\big]_{i,\boldsymbol{j}} \coloneqq \varphi_{\boldsymbol{j}}(\boldsymbol{X}_i).$$

Using Lemma B.1.1 in Appendix, the problem (4.1) has a unique solution if and only if $\widehat{\Phi}_{\boldsymbol{m}}$ is injective and in that case:

$$\begin{aligned} \hat{\mathbf{a}}^{(\boldsymbol{m})} &= (\widehat{\Phi}_{\boldsymbol{m}}^* \times_1 \widehat{\Phi}_{\boldsymbol{m}})^{-1} \times_p \widehat{\Phi}_{\boldsymbol{m}}^* \times_1 \mathbf{Y} \\ &= \frac{1}{n} \widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} \times_p \widehat{\Phi}_{\boldsymbol{m}}^* \times_1 \mathbf{Y}, \end{aligned}$$

where $[\widehat{\Phi}_{\boldsymbol{m}}^*]_{\boldsymbol{j},i} = [\widehat{\Phi}_{\boldsymbol{m}}]_{i,\boldsymbol{j}}$ and where $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is the Gram hypermatrix of $(\varphi_{\boldsymbol{j}})_{\boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}}$ relatively to the empirical inner product $\langle \cdot, \cdot \rangle_n$:

$$\forall\, \boldsymbol{j}, \boldsymbol{k} \leq \boldsymbol{m} - \mathbf{1}, \quad \big[\widehat{\mathbf{G}}_{\boldsymbol{m}}\big]_{\boldsymbol{j},\boldsymbol{k}} \coloneqq \langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_n.$$

Notice that $\widehat{\Phi}_{\boldsymbol{m}}$ is injective if and only if $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible, that is if and only if $\|\cdot\|_n$ is a norm on $S_{\boldsymbol{m}}$.

4.3 Bound on the risk of the estimator

Let us start with the classical bias-variance decomposition of the empirical risk. In our context this result is given by the next Proposition.

Proposition 4.3.1. *If $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible, then we have:*

$$\mathbb{E}_{\boldsymbol{X}} \|b - \hat{b}_{\boldsymbol{m}}\|_n^2 = \inf_{t \in S_m} \|b - t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

As a consequence, if $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible a.s, then we have:

$$\mathbb{E} \|b - \hat{b}_{\boldsymbol{m}}\|_n^2 \leq \inf_{t \in S_m} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

If we want to obtain a similar result for the μ-norm, we need to understand how the empirical norm can deviate from the μ-norm. More generally, we need

to understand the relations between the different norms we have on the subspace $S_{\boldsymbol{m}}$ ($\|\cdot\|_n$, $\|\cdot\|_\mu$, $\|\cdot\|_\nu$ and $\|\cdot\|_\infty$). It is well known that all norms are equivalent on finite dimensional spaces; our question concerns the constants in this equivalence. We introduce the following notation: if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two norms on a space S, we define:

$$K_\beta^\alpha(S) := \sup_{t\in S\setminus\{0\}} \frac{\|t\|_\alpha^2}{\|t\|_\beta^2},$$

and when $S = S_{\boldsymbol{m}}$, we use the notation $K_\beta^\alpha(\boldsymbol{m}) := K_\beta^\alpha(S_{\boldsymbol{m}})$. The next lemma gives the value of $K_\alpha^\beta(S)$ when the norms are Euclidean.

Lemma 4.3.2. *Let $(S, \langle\cdot,\cdot\rangle_\alpha)$ be a d-dimensional Euclidean vector space equipped with an orthonormal basis $(\phi_1,\dots,\phi_d)$. Let $\langle\cdot,\cdot\rangle_\beta$ be another inner product on E and let $\mathbf{G}$ be the Gram matrix of the basis $(\phi_1,\dots,\phi_d)$ relatively to $\langle\cdot,\cdot\rangle_\beta$, that is:*

$$\mathbf{G} := \Big[\langle\phi_j,\phi_k\rangle_\beta\Big]_{1\le j,k\le d}.$$

We have:

$$K_\alpha^\beta(S) = \|\mathbf{G}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{G}), \qquad K_\beta^\alpha(S) = \|\mathbf{G}^{-1}\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(\mathbf{G})}.$$

The proof of Lemma 4.3.2 is identical to the proof of Lemma 3.1 in Baraud (2000), so we leave it out.

The next lemma provides a way to compute $K_\alpha^\infty(S)$ from an orthonormal basis when $\|\cdot\|_\alpha$ is Euclidean. It is essentially the same as Lemma 1 in Birgé & Massart (1998).

Lemma 4.3.3. *Let S be a space of bounded functions on A such that $d := \dim(S)$ is finite. Let $\langle\cdot,\cdot\rangle_\alpha$ be an inner product on S. If $(\psi_1,\dots,\psi_d)$ is an orthonormal basis of S, then we have:*

$$K_\alpha^\infty(S) = \Big\|\sum_{j=1}^d \psi_j^2\Big\|_\infty.$$

The question we are interested in is how close are the norms $\|\cdot\|_n$ and $\|\cdot\|_\mu$ on $S_{\boldsymbol{m}}$. Following a similar idea of Cohen *et al.* (2013), let us define the event:

$$\forall\delta\in(0,1), \qquad \Omega_{\boldsymbol{m}}(\delta) := \left\{\forall t\in S_{\boldsymbol{m}},\ \|t\|_\mu^2 \le \frac{1}{1-\delta}\|t\|_n^2\right\} = \left\{K_n^\mu(\boldsymbol{m}) \le \frac{1}{1-\delta}\right\}. \quad (4.2)$$

The key decomposition of the μ-risk of $\hat{b}_{\boldsymbol{m}}$ is given by the following Proposition.

Proposition 4.3.4. *If $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible, then we have for all $\delta\in(0,1)$:*

$$\begin{aligned}\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \le \left(1+\frac{2}{1-\delta}\left[\frac{K_\mu^\infty(\boldsymbol{m})}{(1-\delta)n}\wedge 1\right]\right)\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_\mu^2 + \frac{2\sigma^2 D_{\boldsymbol{m}}}{(1-\delta)n}\\ + 2\|b\|_\mu^2\,\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] + \mathbb{E}\big[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2\,\mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big].\end{aligned}$$

where $K_n^\mu(\boldsymbol{m})$ and $K_\mu^\infty(\boldsymbol{m})$ are given by Lemmas 4.3.2 and 4.3.3.

We see that we need an upper bound on the probability of the event $\Omega_{\boldsymbol{m}}(\delta)^c$. The following proposition is a consequence of the matrix Chernoff bound of Tropp (2012) (Theorem 1.4.10 in Appendix) .

Proposition 4.3.5. *For all $\delta \in (0,1)$, we have:*

$$\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \leq D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\mu^\infty(\boldsymbol{m})}\right),$$

where $h(\delta) := \delta + (1-\delta)\log(1-\delta)$ and $K_\mu^\infty(\boldsymbol{m})$ is given by Lemma 4.3.3.

Remark 4.3.6. The quantity $K_\mu^\infty(\boldsymbol{m})$ is unknown but we have the following upper bound using Lemmas 4.3.2 and 4.3.3:

$$K_\mu^\infty(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m})\, K_\mu^\nu(\boldsymbol{m}) = \left(\sup_{\boldsymbol{x}\in A} \sum_{\boldsymbol{j}\leq \boldsymbol{m}-\mathbf{1}} \varphi_{\boldsymbol{j}}(\boldsymbol{x})^2\right) \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}.$$

The quantity $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}$ is still unkown but can be estimated by plugging in $\widehat{\mathbf{G}}_{\boldsymbol{m}}$.

For α a positive constant, let us consider the following model collection:

$$\mathscr{M}_{n,\alpha}^{(1)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_\nu^\infty(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\big) \leq \alpha \frac{n}{\log n} \right\}. \tag{4.3}$$

Gathering Propositions 4.3.4 and 4.3.5, we obtain the following bound on the μ-risk of $\hat{b}_{\boldsymbol{m}}$ when $\boldsymbol{m}$ belongs to $\mathscr{M}_{n,\alpha}^{(1)}$.

Theorem 4.3.7. *Let us assume that $b \in \mathrm{L}^{2r}(\mu)$ for some $r \in (1,+\infty]$ and let $r' \in [1,+\infty)$ be the conjugated index of r, that is: $\frac{1}{r} + \frac{1}{r'} = 1$. For all $\alpha \in (0, \frac{1}{2r'+1})$ and for all $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ we have:*

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n(\alpha,r') \inf_{t\in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + C'(\alpha,r')\,\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \frac{C''\big(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \alpha\big)}{n\log n},$$

where the constants $C_n(\alpha,r')$ and $C'(\alpha,r')$ are given by:

$$C_n(\alpha,r') := 1 + \frac{2}{1-\delta(\alpha,r')}\left(\frac{\alpha}{\big(1-\delta(\alpha,r')\big)\log n} \wedge 1\right), \qquad C'(\alpha,r') := \frac{2}{1-\delta(\alpha,r')},$$

and $\delta(\alpha,r') \in (0,1)$ tends to 1 as α tends to $\frac{1}{2r'+1}$.

Remark 4.3.8. Let us make some statements concerning the behavior of $C_n(\alpha,r')$ and $C'(\alpha,r')$:

- $C_n(\alpha,r')$ is bounded relatively to n;
- $C_n(\alpha,r') \geq 1$ and $C'(\alpha,r') \geq 2$;
- as $\alpha \to \frac{1}{2r'+1}$ with n fixed, $C_n(\alpha,r')$ and $C'(\alpha,r')$ tend to $+\infty$;
- as $n \to +\infty$ with α an r' fixed, $C_n(\alpha,r')$ tends to 1.

4.4 Adaptive estimator

We consider the empirical version of the model collection $\mathscr{M}_{n,\alpha}$ defined by (4.3):

$$\widehat{\mathscr{M}}_{n,\beta}^{(1)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; K_\nu^\infty(\boldsymbol{m}) \big(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1 \big) \leq \beta \frac{n}{\log n} \right\},$$

with β a positive constant. We choose $\hat{\boldsymbol{m}}_1 \in \widehat{\mathscr{M}}_{n,\beta}^{(1)}$ by minimizing the following penalized least squares criterion:

$$\hat{\boldsymbol{m}}_1 := \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(1)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right), \qquad \theta > 0. \tag{4.4}$$

Based on a result of Baraud (2000) for fixed design regression, we prove that $\hat{b}_{\hat{\boldsymbol{m}}_1}$ automatically optimizes the bias-variance compromise in empirical norm on $\mathscr{M}_{n,\alpha}$, up to a constant and a remainder term.

Theorem 4.4.1. *If $b \in \mathrm{L}^{2r}(\mu)$ for some $r \in (1, +\infty]$ and if $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$, then there exists a constant $\alpha_{\beta,r'} > 0$ depending on β and r' (the conjugated index of r) such that for all $\alpha \in (0, \alpha_{\beta,r'})$, the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}_1}$ with $\hat{\boldsymbol{m}}_1$ defined by* (4.4) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \leq C(\theta) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

where:

$$C(\theta) := (2 + 8\theta^{-1})(1+\theta), \qquad \Sigma(\theta, q) := C''(\theta, q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)},$$

and where the remainder term is given by:

$$R_n := C'(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},$$

with $\kappa(\alpha, \beta)$ a positive constant satisfying $\frac{\kappa(\alpha,\beta)}{r'} > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ as $\alpha \to \alpha_{\beta,r'}$.

Remark 4.4.2. The term $\Sigma(\theta, q)$ is finite if $q > 6$. Indeed, let $2\epsilon := (\frac{q}{2} - 2) - 1 > 0$, we have:

$$\sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)} = \sum_{d=1}^{+\infty} \mathrm{Card}\big\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\big|\, D_{\boldsymbol{m}} = d \big\} \times d^{-(\frac{q}{2}-2)} \leq \sum_{d=1}^{+\infty} \frac{\mathrm{o}(d^\epsilon)}{d^{1+2\epsilon}} < +\infty,$$

where we use Theorem B.2.2 in Appendix.

Remark 4.4.3. The constant $\alpha_{\beta,r'}$ is increasing with β and goes from 0 to $\frac{1}{2r'+1}$. It is also decreasing with r' (so increasing with r) and tends to 0 as $r' \to +\infty$ (as $r \to 1$).

To transfer the previous adaptive result from the empirical norm into the μ-norm, we use once again concentration inequalities on the matrix $\widehat{\mathbf{G}}_{\boldsymbol{m}}$. However, we need to make a distinction between the compact case and the non-compact case. Indeed, when A is compact, we can make the usual assumption that the density $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is bounded from below and apply the matrix Chernoff bound of Gittens & Tropp (2011), see Lemma 4.6.6. This lemma relies critically on the "bounded from below" assumption so it cannot work in the non-compact case.

To handle the non-compact case, we make use of the matrix Bernstein bound of Tropp (2012) instead (Theorem 1.4.11 in appendix), see Lemma 4.6.7. This inequality is different from the matrix Chernoff bounds we have used so far, so we have to consider smaller model collections to make it work. In the following, we consider two cases:

1. *Compact case.* We assume that there exists $f_0 > 0$ such that for all $x \in A$, $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) > f_0$. In that case, $\mathbf{G}_{\boldsymbol{m}}$ is always invertible and we have:

$$\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\nu^2}{\|t\|_\mu^2} \leq \frac{1}{f_0}. \tag{4.5}$$

2. *General case.* We consider smaller model collections:

$$\mathscr{M}_{n,\alpha}^{(2)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, K_\nu^\infty(\boldsymbol{m}) \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leq \alpha \frac{n}{\log n} \right\},$$
$$\widehat{\mathscr{M}}_{n,\beta}^{(2)} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, K_\nu^\infty(\boldsymbol{m}) \left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leq \beta \frac{n}{\log n} \right\},$$

where α and β are positive constants and we choose $\hat{\boldsymbol{m}}_2 \in \widehat{\mathscr{M}}_{n,\beta}^{(2)}$ as:

$$\hat{\boldsymbol{m}}_2 := \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(2)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right), \quad \theta > 0. \tag{4.6}$$

Theorem 4.4.4. *Let $r \in (1, +\infty]$, let $r' \in [1, +\infty)$ be its conjugated index and let us assume that b belongs to $\mathrm{L}^{2r}(\mu)$ and that $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$.*

• ***Compact case.*** *Let $f_0 > 0$ such that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \geq f_0$ for all $x \in A$, there exists $\beta_{f_0,r'} > 0$ such that for all $\beta \in (0, \beta_{f_0,r'})$, there exists $\alpha_{\beta,r'} > 0$ such that for all $\alpha \in (0, \alpha_{\beta,r'})$, the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}_1}$ with $\hat{\boldsymbol{m}}_1$ defined by* (4.4) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2 \leq C(\theta,\beta,r) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta,r)\sigma^2 \frac{\Sigma(\theta,q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\left(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\right) \left(n^{-\frac{\kappa(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,f_0)} (\log n)^{\frac{p-1}{r'}-1} \right),$$

with $\lambda(\beta, r, f_0) > 1$ *and* $\frac{\kappa(\alpha,\beta)}{r'} > 1$.

• ***General case***. *Let* $B := (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$, *there exists* $\beta_{B,r'} > 0$ *such that for all* $\beta \in (0, \beta_{B,r'})$, *there exists* $\tilde{\alpha}_{\beta,r'} > 0$ *such that for all* $\alpha \in (0, \tilde{\alpha}_{\beta,r'})$, *the following upper bound on the risk of the estimator* $\hat{b}_{\hat{\boldsymbol{m}}_2}$ *with* $\hat{\boldsymbol{m}}_2$ *defined by* (4.6) *holds:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2 \leq C(\theta,\beta,r) \inf_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}^{(2)}} \left(\inf_{t\in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta,r)\sigma^2 \frac{\Sigma(\theta,q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\big(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\big)\left(n^{-\frac{\tilde{\kappa}(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,B)} (\log n)^{\frac{p-1}{r'}-1} \right),$$

with $\lambda(\beta, r, B) > 1$ *and* $\frac{\tilde{\kappa}(\alpha,\beta)}{r'} > 1$.

This result shows that there is a range of values for the constant β that depends on the integrability of b and on f_0 (compact case) or $\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty$ (general case), such that for the μ-norm, the estimator $\hat{b}_{\hat{\boldsymbol{m}}}$ automatically optimizes the bias-variance trade-off (up to a constant and a rest) on $\mathscr{M}_{n,\alpha}$ for all α in a range that depends on β.

Remark 4.4.5. Theorem 4.4.4 improves previous results in the literature:

1. In the compact case, we improve the result of Baraud (2002). Indeed in this article, the model collections considered are built by picking an "envelope model", that is a linear space $\mathscr{S}_n$ with finite dimension N_n, whose all models are a subspace. Their assumptions concern the space $\mathscr{S}_n$: they assume that $K_\nu^\infty(\mathscr{S}_n) \leq C^2 N_n$ for some constant $C > 0$ and they require that $N_n \leq C^{-1}\sqrt{n/(\log n)^3}$. In comparison, our procedure avoids the choice *a priori* of an envelope model, and uses a looser constraint on the dimension of the models.

2. In the non-compact case, we extend the results of Comte & Genon-Catalot (2020b) to the case $p \geq 2$ without losing much on the assumptions: their result requires a moment of order 6 on the noise whereas our result is obtained with a moment of order q, with $q > 6$. We also generalize their result by considering a non i.i.d. design and by using a more general moment assumption on the regression function.

4.5 Numerical illustrations

In this section, we compare our estimator with the Nadaraya–Watson estimator on simulated data in the case $p = 1$ and $p = 2$.

Regression function We consider the following regression functions:

1. $b_1(x) = \exp((x-1)^2) + \exp((x+1)^2)$,
2. $b_2(x) := \frac{1}{1+x^2}$,
3. $b_3(x) := x\cos(x)$,
4. $b_4(x) := |x|$,
5. $b_5(x_1, x_2) := \exp(-\frac{1}{2}[(x_1-1)^2 + (x_2-1)^2]) + \exp(-\frac{1}{2}[(x_1+1)^2 + (x_2+1)^2])$,
6. $b_6(x_1, x_2) := \frac{1}{1+x_1^2+x_2^2}$,
7. $b_7(x_1, x_2) := \cos(x_1)\sin(x_2)$,
8. $b_8(x_1, x_2) := |x_1 x_2|$.

The functions b_2 and b_6 are smooth bounded functions and have a unique maximum at 0, so they should be an easy case. The functions b_1 and b_5 are smooth and bounded with two maximums. The functions b_3 and b_7 are smooth oscillating functions. Finally the functions b_4 and b_8 are not smooth nor bounded, and should be a harder case.

Distribution of the $\boldsymbol{X}_i$s For the sake of simplicity, we consider the case where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. and have a density with respect to Lebesgue measure (i.e. $\nu =$ Leb). For the case $p = 1$, we consider the following distributions for X:

1. $X \sim \mathcal{N}(0,1)$,
2. $X \sim$ Laplace.

Both distributions are symmetric and centered at 0, but the normal distribution is more concentrated around its mean than the Laplace distribution. For the case $p = 2$, we use independent marginals for the distribution of $\boldsymbol{X}$:

1. $\boldsymbol{X} \sim \mathcal{N}(0,1) \otimes \mathcal{N}(0,1)$,
2. $\boldsymbol{X} \sim$ Laplace $\otimes$ Laplace.

Noise distribution We consider the normal distribution: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The variance σ^2 is chosen such that the signal-to-noise ratio is the same for each choice of regression function and distribution of $\boldsymbol{X}$, where we define the signal-to-noise ratio as:

$$\mathrm{SNR} := \frac{\|b\|_\mu^2}{\sigma^2}.$$

We consider the following values of the SNR:

1. *High noise:* SNR = 2,
2. *Low noise:* SNR = 20.

Parameters of the projection estimator Since the distributions of $\boldsymbol{X}$ are supported on $\mathbb{R}$ or $\mathbb{R}^2$, we choose the Hermite basis. The Hermite functions are defined as:

$$\varphi_j(x) \coloneqq c_j\, H_j(x)\, \mathrm{e}^{-\frac{x^2}{2}}, \quad H_j(x) \coloneqq (-1)^j \mathrm{e}^{x^2} \frac{\mathrm{d}^j}{\mathrm{d}x^j}\left[\mathrm{e}^{-x^2}\right], \quad c_j \coloneqq \left(2^j j!\sqrt{\pi}\right)^{-1/2}.$$

and form a basis of $\mathrm{L}^2(\mathbb{R})$. We form a basis of $\mathrm{L}^2(\mathbb{R}^2)$ by tensorizing the Hermite basis as explained in Section 4.2. We choose the parameter $\hat{\boldsymbol{m}}$ with the model selection procedure (4.6). This procedure requires two additional parameters: the constant θ in the penalty and the constant β in the model collection $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$.

We choose β such that the model collection $\widehat{\mathscr{M}}_{n,\beta}^{(2)}$ is not too small, especially for small sample sizes. Indeed, we find that the operator norm $\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}$ can grow very fast with $\boldsymbol{m}$, which can result in model collections with very few models. In our case, we choose $\beta = 10^4$.

The constant $\kappa \coloneqq (1+\theta)$ in front of the penalty is chosen following the "minimum penalty heuristic" (Arlot & Massart, 2009). On several preliminary simulations, we compute the selected dimension $D_{\hat{\boldsymbol{m}}}$ as a function of κ and we find $\kappa_{\min}$ such that for $\kappa < \kappa_{\min}$ the dimension is too high and for $\kappa > \kappa_{\min}$ it is acceptable. Then, we choose $\kappa_\star = 2\kappa_{\min}$. In our case, we find $\kappa_\star = 2$ when $p = 1$ and $p = 2$.

Nadaraya–Watson estimator Let us define the Nadaraya–Watson estimator in the case $p = 1$. For all $h \in (0,1)$, let K_h be the pdf of the $\mathcal{N}(0,h)$ distribution. The Nadaraya–Watson estimator is defined as:

$$\forall x \in \mathbb{R}, \qquad \hat{b}_h^{\mathrm{NW}}(x) \coloneqq \frac{\sum_{i=1}^n Y_i\, K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}.$$

The bandwidth h is selected by leave-one-out cross validation, that is:

$$\hat{h} \coloneqq \operatorname*{arg\,min}_h \sum_{i=1}^n \left(Y_i - \hat{b}_{h,-i}^{\mathrm{NW}}(X_i)\right)^2,$$

where $\hat{b}_{h,-i}^{\mathrm{NW}}$ is the Nadaraya–Watson estimator computed from the data set:

$$\left\{(X_j, Y_j) : j \in \{1, \dots n\} \setminus \{i\}\right\}.$$

In the case $p = 2$, the definition of the estimator is the same but with a couple of bandwidths $\boldsymbol{h} = (h_1, h_2) \in (0,1)^2$, and with $K_{\boldsymbol{h}}$ the pdf of the $\mathcal{N}_2(\mathbf{0}, \mathbf{H})$ distribution, where $\mathbf{H} \coloneqq \mathrm{diag}(h_1, h_2)$.

Computation of the risk We consider samples of size $n = 250$ and $n = 1000$ in the case $p = 1$, and samples of size $n = 500$ and $n = 2000$ in the case $p = 2$. For each choice of regression function, distribution of $\boldsymbol{X}$ and SNR, we generate

$N = 100$ samples of size n. For each sample, we compute the Hermite projection estimator and the Nadaraya–Watson estimator, then we compute the relative μ-error of the estimators, that is:

$$\text{relative error} := \frac{\|\hat{b} - b\|_\mu^2}{\|b\|_\mu^2} = \frac{\int_{\mathbb{R}^p} |\hat{b}(\boldsymbol{x}) - b(\boldsymbol{x})|^2 f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^p} b(\boldsymbol{x})^2 f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}},$$

where f is the density of the distribution μ. We compute an approximation of these integrals: we consider a compact domain $I \times I$ with I an interval such that $\mathbb{P}[X \in I] = 95\%$ in the case $p = 1$ and $\mathbb{P}[\boldsymbol{X} \in I \times I] = 95\%$ in the case $p = 2$. Then, we consider a discretization with 200 points of I. In the case $p = 1$, we use Simpson's rule with this discretization of I to approximate the integrals. In the case $p = 2$, we approximate the integrals by a sum over the grid of $I \times I$:

$$\iint_{\mathbb{R}^2} \left|\hat{b}(\boldsymbol{x}) - b(\boldsymbol{x})\right|^2 f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \sum_{i=1}^{200}\sum_{j=1}^{200} \left|\hat{b}(x_{1,i}, x_{2,j}) - b(x_{1,i}, x_{2,j})\right|^2 f(x_{1,i}, x_{2,j})\Delta^2,$$

where Δ is the discretization step.

Results In the case $p = 1$, we show our results on Table 4.1 (Normal distribution) and Table 4.2 (Laplace distribution).

First of all, we see that the results are superior when X has a Normal distribution compared to a Laplace distribution. This can be explained by the fact that the Laplace distribution is less concentrated around 0 than the normal distribution, so the X_is are more scattered and the mu-risk covers a larger range. In addition, in the normal setting, we see that the Hermite estimator is better than the Nadaraya–Watson estimator for estimating b_1, b_2 and b_3, and both estimators are equivalent for estimating b_4. In the Laplace setting, the Hermite estimator is still better for b_1 and b_2, but for b_3 it has similar performances as the Nadaraya–Watson estimator. For estimating b_4, the latter is better, although the difference becomes small as n increases.

In the case $p = 2$, we show our results on Table 4.3 (Normal distribution) and Table 4.4 (Laplace distribution). In the normal setting, the Hermite projection estimator is better for estimating b_5, b_6 and b_7. For b_8, its performances are worse than the kernel estimator on small samples but they are equivalent on large samples. In the Laplace setting, our estimator is better for estimating b_5 and b_6, but it is worse for estimating b_7. Moreover, the Hermite estimator has very poor performances for estimating b_8. We think that the functions b_7 and b_8 are hard to approximate with the Hermite basis, so that the Hermite projection estimator performs poorly. This can be seen by looking at the mean selected dimension, which grows quickly as n grows, showing that the estimator needs a large number of coefficients to reconstruct the regression function.

In addition, we observe that the Hermite estimator is faster to compute than the Nadaraya–Watson estimator with leave-one-out cross validation. The difference is small when n is small, but for example, when $n = 2000$ and $p = 2$, the

Reg. fun.	Estimator	SNR = 2		SNR = 20	
		$n = 250$	$n = 1000$	$n = 250$	$n = 1000$
b_1	Hermite	1.23	0.288	0.138	0.034
		[1.22, 1.24]	[0.284, 0.292]	[0.136, 0.140]	[0.034, 0.035]
		4	5	6	6
	NW	1.50	0.468	0.255	0.076
		[1.49, 1.51]	[0.463, 0.472]	[0.253, 0.258]	[0.075, 0.076]
		0.307	0.212	0.724	0.763
b_2	Hermite	1.00	0.362	0.159	0.047
		[0.99, 1.01]	[0.358, 0.366]	[0.157, 0.161]	[0.047, 0.047]
		3	5	6	8
	NW	1.38	0.475	0.236	0.075
		[1.37, 1.40]	[0.470, 0.480]	[0.234, 0.238]	[0.074, 0.076]
		0.281	0.214	0.161	0.126
b_3	Hermite	1.77	0.477	0.206	0.050
		[1.76, 1.79]	[0.472, 0.482]	[0.204, 0.208]	[0.049, 0.050]
		10	12	11	13
	NW	2.80	0.823	0.808	0.160
		[2.78, 2.82]	[0.817, 0.829]	[0.799, 0.818]	[0.160, 0.161]
		0.138	0.107	0.088	0.066
b_4	Hermite	1.94	0.532	0.288	0.116
		[1.92, 1.97]	[0.528, 0.536]	[0.286, 0.290]	[0.115, 0.116]
		9	12	11	13
	NW	1.86	0.585	0.344	0.108
		[1.84, 1.88]	[0.581, 0.590]	[0.341, 0.347]	[0.107, 0.108]
		0.216	0.162	0.120	0.096

Table 4.1: **Normal distribution,** $p = 1$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator, when X follows the normal distribution. For each regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected model, and for the Nadaraya–Watson estimator, we display the mean selected bandwidth.

Reg. fun.	Estimator	SNR = 2		SNR = 20	
		$n = 250$	$n = 1000$	$n = 250$	$n = 1000$
b_1	Hermite	1.81	0.400	0.162	0.047
		[1.78, 1.83]	[0.394, 0.405]	[0.159, 0.164]	[0.046, 0.047]
		5	6	6	7
	NW	2.20	0.686	0.335	0.104
		[2.18, 2.23]	[0.681, 0.691]	[0.332, 0.338]	[0.103, 0.105]
		0.347	0.260	0.182	0.147
b_2	Hermite	1.45	0.426	0.202	0.064
		[1.43, 1.47]	[0.421, 0.430]	[0.199, 0.204]	[0.063, 0.064]
		3	5	7	9
	NW	1.94	0.725	0.0337	0.113
		[1.92, 1.95]	[0.720, 0.731]	[0.334, 0.339]	[0.112, 0.114]
		0.315	0.249	0.180	0.145
b_3	Hermite	4.56	0.985	1.39	0.121
		[4.49, 4.63]	[0.979, 0.991]	[1.32, 1.47]	[0.120, 0.123]
		19	27	20	29
	NW	3.57	0.974	1.09	0.258
		[3.52, 3.61]	[0.968, 0.980]	[1.06, 1.11]	[0.254, 0.261]
		0.225	0.184	0.155	0.137
b_4	Hermite	8.61	1.04	1.59	0.177
		[8.23, 8.98]	[1.04, 1.05]	[1.53, 1.65]	[0.175, 0.180]
		19	28	20	29
	NW	2.30	0.729	0.454	0.133
		[2.28, 2.33]	[0.724, 0.733]	[0.451, 0.457]	[0.133, 0.134]
		0.294	0.224	0.171	0.127

Table 4.2: **Laplace distribution,** $p = 1$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator, when X follows the Laplace distribution. For each regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected model, and for the Nadaraya–Watson estimator, we display the mean selected bandwidth.

Reg. fun.	Estimator	SNR = 2		SNR = 20	
		$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
b_5	Hermite	1.69	0.587	0.294	0.067
		[1.68, 1.71]	[0.583, 0.591]	[0.191, 0.196]	[0.066, 0.067]
		12	16	21	25
	NW	2.31	0.845	0.566	0.217
		[2.29, 2.32]	[0.841, 0.848]	[0.564, 0.568]	[0.216, 0.218]
		(0.382, 0.388)	(0.295, 0.297)	(0.231, 0.238)	(0.190, 0.188)
b_6	Hermite	1.41	0.732	0.333	0.094
		[1.40, 1.43]	[0.728, 0.735]	[0.331, 0.336]	[0.094, 0.095]
		5	14	26	29
	NW	2.80	1.10	0.630	0.249
		[2.78, 2.81]	[1.09, 1.10]	[0.628, 0.633]	[0.248, 0.250]
		(0.327, 0.356)	(0.273, 0.272)	(0.213, 0.210)	(0.172, 0.172)
b_7	Hermite	3.32	0.916	0.650	0.123
		[3.29, 3.35]	[0.912, 0.919]	[0.645, 0.654]	[0.123, 0.124]
		26	35	43	59
	NW	3.72	1.45	1.29	0.420
		[3.70, 3.74]	[1.45, 1.46]	[1.28, 1.29]	[0.419, 0.421]
		(0.280, 0.285)	(0.229, 0.225)	(0.181, 0.192)	(0.151, 0.147)
b_8	Hermite	9.00	2.01	4.80	0.847
		[8.89, 9.12]	[2.00, 2.02]	[3.66, 4.93]	[0.841, 0.853]
		50	67	51	70
	NW	5.47	2.08	2.56	0.769
		[5.44, 5.49]	[2.07, 2.08]	[2.55, 2.57]	[0.767, 0.771]
		(0.255, 0.250)	(0.197, 0.197)	(0.179, 0.174)	(0.138, 0.137)

Table 4.3: **Normal distribution,** $p = 2$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator, when $\boldsymbol{X}$ follows the normal distribution. For each regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected dimension, and for the Nadaraya–Watson estimator, we display the mean selected bandwidths.

Reg. fun.	Estimator	SNR = 2		SNR = 20	
		$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
b_5	Hermite	1.91	0.703	0.366	0.076
		[1.90, 1.93]	[0.698, 0.708]	[0.359, 0.373]	[0.076, 0.077]
		12	17	21	27
	NW	3.79	1.66	1.01	0.404
		[3.77, 3.80]	[1.66, 1.67]	[1.01, 1.02]	[0.403, 0.405]
		(0.451, 0.441)	(0.354, 0.357)	(0.252, 0.254)	(0.212, 0.208)
b_6	Hermite	2.09	0.962	0.416	0.172
		[2.07, 2.11]	[0.956, 0.968]	[0.412, 0.420]	[0.171, 0.173]
		7	18	27	39
	NW	4.21	1.80	0.944	0.401
		[4.19, 4.22]	[1.79, 1.80]	[0.941, 0.947]	[0.400, 0.402]
		(0.422, 0.403)	(0.324, 0.339)	(0.231, 0.236)	(0.203, 0.199)
b_7	Hermite	10.3	5.56	14.3	1.49
		[10.1, 10.5]	[5.50, 5.62]	[13.9, 14.6]	[1.46, 1.52]
		30	115	76	128
	NW	7.43	2.80	3.02	0.931
		[7.40, 7.46]	[2.80, 2.81]	[3.01, 3.03]	[0.929, 0.933]
		(0.350, 0.391)	(0.292, 0.235)	(0.230, 0.201)	(0.187, 0.167)
b_8	Hermite	415	74.1	330	71.2
		[406, 424]	[72.1, 76.0]	[322, 338]	[69.5, 72.9]
		77	136	79	135
	NW	9.59	3.34	6.20	1.75
		[9.55, 9.64]	[3.33, 3.35]	[6.17, 6.23]	[1.74, 1.76]
		(0.351, 0.356)	(0.284, 0.275)	(0.257, 0.264)	(0.211, 0.209)

Table 4.4: **Laplace distribution,** $p = 2$**.** Table showing the relative μ-risks of the Hermite projection estimator and the Nadaraya–Watson estimator, when $\boldsymbol{X}$ follows the Laplace distribution. For each regression function, SNR and n, we display the estimated relative μ-risk over $N = 100$ samples with a 95% confidence interval, multiplied by 100. For the projection estimator, we display the mean selected dimension, and for the Nadaraya–Watson estimator, we display the mean selected bandwidths.

Hermite estimator is about 3 time faster. In conclusion, the Hermite projection estimator is a good alternative to the Nadaraya–Watson estimator.

4.6 Proofs

4.6.1 Proofs of Section 4.2

Proof of Proposition 4.3.1. Let $\Pi_{\boldsymbol{m}}^{(n)}$ be the projector on $S_{\boldsymbol{m}}$ for the empirical inner product. We have the decomposition:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\|b-\hat{b}_{\boldsymbol{m}}\|_n^2 &= \|b-\Pi_{\boldsymbol{m}}^{(n)}b\|_n^2 + \mathbb{E}_{\boldsymbol{X}}\|\hat{b}_{\boldsymbol{m}}-\Pi_{\boldsymbol{m}}^{(n)}b\|_n^2 \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \mathbb{E}_{\boldsymbol{X}}\|\Pi_{\boldsymbol{m}}^{(n)}\boldsymbol{\varepsilon}\|_n^2 \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \sigma^2\frac{\operatorname{Tr}\left(\Pi_{\boldsymbol{m}}^{(n)}\right)}{n} \\
&= \inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}.
\end{aligned}$$

Taking the expected value in this equality, we obtain:

$$\begin{aligned}
\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_n^2 &= \mathbb{E}\left[\inf_{t\in S_{\boldsymbol{m}}}\|b-t\|_n^2\right] + \sigma^2\frac{D_{\boldsymbol{m}}}{n} \\
&\le \inf_{t\in S_{\boldsymbol{m}}}\mathbb{E}\|b-t\|_n^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n} \\
&= \inf_{t\in S_{\boldsymbol{m}}}\mathbb{E}\|b-t\|_\mu^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}.
\end{aligned}$$

□

4.6.2 Proofs of Section 4.3

Proof of Lemma 4.3.3. Let $x \in A$ and let $t = \sum_{j=1}^d a_j\psi_j \in S$. The family of functions $(\psi_1,\cdots,\psi_d)$ is orthonormal with respect to $\langle\cdot,\cdot\rangle_\alpha$, so by the Cauchy–Schwarz inequality we have:

$$t^2(x) = \left(\sum_{j=1}^d a_j\psi_j(x)\right)^2 \le \left(\sum_{j=1}^d a_j^2\right)\left(\sum_{j=1}^d \psi_j^2(x)\right) = \|t\|_\alpha^2\sum_{j=1}^d\psi_j^2(x),$$

with equality if $(\alpha_1,\ldots,\alpha_d)$ is proportional to $(\psi_1(x),\ldots,\psi_d(x))$. Hence we have:

$$\sum_{j=1}^d\psi_j^2(x) = \sup_{t\in S\setminus\{0\}}\frac{t^2(x)}{\|t\|_\alpha^2}.$$

Taking the supremum for $x \in A$, we obtain:

$$\sup_{x\in A}\sum_{j=1}^d\psi_j^2(x) = \sup_{x\in A}\sup_{t\in S\setminus\{0\}}\frac{t^2(x)}{\|t\|_\alpha^2} = \sup_{t\in S\setminus\{0\}}\frac{\sup_{x\in A}t^2(x)}{\|t\|_\alpha^2},$$

that is:

$$\Big\| \sum_{j=1}^{d} \psi_j^2 \Big\|_\infty = \sup_{t \in S \setminus \{0\}} \frac{\|t\|_\infty^2}{\|t\|_\alpha^2} =: K_\alpha^\infty(S). \qquad \square$$

To prove Proposition 4.3.5 and Theorem 4.4.1, we need the following lemma.

Lemma 4.6.1. *Let $(\psi_1, \dots, \psi_{D_m})$ be an orthnormal basis of $S_{\boldsymbol{m}}$ relatively to an inner product $\langle \cdot, \cdot \rangle_\alpha$. Let $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ be the Gram matrix of this basis relatively to the empirical inner product and let $\mathbf{H}_{\boldsymbol{m}} := \mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}]$, that is:*

$$\forall\, j,k \in \{1,\dots,D_{\boldsymbol{m}}\}, \quad \big[\widehat{\mathbf{H}}_{\boldsymbol{m}}\big]_{j,k} := \langle \psi_j, \psi_k \rangle_n \text{ and } \big[\mathbf{H}_{\boldsymbol{m}}\big]_{j,k} := \langle \psi_j, \psi_k \rangle_\mu.$$

For all $\delta \in (0,1)$ we have:

$$\mathbb{P}\big[\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \le (1-\delta)\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})\big] \le D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})}{K_\alpha^\infty(\boldsymbol{m})}\right),$$

with $h(\delta) := \delta + (1-\delta)\log(1-\delta)$ and where $K_\alpha^\infty(\boldsymbol{m})$ is given by Lemma 4.3.3.

Proof. We use Theorem 1.4.10 in Appendix. Indeed, $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ can be written as a sum $\mathbf{Z}_1 + \ldots + \mathbf{Z}_n$ where:

$$\forall\, j,k \in \{1,\dots,D_{\boldsymbol{m}}\}, \quad \big[\mathbf{Z}_i\big]_{j,k} := \frac{1}{n}\psi_j(\boldsymbol{X}_i)\psi_k(\boldsymbol{X}_i),$$

so we have using Lemma 4.3.3:

$$\lambda_{\max}(\mathbf{Z}_i) = \|\mathbf{Z}_i\|_{\mathrm{op}} = \frac{1}{n}\sum_{k=1}^{D_{\boldsymbol{m}}} \psi_k(\boldsymbol{X}_i)^2 \le \frac{1}{n}\Big\| \sum_{k=1}^{D_{\boldsymbol{m}}} \psi_k^2 \Big\|_\infty = \frac{1}{n} K_\alpha^\infty(\boldsymbol{m}).$$

Hence, applying inequality (1.25) of Theorem 1.4.10 with $\mu_{\min} = \lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})$ and $R = \frac{1}{n}K_\alpha^\infty(\boldsymbol{m})$ yields:

$$\mathbb{P}\big[\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \le (1-\delta)\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})\big] \le D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}})}{K_\alpha^\infty(\boldsymbol{m})}\right). \qquad \square$$

Proof of Proposition 4.3.5. Let $\psi_1, \dots, \psi_{D_m}$ be an orthonormal basis of $S_{\boldsymbol{m}}$ relatively to the inner product $\langle \cdot, \cdot \rangle_\mu$. Let $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ be their Gram matrix relatively to the empirical inner product. According to Lemma 4.3.2, we have $K_n^\mu(\boldsymbol{m}) = \|\widehat{\mathbf{H}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}})^{-1}$ and we have $\mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}] = \mathbf{I}_{\boldsymbol{m}}$ because $(\psi_1, \dots, \psi_{D_m})$ is orthonormal for the inner product associated with μ, so the event $\Omega_{\boldsymbol{m}}(\delta)^c$ can be written as:

$$\Omega_{\boldsymbol{m}}(\delta)^c = \big\{\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \le 1-\delta\big\} = \big\{\lambda_{\min}(\widehat{\mathbf{H}}_{\boldsymbol{m}}) \le (1-\delta)\lambda_{\min}(\mathbb{E}[\widehat{\mathbf{H}}_{\boldsymbol{m}}])\big\}.$$

Applying Lemma 4.6.1 yields the result. $\square$

Proof of Proposition 4.3.4. We start with the decomposition:

$$\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 = \mathbb{E}\Big[\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] + \mathbb{E}\Big[\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big]. \tag{4.7}$$

We consider these two terms separately. The expectation of the first term is controlled as in Theorem 3 in Cohen *et al.* (2013). On the event $\Omega_{\boldsymbol{m}}(\delta)$ we have $(1-\delta)\|t\|_\mu^2 \le \|t\|_n^2$ for all $t \in S_{\boldsymbol{m}}$, so if $b_{\boldsymbol{m}}^{(\mu)}$ is the projection of b on $S_{\boldsymbol{m}}$ for the norm $\|\cdot\|_\mu$, we have:

$$\begin{aligned}
\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} &\le \|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \|\hat{b}_{\boldsymbol{m}}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \\
&\le \|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + 2\|\hat{b}_{\boldsymbol{m}}-b_{\boldsymbol{m}}^{(n)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} + 2\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \\
&\le \|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \frac{2}{1-\delta}\|\hat{b}_{\boldsymbol{m}}-b_{\boldsymbol{m}}^{(n)}\|_n^2 + 2\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}
\end{aligned}$$

Taking the expectation, we obtain:

$$\mathbb{E}\Big[\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 + \frac{2}{1-\delta}\sigma^2\frac{D_{\boldsymbol{m}}}{n} + 2\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big]. \tag{4.8}$$

We give an upper bound on the last term in two ways. Firstly, we have:

$$\begin{aligned}
\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] &\le \mathbb{E}\Big[K_n^\mu(\boldsymbol{m})\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \\
&\le \frac{1}{1-\delta}\mathbb{E}\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_n^2
\end{aligned}$$

since $K_n^\mu(\boldsymbol{m}) \le \frac{1}{1-\delta}$ on the event $\Omega_{\boldsymbol{m}}(\delta)$, see (4.2). Let $\Pi_{\boldsymbol{m}}^{(n)}$ be the empirical projector on $S_{\boldsymbol{m}}$, we have:

$$\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_n^2 = \Big\|\Pi_{\boldsymbol{m}}^{(n)}\big(b-b_{\boldsymbol{m}}^{(\mu)}\big)\Big\|_n^2 \le \|b-b_{\boldsymbol{m}}^{(\mu)}\|_n^2.$$

Thus, we have shown:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \frac{1}{1-\delta}\mathbb{E}\|b-b_{\boldsymbol{m}}^{(\mu)}\|_n^2 = \frac{1}{1-\delta}\|b-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2. \tag{4.9}$$

Secondly, let $g := b-b_{\boldsymbol{m}}^{(\mu)}$ and let $\Pi_{\boldsymbol{m}}^{(n)}$ be the empirical projector on $S_{\boldsymbol{m}}$ we have:

$$\mathbb{E}\Big[\|b_{\boldsymbol{m}}^{(n)}-b_{\boldsymbol{m}}^{(\mu)}\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] = \mathbb{E}\Big[\|\Pi_{\boldsymbol{m}}^{(n)} g\|_\mu^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big].$$

Let $(\psi_1,\dots,\psi_{D_{\boldsymbol{m}}})$ be an orthonormal basis of $S_{\boldsymbol{m}}$ for the inner product $\langle\cdot,\cdot\rangle_\mu$, we have:

$$\Pi_{\boldsymbol{m}}^{(n)} g = \operatorname*{argmin}_{t\in S_{\boldsymbol{m}}}\|g-t\|_n^2 = \sum_{j=1}^{D_{\boldsymbol{m}}} c_j^\star \psi_j, \quad \mathbf{c}^\star := \operatorname*{argmin}_{\mathbf{c}\in\mathbb{R}^{D_{\boldsymbol{m}}}}\|\mathbf{g}-\Psi_{\boldsymbol{m}}\mathbf{c}\|_{\mathbb{R}^n}^2,$$

where $\Psi_{\boldsymbol{m}} \in \mathbb{R}^{n\times D_{\boldsymbol{m}}}$ is the matrix defined by $[\Psi_{\boldsymbol{m}}]_{i,j} := \psi_j(\boldsymbol{X}_i)$, and where $\mathbf{g}$ is the vector $\big(g(\boldsymbol{X}_1),\dots,g(\boldsymbol{X}_n)\big) \in \mathbb{R}^n$. By Lemma B.1.1, $\mathbf{c}^\star$ is given by:

$$\mathbf{c}^\star = (\Psi_{\boldsymbol{m}}^*\Psi_{\boldsymbol{m}})^{-1}\Psi_{\boldsymbol{m}}^*\mathbf{g} = \frac{1}{n}\mathbf{H}_{\boldsymbol{m}}^{-1}\Psi_{\boldsymbol{m}}^*\mathbf{g},$$

where $\mathbf{H}_{\boldsymbol{m}}$ is the Gram matrix of $(\psi_1,\dots,\psi_{D_{\boldsymbol{m}}})$ relatively to the empirical inner product. Using Lemma 4.3.2, we get:

$$\|\Pi^{(n)}_{\boldsymbol{m}} g\|^2_\mu = \|\mathbf{c}^\star\|^2_{\mathbb{R}^{D_{\boldsymbol{m}}}} \le \|\mathbf{H}_{\boldsymbol{m}}^{-1}\|^2_{\mathrm{op}} \left\| \frac{1}{n}\Psi^*_{\boldsymbol{m}}\mathbf{g}\right\|^2_{\mathbb{R}^{D_{\boldsymbol{m}}}} = K^\mu_n(\boldsymbol{m})^2 \sum_{j=1}^{D_{\boldsymbol{m}}} \langle g,\psi_j\rangle_n^2.$$

Hence, on the event $\Omega_{\boldsymbol{m}}(\delta)$ we obtain:

$$\|\Pi^{(n)}_{\boldsymbol{m}} g\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)} \le \frac{1}{(1-\delta)^2}\sum_{j=1}^{D_{\boldsymbol{m}}} \langle g,\psi_j\rangle_n^2.$$

Since $g = b - b^{(\mu)}_{\boldsymbol{m}}$ is orthogonal to $\psi_1,\dots,\psi_{D_{\boldsymbol{m}}}$ relatively to the inner product $\langle\cdot,\cdot\rangle_\mu$, we have $\mathbb{E}[\langle g,\psi_j\rangle_n] = \langle g,\psi_j\rangle_\mu = 0$, so we get:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{D_{\boldsymbol{m}}}\langle g,\psi_k\rangle_n^2\right] &= \sum_{k=1}^{D_{\boldsymbol{m}}} \mathrm{Var}\big(\langle g,\psi_k\rangle_n\big)\\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^{D_{\boldsymbol{m}}}\mathrm{Var}\big(g(\boldsymbol{X}_i)\psi_j(\boldsymbol{X}_i)\big)\\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[g(\boldsymbol{X}_i)^2\sum_{j=1}^{D_{\boldsymbol{m}}}\psi_j(\boldsymbol{X}_i)^2\right]\\
&\le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\big[g(\boldsymbol{X}_i)^2\big]\sup_{x\in A}\sum_{j=1}^{D_{\boldsymbol{m}}}\psi_j(x)^2\\
&= \frac{1}{n}\|g\|^2_\mu K^\infty_\mu(\boldsymbol{m}) = \frac{K^\infty_\mu(\boldsymbol{m})}{n}\|b-b^{(\mu)}_{\boldsymbol{m}}\|^2_\mu,
\end{aligned}$$

where the last equality comes from Lemma 4.3.3. Hence we have shown:

$$\mathbb{E}\Big[\|b^{(n)}_{\boldsymbol{m}} - b^{(\mu)}_{\boldsymbol{m}}\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \frac{1}{(1-\delta)^2}\frac{K^\infty_\mu(\boldsymbol{m})}{n}\|b-b^{(\mu)}_{\boldsymbol{m}}\|^2_\mu. \tag{4.10}$$

Combining (4.9) and (4.10) yields:

$$\mathbb{E}\Big[\|b^{(n)}_{\boldsymbol{m}} - b^{(\mu)}_{\boldsymbol{m}}\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)}\Big] \le \frac{1}{1-\delta}\|b-b^{(\mu)}_{\boldsymbol{m}}\|^2_\mu\left(1\wedge\frac{K^\infty_\mu(\boldsymbol{m})}{(1-\delta)n}\right). \tag{4.11}$$

For the second term in (4.7), we have:

$$\mathbb{E}\Big[\|b-\hat{b}_{\boldsymbol{m}}\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big] \le 2\|b\|^2_\mu\,\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] + 2\,\mathbb{E}\Big[\|\hat{b}_{\boldsymbol{m}}\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big].$$

We have the following upper bound on $\|\hat{b}_{\boldsymbol{m}}\|^2_\mu$:

$$\|\hat{b}_{\boldsymbol{m}}\|^2_\mu \le K^\mu_n(\boldsymbol{m})\,\|\hat{b}_{\boldsymbol{m}}\|^2_n \le K^\mu_n(\boldsymbol{m})\,\|\mathbf{Y}\|^2_n, \tag{4.12}$$

where the last inequality comes from the fact that $\hat{b}_{\boldsymbol{m}}$ is the empirical projection of $\mathbf{Y}$. Hence, we get:

$$\mathbb{E}\Big[\|b-\hat{b}_{\boldsymbol{m}}\|^2_\mu \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big] \le 2\|b\|^2_\mu\,\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] + 2\,\mathbb{E}\big[K^\mu_n(\boldsymbol{m})\,\|\mathbf{Y}\|^2_n \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big]. \tag{4.13}$$

The inequality of Proposition 4.3.4 is obtained using (4.8), (4.11) and (4.13) in (4.7). □

Proof of Theorem 4.3.7. Let $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ and let $\delta \in (0,1)$ (we choose it later in the proof). By Remark 4.3.6, we have by definition of $\mathscr{M}_{n,\alpha}^{(1)}$:

$$K_\mu^\infty(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m}) \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \alpha \frac{n}{\log n}, \tag{4.14}$$

so Proposition 4.3.4 yields:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n(\delta,\alpha) \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + C'(\delta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + R_n,$$

with $C_n(\alpha,\delta) := \left(1 + \frac{2}{1-\delta}\left[\frac{\alpha}{(1-\delta)\log n} \wedge 1\right]\right)$, $C'(\delta) := \frac{2}{1-\delta}$ and:

$$R_n := 2\|b\|_\mu^2 \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] + \mathbb{E}\big[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big].$$

For the first term in R_n, we apply Proposition 4.3.5 with (4.14):

$$\mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \leq D_{\boldsymbol{m}}\, n^{-\frac{h(\delta)}{\alpha}} \leq n^{-\frac{h(\delta)}{\alpha}+1}. \tag{4.15}$$

For the second term in R_n, since $\|\cdot\|_\mu \leq \|\cdot\|_\infty$ and $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}$ we have:

$$K_n^\mu(\boldsymbol{m}) \leq K_\nu^\mu(\boldsymbol{m})\, K_n^\nu(\boldsymbol{m}) \leq K_\nu^\infty(\boldsymbol{m})\, \|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leq \alpha \frac{n}{\log n}, \tag{4.16}$$

and we have using the independence of $(\boldsymbol{X}_i)_{1\leq i\leq n}$ and $(\varepsilon_i)_{1\leq i\leq n}$:

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big(b(\boldsymbol{X}_i) + \varepsilon_i\big)^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\Big] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\right] + \sigma^2 \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big].\end{aligned}$$

We apply Hölder's inequality with $r, r' \in (1,+\infty)$ such that $\frac{1}{r} + \frac{1}{r'} = 1$:

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] &\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2\right)^r\right]^{\frac{1}{r}} \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big]^{\frac{1}{r'}} + \sigma^2 \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \\ &\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^{2r}\right]^{\frac{1}{r}} \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big]^{\frac{1}{r'}} + \sigma^2 \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \\ &\leq \|b\|_{\mathrm{L}^{2r}(\mu)}^2\, n^{-\frac{h(\delta)}{\alpha r'}+\frac{1}{r'}} + \sigma^2\, n^{-\frac{h(\delta)}{\alpha}+1},\end{aligned}$$

and if $b \in \mathrm{L}^\infty(\mu)$, the last inequality also holds for $r = \infty$ and $r' = 1$ (just take the limit as $r \to +\infty$). Hence, we obtain:

$$\mathbb{E}\big[K_n^\mu(\boldsymbol{m})\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Omega_{\boldsymbol{m}}(\delta)^c}\big] \leq \frac{\alpha}{\log n}\left(\|b\|_{\mathrm{L}^{2r}(\mu)}^2\, n^{-\frac{h(\delta)}{\alpha r'}+\frac{1}{r'}+1} + \sigma^2\, n^{-\frac{h(\delta)}{\alpha}+2}\right). \tag{4.17}$$

If we choose δ such that $h(\delta) \geq (2r'+1)\alpha$, then all the exponents of n in (4.15) and (4.17) are less than -1. The function h is an increasing function from $[0,1]$ to

itself so it is invertible on $[0,1]$. Since $\alpha \in (0, \frac{1}{2r'+1})$, we can choose $\delta = \delta(\alpha, r') := h^{-1}((2r'+1)\alpha)$. For this choice, we obtain:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leq C_n\big(\delta(\alpha,r'),\alpha\big) \inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + C'\big(\delta(\alpha,r')\big)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \frac{C''\big(\alpha, \|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2\big)}{n \log n},$$

where $C_n(\delta,\alpha)$ and $C'(\delta)$ were defined at the beginning of the proof. □

4.6.3 Proof of Theorem 4.4.1

The proof of Theorem 4.4.1 is based on a result for fixed design regression of Baraud (2000). Let $\widehat{\mathscr{M}_n}$ be a finite collection of models, that may depend on $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, such that for all $\boldsymbol{m} \in \widehat{\mathscr{M}_n}$, $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ is invertible. Let $\hat{\boldsymbol{m}} \in \widehat{\mathscr{M}_n}$ be the minimizer of the following penalized least squares criterion:

$$\hat{\boldsymbol{m}} := \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} \big(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + \mathrm{pen}(\boldsymbol{m})\big), \quad \mathrm{pen}(\boldsymbol{m}) := (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n}, \quad \theta > 0. \tag{4.18}$$

Theorem 4.6.2 (Corollary 3.1 in Baraud (2000))**.** *If $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 4$, then the following upper bound on the risk of the estimator $\hat{b}_{\hat{\boldsymbol{m}}}$ with $\hat{\boldsymbol{m}}$ defined by* (4.18) *holds:*

$$\mathbb{E}_{\boldsymbol{X}}\|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_n^2 \leq C(\theta) \inf_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma_n(\theta,q)}{n},$$

with:

$$\Sigma_n(\theta,q) := C'(\theta,q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \widehat{\mathscr{M}_n}} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)},$$

where $C(\theta) := (2 + 8\theta^{-1})(1+\theta)$ and $C'(\theta,q)$ is a positive constant.

Proof of Theorem 4.4.1. Let $\Delta_{n,\alpha,\beta} := \{\mathscr{M}_{n,\alpha}^{(1)} \subseteq \widehat{\mathscr{M}}_{n,\beta}^{(1)}\}$, we have:

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 = \mathbb{E}\big[\mathbb{E}_{\boldsymbol{X}}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}}\big] + \mathbb{E}\Big[\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}\Big].$$

For the first term, on $\Delta_{n,\alpha,\beta}$ we have $\inf_{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(1)}}(\ldots) \leq \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(1)}}(\ldots)$ so by applying Theorem 4.6.2 we obtain:

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}_{\boldsymbol{X}}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}}\big] &\leq \mathbb{E}\left[C(\theta) \inf_{m \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta,q)}{n} \right] \\
&\leq C(\theta) \inf_{m \in \mathscr{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta,q)}{n}.
\end{aligned}$$

For the second term, we have:

$$\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c} \leq 2\|b\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c} + 2\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta_{n,\alpha,\beta}^c}.$$

Using Hölder's inequality with $r, r' \in (1,\infty)$ such that $\frac{1}{r} + \frac{1}{r'} = 1$, we obtain:

$$\mathbb{E}\Big[\|b\|_n^2 \mathbf{1}_{\Delta^c_{n,\alpha,\beta}}\Big] \le \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2\right)^r\right]^{1/r} \mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big]^{1/r'} \le \|b\|^2_{\mathrm{L}^{2r}(\mu)} \mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big]^{1/r'},$$

and if $b \in \mathrm{L}^\infty(\mu)$, the inequality also holds for $r = \infty$ and $r' = 1$. Since $\hat{b}_{\hat{\boldsymbol{m}}_1}$ is the empirical projection of $\mathbf{Y}$ on $S_{\hat{\boldsymbol{m}}_1}$, we have $\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \le \|\mathbf{Y}\|_n^2$. Hence, we get:

$$\begin{aligned}\mathbb{E}\Big[\|\hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \mathbf{1}_{\Delta^c_{n,\alpha,\beta}}\Big] \le \mathbb{E}\Big[\|\mathbf{Y}\|_n^2 \mathbf{1}_{\Delta^c_{n,\alpha,\beta}}\Big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n b(\boldsymbol{X}_i)^2 \mathbf{1}_{\Delta^c_{n,\alpha,\beta}}\right] + \sigma^2 \mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big] \\ &\le \|b\|^2_{\mathrm{L}^{2r}(\mu)} \mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big]^{\frac{1}{r'}} + \sigma^2 \mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big]. \quad (4.19)\end{aligned}$$

To conclude, we give an upper bound on $\mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big]$:

$$\begin{aligned}&\mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big] = \mathbb{P}\Big[\exists \boldsymbol{m} \in \mathscr{M}^{(1)}_{n,\alpha},\, \boldsymbol{m} \notin \widehat{\mathscr{M}}^{(1)}_{n,\beta}\Big] \\ &\le \sum_{\boldsymbol{m}\in\mathscr{M}^{(1)}_{n,\alpha}} \mathbb{P}\left[\left\{K_V^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\right) \le \alpha\frac{n}{\log n}\right\} \cap \left\{K_V^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\right) \ge \beta\frac{n}{\log n}\right\}\right] \\ &\le \sum_{\boldsymbol{m}\in\mathscr{M}^{(1)}_{n,\alpha}} \mathbb{P}\left[\frac{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \ge \frac{\beta}{\alpha}\right] \le \sum_{\boldsymbol{m}\in\mathscr{M}^{(1)}_{n,\alpha}} \mathbb{P}\Big[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \le \frac{\alpha}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\Big]. \quad (4.20)\end{aligned}$$

Using Lemma 4.6.1 with the inequality $K_V^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \le \alpha\frac{n}{\log n}$ for $\boldsymbol{m} \in \mathscr{M}^{(1)}_{n,\alpha}$, we obtain:

$$\begin{aligned}\forall \boldsymbol{m} \in \mathscr{M}^{(1)}_{n,\alpha},\ \mathbb{P}\Big[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}}) \le \frac{\alpha}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\Big] &\le D_{\boldsymbol{m}} \exp\left(h(1-\tfrac{\alpha}{\beta})\frac{n}{K_V^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right) \\ &\le D_{\boldsymbol{m}}\, n^{-h(1-\frac{\alpha}{\beta})/\alpha}.\end{aligned}$$

Hence, we get:

$$\mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big] \le \sum_{\boldsymbol{m}\in\mathscr{M}^{(1)}_{n,\alpha}} D_{\boldsymbol{m}}\, n^{-h(1-\frac{\alpha}{\beta})/\alpha} \le \mathrm{Card}(\mathscr{M}^{(1)}_{n,\alpha})\, n^{1-h(1-\frac{\alpha}{\beta})/\alpha}.$$

Using Proposition B.2.1 in appendix, we obtain:

$$\mathbb{P}\Big[\Delta^c_{n,\alpha,\beta}\Big] \le n^{2-h(1-\frac{\alpha}{\beta})/\alpha} H_n^{p-1} = n^{-\kappa(\alpha,\beta)} H_n^{p-1},$$

with $H_n := \sum_{k=1}^n \frac{1}{k}$ and $\kappa(\alpha,\beta) := \frac{h(1-\frac{\alpha}{\beta})}{\alpha} - 2$. We know that $H_n \sim \log n$, so we want a condition on α such that the $\kappa(\alpha,\beta)$ is strictly greater than r'. Let $x := \frac{\beta}{\alpha} \ge 1$, we

have:

$$\begin{aligned}\kappa(\alpha,\beta) > r' &\iff h\left(1-\frac{\alpha}{\beta}\right) > (2+r')\alpha \\ &\iff 1-\frac{\alpha}{\beta}+\frac{\alpha}{\beta}\log\left(\frac{\alpha}{\beta}\right) > (2+r')\alpha \\ &\iff 1-\frac{1+\log(x)}{x} > \frac{(2+r')\beta}{x} \\ &\iff \frac{1+(2+r')\beta+\log(x)}{x} < 1. \end{aligned} \tag{4.21}$$

The function:

$$f_{\beta,r'}(x) := \frac{1+(2+r')\beta+\log(x)}{x},$$

is decreasing on $[1,+\infty)$, we have $f_{\beta,r'}(1) > 1$ and $f_{\beta,r'}(x) \to 0$ when $x \to +\infty$, so there exists a unique $x_{\beta,r'} \in (1,+\infty)$ such that $f_{\beta,r'}(x_{\beta,r'}) = 1$. Thus, we have:

$$(4.21) \iff x \in (x_{\beta,r'},+\infty) \iff \alpha \in (0,\alpha_{\beta,r'}),$$

where $\alpha_{\beta,r'} := \frac{\beta}{x_{\beta,r'}}$. Hence, if $\alpha \in (0,\alpha_{\beta,r'})$ then we have:

$$\mathbb{P}\left[\Delta^c_{n,\alpha,\beta}\right]^{1/r'} \leq n^{-\frac{\kappa(\alpha,\beta)}{r'}} H_n^{\frac{p-1}{r'}},$$

with $\frac{\kappa(\alpha,\beta)}{r'} > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ as $\alpha \to \alpha_{\beta,r'}$. □

Remark 4.6.3. If we use the collections $\mathscr{M}^{(2)}_{n,\alpha}$ and $\widehat{\mathscr{M}}^{(2)}_{n,\beta}$ instead, we obtain the inequality (4.20) with α and β replaced by $\alpha' := \sqrt{\alpha}$ and $\beta' := \sqrt{\beta}$. The rest of the proof is unchanged.

Proof of Remark 4.4.3. We have $\alpha_{\beta,r'} := \frac{\beta}{x_{\beta,r'}}$ where $x_{\beta,r'}$ is the unique solution in $(1,+\infty)$ of the equation $f_{\beta,r'}(x) = 1$ with:

$$f_{\beta,r'}(x) := \frac{1+(2+r')\beta+\log x}{x}.$$

Hence, x_β satisfies the relation:

$$x_{\beta,r'} - \log x_{\beta,r'} = 1+(2+r')\beta. \tag{4.22}$$

Since the functions $f_{\beta,r'}$ are decreasing on $(1,+\infty)$ and since $\forall x$, $f_{\beta,r'}(x)$ is increasing with β and r', we see that $x_{\beta,r'}$ is increasing with β and r'. Thus, the limits of $x_{\beta,r'}$ when $\beta \to 0$ and $\beta \to +\infty$ exist. Using the relation (4.22), we obtain:

$$\lim_{\beta\to 0} x_{\beta,r'} = 1, \qquad \lim_{\beta\to+\infty} x_{\beta,r'} = +\infty, \qquad \lim_{r'\to\infty} x_{\beta,r'} = +\infty,$$

and we have $x_{\beta,r'} \sim (2+r')\beta$ when $\beta \to +\infty$. Thus, the limits of $\alpha_{\beta,r'}$ are:

$$\lim_{\beta\to 0} \alpha_{\beta,r'} = 0, \qquad \lim_{\beta\to+\infty} \alpha_{\beta,r'} = \frac{1}{2+r'}, \qquad \lim_{r'\to+\infty} \alpha_{\beta,r'} = 0.$$

Since $x_{\beta,r'}$ is increasing with r', we see that $\alpha_{\beta,r'}$ is deacreasing with r'. Finally, using the relation (4.22) again, we have:

$$\alpha_{\beta,r'} = \frac{\beta}{x_{\beta,r'}} = \frac{1}{2+r'}\left(1 - \frac{1}{x_{\beta,r'}} - \frac{\log x_{\beta,r'}}{x_{\beta,r'}}\right).$$

It is easy to see that the function $x \mapsto 1 - \frac{1}{x} - \frac{\log x}{x}$ is increasing on $[1,+\infty)$ so $\alpha_{\beta,r'}$ is also increasing with β. □

4.6.4 Proof of Theorem 4.4.4

Before proving Theorem 4.4.4, we need some preliminary results.

Lemma 4.6.4. *For all $x > 0$ and all $\boldsymbol{m} \in \mathbb{N}_+^p$ we have:*

$$\begin{aligned}\mathbb{P}\big[\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq x\big] &\leq D_{\boldsymbol{m}} \exp\left(\frac{-nx^2/2}{K_v^\infty(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} + \frac{2}{3}x\big)}\right) \\ &\leq D_{\boldsymbol{m}} \exp\left(\frac{-nx^2/2}{K_v^\infty(\boldsymbol{m})\big(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3}x\big)}\right).\end{aligned}$$

Proof. The set $\{\varphi_{\boldsymbol{j}} : \boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}\}$ has cardinality $D_{\boldsymbol{m}}$ so let $\{\phi_1, \dots, \phi_{D_{\boldsymbol{m}}}\}$ be its elements. We define the matrix $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ as:

$$\forall\, j,k \in \{1,\dots,D_{\boldsymbol{m}}\}, \quad \big[\widehat{\mathbf{H}}_{\boldsymbol{m}}\big]_{j,k} := \langle \phi_j, \phi_k \rangle_n,$$

and we denote its expectation $\mathbf{H}_{\boldsymbol{m}}$, of which the components are $\langle \phi_j, \phi_k \rangle_\mu$. In other words, we have reshaped the hypermatrices $\widehat{\mathbf{G}}_{\boldsymbol{m}}$ and $\mathbf{G}_{\boldsymbol{m}}$ into $D_{\boldsymbol{m}} \times D_{\boldsymbol{m}}$ matrices. Moreover, this operation preserves the operator norm:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \|\mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}}.$$

Indeed, let $d := D_{\boldsymbol{m}}$, we have:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{\boldsymbol{m}} \\ \|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}=1}} \|\mathbf{G}_{\boldsymbol{m}} \times_p \mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}^2 = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{\boldsymbol{m}} \\ \|\mathbf{a}\|_{\mathbb{R}^{\boldsymbol{m}}}=1}} \sum_{\boldsymbol{\ell}\leq \boldsymbol{m}-\mathbf{1}} \left(\sum_{\boldsymbol{k}\leq \boldsymbol{m}-\mathbf{1}} \langle \varphi_{\boldsymbol{\ell}}, \varphi_{\boldsymbol{k}} \rangle a_{\boldsymbol{k}}\right)^2,$$

$$\|\mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{d} \\ \|\mathbf{a}\|_{\mathbb{R}^{d}}=1}} \|\mathbf{H}_{\boldsymbol{m}}\mathbf{a}\|_{\mathbb{R}^d}^2 = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{d} \\ \|\mathbf{a}\|_{\mathbb{R}^{d}}=1}} \sum_{j=1}^{d} \left(\sum_{i=1}^{d} \langle \psi_j, \psi_i \rangle a_i\right)^2.$$

Since the sets $\{\varphi_{\boldsymbol{j}} : \boldsymbol{j} \leq \boldsymbol{m} - \mathbf{1}\}$ and $\{\phi_1, \dots, \phi_d\}$ are equal, these two quantities are also equal. Hence we have:

$$\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \|\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}},$$

so we work on $\widehat{\mathbf{H}}_{\boldsymbol{m}}$ and $\mathbf{H}_{\boldsymbol{m}}$ from now on. We write:

$$\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}} = \sum_{i=1}^{n} \mathbf{Z}_i, \quad \mathbf{Z}_i := \frac{1}{n}\left(\mathbf{V}_i \mathbf{V}_i^\top - \mathbb{E}\left[\mathbf{V}_i \mathbf{V}_i^\top\right]\right), \quad \mathbf{V}_i := \begin{bmatrix} \phi_1(\boldsymbol{X}_i) \\ \vdots \\ \phi_{D_{\boldsymbol{m}}}(\boldsymbol{X}_i) \end{bmatrix},$$

and we use the Matrix Bernstein bound (Theorem 1.4.11 in appendix).

1. Bound on $\|\mathbf{Z}_i\|_{\mathrm{op}}$:

$$\frac{1}{n}\|\mathbf{V}_i \mathbf{V}_i^\top\|_{\mathrm{op}} = \frac{1}{n}\|\mathbf{V}_i\|^2 = \frac{1}{n}\sum_{j=1}^{D_{\boldsymbol{m}}} \phi_j(\boldsymbol{X}_i)^2 \le \frac{K_\nu^\infty(\boldsymbol{m})}{n},$$

   where the last inequality comes from Lemma 4.3.3. Hence, $\|\mathbf{Z}_i\|_{\mathrm{op}} \le R$, with $R := \frac{K_\nu^\infty(\boldsymbol{m})}{n}$.

2. Bound on $\|\sum_{i=1}^n \mathbb{E}\left[\mathbf{Z}_i^2\right]\|_{\mathrm{op}}$:

$$\begin{aligned}
\left\|\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{Z}_i^2\right]\right\|_{\mathrm{op}} &= \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{Z}_i\,\mathbf{a}\|^2\right] \\
&= \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n} \sum_{j=1}^{D_{\boldsymbol{m}}} \mathbb{E}\left[(\mathbf{Z}_i\,\mathbf{a})_j^2\right] \\
&= \sup_{\|\mathbf{a}\|=1} \sum_{i=1}^{n} \sum_{j=1}^{D_{\boldsymbol{m}}} \mathrm{Var}\left[(\mathbf{Z}_i\,\mathbf{a})_j\right],
\end{aligned}$$

   since $\mathbb{E}\mathbf{Z}_i = \mathbf{0}$. We compute the variance:

$$\begin{aligned}
\mathrm{Var}\left[(\mathbf{Z}_i\,\mathbf{a})_j\right] &= \mathrm{Var}\left[\frac{1}{n}\phi_j(\boldsymbol{X}_i)\sum_{k=1}^{D_{\boldsymbol{m}}} \phi_k(\boldsymbol{X}_i)\,a_k\right] \\
&\le \frac{1}{n^2}\mathbb{E}\left[\left(\phi_j(\boldsymbol{X}_i)\sum_{k=1}^{D_{\boldsymbol{m}}} \phi_k(\boldsymbol{X}_i)\,a_k\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\phi_j(\boldsymbol{X}_i)^2\, t_{\mathbf{a}}(\boldsymbol{X}_i)^2\right],
\end{aligned}$$

   where $t_{\mathbf{a}} := \sum_{k=1}^{D_{\boldsymbol{m}}} a_k \phi_k$. Using Lemmas 4.3.2 and 4.3.3 yields:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{D_{\boldsymbol{m}}} \mathrm{Var}\left[(\mathbf{Z}_i\,\mathbf{a})_j\right] &\le \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{D_{\boldsymbol{m}}} \phi_j(\boldsymbol{X}_i)^2\, t_{\mathbf{a}}(\boldsymbol{X}_i)^2\right] \\
&\le \frac{1}{n} K_\nu^\infty(\boldsymbol{m})\,\|t_{\mathbf{a}}\|_\mu^2 \\
&\le \frac{1}{n} K_\nu^\infty(\boldsymbol{m})\,K_\nu^\mu(\boldsymbol{m})\,\|t_{\mathbf{a}}\|_\nu^2 \\
&= \frac{1}{n} K_\nu^\infty(\boldsymbol{m})\,\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}\,\|\mathbf{a}\|^2.
\end{aligned}$$

   Hence, $\|\sum_{i=1}^n \mathbb{E}\left[\mathbf{Z}_i^2\right]\|_{\mathrm{op}} \le \frac{1}{n} K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} =: \upsilon$.

Applying Theorem 1.4.11 yields:

$$\mathbb{P}\big[\|\widehat{\mathbf{H}}_{\boldsymbol{m}} - \mathbf{H}_{\boldsymbol{m}}\|_{\mathrm{op}} \geq x\big] \leq D_{\boldsymbol{m}} \exp\left(-\frac{nx^2/2}{K_\nu^\infty(\boldsymbol{m})\big(\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} + \frac{2}{3}x\big)}\right),$$

which is the first inequality of Lemma 4.6.4. The second inequality follows from the following upper bound on $\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}$:

$$\|\mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} = \sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\mu^2}{\|t\|_\nu^2} \leq \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty. \qquad \square$$

In order to prove Theorem 4.4.4, let us consider the events:

$$\Lambda_n^{(\iota)}(\beta,\gamma) := \left\{ \widehat{\mathscr{M}}_{n,\beta}^{(\iota)} \subseteq \mathscr{M}_{n,\gamma}^{(\iota)} \right\}, \qquad \widetilde{\Omega}_n^{(\iota)}(\delta,\gamma) := \bigcap_{\boldsymbol{m} \in \mathscr{M}_{n,\gamma}^{(\iota)}} \Omega_{\boldsymbol{m}}(\delta), \qquad \iota \in \{1,2\}, \tag{4.23}$$

where $\Omega_{\boldsymbol{m}}(\delta)$ is defined by (4.2).

Lemma 4.6.5. *For $\iota \in \{1,2\}$, we have for all $\delta \in (0,1)$ and all $\gamma > 0$:*

$$\mathbb{P}\big[\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)^c\big] \leq n^{-\frac{h(\delta)}{\gamma}+2} H_n^{p-1},$$

where $H_n := \sum_{k=1}^n \frac{1}{k}$ is the n-th harmonic number.

Proof. We use Proposition 4.3.5 with Remark 4.3.6:

$$\begin{aligned}
\mathbb{P}\big[\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)^c\big] &\leq \sum_{\boldsymbol{m} \in \mathscr{M}_{n,\gamma}^{(\iota)}} \mathbb{P}\big[\Omega_{\boldsymbol{m}}(\delta)^c\big] \\
&\leq \sum_{\boldsymbol{m} \in \mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\mu^\infty(\boldsymbol{m})}\right) \\
&\leq \sum_{\boldsymbol{m} \in \mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{K_\nu^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right) \\
&\leq \sum_{\boldsymbol{m} \in \mathscr{M}_{n,\gamma}^{(\iota)}} D_{\boldsymbol{m}}\, n^{-\frac{h(\delta)}{\gamma}} \leq n^{-\frac{h(\delta)}{\gamma}+2} H_n^{p-1},
\end{aligned}$$

where the last inequality comes from Proposition B.2.1. $\square$

Lemma 4.6.6 (Compact case)**.** *We have for all $\gamma > \beta > 0$:*

$$\mathbb{P}\big[\Lambda_n^{(1)}(\beta,\gamma)^c\big] \leq n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1} H_n^{p-1},$$

where $h(\delta) = \delta + (1-\delta)\log(1-\delta)$, $f_0 > 0$ is such that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \geq f_0$ for all $x \in A$ and $H_n := \sum_{k=1}^n \frac{1}{k}$.

Proof. We start with a union bound:

$$\begin{aligned}\mathbb{P}\big[\Lambda_n^{(1)}(\beta,\gamma)^c\big] &= \mathbb{P}\Big[\exists \boldsymbol{m}\in\mathbb{N}_+^p,\ \boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\Big]\\ &\leq \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p\\ K_v^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}} \mathbb{P}\Big[\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\Big].\end{aligned}$$

We have the following inclusion of events:

$$\begin{aligned}&\Big\{\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(1)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(1)}\Big\}\\ &\subseteq\left\{K_v^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right)\leq\beta\frac{n}{\log n}\right\}\cap\left\{K_v^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\vee 1\right)\geq\gamma\frac{n}{\log n}\right\}\\ &\subseteq\left\{\frac{\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\geq\frac{\gamma}{\beta}\right\}\subseteq\left\{\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}})\geq\frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right\},\end{aligned}$$

hence we obtain:

$$\mathbb{P}\big[\Lambda_n^{(1)}(\beta,\gamma)^c\big]\leq\sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p\\ K_v^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}}\mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}})\geq\frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right].$$

We apply inequality (1.26) of Theorem 1.4.10 with $R=\frac{1}{n}K_v^\infty(\boldsymbol{m})$:

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\mathbf{G}}_{\boldsymbol{m}})\geq\frac{\gamma}{\beta}\lambda_{\min}(\mathbf{G}_{\boldsymbol{m}})\right]\leq\exp\left(-h\left(1-\frac{\gamma}{\beta}\right)\frac{n}{K_v^\infty(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right).$$

In the compact case, we have $\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\leq\frac{1}{f_0}$, see (4.5). Using Proposition B.2.1, we obtain:

$$\mathbb{P}\big[\Lambda_n^{(1)}(\beta,\gamma)^c\big]\leq\sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p\\ K_v^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}}n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}}\leq n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1}H_n^{p-1}.\qquad\square$$

Lemma 4.6.7 (General case)**.** *We have for all $\gamma>\beta>0$:*

$$\mathbb{P}\big[\Lambda_n^{(2)}(\beta,\gamma)^c\big]\leq n^{-C(\beta,\gamma)\frac{B}{2\beta}+2}H_n^{p-1},$$

where $C(\beta,\gamma):=\left(1-\sqrt{\beta/\gamma}\right)^2$, $B:=\left(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty+\frac{2}{3}\right)^{-1}$ and $H_n:=\sum_{k=1}^n\frac{1}{k}$.

Proof. We start with a union bound:

$$\begin{aligned}&\mathbb{P}\big[\Lambda_n^{(2)}(\beta,\gamma)^c\big]=\mathbb{P}\Big[\exists\boldsymbol{m}\in\mathbb{N}_+^p,\ \boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(2)}\Big]\\ &\leq\sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p\\ K_v^\infty(\boldsymbol{m})\leq\beta\frac{n}{\log n}}}\mathbb{P}\Big[\boldsymbol{m}\in\widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m}\notin\mathscr{M}_{n,\gamma}^{(2)}\Big].\end{aligned}$$

We have the following inclusion of events:

$$\begin{aligned}
&\left\{\boldsymbol{m} \in \widehat{\mathscr{M}}_{n,\beta}^{(2)} \text{ and } \boldsymbol{m} \notin \mathscr{M}_{n,\gamma}^{(2)}\right\} \\
&\subseteq \left\{K_v^\infty(\boldsymbol{m})\left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1\right) \le \beta\frac{n}{\log n}\right\} \cap \left\{K_v^\infty(\boldsymbol{m})\left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1\right) \ge \gamma\frac{n}{\log n}\right\} \\
&\subseteq \left\{K_v^\infty(\boldsymbol{m})\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \le \beta\frac{n}{\log n}\right\} \cap \left\{K_v^\infty(\boldsymbol{m})\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} - \mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \ge \left(\sqrt{\gamma}-\sqrt{\beta}\right)^2\frac{n}{\log n}\right\} \\
&\subseteq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \le \frac{\beta}{K_v^\infty(\boldsymbol{m})}\frac{n}{\log n}\right\} \cap \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} - \mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \ge \left(\sqrt{\frac{\gamma}{\beta}}-1\right)\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\}.
\end{aligned}$$

Let $\eta := \sqrt{\frac{\gamma}{\beta}} - 1$ and let $\epsilon \in (0,1)$. We consider the following decomposition:

$$\left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} - \mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \ge \eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\} = E_1 \cup E_2,$$

with:

$$\begin{aligned}
E_1 &:= \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} - \mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \ge \eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\} \cap \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}})\|_{\mathrm{op}} < \epsilon\right\}, \\
E_2 &:= \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} - \mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \ge \eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\} \cap \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}})\|_{\mathrm{op}} \ge \epsilon\right\}.
\end{aligned}$$

- For E_1, we apply Lemma B.1.2 with $\mathbf{A} := \widehat{\mathbf{G}}_{\boldsymbol{m}}$ and $\mathbf{B} := \mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}$:

$$\begin{aligned}
E_1 &\subseteq \left\{\frac{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}}}{1 - \|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}})\|_{\mathrm{op}}} \ge \eta\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\right\} \cap \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}})\|_{\mathrm{op}} < \epsilon\right\} \\
&\subseteq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge (1-\epsilon)\eta\right\}.
\end{aligned}$$

- For E_2, we have directly:

$$E_2 \subseteq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}(\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}})\|_{\mathrm{op}} \ge \epsilon\right\} \subseteq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|\,\|\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge \epsilon\right\}.$$

Thus, we obtain:

$$\forall \epsilon \in (0,1), \qquad E_1 \cup E_2 \subseteq \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}\|\mathbf{G}_{\boldsymbol{m}} - \widehat{\mathbf{G}}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge (1-\epsilon)\eta \wedge \epsilon\right\}.$$

We now choose ϵ maximizing $(1-\epsilon)\eta \wedge \epsilon$. This maximum is achieved when $\epsilon = (1-\epsilon)\eta$, that is:

$$\epsilon = \frac{\eta}{1-\eta} = 1 - \sqrt{\beta/\gamma} =: c(\beta,\gamma) \in (0,1).$$

Thus, we obtain:

$$\begin{aligned}
&\mathbb{P}\left[\Lambda_n^{(2)}(\beta,\gamma)^c\right] \\
&\le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_v^\infty(\boldsymbol{m}) \le \beta\frac{n}{\log n}}} \mathbb{P}\left[\left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \le \frac{\beta}{K_v^\infty(\boldsymbol{m})}\frac{n}{\log n}\right\} \cap \left\{\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge \frac{c(\beta,\gamma)}{\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right\}\right] \\
&\le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_v^\infty(\boldsymbol{m}) \le \beta\frac{n}{\log n}}} \mathbb{P}\left[\|\widehat{\mathbf{G}}_{\boldsymbol{m}} - \mathbf{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \ge c(\beta,\gamma)\sqrt{\frac{K_v^\infty(\boldsymbol{m})}{\beta}\frac{\log n}{n}}\right].
\end{aligned}$$

Let $x \coloneqq c(\beta,\gamma)\sqrt{\frac{K_\nu^\infty(\boldsymbol{m})}{\beta}\frac{\log n}{n}}$ and notice that $x \le 1$ if $K_\nu^\infty(\boldsymbol{m}) \le \beta\frac{n}{\log n}$. We apply Lemma 4.6.4 and Proposition B.2.1:

$$\begin{aligned}
&\mathbb{P}\big[\Lambda_n^{(2)}(\beta,\gamma)^c\big] \\
&\le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m})\le\beta\frac{n}{\log n}}} D_{\boldsymbol{m}} \exp\left(-\frac{n}{2}c^2(\beta,\gamma)\frac{K_\nu^\infty(\boldsymbol{m})}{\beta}\frac{\log n}{n}\left[K_\nu^\infty(\boldsymbol{m})\left(\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_\infty + \frac{2}{3}x\right)\right]^{-1}\right) \\
&\le \sum_{\substack{\boldsymbol{m}\in\mathbb{N}_+^p \\ K_\nu^\infty(\boldsymbol{m})\le\beta\frac{n}{\log n}}} D_{\boldsymbol{m}}\, n^{-c^2(\beta,\gamma)\frac{B}{2\beta}} \le n^{-c^2(\beta,\gamma)\frac{B}{2\beta}+2} H_n^{p-1},
\end{aligned}$$

where $B \coloneqq (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$. □

Now we can prove Theorem 4.4.4.

Proof of Theorem 4.4.4. Let $\delta \in (0,1)$ and $\gamma > \beta$ be constants to be chosen later. Let us introduce the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta) \coloneqq \Lambda_n^{(\iota)}(\beta,\gamma) \cap \widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)$ where $\Lambda_n^{(\iota)}(\beta,\gamma)$ and $\widetilde{\Omega}_n^{(\iota)}(\delta,\gamma)$ are defined by (4.23). On the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta)$, for all $\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(\iota)}$, for all $t \in S_{\boldsymbol{m}}$ we have:

$$\begin{aligned}
\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 &\le 2\|b-t\|_\mu^2 + 2\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}-t\|_\mu^2 \\
&\le 2\|b-t\|_\mu^2 + \frac{2}{1-\delta}\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}-t\|_n^2 \\
&\le 2\|b-t\|_\mu^2 + \frac{4}{1-\delta}\|b-t\|_n^2 + \frac{4}{1-\delta}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_n^2.
\end{aligned}$$

Taking the expectation yields for all $t \in S_{\boldsymbol{m}}$:

$$\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 \mathbf{1}_{\Xi_n^{(\iota)}(\beta,\gamma,\delta)}\Big] \le \left(2+\frac{4}{1-\delta}\right)\|b-t\|_\mu^2 + \frac{4}{1-\delta}\mathbb{E}\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_n^2. \tag{4.24}$$

On the event $\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c$, we use inequalities (4.12) and (4.16):

$$\begin{aligned}
\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 &\le 2\,\|b\|_\mu^2 + 2\,\|\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 \\
&\le 2\,\|b\|_\mu^2 + 2\,K_n^\mu(\hat{\boldsymbol{m}}_\iota)\|\mathbf{Y}\|_n^2 \\
&\le 2\,\|b\|_\mu^2 + 2\,K_\nu^\infty(\hat{\boldsymbol{m}}_\iota)\|\widehat{\mathbf{G}}_{\hat{\boldsymbol{m}}_\iota}^{-1}\|_{\mathrm{op}}\|\mathbf{Y}\|_n^2 \\
&\le 2\,\|b\|_\mu^2 + 4\beta\frac{n}{\log n}\|\mathbf{Y}\|_n^2.
\end{aligned}$$

Using Hölder's inequality as we did in (4.19), we obtain:

$$\begin{aligned}
&\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_\iota}\|_\mu^2 \mathbf{1}_{\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c}\Big] \le 2\,\|b\|_\mu^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big] \\
&+ 8\beta\frac{n}{\log n}\Big(\|b\|_{\mathrm{L}^{2r}(\mu)}^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]^{1/r'} + \sigma^2\,\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]\Big).
\end{aligned} \tag{4.25}$$

We see we need to control $\mathbb{P}\big[\Xi_n^{(\iota)}(\beta,\gamma,\delta)^c\big]$ by a term of order $n^{-2r'}$.

We have decomposed the risk as the sum of (4.24) and (4.25). We give different upper bounds on these two terms depending on whether we are in the compact case or the general case.

• *Compact case.* In equation (4.24), we apply Theorem 4.4.1: for all $\alpha \in (0, \alpha_{\beta,r'})$ we have:

$$\mathbb{E}\Big[\|b-\hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2 \mathbf{1}_{\Xi_n^{(1)}(\beta,\gamma,\delta)}\Big] \le \left(2+\frac{4}{1-\delta}\big(1+C(\theta)\big)\right) \inf_{\boldsymbol{m}\in\mathscr{M}_{n,\alpha}} \left(\inf_{t\in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}\right) + \frac{4\sigma^2}{1-\delta}\frac{\Sigma(\theta,q)}{n} + \frac{4}{1-\delta} C'\big(\|b\|^2_{\mathrm{L}^{2r}(\mu)}, \sigma^2\big) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},$$

with $\frac{\kappa(\alpha,\beta)}{r'} > 1$. To obtain an upper bound on (4.25), we apply Lemmas 4.6.5 and 4.6.6:

$$\begin{aligned}\mathbb{P}\big[\Xi_n^{(1)}(\beta,\gamma,\delta)^c\big] &\le \mathbb{P}\big[\widetilde{\Omega}_n^{(1)}(\delta,\gamma)^c\big] + \mathbb{P}\big[\Lambda_n^{(1)}(\beta,\gamma)^c\big] \\ &\le \left(n^{-\frac{h(\delta)}{\gamma}+2} + n^{-h(1-\frac{\gamma}{\beta})\frac{f_0}{\beta}+1}\right) H_n^{p-1},\end{aligned}$$

where $h(\delta) \coloneqq \delta + (1-\delta)\log(1-\delta)$ and $H_n \coloneqq \sum_{k=1}^n \frac{1}{k}$. In order to obtain a term of order $n^{-2r'}$, we need:

$$\begin{cases} \dfrac{h(\delta)}{\gamma} - 2 > 2r', \\ h\left(1-\dfrac{\gamma}{\beta}\right)\dfrac{f_0}{\beta} - 1 > 2r', \end{cases} \iff \begin{cases} h(\delta) > 2(1+r')\gamma, \\ h\left(1-\dfrac{\gamma}{\beta}\right) > (2r'+1)\dfrac{\beta}{f_0}, \end{cases}$$

$$\iff \begin{cases} \delta > h^{-1}\big(2(1+r')\gamma\big), \\ \gamma < \dfrac{1}{2(1+r')}, \\ h\left(1-\dfrac{\gamma}{\beta}\right) > (2r'+1)\dfrac{\beta}{f_0}. \end{cases}$$

Let us work on the last two conditions. Let $x \coloneqq \frac{\gamma}{\beta} > 1$, the conditions on (β,γ) become:

$$\begin{cases} x < \dfrac{1}{2(1+r')\beta}, \\ x\log x - x + 1 > (2r'+1)\dfrac{\beta}{f_0}. \end{cases}$$

The function $x \mapsto x\log x - x + 1$ is increasing on $(1,+\infty)$ and ranges from 0 to $+\infty$, so there exists $x_{f_0,\beta} > 1$ such that for all $x > x_{f_0,\beta}$ we have $x\log x - x + 1 > (2r'+1)\frac{\beta}{f_0}$. Hence we need to choose x such that:

$$x_{f_0,\beta} < x < \frac{1}{(2r'+2)\beta}. \tag{4.26}$$

This is possible only if $x_{f_0,\beta} < \frac{1}{(2r'+2)\beta}$, that is if:

$$(2r'+1)\frac{\beta}{f_0} < \frac{1}{(2r'+2)\beta}\log\left(\frac{1}{(2r'+2)\beta}\right) - \frac{1}{(2r'+2)\beta} + 1.$$

Let us introduce a new variable $y := (2r'+2)\beta$ and let $R = \frac{2r'+1}{2r'+2}$, the last inequality becomes:

$$\frac{R}{f_0} y + \frac{1+\log y}{y} < 1. \tag{4.27}$$

The function $y \mapsto \frac{R}{f_0} y + \frac{1+\log y}{y}$ is increasing on $(0,1)$, it tends to $-\infty$ at 0 and for $y = 1$ it is greater that 1, so there exists $y_{f_0,r'} \in (0,1)$ such that the condition (4.27) is satisfied on $(0, y_{f_0,r'})$. To sum up, we have shown that there exists $\beta_{f_0,r'} \in (0, \frac{1}{2r'+2})$ such that for every $\beta < \beta_{f_0,r'}$, the condition (4.26) is not empty. We choose:

$$\gamma := \beta x, \qquad x \text{ satisfying (4.26)}, \qquad \delta := \frac{1 + h^{-1}\big(2(1+r')\gamma\big)}{2},$$

and we obtain that:

$$\mathbb{E}\Big[\| b - \hat{b}_{\hat{\boldsymbol{m}}_1} \|_\mu^2 \mathbf{1}_{\Xi_n^{(1)}(\beta,\gamma,\delta)^c} \Big] \le C''(\|b\|_{\mathrm{L}^{2r}(\mu)}, \beta, \sigma^2)\, n^{-\lambda(\beta,r,f_0)} (\log n)^{\frac{p-1}{r'}-1},$$

where $\lambda(\beta, r, f_0) > 1$.

• *General case.* In equation (4.24), if we follow the proof of Theorem 4.4.1 (see Remark 4.6.3), we see that if $\alpha \in (0, \alpha^2_{\beta^{1/2},r'})$ then we have:

$$\mathbb{E}\| b - \hat{b}_{\hat{\boldsymbol{m}}_2} \|_n^2 \le C(\theta) \| b - t \|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \sigma^2 \frac{\Sigma(\theta,q)}{n} + C'\big(\|b\|^2_{\mathrm{L}^{2r}(\mu)}, \sigma^2\big) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha^{\frac12},\beta^{\frac12})/r'}},$$

with $\frac{\kappa(\alpha^{\frac12},\beta^{\frac12})}{r'} > 1$. Thus, we obtain:

$$\begin{aligned}\mathbb{E}\Big[\| b - \hat{b}_{\hat{\boldsymbol{m}}_2} \|_\mu^2 \mathbf{1}_{\Xi_n^{(2)}(\beta,\gamma,\delta)} \Big] \le{}& \Big(2 + \frac{4}{1-\delta}\big(1 + C(\theta)\big)\Big) \inf_{\boldsymbol{m} \in \mathscr{M}_{n,\alpha}^{(2)}} \Big(\inf_{t \in S_{\boldsymbol{m}}} \| b - t \|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \Big) \\ &+ \frac{4\sigma^2}{1-\delta} \frac{\Sigma(\theta,q)}{n} + \frac{4}{1-\delta} C'\big(\|b\|^2_{\mathrm{L}^{2r}(\mu)}, \sigma^2\big) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha^{\frac12},\beta^{\frac12})/r'}}.\end{aligned}$$

To obtain an upper bound on (4.25), we apply Lemmas 4.6.5 and 4.6.7:

$$\begin{aligned}\mathbb{P}\big[\Xi_n^{(2)}(\beta,\gamma,\delta)^c\big] &\le \mathbb{P}\big[\widetilde{\Omega}_n^{(2)}(\delta,\gamma)^c\big] + \mathbb{P}\big[\Lambda_n^{(2)}(\beta,\gamma)^c\big] \\ &\le \Big(n^{-\frac{h(\delta)}{\gamma}+2} + n^{-C(\beta,\gamma)\frac{B}{2\beta}+2} \Big) H_n^{p-1},\end{aligned}$$

where $C(\beta,\gamma) := \big(1 - \sqrt{\beta/\gamma}\big)^2$, $B := (\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \|_\infty + \frac{2}{3})^{-1}$ and $H_n := \sum_{k=1}^n \frac{1}{k}$. To obtain a term of order $n^{-2r'}$, we need:

$$\begin{cases} \dfrac{h(\delta)}{\gamma} - 2 > 2r', \\ C(\beta,\gamma)\dfrac{B}{2\beta} - 2 > 2r', \end{cases} \iff \begin{cases} h(\delta) > 2(1+r')\gamma, \\ C(\beta,\gamma)\dfrac{B}{2} > 2(1+r')\beta, \end{cases} \iff \begin{cases} \delta > h^{-1}\big(2(1+r')\gamma\big), \\ \gamma < \dfrac{1}{2(1+r')}, \\ \dfrac{C(\beta,\gamma)B}{4(1+r')} > \beta. \end{cases}$$

Let $x := \sqrt{\beta/\gamma} \in (0,1)$, the conditions on (β,γ) can be rewritten as:

$$\begin{cases} \dfrac{\beta}{x^2} < \dfrac{1}{2(1+r')}, \\ \beta < (1-x)^2 \dfrac{B}{4(1+r')}, \end{cases} \iff \beta < \frac{1}{2(1+r')}\left(x^2 \wedge (1-x)^2 \frac{B}{2}\right).$$

We choose x maximizing this bound. This maximum is achieved when $x^2 = (1-x)^2 \frac{B}{2}$, that is $x = \frac{\sqrt{B/2}}{1+\sqrt{B/2}}$. Finally we choose:

$$x := \frac{\sqrt{B/2}}{1+\sqrt{B/2}}, \qquad \gamma := \frac{\beta}{x^2}, \qquad \delta := \frac{1+h^{-1}\big(2(1+r')\gamma\big)}{2},$$

and we obtain that for all $\beta \in (0, \beta_{B,r'})$ with:

$$\beta_{B,r'} := \frac{1}{2(1+r')}\left(\frac{\sqrt{B/2}}{1+\sqrt{B/2}}\right)^2,$$

we have:

$$\mathbb{E}\Big[\|b - \hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2 \mathbf{1}_{\Xi_n^{(2)}(\beta,\gamma,\delta)^c}\Big] \le C''(\|b\|_{\mathrm{L}^{2r}(\mu)}, \beta, \sigma^2)\, n^{-\lambda(\beta,r,B)} (\log n)^{\frac{p-1}{r'}-1},$$

where $\lambda(\beta,r,B) > 1$. □

Appendix A

Laguerre functions

In this chapter, we gather different useful facts on the Laguerre functions.

A.1 Laguerre polynomials

The following formulas can be found in Chapter 22 of the book of Abramowitz & Stegun (1972). The Laguerre polynomials are orthogonal polynomials on $\mathbb{R}_+$ with respect to the weight function e^{-x}:

$$\int_0^{+\infty} L_k(x)\, L_j(x)\, \mathrm{e}^{-x}\, \mathrm{d}x = \delta_{k,j},$$

where $\delta_{k,j}$ is the Kronecker delta. Hence, the Laguerre functions $(\psi_k)_{k\geq 0}$ defined as:

$$\forall x \in \mathbb{R}_+, \quad \psi_k(x) := \sqrt{2}\, L_k(2x)\, \mathrm{e}^{-x},$$

form an orthonormal family in $\mathrm{L}^2(\mathbb{R}_+)$. One can show that this family is total, hence it is an orthonormal basis of $\mathrm{L}^2(\mathbb{R}_+)$. The Laguerre polynomials are given by the following closed expression:

$$L_k(x) = \sum_{j=0}^{k} \binom{k}{j} \frac{(-x)^j}{j!}.$$

These polynomials satisfy the following recursive relation:

$$(k+1)L_{k+1}(x) = (2k+1-x)L_k(x) - kL_{k-1}(x), \tag{A.1}$$

hence the Laguerre functions satisfy the same relation:

$$(k+1)\psi_{k+1}(x) = (2k+1-x)\psi_k(x) - k\psi_{k-1}(x).$$

We use this relation to compute the Laguerre functions in our simulations. The Laguerre polynomial L_k is also the solution of the Laguerre equation:

$$xL_k'' + (1-x)L_k' + kL_k = 0, \tag{A.2}$$

with initial conditions $L_k(0) = 1$ and $L'_k(0) = -k$. We have the following relation between the derivative of L_k and the polynomials L_k and L_{k-1}:

$$xL'_k(x) = k\big(L_k(x) - L_{k-1}(x)\big). \tag{A.3}$$

Finally, the following bound on Laguerre polynomials holds:

$$|L_k(x)| \leq \mathrm{e}^{\frac{x^2}{2}},$$

hence the Laguerre functions are bounded by $\sqrt{2}$, which is their common value at 0.

A.2 Sobolev–Laguerre spaces

The Sobolev–Laguerre spaces were initially defined as regularity spaces associated with the Laguerre differential operator by Bongioanni & Torrea (2009). The equivalent definition from the Laguerre coefficients comes from Section 7 of the paper of Comte & Genon-Catalot (2015).

Definition A.2.1. Let $s \in (0, +\infty)$, we define the Sobolev–Laguerre space of regularity s as:

$$\mathrm{W}^s(\mathbb{R}_+) := \left\{ f \in \mathrm{L}^2(\mathbb{R}_+) \,\middle|\, \sum_{k=0}^{+\infty} \langle f, \varphi_k \rangle^2 \, k^s < +\infty \right\}.$$

For $L > 0$, we define the Sobolev–Laguerre ball of radius L as:

$$\mathrm{W}^s(\mathbb{R}_+, L) := \left\{ f \in \mathrm{L}^2(\mathbb{R}_+) \,\middle|\, \sum_{k=0}^{+\infty} \langle f, \varphi_k \rangle^2 \, k^s \leq L \right\}.$$

Theorem A.2.2 (Comte & Genon-Catalot (2015))**.** *Let $s \in \mathbb{N}_+$. A function f belongs to $\mathrm{W}^s(\mathbb{R}_+)$ if and only if the two following conditions hold:*

1. *The function f admits derivatives up to order $s-1$ and $f^{(s-1)}$ is absolutely continuous.*
2. *For all $k \in \{0, \dots, s-1\}$, we have $x^{\frac{k+1}{2}} \sum_{j=0}^{k+1} \binom{k+1}{j} f^{(j)} \in \mathrm{L}^2(\mathbb{R}_+)$.*

A.3 Primitives of the Laguerre functions

Let Ψ_k be the primitive of the Laguerre function ψ_k that vanishes at 0:

$$\Psi_k(x) := \int_0^x \psi_k(t) \, \mathrm{d}t.$$

From (A.1) and (A.3), one can prove:

$$(x\psi_k)' = \frac{k+1}{2}\psi_{k+1} + \frac{1}{2}\psi_k - \frac{k}{2}\psi_{k-1}.$$

By integrating the last equation, we obtain the recurrence relation:

$$(k+1)\Psi_k(x) = 2x\psi_k(x) - \Psi_k(x) + k\Psi_{k-1}(x),$$

which can be used to compute exactly these functions. In this section remaining, we prove that the primitives of the Laguerre function are uniformly bounded. The sketch of the proof comes from fedja (2021).

Theorem A.3.1. *The functions $(\Psi_k)_{k\geq 0}$ are uniformly bounded.*

Proof. Let $u_k(x) \coloneqq L_k(x)\mathrm{e}^{-x/2}$. We first notice that the complete integral of u_k is uniformly bounded:

$$\begin{aligned}\int_0^{+\infty} u_k(x)\,\mathrm{d}x &= 2\int_0^{+\infty} L_k(2x)\mathrm{e}^{-x}\,\mathrm{d}x \\ &= 2\sum_{j=0}^{k}\binom{k}{j}\frac{(-2)^j}{j!}\int_0^{+\infty} x^j\mathrm{e}^{-x}\,\mathrm{d}x \\ &= 2\sum_{j=0}^{k}\binom{k}{j}(-2)^j = 2(-1)^k.\end{aligned}$$

We will show that $|\int_0^x u_k| \leq C|\int_0^\infty u_k| = 2C$ for an absolute constant $C > 0$. Since the k-th Laguerre polynomial satisfies the ODE (A.2), the function u_k satisfies:

$$xu_k'' + u_k' + \left(k + \frac{1}{2} - \frac{1}{4}x\right)u_k = 0 \tag{A.4}$$

To kill the first derivative, we consider $v_k(x) \coloneqq u_k(x^{3/2})x^{1/2}$. The functions u_k and v_k have the same partial integrals (up to the constant 2/3):

$$\forall x \geq 0, \quad \int_0^x v_k(t)\,\mathrm{d}t = \frac{2}{3}\int_0^{x^{2/3}} u_k(t)\,\mathrm{d}t.$$

The first two derivatives of v_k are:

$$\begin{aligned}v_k'(x) &= \frac{3}{2}xu_k'(x^{3/2}) + \frac{1}{2}x^{-1/2}u(x^{3/2}) \\ v_k''(x) &= \frac{9}{4}x^{3/2}u''(x^{3/2}) + \frac{9}{4}u'(x^{3/2}) - \frac{1}{4}x^{-3/2}u(x^{3/2})\end{aligned}$$

so using the ODE (A.4) for u yields the following ODE for v:

$$v_k'' + \Phi_k v_k = 0$$

where Φ_k is given by:

$$\Phi_k(x) \coloneqq \frac{9}{4}\left(\frac{k+\frac{1}{2}}{x^{1/2}} - \frac{1}{4}x\right) + \frac{1}{4x^2}.$$

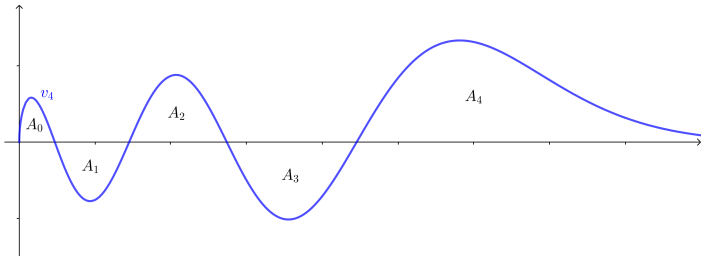

Figure A.1: Graph of v_4. We see that the area of the excursions is increasing.

The important properties of this function are that it is convex and decreasing.

Since the Laguerre polynomials have simple zeros, the function v_k has $(k+1)$ simple zeros (the zeros of the k-th Laguerre polynomial, and 0) so the integral of v_k can be decomposed as an alternating sum $\int_0^\infty v_k =: I_k = A_0 - A_1 + A_2 - \cdots + (-1)^k A_k$, where the A_i's are the unsigned areas of the excursions of v_k (see Figure A.1). Based on the following lemma (proven later), we claim that $A_0 < A_1 < \cdots < A_k$.

Lemma A.3.2. *Let F_1, F_2 be two $\mathscr{C}^1$ functions defined on an open interval I containing 0, and let y_1, y_2 be the solutions of the following ODE:*

$$\begin{cases} y_1'' + F_1(x)\,y_1 = 0 \\ y_2'' + F_2(x)\,y_2 = 0 \\ y_1(0) = y_2(0) = 0 \\ y_2'(0) \geq y_1'(0) > 0. \end{cases}$$

Let $M > 0$ such that y_1 and y_2 are positive on $J := (0, M) \subseteq I$. If $F_1 > F_2$ on J, then we have $y_1 \leq y_2$ on J.

Indeed, let $z > 0$ be a zero of v_k and assume w.l.o.g. that v_k is positive after z and negative before. Let $y_1(x) := -v_k(z-x)$ and $y_2(x) := v_k(z+x)$ (i.e. y_1 is the central inversion of v_k with respect to z). Thus, y_1 and y_2 satisfy the ODE:

$$\begin{cases} y_1'' + \Phi_k(z-x)\,y_1 = 0 \\ y_2'' + \Phi_k(z+x)\,y_2 = 0 \end{cases}$$

with the initial conditions $y_1(0) = y_2(0) = v_k(z) = 0$ and $y_1'(0) = y_2'(0) = v_k'(z) > 0$. Since Φ_k is decreasing, we have $\Phi_k(z-x) > \Phi_k(z+x)$, so Lemma A.3.2 yields that $y_1(x) \leq y_2(x)$ for $x > 0$ as long as $y_1(x)$ and $y_2(x)$ are positive. Hence, the area of the excursion preceding z is smaller than the area of the excursion following z.

Now let z be the last zero of v_k, and let us assume w.l.o.g. that v_k is positive after z (otherwise we consider $-v_k$, it satisfies the same ODE than v_k). In this case, $I_k = A_0 - A_1 + \ldots - A_{k-1} + A_k$ with $A_0 < \ldots < A_{k-1} < A_k$. Thus the maximum value of $|\int_0^x v_k|$ is attained either by taking the complete integral (it is the maximum of $\int_0^x v_k$), either by leaving out the last excursion (it is the minimum

of $\int_0^x v_k$). We show that the second option is dominated by the first one: there exists an absolute constant $C > 0$ such that $A_k - I_k \leq C I_k$. To do that it suffices to show that $A_{k-1} \leq c A_k$ for absolute constant $c \in (0,1)$, hence $A_k - I_k \leq \frac{c}{1-c} I_k$.

The strategy is to compare the function v_k to the Airy function of the first kind. This function is solution of the ODE:

$$\begin{cases} y'' - xy = 0 \\ \lim\limits_{x \to +\infty} y(x) = 0. \end{cases} \tag{A.5}$$

Let z_ℓ^* and z_p^* be respectively the last and the penultimate zeros of Ai. We recall that z_ℓ^* is negative (all the zeros of Ai are negative), and that Ai is negative on (z_p^*, z_ℓ^*) and positive on $(z_\ell^*, +\infty)$.

Lemma A.3.3. *Let* Ai *be the Airy function of the first kind, and let z be the last zero of v_k. There exists a decreasing linear function Υ_k satisfying $\Upsilon_k(z) = \Phi_k(z)$, and there exists $a > 0$ and $b \in \mathbb{R}$, such that the function* $\mathrm{Ai}(ax+b)$ *satisfies:*

$$\begin{cases} y'' + \Upsilon_k(x)\, y = 0 \\ y(z) = 0 \end{cases}$$

and such that it stays positive on $(z, +\infty)$ and tends to 0 *at* $+\infty$.

Let $w_k(x) := \varepsilon \mathrm{Ai}(ax+b)$ where a and b are given by Lemma A.3.3, and $\varepsilon > 0$ is small enough such that $0 < w_k'(z) < v_k'(z)$. Let z_ℓ and z_p be respectively the last and the penultimate zeros of w_k. By definition of w_k, its last zero is the same than v_k's, that is $z_\ell = z$. Moreover, the zeros of w_k and Ai are linked by the linear transformation $x \mapsto ax + b$: we have $z_\ell^* = a z_\ell + b$ and $z_p^* = a z_p + b$.

Let us consider:

$$\mathscr{W} := \det \begin{pmatrix} v_k & w_k \\ v_k' & w_k' \end{pmatrix} = v_k w_k' - v_k' w_k,$$

the Wronskian of v_k and w_k. Since $\mathscr{W}$ vanishes at z and $+\infty$, we have:

$$0 = \int_z^{+\infty} \mathscr{W}'(x)\,\mathrm{d}x = \int_z^{+\infty} \big(\Phi_k(x) - \Upsilon_k(x)\big) \underbrace{v_k(x)\, w_k(x)}_{>0}\,\mathrm{d}x,$$

so the sign of $\Phi_k - \Upsilon_k$ must change on $(z, +\infty)$. Since $\Phi_k - \Upsilon_k$ is convex and $(\Phi_k - \Upsilon_k)(z) = 0$, this is possible only if it is negative and then positive. Hence, $\mathscr{W}$ is first decreasing and then increasing on $(z, +\infty)$. Since it starts from zero at z and tends to zero at $+\infty$, we conclude that $\mathscr{W}$ is negative on $(z, +\infty)$.

It follows that $w_k < v_k$ on $(z, +\infty)$. Indeed, by contradiction if $x_0 := \inf\{x \in (z, +\infty) \mid w_k(x) \geq v_k(x)\}$ existed and was finite, we would have $v_k(x_0) = w_k(x_0) > 0$ and $v_k'(x_0) \leq w_k'(x_0)$. Thus, we would have $\mathscr{W}(x_0) = v_k(x_0) w_k'(x_0) - v_k'(x_0) w_k(x_0) \geq 0$; contradiction. We conclude that the area of the last excursion of w_k is less than v_k's: $\int_{z_\ell}^{\infty} w_k \leq A_k$.

We have seen that $\Phi_k - \Upsilon_k$ was negative then positive on $(z, +\infty)$. By convexity, it has to be positive on $(0, z)$, i.e. $\Phi_k > \Upsilon_k$ on the left of z. We apply Lemma

A.3.2 to $-v_k(z-x)$ and $-w_k(z-x)$, and we conclude that the area of the penultimate excursion of v_k is less than w_k's: $A_{k-1} \leq |\int_{z_p}^{z_\ell} w_k|$.

We conclude that A_{k-1}/A_k is bounded above by the ratio of the areas of the penultimate excursion to the last excursion of w_k. By a linear change of variable, this ratio is equal to the ratio of the areas of the penultimate excursion to the last excursion of the Airy function:

$$\frac{A_{k-1}}{A_k} \leq \frac{\left|\int_{z_p}^{z_\ell} w_k(x)\,\mathrm{d}x\right|}{\int_{z_\ell}^{+\infty} w_k(x)\,\mathrm{d}x} = \frac{\left|\int_{z_p^*}^{z_\ell^*} \mathrm{Ai}(x)\,\mathrm{d}x\right|}{\int_{z_\ell^*}^{+\infty} \mathrm{Ai}(x)\,\mathrm{d}x} =: c.$$

This is an absolute constant, we just need to prove it is smaller than 1 to end the proof. The function Ai satisfies an ODE of the form $y'' + F(x)y = 0$ with $F(x) = -x$ a decreasing function, thus by considering the functions $y_1(x) = -\mathrm{Ai}(z_\ell^* - x)$ and $y_2 := \mathrm{Ai}(z_\ell^* + x)$, we can apply again Lemma A.3.2 (as we did with v_k), to conclude that $y_1(x) \leq y_2(x)$ for $x > 0$ as long as $y_1(x)$ is positive (y_2 being positive for every $x > 0$). This proves that $|\int_{z_p^*}^{z_\ell^*} \mathrm{Ai}| < \int_{z_\ell^*}^{\infty} \mathrm{Ai}$, that is $c < 1$. □

Proof of Lemma A.3.2. First, let us consider the case $y_2'(0) > y_1'(0)$. Let $\mathscr{W} := y_1 y_2' - y_1' y_2$ be the Wronskian of y_1 and y_2. Then $\mathscr{W}' = (F_1 - F_2) y_1 y_2$ is positive on J and $\mathscr{W}(0) = 0$, so $\mathscr{W} > 0$ on J.

By contradiction, suppose there exists $x \in J$ such that $y_1(x) \geq y_2(x) > 0$ and consider $x_0 := \inf\{x \in J \,|\, y_1(x) \geq y_2(x) > 0\}$. Since $y_2'(0) > y_1'(0)$, we know that y_1 is below y_2 on a right neighborhood of 0, so $x_0 > 0$. By continuity, we have $y_1(x_0) = y_2(x_0) > 0$, so we must have $y_1'(x_0) \geq y_2'(x_0)$; otherwise, we would have $y_1(x) \geq y_2(x) > 0$ on a left neighborhood of x_0, which is impossible if $x_0 > 0$. Thus $\mathscr{W}(x_0) = y_1(x_0)[y_2'(x_0) - y_1'(x_0)] \leq 0$ with $x_0 \in J$; contradiction.

Now consider the case $y_2'(0) = y_1'(0)$. For $\varepsilon > 0$ small enough, we consider y_ε the solution of:

$$\begin{cases} y_\varepsilon'' + F_1(x) y_\varepsilon = 0 \\ y_\varepsilon(0) = 0 \\ y_\varepsilon'(0) = y_2'(0) - \varepsilon > 0. \end{cases}$$

Applying the first case, we have $y_\varepsilon \leq y_2$ while y_ε and y_2 are positive. Taking $\varepsilon \to 0$, we obtain $y_1 \leq y_2$. □

Proof of Lemma A.3.3. The Airy function is solution of (A.5), so $\mathrm{Ai}(ax+b)$ (with $a > 0$) satisfies:

$$\begin{cases} y'' - a^2(ax+b)y = 0 \\ \lim\limits_{x \to +\infty} y(x) = 0. \end{cases}$$

We need to determine (a, b) such that the two following conditions hold:

1. Let $\Upsilon_k(x) := -a^2(ax+b)$, we need to choose a, b such that $\Upsilon_k(z) = \Phi_k(z)$.
2. Let z_ℓ^* be the last zero of Ai. We know that $z_\ell^* < 0$ and that Ai is positive on $(z_\ell^*, +\infty)$, thus we need to choose a, b such that $az + b = z_\ell^*$ so $\mathrm{Ai}(ax+b)$ stays positive on $(z, +\infty)$ and vanishes at z.

Thus, $(a,b) \in \mathbb{R}_+^* \times \mathbb{R}$ must be solution of:

$$\begin{cases} -a^2(az+b) = \Phi_k(z) \\ az+b = z_\ell^* \end{cases} \iff \begin{cases} a^2 = -\frac{\Phi_k(z)}{z_\ell^*} \\ az+b = z_\ell^* \end{cases} \tag{A.6}$$

Since $z_\ell^* < 0$, this system has a solution iff $\Phi_k(z) > 0$. By contradiction, if we had $\Phi_k(z) \leq 0$, then Φ_k would be negative on $(z,+\infty)$. Since the function v_k is positive after z and satisfies $v_k''(x) = -\Phi_k(x)v_k(x)$, then v_k would be strictly convex on $(z,+\infty)$. But the function v_k starts from zero at z with a positive derivative, stays positive, and tends to 0 at $+\infty$, so it cannot be strictly convex on $(z,+\infty)$; contradiction. Thus, the system (A.6) has a solution. □

Appendix B

Miscellaneous results

B.1 Linear Algebra

Lemma B.1.1. *Let E be a Euclidean vector space and let $\ell : E \to \mathbb{R}^n$ be an injective linear map. For $y \in \mathbb{R}^n$, the solution of the problem:*

$$\hat{a} := \underset{a \in E}{\arg\min} \, \| y - \ell(a) \|_{\mathbb{R}^n}^2$$

is given by:

$$\hat{a} = \left[(\ell^* \circ \ell)^{-1} \circ \ell^* \right](y),$$

where $\ell^* : \mathbb{R}^n \to E$ is characterized by the relation $\langle y, \ell(a) \rangle_{\mathbb{R}^n} = \langle \ell^*(y), a \rangle_E$.

Lemma B.1.2. *Let $\mathbf{A}$, $\mathbf{B}$ be square matrices. If $\mathbf{A}$ is invertible and $\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}} < 1$, then $\mathbf{A}+\mathbf{B}$ is invertible and it holds:*

$$\left\| (\mathbf{A}+\mathbf{B})^{-1} - \mathbf{A}^{-1} \right\|_{\text{op}} \leq \frac{\|\mathbf{A}^{-1}\|_{\text{op}}^2 \|\mathbf{B}\|_{\text{op}}}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}.$$

Proof. Since $\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}} < 1$, its Neumann series is normally convergent and we have:

$$\sum_{k=0}^{+\infty} (-1)^k (\mathbf{A}^{-1}\mathbf{B})^k = (\mathbf{I}_m + \mathbf{A}^{-1}\mathbf{B})^{-1}.$$

Hence:

$$\begin{aligned}
\left\| (\mathbf{A}+\mathbf{B})^{-1} - \mathbf{A}^{-1} \right\|_{\text{op}} &\leq \left\| \mathbf{A}^{-1} \right\|_{\text{op}} \left\| (\mathbf{I}_m + \mathbf{A}^{-1}\mathbf{B})^{-1} - \mathbf{I}_m \right\|_{\text{op}} \\
&\leq \left\| \mathbf{A}^{-1} \right\|_{\text{op}} \sum_{k=1}^{+\infty} \left\| \mathbf{A}^{-1}\mathbf{B} \right\|_{\text{op}}^k \\
&= \left\| \mathbf{A}^{-1} \right\|_{\text{op}} \frac{\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}} \leq \frac{\|\mathbf{A}^{-1}\|_{\text{op}}^2 \|\mathbf{B}\|_{\text{op}}}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}.
\end{aligned}$$

□

B.2 Combinatorics

Proposition B.2.1. *For $n \geq 1$ and $p \geq 2$ we have:*

$$\mathrm{Card}\{\boldsymbol{m} \in \mathbb{N}_+^p \,|\, m_1 \cdots m_p \leq n\} \leq n\, H_n^{p-1},$$

where $H_n := \sum_{k=1}^n \frac{1}{k}$ is the n-th harmonic number.

Proof. We compute:

$$\begin{aligned}
\mathrm{Card}\{\boldsymbol{m} \in \mathbb{N}_+^p \,|\, D_{\boldsymbol{m}} \leq n\} &= \sum_{m_1=1}^n \cdots \sum_{m_p=1}^n \mathbf{1}_{m_1 \cdots m_p \leq n} \\
&= \sum_{m_1=1}^n \cdots \sum_{m_p=1}^n \mathbf{1}_{m_p \leq \frac{n}{m_1 \cdots m_{p-1}}} \\
&= \sum_{m_1=1}^n \cdots \sum_{m_{p-1}=1}^n \left\lceil \frac{n}{m_1 \cdots m_{p-1}} \right\rceil \\
&\leq \sum_{m_1=1}^n \cdots \sum_{m_{p-1}=1}^n \frac{n}{m_1 \cdots m_{p-1}} = n\, H_n^{p-1}.
\end{aligned}$$

□

Theorem B.2.2 (Divisor bound)**.** *Let $N \in \mathbb{N}_+$ and let* $\mathrm{div}(N)$ *be the set of divisors of N. We have for all $\epsilon > 0$:*

$$\mathrm{Card}\big(\mathrm{div}(N)\big) = \mathrm{o}(N^\epsilon).$$

As a consequence, we have for all $\epsilon > 0$:

$$\mathrm{Card}\{\boldsymbol{m} \in \mathbb{N}_+^p \,|\, m_1 \cdots m_p = N\} \leq \mathrm{Card}\big(\mathrm{div}(N)\big)^p = \mathrm{o}(N^\epsilon).$$

A proof of this result can be found in Tao (2008).

Bibliographie

M. ABRAMOWITZ & I. A. STEGUN : *Handbook of Mathematical Functions : with Formulas, Graphs, and Mathematical Tables*. Num. 55 de Applied Mathematics Series. National Bureau of Standards, New York, NY, 10e édn, 1972.

H. AKAIKE : Information Theory and an Extension of the Maximum Likelihood Principle. *In* B. N. PETROV & F. CSAKI, éds : *Proceedings of the 2nd International Symposium on Information Theory*, p. 267–281, Budapest, 1973. Akademia Kiado.

S. ARLOT & P. MASSART : Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research*, 10(10):245–279, 2009.

R. ASKEY & S. WAINGER : Mean Convergence of Expansions in Laguerre and Hermite Series. *American Journal of Mathematics*, 87(3):695–708, 1965.

S. ASMUSSEN & H. ALBRECHER : *Ruin probabilities*, vol. 14 de *Advanced series on statistical science and applied probability*. World Scientific, Singapore ; New Jersey, 2nd édn, 2010.

Y. BARAUD : Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4):467–493, 2000.

Y. BARAUD : Model selection for regression on a random design. *ESAIM : Probability and Statistics*, 6:127–146, 2002.

A. BARRON, L. BIRGÉ & P. MASSART : Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

J.-P. BAUDRY, C. MAUGIS & B. MICHEL : Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

D. BELOMESTNY, F. COMTE & V. GENON-CATALOT : Nonparametric Laguerre estimation in the multiplicative censoring model. *Electronic Journal of Statistics*, 10(2), 2016.

D. BELOMESTNY, F. COMTE & V. GENON-CATALOT : Sobolev-Hermite versus Sobolev nonparametric density estimation on $\mathbb{R}$. *Annals of the Institute of Statistical Mathematics*, 71(1):29–62, 2019.

R. BENHADDOU, M. PENSKY & R. RAJAPAKSHAGE : Anisotropic functional Laplace deconvolution. *Journal of Statistical Planning and Inference*, 199:271–285, 2019.

L. BIRGÉ & P. MASSART : Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

L. BIRGÉ & P. MASSART : From Model Selection to Adaptive Estimation. *In* D. POLLARD, E. TORGERSEN & G. L. YANG, éds : *Festschrift for Lucien Le Cam*, p. 55–87. Springer New York, New York, NY, 1997.

L. BIRGÉ & P. MASSART : Minimum Contrast Estimators on Sieves : Exponential Bounds and Rates of Convergence. *Bernoulli*, 4(3):329–375, 1998.

L. BIRGÉ & P. MASSART : Minimal Penalties for Gaussian Model Selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2007.

B. BONGIOANNI & J. L. TORREA : Sobolev spaces associated to the harmonic oscillator. *Proceedings of the Indian Academy of Sciences - Section A*, 116(3):337–360, 2006.

B. BONGIOANNI & J. L. TORREA : What is a Sobolev space for the Laguerre function systems ? *Studia Mathematica*, 192(2):147–172, 2009.

C. BUTUCEA : Deconvolution of supersmooth densities with smooth noise. *Canadian Journal of Statistics*, 32(2):181–192, 2004.

C. BUTUCEA & A. B. TSYBAKOV : Sharp Optimality in Density Deconvolution with Dominating Bias. I. *Theory of Probability & Its Applications*, 52(1):24–39, 2008a.

C. BUTUCEA & A. B. TSYBAKOV : Sharp Optimality in Density Deconvolution with Dominating Bias. II. *Theory of Probability & Its Applications*, 52(2):237–249, 2008b.

A. BÖTTCHER & S. M. GRUDSKY : *Toeplitz Matrices, Asymptotic Linear Algebra, and Functional Analysis*. Birkhäuser Basel, Basel, 2000.

A. BÖTTCHER & S. M. GRUDSKY : *Spectral properties of banded Toeplitz matrices*. Society for Industrial and Applied Mathematics, Philadelphia, 2005.

R. J. CARROLL & P. HALL : Optimal Rates of Convergence for Deconvolving a Density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

G. CHAGNY : *Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie*. Thèse de doctorat, Paris Descartes, Paris, 2013a.

G. CHAGNY : Penalization versus Goldenshluger–Lepski strategies in warped bases regression. *ESAIM : Probability and Statistics*, 17:328–358, 2013b.

G. CHAGNY : Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 22(4):253–282, 2013c.

R. Y. CHEN, A. GITTENS & J. A. TROPP : The masked sample covariance estimator : an analysis using matrix concentration inequalities. *Information and Inference*, 1(1):2–20, 2012.

A. COHEN, M. A. DAVENPORT & D. LEVIATAN : On the Stability and Accuracy of Least Squares Approximations. *Foundations of Computational Mathematics*, 13(5):819–834, 2013.

F. COMTE, C. DUVAL & O. SACKO : Optimal Adaptive Estimation on $\mathbb{R}$ or $\mathbb{R}_+$ of the Derivatives of a Density. *Mathematical Methods of Statistics*, 29(1):1–31, 2020.

F. COMTE & V. GENON-CATALOT : Regression function estimation on non compact support in an heteroscesdastic model. *Metrika*, 83(1):93–128, 2020a.

F. COMTE : *Estimation non-paramétrique*. Spartacus supérieur. Spartacus IDH, Paris, 2e éd édn, 2017.

F. COMTE, C.-A. CUENOD, M. PENSKY & Y. ROZENHOLC : Laplace deconvolution on the basis of time domain data and its application to dynamic contrast-enhanced imaging. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(1):69–94, 2017.

F. COMTE & V. GENON-CATALOT : Adaptive Laguerre density estimation for mixed Poisson models. *Electronic Journal of Statistics*, 9(1):1113–1149, 2015.

F. COMTE & V. GENON-CATALOT : Laguerre and Hermite bases for inverse problems. *Journal of the Korean Statistical Society*, 47(3):273–296, 2018.

F. COMTE & V. GENON-CATALOT : Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics*, 72(4):1023–1054, 2020b.

F. COMTE & C. LACOUR : Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(2):569–609, 2013.

F. COMTE & G. MABON : Laguerre deconvolution with unknown matrix operator. *Mathematical Methods of Statistics*, 26(4):237–266, 2017.

F. COMTE & N. MARIE : On a Nadaraya-Watson estimator with two bandwidths. *Electronic Journal of Statistics*, 15(1), 2021.

F. COMTE, Y. ROZENHOLC & M.-L. TAUPIN : Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 34(3):431–452, 2006.

K. CROUX & N. VERAVERBEKE : Nonparametric estimators for the probability of ruin. *Insurance : Mathematics and Economics*, 9(2-3):127–130, 1990.

A. DELAIGLE & I. GIJBELS : Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1):19–47, 2004.

R. A. DEVORE & G. G. LORENTZ : *Constructive approximation*. Num. 303 de Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin ; New York, 1993.

D. L. DONOHO & I. M. JOHNSTONE : Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

D. L. DONOHO & I. M. JOHNSTONE : Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

D. L. DONOHO & I. M. JOHNSTONE : Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3), 1998.

D. L. DONOHO, I. M. JOHNSTONE, G. KERKYACHARIAN & D. PICARD : Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2), 1996.

S. Y. EFROMOVICH : Nonparametric Estimation of a Density of Unknown Smoothness. *Theory of Probability & Its Applications*, 30(3):557–568, 1986.

S. EFROMOVICH : *Nonparametric curve estimation : methods, theory and applications*. Springer series in statistics. Springer, New York, 1999.

J. FAN : On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *The Annals of Statistics*, 19(3):1257–1272, 1991.

FEDJA : Proving that the primitives of the laguerre functions are uniformly bounded. MathOverflow, fév. 2021. URL `https://mathoverflow.net/q/383022`.

E. W. FREES : Nonparametric Estimation of the Probability of Ruin. *ASTIN Bulletin*, 16(S1):S81–S90, 1986.

H. U. GERBER & E. S. SHIU : On the Time Value of Ruin. *North American Actuarial Journal*, 2(1):48–72, 1998.

A. GITTENS & J. A. TROPP : Tail bounds for all eigenvalues of a sum of random matrices. *arXiv :1104.4513 [math]*, 2011. arXiv : 1104.4513.

A. GOLDENSHLUGER & O. LEPSKI : Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4), 2008.

A. GOLDENSHLUGER & O. LEPSKI : Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probability Theory and Related Fields*, 143(1):41–71, 2009.

A. GOLDENSHLUGER & O. LEPSKI : Bandwidth selection in kernel density estimation : Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

L. GYÖRFI, M. KOHLER, A. KRZYŻAK & H. WALK : *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, New York, NY, 2002.

W. HARDLE & J. S. MARRON : Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *The Annals of Statistics*, 13(4), 1985.

M. L. HAZELTON & B. A. TURLACH : Nonparametric density deconvolution by weighted kernel estimators. *Statistics and Computing*, 19(3):217–228, 2009.

M. L. HAZELTON & B. A. TURLACH : Semiparametric Density Deconvolution. *Scandinavian Journal of Statistics*, 37(1):91–108, 2010.

C. HIPP : Estimators and bootstrap confidence intervals for ruin probabilities. *ASTIN Bulletin*, 19:57–70, 1989.

L. HOGBEN, éd. *Handbook of Linear Algebra*, chap. 15. Chapman and Hall/CRC, 2nd édn, 2013.

T. KLEIN & E. RIO : Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

M. KÖHLER, A. SCHINDLER & S. SPERLICH : A Review and Comparison of Bandwidth Selection Methods for Kernel Regression : Review of Bandwidth Selection for Regression. *International Statistical Review*, 82(2):243–274, 2014.

C. LACOUR & P. MASSART : Minimal penalty for Goldenshluger–Lepski method. *In Memoriam : Evarist Giné*, 126(12):3774–3789, 2016.

C. LACOUR, P. MASSART & V. RIVOIRARD : Estimator Selection : a New Method with Applications to Kernel Density Estimation. *Sankhya A*, 79(2):298–335, 2017.

M. LEDOUX : On Talagrand's deviation inequalities for product measures. *ESAIM : Probability and Statistics*, 1:63–87, 1997.

O. LEPSKI & T. WILLER : Oracle inequalities and adaptive estimation in the convolution structure density model. *The Annals of Statistics*, 47(1):233–287, 2019.

G. MABON : Adaptive Deconvolution on the Non-negative Real Line : Adaptive deconvolution on $\mathbb{R}_+$. *Scandinavian Journal of Statistics*, 44(3):707–740, 2017.

C. L. MALLOWS : Some Comments on C_p. *Technometrics*, 15(4):661–675, 1973.

E. MASIELLO : On semiparametric estimation of ruin probabilities in the classical risk model. *Scandinavian Actuarial Journal*, 2014(4):283–308, 2014.

E. MASRY : Multivariate probability density deconvolution for stationary random processes. *IEEE Transactions on Information Theory*, 37(4):1105–1115, 1991.

P. MASSART : The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

P. MASSART : *Concentration Inequalities and Model Selection : École d'Été de Probabilités de Saint-Flour XXXIII - 2003*. Num. 1896 de Lecture Notes in Mathematics. Springer-Verlag, Berlin ; New York, 2007.

A. MEISTER : *Deconvolution Problems in Nonparametric Statistics*, vol. 193 de *Lecture Notes in Statistics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Y. MEYER : *Ondelettes et opérateurs*, vol. 1. Hermann, Paris, 1990.

R. MNATSAKANOV, L. L. RUYMGAART & F. H. RUYMGAART : Nonparametric estimation of ruin probabilities given a random sample of claims. *Mathematical Methods of Statistics*, 17(1):35–43, 2008.

E. A. NADARAYA : On Estimating Regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

M. PENSKY & B. VIDAKOVIC : Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27(6):2033–2053, 1999.

S. M. PITTS : Nonparametric estimation of compound distributions with applications in insurance. *Annals of the Institute of Statistical Mathematics*, 46(3):537–555, 1994.

K. POLITIS : Semiparametric Estimation for Non-Ruin Probabilities. *Scandinavian Actuarial Journal*, 2003(1):75–96, 2003.

G. REBELLES : Structural adaptive deconvolution under Lp-losses. *Mathematical Methods of Statistics*, 25(1):26–53, 2016.

O. SACKO : Hermite density deconvolution. *Latin American Journal of Probability and Mathematical Statistics*, 17(1):419–443, 2020.

G. SCHWARZ : Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2):461–464, 1978.

Y. SHIMIZU : Estimation of the expected discounted penalty function for Lévy insurance risks. *Mathematical Methods of Statistics*, 20(2):125–149, 2011.

Y. SHIMIZU : Non-parametric estimation of the Gerber–Shiu function for the Wiener–Poisson risk model. *Scandinavian Actuarial Journal*, 2012(1):56–69, 2012.

Y. SHIMIZU & Z. ZHANG : Estimating Gerber–Shiu functions from discretely observed Lévy driven surplus. *Insurance : Mathematics and Economics*, 74:84–98, 2017.

L. A. STEFANSKI & R. J. CARROLL : Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

W. SU, B. SHI & Y. WANG : Estimating the Gerber-Shiu function under a risk model with stochastic income by Laguerre series expansion. *Communications in Statistics - Theory and Methods*, 49(23):5686–5708, 2020.

W. SU, Y. YONG & Z. ZHANG : Estimating the Gerber–Shiu function in the perturbed compound Poisson model by Laguerre series expansion. *Journal of Mathematical Analysis and Applications*, 469(2):705–729, 2019.

W. SU & W. YU : Asymptotically Normal Estimators of the Gerber-Shiu Function in Classical Insurance Risk Model. *Mathematics*, 8(10):1638, 2020.

M. TALAGRAND : New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

T. TAO : The divisor bound, sept. 2008. URL `https://terrytao.wordpress.com/2008/09/23/the-divisor-bound/`.

J. A. TROPP : User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A. B. TSYBAKOV : *Introduction to nonparametric estimation*. Springer series in statistics. Springer, New York ; London, 2009.

T. VARESCHI : Noisy Laplace deconvolution with error in the operator. *Journal of Statistical Planning and Inference*, 157-158:16 – 35, 2015.

G. S. WATSON : Smooth Regression Analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

E. YOUNDJÉ & M. T. WELLS : Optimal bandwidth selection for multivariate kernel deconvolution density estimation. *TEST*, 17(1):138–162, 2008.

Z. ZHANG : Estimating the Gerber–Shiu function by Fourier–Sinc series expansion. *Scandinavian Actuarial Journal*, 2017(10):898–919, 2017.

Z. ZHANG & W. SU : A new efficient method for estimating the Gerber–Shiu function in the classical risk model. *Scandinavian Actuarial Journal*, 2018(5):426–449, 2018.

Z. ZHANG & W. SU : Estimating the Gerber–Shiu function in a Lévy risk model by Laguerre series expansion. *Journal of Computational and Applied Mathematics*, 346:133–149, 2019.