ADAPTIVE ESTIMATION OF THE CONDITIONAL INTENSITY OF MARKER-DEPENDENT COUNTING PROCESSES

F. COMTE[(1)], S. GAÏFFAS[(2)] & A. GUILLOUX[(3)]

Abstract. We propose in this work an original estimator of the conditional intensity of a marker-dependent counting process, that is, a counting process with covariates. We use model selection methods and provide a non asymptotic bound for the risk of our estimator on a compact set. We show that our estimator reaches automatically a convergence rate over a functional class with a given (unknown) anisotropic regularity. Then, we prove a lower bound which establishes that this rate is optimal. Lastly, we provide a short illustration of the way the estimator works in the context of conditional hazard estimation.

October 23, 2008

AMS (2000) subject classification. 62N02, 62G05.

Keywords. Marker-dependent counting process. Conditional intensity. Model selection. Adaptive estimation. Minimax and Nonparametric methods. Censored data. Conditional hazard function.

1. Introduction

As counting processes can model a great diversity of observations, especially in medicine, actuarial science or economics, their statistical inference has received a continuous attention since half a century - see Andersen et al. (1993) for the most detailed presentation on the subject. In this paper, we propose a new strategy, based on model selection, for the inference for counting processes in presence of covariates. The model considered can be described as follows.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ a filtration satisfying the usual conditions. Let N be a marker-dependent counting process, with compensator Λ with respect to $(\mathcal{F}_t)_{t\geq 0}$, such that $N - \Lambda = M$, where M is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. We assume that N is a marker-dependent counting process satisfying the Aalen multiplicative intensity model in the sense that :

$$\Lambda(t) = \int_0^t \alpha(X, z)Y(z)dz, \text{ for all } t \geq 0 \tag{1}$$

where X is a vector of covariates in $\mathbb{R}^d$ which is $\mathcal{F}_0$-measurable, the process Y is nonnegative and predictable and α is an unknown deterministic function called intensity.

The purpose of this paper is to estimate the intensity function α on the basis of the observation of a n-sample $(X_i, N^i(z), Y^i(z), z \leq \tau)$ for $i = 1, \ldots, n$, where $\tau < +\infty$.

(1) MAP5, University Paris Descartes, France. email: fabienne.comte@parisdescartes.fr,
(2) LSTA University Pierre et Marie Curie, France. email: stephane.gaiffas@upmc.fr,
(3) LSTA University Pierre et Marie Curie, France. email: agathe.guilloux@upmc.fr.

There are many examples, crucial in practice, which fulfill this model. For the seek of conciseness, we restrict our presentation to the three following ones.

Example 1 (Regression model for right-censored data). Let T be a nonnegative random variable (r.v.) and X a vector of covariates in $\mathbb{R}^d$, with respective cumulative distribution functions (c.d.f.) F_T and F_X. We consider in addition that T can be censored. We introduce the nonnegative r.v. C, with c.d.f. G, such that the observable r.v. are $Z = T \wedge C$, $\delta = \mathbb{1}(T \leq C)$ and X. We assume that:

$$(\mathcal{C}) : \ T \text{ and } C \text{ are independent conditionally to } X.$$

In this case, the processes to consider (see e.g. Andersen et al. (1993)) are given, for $i = 1, \ldots, n$ and $z \geq 0$, by:

$$N^i(z) = \mathbb{1}(Z_i \leq z, \delta_i = 1) \text{ and } Y^i(z) = \mathbb{1}(Z_i \geq z).$$

The unknown intensity function α to be estimated is the conditional hazard rate of the r.v. T given $X = x$ defined, for all $z > 0$ by:

$$\alpha(x, z) = \alpha_{T|X}(x, z) = \frac{f_{T|X}(x, z)}{1 - F_{T|X}(x, z)},$$

where $f_{T|X}$ and $F_{T|X}$ are respectively the conditional probability density function (p.d.f.) and the conditional c.d.f. of Y given X.

Nonparametric estimation of the hazard rate in presence of covariates was initiated by Beran (1981). Stute (1986), Dabrowska (1987), McKeague and Utikal (1990) and Li and Doss (1995) extended his results. Many authors have considered semiparametric estimation of the hazard rate, beginning with Cox (1972), see Andersen et al. (1993) for a review of the enormous literature on semiparametric models. We refer to Huang (1999) and Linton et al. (2003) for some recent developments.

As far as we know, adaptive nonparametric estimation for censored data in presence of covariates has only been considered in Brunel et al. (2007), who constructed an optimal adaptive estimator of the conditional density.

Example 2 (Cox processes). Let η^i, for $i = 1, \ldots, n$, be a Cox process (see Kaar (1986)) on $\mathbb{R}_+$ with random mean-measure Λ^i given by :

$$\Lambda^i(t) = \int_0^t \alpha(X_i, z) dz,$$

where X_i is a vector of covariates in $\mathbb{R}^d$. In this context the predictable process Y of Equation (1) constantly equals 1. As a consequence, these processes can be seen as generalizations of nonhomogeneous Poisson processes on $\mathbb{R}_+$ with random intensities. This is a particular case of longitudinal data, see e.g. Example VII.2.15 in Andersen et al. (1993). The nonparametric estimation of the intensity of Poisson processes without covariates has been considered in several papers. We refer to Reynaud-Bouret (2003) and Baraud and Birgé (2006) for the adaptive estimation of the intensity of nonhomogeneous Poisson processes in general spaces.

Example 3 (Regression model for transition intensities of Markov processes). Consider a n-sample of nonhomogeneous time-continuous Markov processes $P^1, \ldots, P^n$ with finite state space $\{1, \ldots, k\}$ and denote by α_{jl} the transition intensity from state j to state l. For

individual i with covariate X_i, let $N^i_{jl}(t)$ be the number of observed direct transitions from j to l before time t (we allow the possibility of right-censoring for example). Conditionally on the initial state, the counting process N^i_{jl} verifies the following Aalen multiplicative intensity model:

$$N^i_{jl}(t) = \int_0^t \alpha_{jl}(X_i, z) Y^i_j(z) dz + M^i(t) \text{ for all } t \geq 0,$$

where $Y^i_j(t) = \mathbb{1}\{P^i(t-) = j\}$ for all $t \geq 0$, see Andersen et al. (1993) or Jacobsen (1982). This setting is discussed in Andersen et al. (1993), see Example VII.11 on mortality and nephropathy for insulin dependent diabetics.

We finally cite three papers, where different strategies for the estimation of the intensity of counting processes is considered, gathering as a consequence all the previous examples, but in none of them the presence of covariates was considered. Ramlau-Hansen (1983) proposed a kernel-type estimator, Grégoire (1993) studied cross-validation for these estimators. More recently, Reynaud-Bouret (2006) considered adaptive estimation by model selection.

Our aim in this work is to provide an optimal adaptive nonparametric estimator of the conditional intensity. Our estimation procedure involves the minimization of a so-called contrast. To achieve that purpose, we proceed as follows. In Section 2, we describe the estimation procedure: we explain how the contrast is built, on which collections of spaces the estimators are defined and how the relevant space is selected via a data driven penalized criterion. In Section 3, we state an oracle inequality for our estimator (see Theorem 1), a resulting upper bound (see Corollary 1) and a lower bound (see Theorem 2), the latter asserts the optimality in the minimax sense. An auxiliary estimation of the density of the reference measure is also studied. The examples of Section 4 are taken in the setting of Example 1, in order to provide a short illustration of the practical properties of our estimator. Lastly, proofs are gathered in Sections 5-6-7. We mention that the deviation inequalities proved in Section 6 may be of intrinsic interest.

Remark 1. An inherent remark about this model is that there is no reason for the conditional intensity $\alpha(x, z)$ to have the same behavior with respect to the z (time) and x (covariates) variables. This is the reason why it is mandatory in our purely nonparametric setting to consider anisotropic regularity for α. Think for instance of the very popular case of proportional hazards Cox model, see Cox (1972), it is assumed that $\alpha(x, z) = \alpha_0(z) \exp(\beta^\top x)$ for some unknown function α_0 and unknown vector $\beta \in \mathbb{R}^d$. Of course, in this model, the smoothness in the x direction is higher than in the z direction.

For the sake of simplicity, we will assume in the following that the covariate X is one-dimensional. Similar procedures and results for multivariate covariates are an almost effortless extension, as discussed in Remark 3.

2. Description of the procedure

Our estimation procedure involves the minimization of a contrast. This contrast is tuned to the problem considered in this paper, as explained in the next section.

2.1. **Definition of the contrast.** Let $A = A_1 \times A_2$ be a compact set on $\mathbb{R} \times \mathbb{R}_+$ on which the function α will be estimated. Without loss of generality, we set $A = [0,1] \times [0,1]$, and in particular $\tau = 1$. Let h be a function in $(L^2 \cap L^\infty)(A)$. Define the contrast function:

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^n \int_0^1 h^2(X_i,z)Y^i(z)dz - \frac{2}{n}\sum_{i=1}^n \int_0^1 h(X_i,z)dN^i(z). \tag{2}$$

This contrast is of least-squares type adapted to the problem considered here. Since each N^i admits a Doob-Meyer decomposition ($N^i = \Lambda^i + M^i$), we have:

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^n \int_0^1 h^2(X_i,z)Y^i(z)dz - \frac{2}{n}\sum_{i=1}^n \int_0^1 h(X_i,z)d\Lambda^i(z) - \frac{2}{n}\sum_{i=1}^n \int_0^1 h(X_i,z)dM^i(z),$$

so that:

$$\mathbb{E}\big(\gamma_n(h)\big) = \mathbb{E}\big(\int_0^1 h^2(X,z)Y(z)dz\big) - \mathbb{E}\big(2\int_0^1 h(X,z)d\Lambda(z)\big).$$

Let F_X denote the c.d.f. of the covariate X and $\|\cdot\|_\mu$ the norm defined by:

$$\|h\|_\mu^2 \;:=\; \mathbb{E}\big(\int_0^1 h^2(X,z)Y(z)dz\big) = \iint_A h^2(x,z)d\mu(x,z),$$

where $d\mu(x,z) := \mathbb{E}(Y(z)|X=x)F_X(dx)dz$. By the Aalen multiplicative intensity model, see Equation (1), we get:

$$\mathbb{E}\big(\gamma_n(h)\big) = \|h\|_\mu^2 - 2\iint h(x,z)\alpha(x,z)\mathbb{E}(Y(z)|X=x)F_X(dx)dz = \|h-\alpha\|_\mu^2 - \|\alpha\|_\mu^2.$$

This explains why minimizing $\gamma_n(\cdot)$ over an appropriate set of functions described below, is a relevant strategy to estimate α.

Example 1 continued. In the particular case of regression for right-censored data, the conditional hazard function is estimated and the contrast function has the following form:

$$\gamma_n(h) \;=\; \frac{1}{n}\sum_{i=1}^n \int_0^1 h^2(X_i,z)\mathbb{1}(Z_i \geq z)dz - \frac{2}{n}\sum_{i=1}^n \delta_i h(X_i,Z_i).$$

We have in addition an explicit formula for $d\mu(x,z)$:

$$d\mu(x,z) = (1 - L_{Z|X}(z,x))F_X(dx)dz,$$

where

$$1 - L_{Z|X}(z,x) := \mathbb{P}(Z \geq z|X=x) = (1 - F_{T|X}(x,z))(1 - G_{C|X}(x,z))$$

and $G_{C|X}$ is the conditional c.d.f. of C given X.

Remark 2. In our setting, it is possible to let the censoring depend on the covariates, as in Dabrowska (1989) or, more recently Heuchenne and Van Keilegom (2006). Assumption ($\mathcal{C}$) above is weaker than the assumption: T and C are independent and $\mathbb{P}(T \leq C|X,Y) = \mathbb{P}(T \leq C|Y)$ in Stute (1996).

2.2. **Assumptions and notations.** Before defining the estimation procedure, we need to introduce some assumptions and notations. Define the norms

$$\|h\|^2 := \iint h^2(x,z)dxdz, \|h\|_A^2 := \iint_A h^2(x,z)dxdz \text{ and } \|h\|_{\infty,A} := \sup_{(x,z)\in A} |h(x,z)|,$$

and assume that the following holds:

- $(\mathcal{A}1)$ The covariates X_i admit a p.d.f. f_X such that $\sup_{A_2} |f_X| < +\infty$.

Assumption $(\mathcal{A}1)$ implies that μ admits a density w.r.t. the Lebesgue measure. We denote by f this density:

$$d\mu(x,z) = f(x,z)dxdz \text{ where } f(x,z) = \mathbb{E}(Y(z)|X=x)f_X(x). \tag{3}$$

We also assume:

- $(\mathcal{A}2)$ There exists $f_0 > 0$, such that $\forall (x,z) \in A_1 \times A_2,\ f(x,z) \geq f_0$.
- $(\mathcal{A}3)$ $\forall (x,z) \in A_1 \times A_2,\ \alpha(x,z) \leq \|\alpha\|_{\infty,A} < +\infty$.
- $(\mathcal{A}4)$ $\forall i, \forall t,\ Y^i(t) \leq C_Y$ where C_Y is a known fixed constant.

Note that in the examples described in Section 1, Assumption $(\mathcal{A}4)$ is clearly fulfilled with $C_Y = 1$. We will set $C_Y = 1$ in the following for simplicity.

2.3. **Definition of the estimator.** We use the usual model selection paradigm (see, for instance, Massart (2007)): first minimize the contrast $\gamma_n(\cdot)$ over a finite-dimensional function space S_m, then select the appropriate space by penalization. We introduce a collection $\{S_m, m \in \mathcal{M}_n\}$ of projection spaces: S_m is called a model and $\mathcal{M}_n$ is a set of multi-indexes (see the examples in Section 2.4). For each $m = (m_1, m_2)$, the space S_m of functions with support in $A = A_1 \times A_2$ is defined by:

$$S_m = F_{m_1} \otimes H_{m_2} = \Big\{h, \quad h(x,z) = \sum_{j\in J_m} \sum_{k\in K_m} a_{j,k}^m \varphi_j^m(x)\psi_k^m(z),\ a_{j,k}^m \in \mathbb{R}\Big\},$$

where F_{m_1} and H_{m_2} are subspaces of $(L^2 \cap L^\infty)(\mathbb{R})$ respectively spanned by two orthonormal bases $(\varphi_j^m)_{j\in J_m}$ with $|J_m| = D_{m_1}$ and $(\psi_k^m)_{k\in K_m}$ with $|K_m| = D_{m_2}$. For all j and all k, the supports of φ_j^m and ψ_k^m are respectively included in A_1 and A_2. Here j and k are not necessarily integers, they can be couples of integers, as in the case of a piecewise polynomial space, see Section 2.4.

Remark 3. From a theoretical point of view, we could consider that the covariates X are in $\mathbb{R}^d$ and even that their density has an anisotropic regularity. For this end, we would have to consider models of the form $S_m = F_{m_1} \otimes H_{m_2} \otimes \cdots \otimes H_{m_{d+1}}$. However, this would make the proofs more intricate. Notice also the convergence rate would be slower because of the curse of dimensionality. For the sake of clarity, we deliberately restrict ourselves to $X \in \mathbb{R}$.

The first step would be to define $\hat{\alpha}_m = \text{argmin}_{h\in S_m} \gamma_n(h)$. To that end, let $h(x,y) = \sum_{j\in J_m} \sum_{k\in K_m} a_{j,k}\varphi_j^m(x)\psi_k^m(y)$ be a function in S_m. To compute $\hat{\alpha}_m$, we have to solve:

$$\forall j_0 \forall k_0, \quad \frac{\partial \gamma_n(h)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow G_m A_m = \Upsilon_m,$$

where A_m denotes the matrix $(a_{j,k})_{j\in J_m,k\in K_m}$,

$$G_m := \Big(\frac{1}{n}\sum_{i=1}^n \varphi_j^m(X_i)\varphi_l^m(X_i)\int \psi_k^m(z)\psi_p^m(z)Y^i(z)dz\Big)_{(j,k),(l,p)\in J_m\times K_m}$$

and

$$\Upsilon_m := \Big(\frac{1}{n}\sum_{i=1}^n \varphi_j^m(X_i)\int \psi_k^m(z)dN^i(z)\Big)_{j\in J_m,k\in K_m}.$$

Unfortunately G_m may not be invertible. To overcome this problem, we modify the definition of $\hat\alpha_m$ in the following way:

$$\hat\alpha_m := \begin{cases} \operatorname{argmin}_{h\in S_m}\gamma_n(h) & \text{on } \hat\Gamma_m \\ 0 & \text{on } \hat\Gamma_m^{\complement} \end{cases}, \tag{4}$$

where

$$\hat\Gamma_m := \Big\{\min \mathrm{Sp}(G_m) \geq \max(\hat f_0/3, n^{-1/2})\Big\}$$

where $\mathrm{Sp}(G_m)$ denotes the spectrum of G_m i.e. the set of the eigenvalues of the matrix G_m (it is easy to see that they are nonnegative). The estimator $\hat f_0$ of f_0 (the minimum of the density f, see $(\mathcal{A}2)$) is required to fulfill the following assumption:

- $(\mathcal{A}5)$ For any integer $k\geq 1$, $\mathbb{P}(|\hat f_0 - f_0| > f_0/2) \leq C_k/n^k$.

An estimator satisfying $(\mathcal{A}5)$ is defined in Section 3.4. In fact, $k=7$ is enough for the proofs. We refer the reader to the proof of Lemma 1, see Section 7, for an explanation of the presence of $n^{1/2}$ in the definition of $\hat\Gamma_m$. In practice, this constraint is generally not used (the matrix is invertible, otherwise another model is considered).

The final step is to select the relevant space via the penalized criterion:

$$\hat m = \operatorname*{argmin}_{m\in\mathcal{M}_n}\Big(\gamma_n(\hat\alpha_m) + \mathrm{pen}(m)\Big), \tag{5}$$

where $\mathrm{pen}(m)$ is defined in Theorem 1 below, see Section 3. Our estimator of α on A is then $\hat\alpha_{\hat m}$.

2.4. **Assumptions on the models and examples.** Let us introduce the following set of assumptions on the models $\{S_m : m\in\mathcal{M}_n\}$, which are usual in model selection techniques.

- $(\mathcal{M}1)$ For $i=1,2$, $\mathcal{D}_n^{(i)} := \max_{m\in\mathcal{M}_n} D_{m_i} \leq n^{1/4}/\sqrt{\log n}$.
- $(\mathcal{M}2)$ There exist positive reals ϕ_1,ϕ_2 such that, for all u in F_{m_1} and for all v in H_{m_2}, we have

$$\sup_{x\in A_1}|u(x)|^2 \leq \phi_1 D_{m_1}\int_{A_1} u^2 \text{ and } \sup_{x\in A_2}|v(x)|^2 \leq \phi_2 D_{m_2}\int_{A_2} v^2.$$

  By letting $\phi_0 = \sqrt{\phi_1\phi_2}$, that leads to

$$\forall h\in S_m \qquad \|h\|_{\infty,A} \leq \phi_0\sqrt{D_{m_1}D_{m_2}}\|h\|_A. \tag{6}$$

- $(\mathcal{M}3)$ Nesting condition:

$$D_{m_1}\leq D_{m_1'} \Rightarrow F_{m_1}\subset F_{m_1'} \text{ and } D_{m_2}\leq D_{m_2'} \Rightarrow H_{m_2}\subset H_{m_2'}.$$

  Moreover, there exists a global nesting space $\mathcal{S}_n$ in the collection, such that $\forall m\in\mathcal{M}_n, S_m\subset\mathcal{S}_n$ and $\dim(\mathcal{S}_n) := N_n \leq \sqrt{n/\log n}$.

Assumptions ($\mathcal{M}1$)–($\mathcal{M}3$) are not too restrictive. Indeed, they are verified for the spaces F_{m_1} (and H_{m_2}) on $A_1 = [0,1]$ spanned by the following bases (see Barron et al. (1999)):

- $[T]$ Trigonometric basis: $\text{span}(\varphi_0, \dots, \varphi_{m_1-1})$ with $\varphi_0 = \mathbb{1}([0,1])$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\ \mathbb{1}([0,1])(x)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{1}([0,1])(x)$ for $j \geq 1$. For this model $D_{m_1} = m_1$ and $\phi_1 = 2$ hold.
- $[DP]$ Regular piecewise polynomial basis: polynomials of degree $0, \dots, r$ (where r is fixed) on each interval $[(l-1)/2^D, l/2^D[$ with $l = 1, \dots, 2^D$. In this case, we have $m_1 = (D, r)$, $J_m = \{j = (l,d),\ 1 \leq l \leq 2^D, 0 \leq d \leq r\}$, $D_{m_1} = (r+1)2^D$ and $\phi_1 = \sqrt{r+1}$.
- $[W]$ Regular wavelet basis: $\text{span}(\Psi_{lk}, l = -1, \dots, m_1, k \in \Lambda(l))$ where $\Psi_{-1,k}$ is the translates of the father wavelet Ψ_{-1} and $\Psi_{lk}(x) = 2^{l/2}\Psi(2^l x - k)$ where Ψ is the mother wavelet. We assume that the supports of the wavelets are included in A_1 and that Ψ_{-1} belongs to the Sobolev space W_2^r, see Härdle et al. (1998).
- $[H]$ Histogram basis: for $A_1 = [0,1]$, $\text{span}(\varphi_1, \dots, \varphi_{2^{m_1}})$ with $\varphi_j = 2^{m_1/2}\mathbb{1}([(j-1)/2^{m_1}, j/2^{m_1}[)$ for $j = 1, \dots, 2^{m_1}$. Here $D_{m_1} = 2^{m_1}$, $\phi_1 = 1$. Notice that $[H]$ is a particular case of both $[DP]$ and $[W]$.

Remark 4. The first assumption prevents the dimension to be too large compared to the number of observations. We can lighten considerably this constraint for localized basis: for histogram basis, piecewise polynomial basis and wavelets, ($\mathcal{M}1$) reduces to $\mathcal{D}_n^{(i)} \leq \sqrt{n/\log n}$. Analogously in ($\mathcal{M}3$), we would get $N_n \leq n/\log n$. The condition ($\mathcal{M}2$) implies a useful link between the L^2 norm and the infinite norm. The third assumption ($\mathcal{M}3$) implies in particular that $\forall m, m' \in \mathcal{M}_n$, $S_m + S_{m'} \subset \mathcal{S}_n$. This condition is useful for the chaining argument used in the proofs, see Section 6.

3. Main results

3.1. **Oracle inequality.** For a function h and a space S, let

$$d(h, S) = \inf_{g \in S} \|h - g\| = \inf_{g \in S} \Big(\iint |h(x,y) - g(x,y)|^2 dxdy \Big)^{1/2}.$$

The estimator $\hat{\alpha}_{\hat{m}}$ where $\hat{\alpha}_m$ is given respectively by (4) and $\hat{m}$ is given by (5) satisfies the following oracle inequality.

Theorem 1. *Let* ($\mathcal{A}1$) – ($\mathcal{A}5$) *and* ($\mathcal{M}1$) – ($\mathcal{M}3$) *hold. Define the following penalty*:

$$\text{pen}(m) := K_0(1 + \|\alpha\|_{\infty,A})\frac{D_{m_1}D_{m_2}}{n}, \tag{7}$$

where K_0 is a numerical constant. We have

$$\mathbb{E}(\|\alpha\mathbb{1}(A) - \hat{\alpha}_{\hat{m}}\|^2) \leq C \inf_{m \in \mathcal{M}_n} \{d^2(\alpha\mathbb{1}(A), S_m) + \text{pen}(m)\} + \frac{C'}{n} \tag{8}$$

where $C = C(f_0, \|f\|_{A,\infty})$ and C' is a constant depending on $\phi_1, \phi_2, \|\alpha\|_{\infty,A}, f_0$.

The proof of Theorem 1 involves a deviation inequality for the empirical process

$$\nu_n(h) := \frac{1}{n}\sum_{i=1}^n \int_0^1 h(X_i, z)dM^i(z),$$

where $M^i(t) = N^i(t) - \int_0^t \alpha(X_i, z)Y^i(z)dz$ are martingales, see Section 1, and a $L^2 - L^\infty$ chaining argument.

Remark 5. The penalty involves the unknown quantity $\|\alpha\|_{\infty,A}$. This is a usual situation, and the solution is to replace it by an estimator $\|\hat{\alpha}_{m_n}\|_{\infty,A}$ where $\hat{\alpha}_{m_n}$ is an estimator of the collection, chosen on a space S_{m_n} which is arbitrary, generally middle sized. Note that, by doing this, the penalty function becomes random. For details, we refer to Lacour (2007), Theorem 2.2.

3.2. **Upper bound for the rate.** From Theorem 1, we can derive the rate of convergence of $\hat{\alpha}_{\hat{m}}$ over anisotropic Besov spaces. We recall that anisotropy is almost mandatory in this context, see Remark 1. For that purpose, assume that α restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ on A with regularity $\boldsymbol{\beta} = (\beta_1, \beta_2)$. Let us recall the definition of $B^{\boldsymbol{\beta}}_{2,\infty}(A)$. Let $\{e_1, e_2\}$ the canonical basis of $\mathbb{R}^2$ and take $A^r_{h,i} := \{x \in \mathbb{R}^2; x, x + he_i, \dots, x + rhe_i \in A\}$, for $i = 1, 2$. For $x \in A^r_{h,i}$, let

$$\Delta^r_{h,i} g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i)$$

be the rth difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g, t) = \sup_{|h| \le t} \Big(\int_{A^{r_i}_{h,i}} |\Delta^{r_i}_{h,i} g(x)|^2 dx \Big)^{1/2}.$$

We say that g is in the Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ if $\sup_{t>0} \sum_{i=1}^2 t^{-\beta_i} \omega_{r_i,i}(g, t) < \infty$ for r_i integers larger than β_i. More details concerning Besov spaces can be found in Triebel (2006). The next corollary shows that $\hat{\alpha}_{\hat{m}}$ adapts to the unknown anisotropic smoothness of α.

Corollary 1. *Assume that α restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ with regularity $\boldsymbol{\beta} = (\beta_1, \beta_2)$ such that $\beta_1 > 1/2$ and $\beta_2 > 1/2$. We consider the piecewise polynomial or wavelet spaces described in Subsection 2.4 (with the regularity r of the polynomials and the wavelets larger than $\beta_i - 1$). Then, under the assumptions of Theorem 1, we have*

$$\mathbb{E}\|\alpha - \hat{\alpha}_{\hat{m}}\|^2_A = O(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+2}}).$$

where $\bar{\beta}$ is the harmonic mean of β_1 and β_2 (i.e. $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$).

The rate of convergence achieved by $\hat{\alpha}_{\hat{m}}$ in Corollary 1 is optimal in the minimax sense as proved in Theorem 2 below. For trigonometric spaces, the result also holds, but for $\beta_1 > 3/2$ and $\beta_2 > 3/2$ (because of $(\mathcal{M}1)$).

Moreover, assuming for example that $\beta_2 > \beta_1$, one can see in the proof of Corollary 1 that the estimator chooses a space of dimension $D_{\hat{m}_2} = D_{\hat{m}_1}^{\beta_1/\beta_2} < D_{\hat{m}_1}$. This shows that the estimator is adaptive with respect to the approximation space for each directional regularity.

3.3. **Lower bound.** In the next Theorem, we prove that the rate $n^{-2\bar{\beta}/(2\bar{\beta}+2)}$ is optimal over $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ where we recall that $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$. Since the lower bound stated in Theorem 2 is uniform over $B^{\boldsymbol{\beta}}_{2,\infty}(A)$, we need to introduce the ball

$$B^{\boldsymbol{\beta}}_{2,\infty}(A, L) = \{\alpha \in B^{\boldsymbol{\beta}}_{2,\infty}(A) : \|\alpha\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \leq L\},$$

where

$$\|\alpha\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} := \|\alpha\|_A + |\alpha|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} = \|\alpha\|_A + \sup_{t>0} \sum_{i=1}^{2} t^{-\beta_i} \omega_{r_i,i}(g,t). \tag{9}$$

Let us denote by E_α the integration w.r.t. the joint law P^n_α, when the intensity is α, of the n-sample $(X_i, N^i(z), Y^i(z); z \leq 1, i = 1, \ldots, n)$.

Theorem 2. *There is a positive constant C_L such that*

$$\inf_{\tilde{\alpha}} \sup_{\alpha \in B^{\boldsymbol{\beta}}_{2,\infty}(A,L)} \mathbb{E}_\alpha \|\tilde{\alpha} - \alpha\|^2_A \geq C_L n^{-2\bar{\beta}/(2\bar{\beta}+2)}$$

for n large enough, where the infimum is taken among all estimators and where C_L is a constant that depends on $\boldsymbol{\beta}, L$ and A only.

3.4. **Estimation of f and f_0.** We recall that f is the density of μ, which is defined in Equation (3). We define

$$\hat{f}_m = \operatorname*{argmin}_{h \in S_m} \upsilon_n(h) \text{ where } \upsilon_n(h) = \|h\|^2 - \frac{2}{n} \sum_{i=1}^{n} \int_0^1 h(X_i, z) Y^i(z) dz. \tag{10}$$

This estimator admits a simple explicit formulation:

$$\hat{f}_m = \sum_{(j,k) \in J_m \times K_m} \hat{b}_{j,k} \varphi^m_j(x) \psi^m_k(y), \text{ with } \hat{b}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \varphi^m_j(X_i) \int \psi^m_k(z) Y^i(z) dz. \tag{11}$$

As before, we consider estimation of f over the compact set $A = [0,1] \times [0,1]$. We choose the space H_{m_2} as the space with maximal dimension, as explained below. Let us denote it by $\mathcal{H}_n$, by $\mathcal{D}^{(2)}_n = \dim(\mathcal{H}_n)$ its dimension (see $(\mathcal{M}1)$) and by ℓ_n its index so that $H_{\ell_n} = \mathcal{H}_n$. Hence, we consider, instead of a general $\hat{f}_m$, the estimator

$$\hat{f}_{m_1} := \operatorname*{argmin}_{h \in F_{m_1} \times \mathcal{H}_n} \upsilon_n(h).$$

We are now in a position to define an estimator of f_0 by considering any $\inf_{(x,z)\in A} \hat{f}_{m_1}(x,z)$ with a given m_1. Indeed, an arbitrary choice is sufficient for our estimation problem concerning f_0. In our setting, only a rough estimation of the lower bound on f is useful. Therefore, for the purpose of estimating α, we can define

$$\hat{f}_0 := \inf_{(x,z)\in A} \hat{f}_{m_1^*}(x,z) \text{ with } m_1^* = (D_{m_1^*}, \mathcal{D}^{(2)}_n). \tag{12}$$

Then, the following result holds:

Proposition 1. *Consider $\hat{f}_0$ defined by (12) in the basis* [T]*, with $\log n \leq D_{m_1^*} \leq n^{1/4}/\sqrt{\log n}$ and $\mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log n}$. Assume that $f \in \mathcal{B}_{2,\infty}^{(\tilde{\beta}_1,\tilde{\beta}_2)}(A)$ with $\bar{\tilde{\beta}} > 1$, then $\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C'_k/n^k$, for any integer k, where C_k is a constant and therefore $\hat{f}_0$ fulfills assumption ($\mathcal{A}5$).*

The proof of this result is given in Section 7.

Hereafter, we develop a remark concerning the estimation of f in order to explain why we have selected the second dimension D_{m_2} the largest as possible. Let f_{m_1} be the orthogonal projection of the restriction of f to A on the space $F_{m_1} \times \mathcal{H}_n$, i.e. for $m_n = (m_1, \ell_n)$, $f_{m_1} = \sum_{(j,k)\in J_{m_1}\times\mathcal{K}_n} b_{j,k}\varphi_j^{m_n}\psi_k^{m_n}$, with $|J_{m_1}| = D_{m_1}$ and $|\mathcal{K}_n| = \mathcal{D}_n^{(2)}$. We obtain the following bias-variance decomposition.

Proposition 2. *Under ($\mathcal{M}1$), ($\mathcal{M}2$), ($\mathcal{A}1$) and ($\mathcal{A}4$), we have*

$$\mathbb{E}(\|\hat{f}_{m_1} - f\|_A^2) \leq \|f_{m_1} - f\|_A^2 + \frac{\ell(A_2)\phi_1 D_{m_1}}{n}, \tag{13}$$

where $\ell(A_2)$ is the Lebesgue measure of A_2.

Proof. We clearly have

$$\|\hat{f}_{m_1} - f\|_A^2 = \|f_{m_1} - f\|_A^2 + \|\hat{f}_{m_1} - f_{m_1}\|_A^2, \tag{14}$$

where the first term is the bias term and $\|\hat{f}_{m_1} - f_{m_1}\|_A^2 = \sum_{(j,k)\in J_{m_1}\times\mathcal{K}_n}(\hat{b}_{j,k} - b_{j,k})^2$ is the variance term. In view of (11), we have $\mathbb{E}(\hat{b}_{j,k}) = b_{j,k}$, and, as a consequence:

$$\begin{aligned}
\mathbb{E}(\|\hat{f}_{m_1} - f_{m_1}\|_A^2) &= \sum_{(j,k)\in J_{m_1}\times\mathcal{K}_n} \mathrm{Var}(\hat{b}_{j,k}) \\
&= \sum_{(j,k)\in J_{m_1}\times\mathcal{K}_n} \frac{1}{n}\mathrm{Var}\Big(\varphi_j^{m_n}(X_1)\int_{A_2}\psi_k^{m_n}(z)Y^1(z)dz\Big) \\
&\leq \sum_{(j,k)\in J_{m_1}\times\mathcal{K}_n} \frac{1}{n}\mathbb{E}\Big([\varphi_j^{m_n}(X_1)]^2\Big[\int_{A_2}\psi_k^{m_n}(z)Y^1(z)dz\Big]^2\Big)
\end{aligned}$$

Now, we note that for any A_2-square integrable function ξ,

$$\sum_{k\in\mathcal{K}_n}\Big[\int_{A_2}\psi_k^{m_n}(z)\xi(z)dz\Big]^2 \leq \int_{A_2}\xi^2(z)dz$$

by a simple projection argument (the left-hand-side term is the squared norm of the projection of ξ on $\mathcal{H}_n$), and thus under assumption ($\mathcal{A}4$),

$$\mathbb{E}(\|\hat{f}_{m_1} - f_{m_1}\|_A^2) \leq \frac{\ell(A_2)}{n}\sum_{j\in J_{m_1}}\mathbb{E}\Big([\varphi_j^{m_1}(X_1)]^2\Big) \leq \frac{\ell(A_2)\phi_1 D_{m_1}}{n}.$$

Gathering the terms, the risk of the estimator is bounded as in (13). □

Let us discuss the asymptotic rate of estimation of f_A, the restriction of f to A, using the above procedure. For that purpose, assume that f_A belongs to $B_{2,\infty}^{\tilde{\boldsymbol{\beta}}}(A)$ with regularity

$\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2)$. Now, consider the collection of trigonometric polynomials for φ_j, ψ_k, and apply lemma of Lacour (2007) (see Section 5 below). The bias term is bounded by

$$\|f_{m_1} - f\|_A^2 \leq C\{D_{m_1}^{-2\tilde{\beta}_1} + [\mathcal{D}_n^{(2)}]^{-2\tilde{\beta}_2}\}.$$

It is worth noticing that the variance term (i.e. the last term of (13)) does not depend on ℓ_n nor on $\mathcal{D}_n^{(2)}$. This explains why the size of the projection space in the z-direction must be chosen the largest as possible, when the mean square risk is under study. Take $\mathcal{D}_n^{(2)} = \sqrt{n/\log n}$ and assume that $\tilde{\beta}_2 > 1$, then (13) becomes

$$\mathbb{E}(\|\hat{f}_{m_1} - f\|_A^2) \leq C[D_{m_1}^{-2\tilde{\beta}_1} + \frac{\ell(A_2)D_{m_1}}{n}] + \frac{C' \log n}{n}.$$

Therefore, choosing $D_{m_1^*} = n^{1/(2\tilde{\beta}_1+1)}$ gives the rate

$$\mathbb{E}(\|\hat{f}_{m_1} - f_A\|^2) \leq C'' n^{-2\tilde{\beta}_1/(2\tilde{\beta}_1+1)}$$

which is the standard asymptotic rate for a single variable function with regularity $\tilde{\beta}_1$. We could study a model selection procedure and find a penalty function of order D_{m_1}/n, so that a relevant space is chosen in an automatic way. We do not go into further details since a rough estimation of f_0 is sufficient to estimate the conditional intensity α.

4. Illustration

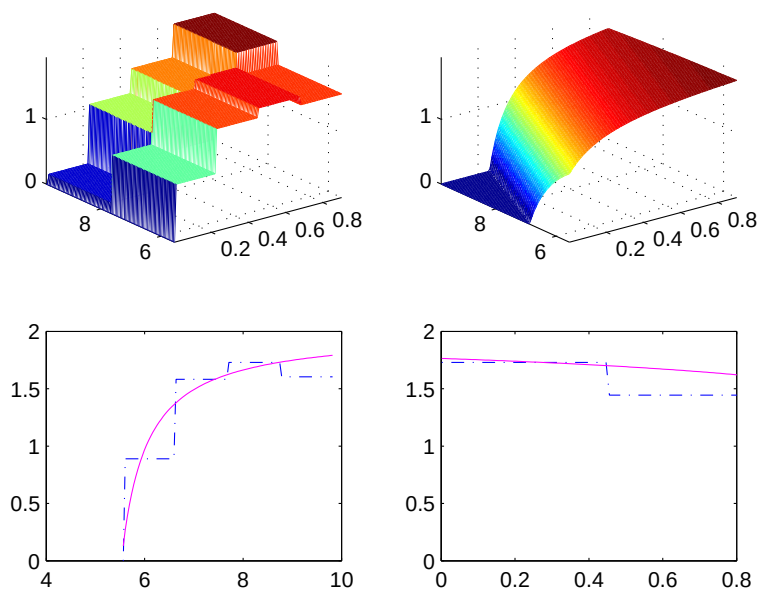

Figure 1. Case (NL) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

In this section, we give a numerical illustration of the adaptive estimator $\hat{\alpha}_{\hat{m}}$, defined in Section 2, computed with the dyadic histogram basis $[H]$. We sample i.i.d. data $(X_1, T_1), \ldots, (X_n, T_n)$ in three particular cases of the regression model of Example 1 from Section 1. For the sake of simplicity, we simulate the covariates X_i with the uniform distribution on $[0, 1]$. The size of the data set is $n = 1000$.

- Case (NL). Non-Linear regression:
$$T_i = b(X_i) + \sigma\varepsilon_i.$$
We simulate ε_i with a $\chi^2(4)$ distribution and $b(x) = 2x+5$. Note that in this case, the hazard function to be estimated is
$$\alpha_{\mathrm{NL}}(x,t) = \frac{1}{\sigma}\alpha_\varepsilon\Big(\frac{t-b(x)}{\sigma}\Big),$$
where α_ε denotes the hazard function of ε.
- Case (AFT). Accelerated Failure Time model:
$$\log(T_i) = a + bX_i + \varepsilon_i,$$
where the ε_i are standard normal and $a = 5$ and $b = 2$. The hazard function to be estimated is then:
$$\alpha_{AFT}(x,t) = \frac{\alpha_\varepsilon(\log(t) - (a+bx))}{t}.$$
- Case (PH). Proportional Hazards model: in this case, the hazard writes
$$\alpha(x,t) = \exp(bx)\alpha_0(t).$$
We take $b = 0.4$ and $\alpha_0(t) = a\lambda t^{a-1}$, which is a Weibull hazard function with $a = 3$ and $\lambda = 1$.

The penalty is taken as
$$\widehat{\mathrm{pen}}(m_1, m_2) = 5\widehat{\|\alpha\|_{\infty,A}}\frac{2^{m_1+m_2}}{n},$$
where $\widehat{\|\alpha\|_{\infty,A}}$ is estimated as the maximal of the estimated histogram coefficients ($\max_{j,k}\hat{a}_{j,k}$) on the largest space which is considered (taken with dimension $\sqrt{n}$).

We can see from Figures 1-3 that the algorithm exploits the opportunity (Figures 1 and 3) of choosing different dimensions in the two directions, and that it captures well the general form of the surfaces.

5. Proofs of the main results

5.1. **Proof of Theorem 1.** We define, for h_1, h_2 in $L^2 \cap L^\infty(A)$, the empirical scalar product

$$\langle h_1, h_2\rangle_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 h_1(X_i,z)h_2(X_i,z)Y^i(z)dz\mathbb{1}(X_i \in [0,1]) \tag{15}$$

and the associated empirical norm $\|h_1\|_n^2 = \langle h_1, h_1\rangle_n$ which is such that

$$\mathbb{E}(\|h_1\|_n^2) = \iint_A h_1^2(x,y)d\mu(x,y) = \iint_A h_1^2(x,y)f(x,y)dxdy = \|h_1\|_\mu^2$$

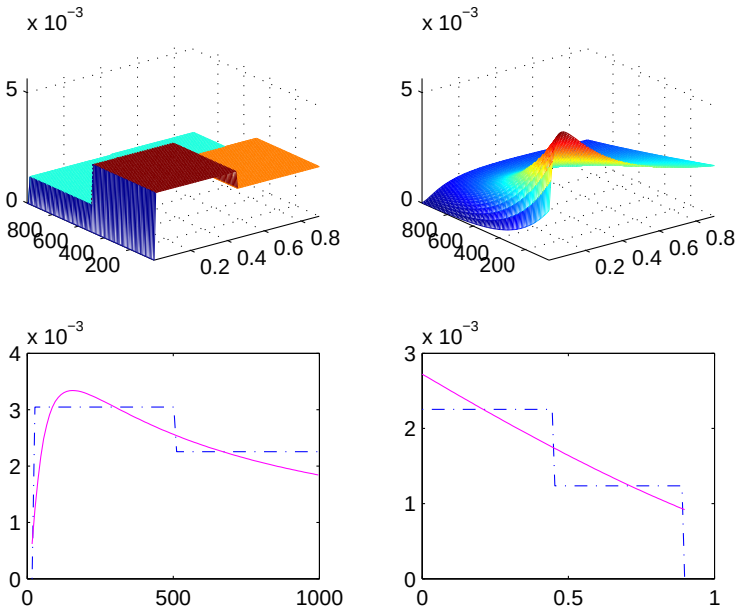

FIGURE 2. Case (AFT) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

where we recall that f denotes the density of μ w.r.t. the Lebesgue measure on A. We shall use the following sets:

$$\hat{\Gamma}_m = \{\min \mathrm{Sp}(G_m) \geq \max(\hat{f}_0/3, n^{-1/2})\}, \quad \hat{\Gamma} := \bigcap_{m \in \mathcal{M}_n} \hat{\Gamma}_m,$$

$$\Delta := \Big\{\forall h \in \mathcal{S}_n : \Big|\frac{\|h\|_n^2}{\|h\|_\mu^2} - 1\Big| \leq \frac{1}{2}\Big\}, \text{ and } \Omega := \Big\{\Big|\frac{\hat{f}_0}{f_0} - 1\Big| \leq \frac{1}{2}\Big\}. \tag{16}$$

For $m \in \mathcal{M}_n$, we denote by α_m the orthogonal projection on S_m of α restricted to A. The following bounds hold:

$$\begin{aligned}
\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\|_A^2) &\leq 2\|\alpha - \alpha_m\|_A^2 + 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_A^2 \mathbb{1}(\Delta \cap \Omega)) \\
&+ 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_A^2 \mathbb{1}(\Delta^{\complement} \cap \Omega)) + 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_A^2 \mathbb{1}(\Omega^{\complement})) \\
&\leq 2\|\alpha - \alpha_m\|_A^2 + 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_A^2 \mathbb{1}(\Delta \cap \Omega)) \\
&+ 4\mathbb{E}((\|\hat{\alpha}_{\hat{m}}\|^2 + \|\alpha\|_A^2)\mathbb{1}(\Delta^{\complement} \cap \Omega)) + 4\mathbb{E}((\|\hat{\alpha}_{\hat{m}}\|^2 + \|\alpha\|_A^2)\mathbb{1}(\Omega^{\complement})).
\end{aligned} \tag{17}$$

We use the following results, whose proofs can be found in Sections 6.2 and 7.

Proposition 3. *We have $\mathbb{E}(\|\hat{\alpha}_{\hat{m}}\|^4) \leq C' n^5$, where C' is a constant.*

Proposition 4. *If $(\mathcal{M}_1)$ is fulfilled, we have $\mathbb{P}(\Delta^{\complement}) \leq C_k/n^k$ for any $k \geq 1$, when n is large enough, where C_k is a constant.*

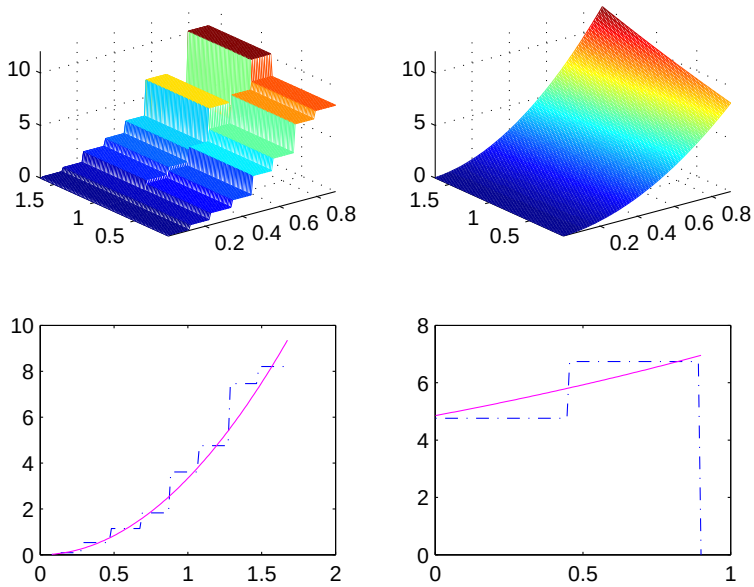

FIGURE 3. Case (PH) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

Moreover, $(\mathcal{A}5)$ ensures that $\mathbb{P}(\Omega^{\complement}) \leq C_k/n^k$ for any integer k. Thus, using Propositions 3 and 4 and Assumption $(\mathcal{A}5)$, we get

$$\begin{aligned} & \mathbb{E}((\|\hat{\alpha}_{\hat{m}}\|^2 + \|\alpha\|_A^2)\mathbb{1}(\Delta^{\complement} \cap \Omega)) + \mathbb{E}((\|\hat{\alpha}_{\hat{m}}\|^2 + \|\alpha\|_A^2)\mathbb{1}(\Omega^{\complement})) \\ \leq \; & \|\alpha\|_A^2(\mathbb{P}(\Omega^{\complement}) + \mathbb{P}(\Delta^{\complement})) + \mathbb{E}^{1/2}(\|\hat{\alpha}_{\hat{m}}\|^4)(\mathbb{P}^{1/2}(\Omega^{\complement}) + \mathbb{P}^{1/2}(\Delta^{\complement})) \\ \leq \; & C_2/n. \end{aligned} \tag{18}$$

Thus it remains to study $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_A^2 \mathbb{1}(\Delta \cap \Omega))$. We state the following Lemma:

Lemma 1. *The following embedding holds:*

$$\Delta \cap \Omega \subset \hat{\Gamma} \cap \Omega.$$

As a consequence, for all $m \in \mathcal{M}_n$, the matrices G_m are invertible on $\Delta \cap \Omega$.

Let us now define the centered empirical process

$$\begin{aligned} \nu_n(h) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\int h(X_i,z)dN^i(z) - \int h(X_i,z)\alpha(X_i,z)Y^i(z)dz\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int h(X_i,z)dM^i(z), \end{aligned} \tag{19}$$

where we use the Doob-Meyer decomposition. For any $h_1, h_2 \in (L^2 \cap L^\infty)(A)$, we have

$$\begin{aligned}\gamma_n(h_1) - \gamma_n(h_2) &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2\rangle_n - \frac{2}{n}\sum_{i=1}^n \int (h_1 - h_2)(X_i, z) dN^i(z) \\ &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 - \alpha\rangle_n - 2\nu_n(h_1 - h_2).\end{aligned}$$

Now, as on $\Delta \cap \Omega$ we have

$$\gamma_n(\hat{\alpha}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\alpha_m) + \mathrm{pen}(m).$$

It follows, from the inequality $2xy \leq x^2/\theta^2 + \theta^2 y^2$, with $x, y, \theta \in \mathbb{R}^+$, that, on $\Delta \cap \Omega$,

$$\begin{aligned}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 &\leq 2\langle \hat{\alpha}_{\hat{m}} - \alpha_m, \alpha - \alpha_m\rangle_n + \mathrm{pen}(m) + 2\nu_n(\hat{\alpha}_{\hat{m}} - \alpha_m) - \mathrm{pen}(\hat{m}) \\ &\leq \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 + 4\|\alpha - \alpha_m\|_n^2 + \mathrm{pen}(m) \\ &\quad + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4 \sup_{h \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(h) - \mathrm{pen}(\hat{m}),\end{aligned}$$

where $B^\mu_{m,m'}(0,1) := \{h \in S_m + S_{m'} : \|h\|_\mu \leq 1\}$. This yields

$$\begin{aligned}\frac{3}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 &\leq 4\|\alpha - \alpha_m\|_n^2 + \mathrm{pen}(m) + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \\ &\quad + 4\Big(\sup_{h \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(h) - p(m, \hat{m})\Big) + 4p(m, \hat{m}) - \mathrm{pen}(\hat{m}).\end{aligned}$$

Now, let us choose the penalty such that

$$\forall m, m',\ 4p(m, m') \leq \mathrm{pen}(m) + \mathrm{pen}(m'), \tag{20}$$

and use the definition of Δ. We obtain on $\Delta \cap \Omega$:

$$\begin{aligned}\frac{1}{2}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 &\leq 4\|\alpha - \alpha_m\|_n^2 + 2\mathrm{pen}(m) \\ &\quad + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4\sum_{m' \in \mathcal{M}_n}\Big(\sup_{h \in B^\mu_{m,m'}(0,1)} \nu_n^2(h) - p(m, m')\Big)\end{aligned}$$

and thus on $\Delta \cap \Omega$:

$$\begin{aligned}\frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 &\leq 4\|\alpha - \alpha_m\|_n^2 + 2\mathrm{pen}(m) \\ &\quad + 4\sum_{m' \in \mathcal{M}_n}\Big(\sup_{h \in B^\mu_{m,m'}(0,1)} \nu_n^2(h) - p(m, m')\Big).\end{aligned}$$

Using the following proposition, we can achieve the proof of Theorem 1.

Proposition 5. *Let*

$$p(m, m') = \kappa(1 + \|\alpha\|_{\infty,A})\frac{D_m + D_{m'}}{n}$$

where C_0 is a numerical constant. Under the assumptions of Theorem 1, we have

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\Big(\sup_{h \in B^\mu_{m,m'}(0,1)} (\nu_n^2(h) - p(m, m'))_+ \mathbb{1}(\Delta)\Big) \leq \frac{C_1}{n}.$$

This proposition entails:

$$\frac{1}{4}\mathbb{E}(\|\hat{\alpha}_{\hat{m}}-\alpha_m\|_\mu^2\mathbb{1}(\Delta\cap\Omega))\leq 4\|\alpha-\alpha_m\|_\mu^2+2\mathrm{pen}(m)+\frac{C_1}{n}. \tag{21}$$

Gathering (17), (18) and (21) leads to

$$\begin{aligned}\mathbb{E}(\|\hat{\alpha}_{\hat{m}}-\alpha\|_A^2) &\leq 2\|\alpha_m-\alpha\|_A^2+\frac{8}{f_0}\Big(4\|\alpha-\alpha_m\|_\mu^2+2\mathrm{pen}(m)+\frac{C_1}{n}\Big)+\frac{C_2}{n}\\ &\leq 2\Big(1+\frac{16\|f_X\|_{A,\infty}}{f_0}\Big)\|\alpha_m-\alpha\|_A^2+\frac{16}{f_0}\mathrm{pen}(m)+\frac{C_3}{n} \qquad (22)\end{aligned}$$

for any $m\in\mathcal{M}_n$. This concludes the proof of Theorem 1. □

5.2. **Proof of Corollary 1.** To control the bias term, we state the following lemma proved in Lacour (2007) and following from Hochmuth (2002) and Nikol'skii (1975):

Lemma. *Lacour (2007) Let s belong to $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ where $\boldsymbol{\beta}=(\beta_1,\beta_2)$. We consider that S'_m is one of the following spaces on A of dimension $D_{m_1}D_{m_2}$:*

- *a space of piecewise polynomials of degrees bounded by $s_i>\beta_i-1$ $(i=1,2)$ based on a partition with rectangles of sidelengthes $1/D_{m_1}$ and $1/D_{m_2}$,*
- *a linear span of $\{\phi_\lambda\psi_\mu,\lambda\in\cup_0^{m_1}\Lambda(j),\mu\in\cup_0^{m_2}M(k)\}$ where $\{\phi_\lambda\}$ and $\{\psi_\mu\}$ are orthonormal wavelet bases of respective regularities $s_1>\beta_1-1$ and $s_2>\beta_2-1$ (here $D_{m_i}=2^{m_i},i=1,2$),*
- *the space of trigonometric polynomials with degree smaller than D_{m_1} in the first direction and smaller than D_{m_2} in the second direction.*

Let s_m be the orthogonal projection of s on S'_m. Then, there exists a positive constant C_0 such that

$$\|s-s_m\|_A=\Big(\int_A|s-s_m|^2\Big)^{1/2}\leq C_0[D_{m_1}^{-\beta_1}+D_{m_2}^{-\beta_2}].$$

If we choose for S_m as one of the S'_ms, we can apply the above lemma to the function α_A, the restriction of α to A. As α_m has been defined as the orthogonal projection of α_A on S_m, we get:

$$\|\alpha-\alpha_m\|_A\leq C_0[D_{m_1}^{-\beta_1}+D_{m_2}^{-\beta_2}].$$

Now, according to Theorem 1, we obtain:

$$\mathbb{E}\|\hat{\alpha}_{\hat{m}}-\alpha\|_A^2\leq C''\inf_{m\in\mathcal{M}_n}\Big\{D_{m_1}^{-2\beta_1}+D_{m_2}^{-2\beta_2}+\frac{D_{m_1}D_{m_2}}{n}\Big\}.$$

In particular, if $m^*=(m_1^*,m_2^*)$ is such that

$$D_{m_1^*}=\lfloor n^{\frac{\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2}}\rfloor \text{ and } D_{m_2^*}=\lfloor (D_{m_1^*})^{\frac{\beta_1}{\beta_2}}\rfloor$$

then

$$\mathbb{E}\|\hat{\alpha}_{\hat{m}}-\alpha\|_A^2\leq C'''\Big\{D_{m_1^*}^{-2\beta_1}+\frac{D_{m_1^*}^{1+\beta_1/\beta_2}}{n}\Big\}=O\Big(n^{-\frac{2\beta_1\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2}}\Big)=O(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+2}}),$$

where the harmonic mean of β_1 and β_2 is $\bar{\beta}=2\beta_1\beta_2/(\beta_1+\beta_2)$. The condition $D_{m_1}\leq n^{1/2}/\log n$ allows this choice of m only if $\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)<1/2$ i.e. if $\beta_1-\beta_2+2\beta_1\beta_2>0$. In the same manner, the condition $\beta_2-\beta_1+2\beta_1\beta_2>0$ must be verified. Both conditions hold if $\beta_1>1/2$ and $\beta_2>1/2$.

5.3. **Proof of Theorem 2.** In order prove Theorem 2, we use the following theorem from Tsybakov (2003), which is a standard tool for the proof of such a lower bound. We say that ∂ is a *semi-distance* on some set Θ if it is symmetric and if it satisfies the triangle inequality and $\partial(\theta,\theta)=0$ for any $\theta\in\Theta$. We consider $K(P,Q):=\int\log(dP/dQ)dP$ the Kullback-Leibler divergence between probability measures P and Q such that $P\ll Q$.

Theorem (Tsybakov (2003))**.** *Let (Θ,∂) be a set endowed with a semi-distance ∂. We suppose that $\{P_\theta:\theta\in\Theta\}$ is a family of probability measures on a measurable space $(\mathcal{X},\mathcal{A})$ and that $v>0$. If there exist $\{\theta_0,\dots,\theta_M\}\subset\Theta$, with $M\geq 2$, such that*

(1) $\partial(\theta_j,\theta_k)\geq 2v \quad \forall\, 0\leq j<k\leq M$
(2) $P_{\theta_j}\ll P_{\theta_0} \quad \forall\, 1\leq j\leq M$,
(3) $\frac{1}{M}\sum_{j=1}^M K(P_{\theta_j},P_{\theta_0})\leq a\log(M)$ *for some* $a\in(0,1/8)$,

then

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}E_\theta[(v^{-1}\partial(\hat\theta,\theta))^2]\geq\frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2a-2\sqrt{\frac{a}{\log(M)}}\right),$$

where the infimum is taken among all estimators.

We construct a family of functions $\{\alpha_0,\dots,\alpha_M\}$ that satisfies points (1)–(3). Let $\alpha_0(x,t)=|B|^{-1}\mathbb{1}(t\in B)$ where B is a compact set such that $A=A_1\times A_2\subset B\times B$ and $|B|\geq 2|A|^{1/2}/L$. As a consequence, we have $\alpha_0(x,t)>0$ for $(x,t)\in A$ and $\|\alpha_0\|_{B^{\boldsymbol\beta}_{2,\infty}(A)}=\|\alpha_0\|_A+|\alpha_0|_{B^{\boldsymbol\beta}_{2,\infty}(A)}\leq L/2$ since $|\alpha_0|_{B^{\boldsymbol\beta}_{2,\infty}(A)}=0$, see (9). We shall denote for short $a_0=|B|^{-1}$ in the following. Let ψ be a very regular wavelet with compact support (the Daubechies's wavelet for instance), and for $j=(j_1,j_2)\in\mathbb{Z}^2$ and $k=(k_1,k_2)\in\mathbb{Z}^2$, let us consider

$$\psi_{j,k}(x,t)=2^{(j_1+j_2)/2}\psi(2^{j_1}t-k_1)\psi(2^{j_2}x-k_2).$$

Let $S_{j,k}$ stands for the support of $\psi_{j,k}$. We consider the maximal set $K_j\subset\mathbb{Z}^2$ such that

$$S_{j,k}\subset A,\forall k\in R_j \text{ and } S_{j,k}\cap S_{j,k'}=\emptyset,\forall k,k'\in R_j,k\neq k'. \tag{23}$$

The cardinality of R_j satisfies $|R_j|=c2^{j_1+j_2}$, where c is a positive constant that depends on A and on the support of ψ only. Consider the set $\Omega_j=\{0,1\}^{|R_j|}$ and define for any $\omega=(\omega_k)\in\Omega_j$

$$\alpha(\cdot;\omega):=\alpha_0+\sqrt{\frac{b}{n}}\sum_{k\in R_j}\omega_k\psi_{j,k},$$

where $b>0$ is some constant to be chosen below. In view of (23) we have

$$\|\alpha(\cdot;\omega)-\alpha(\cdot;\omega')\|_A^2=\frac{b\rho(\omega,\omega')}{n}$$

where

$$\rho(\omega,\omega'):=\sum_{k\in R_j}\mathbb{1}(\omega_k\neq\omega'_k)$$

is the Hamming distance on Ω_j. Using a result of Varshamov-Gilbert - see Tsybakov (2003) - we can find a subset $\{\omega^{(0)},\dots,\omega^{(M_j)}\}$ of Ω_j such that

$$\omega^{(0)}=(0,\dots,0),\quad \rho(\omega^{(p)},\omega^{(q)})\geq|R_j|/8$$

for any $0 \le p < q \le M_j$, where $M_j \ge 2^{|R_j|/8}$. We consider the family $\mathcal{A}_j = \{\alpha_0, \dots, \alpha_{M_j}\}$ where $\alpha_p = \alpha(\cdot, \omega^{(p)})$. This family satisfies for any $0 \le p < q \le M_j$

$$\|\alpha_p - \alpha_q\|_A \ge \Big(\frac{b|R_j|}{8n}\Big)^{1/2} = 2v_j$$

for $v_j := \sqrt{b|R_j|/(32n)}$. This proves point (1). Now, let us gather here some properties for this family of functions. We have

$$\|\alpha(\cdot;\omega) - \alpha_0\|_{\infty,A} \le \sqrt{\frac{b2^{(j_1+j_2)}}{n}}\|\psi\|_\infty^2 \le a_0/3$$

and consequently $\alpha(x,t;\omega) \ge 2a_0/3 > 0$ for any $(x,t) \in A$ and $\omega \in \Omega_j$ whenever

$$\Big(\frac{b2^{j_1+j_2}}{n}\Big)^{1/2} \le \frac{a_0}{3\|\psi\|_\infty^2}. \tag{24}$$

Using Hochmuth (2002), we have for ψ smooth enough that

$$\|\sum_{k\in R_j} \omega_k \psi_{j,k}\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \le (2^{j_1\beta_1} + 2^{j_2\beta_2})\|\sum_{k\in R_j} \omega_k \psi_{j,k}\|_A \le (2^{j_1\beta_1} + 2^{j_2\beta_2})(c2^{j_1+j_2})^{1/2}.$$

Hence, if

$$\frac{(2^{j_1\beta_1} + 2^{j_2\beta_2})(2^{j_1+j_2})^{1/2}}{\sqrt{n}} \le \frac{L}{2\sqrt{bc}}, \tag{25}$$

we have $\|\alpha(\cdot;\omega)\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \le L$, so $\alpha(\cdot;\omega) \in B^{\boldsymbol{\beta}}_{2,\infty}(A,L)$ for any $\omega \in \Omega_j$. This proves that $\mathcal{A}_j \subset B^{\boldsymbol{\beta}}_{2,\infty}(A,L)$.

Points (2) and (3) are derived using Jacod's formula (see Andersen et al. (1993)). Indeed, we can prove that the log-likelihood $\ell(\alpha,\alpha_0) := \log(dP_\alpha/dP_{\alpha_0})$ of N writes

$$\ell(\alpha,\alpha_0) = \int_0^1 (\log\alpha(X,t) - \log\alpha_0(X,t))dN(t) - \int_0^1 (\alpha(X,t) - \alpha_0(X,t))Y(t)dt.$$

For any $\alpha \in \mathcal{A}_j$, we have $\|\alpha - \alpha_0\|_{\infty,A} \le a_0/3 \le \alpha(x,t)/2$ for any $(x,t) \in A$. The Doob-Meyer decomposition allows to write that, under P_{α_0}:

$$\begin{aligned}\ell(\alpha,\alpha_0) = &\int_0^1 \Big(\Phi_{1/\alpha(X,t)}(\alpha(X,t) - \alpha_0(X,t)) - (\alpha(X,t) - \alpha_0(X,t))\Big)Y(t)dt \\ &+ \int_0^1 (\log\alpha(X,t) - \log\alpha_0(X,t))dM(t)\end{aligned}$$

where $\Phi_a(x) := -\log(1-ax)/a$ for $a > 0$ and $x < 1/a$. But since $\Phi_a(x) \le x + ax^2$ for any $x \le 1/(2a)$, we obtain

$$\ell(\alpha,\alpha_0) \le \frac{3}{2a_0}\int_0^1 (\alpha(t,X) - \alpha_0(t,X))^2 Y(t)dt + \int_0^1 (\log\alpha_0(t,X) - \log\alpha(t,X))dM(t)$$

which gives by integration with respect to P_α

$$K(P_\alpha, P_{\alpha_0}) \le \frac{3\|\alpha - \alpha_0\|_\mu^2}{2a_0} \le \frac{3\|f_X\|_\infty\|\alpha - \alpha_0\|_A^2}{2a_0} \le \frac{3b\|f_X\|_\infty|R_j|}{2na_0},$$

for any $\alpha \in \mathcal{A}_j$. Since the counting processes $(N^1, \dots, N^n)$ are independent, we have $K(P^n_\alpha, P^n_{\alpha_0}) = nK(P_\alpha, P_{\alpha_0})$ and

$$\frac{1}{M}\sum_{p=0}^{M} K(P^n_{\alpha_p}, P^n_{\alpha_0}) \leq \frac{3b\|f_X\|_\infty |R_j|}{2a_0} \leq a \log M_j$$

with $a = 12b\|f_X\|_\infty/(a_0 \log 2) \in (0, 1/8)$ for b small enough. It only remains to choose the levels j_1 and j_2 so that (24) and (25) holds, and to compute the corresponding v_j. We take $j = (j_1, j_2)$ such that

$$c_1/2 \leq 2^{j_1} n^{-\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq c_1 \text{ and } c_2/2 \leq 2^{j_2} n^{-\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq c_2$$

where c_1 and c_2 are positive constants satisfying $(c_1^{\beta_1} + c_2^{\beta_2})\sqrt{c_1 c_2} \leq L/(2\sqrt{bc})^{1/2}$. For this choice, $2^{j_1+j_2}/n \leq c_1 c_2 n^{-2\bar{\beta}/(2\bar{\beta}+2)}$ so (24) holds for n large enough and (25) holds and $v_j \geq c_3 n^{-\bar{\beta}/(2\bar{\beta}+2)}$ where $c_3 = \sqrt{bcc_1c_2/128}$. □

6. Deviation and maximal inequalities for the empirical process

Usually, in model selection (see for instance Massart (2007)), the penalty is explained using the so-called Talagrand's deviation inequality for the maximum of empirical processes. Because the empirical process $\nu(\cdot)$ (see Equation (19)) considered here has a particular structure, we cannot use directly Talagrand's inequality. In this Section, we prove Bennett and Bernstein inequalities for $\nu_n(\cdot)$, and derive a maximal bound using the so-called chaining technique which explains the penalty (7).

6.1. **Deviation inequality.**

Lemma 2. *For any positive δ, ϵ and for any function $h \in (L^2 \cap L^\infty)(A)$, we have the following Bennett-type deviation inequality*:

$$\mathbb{P}\big(\nu_n(h) \geq \epsilon, \|h\|_n \leq \delta\big) \leq \exp\Big(-\frac{n\delta^2\|\alpha\|_{\infty,A}}{\|h\|^2_{\infty,A}} g\Big(\frac{\epsilon\|h\|_{\infty,A}}{\|\alpha\|_{\infty,A}\delta^2}\Big)\Big)$$

where $g(x) = (1+x)\log(1+x) - x$ for any $x \geq 0$. As a consequence, we obtain the following Bernstein-type inequalities:

$$\mathbb{P}\big(\nu_n(h) \geq \epsilon, \|h\|_n \leq \delta\big) \leq \exp\Big(-\frac{n\epsilon^2/2}{\|\alpha\|_{A,\infty}\delta^2 + \frac{1}{3}\epsilon\|h\|_{A,\infty}}\Big), \tag{26}$$

and

$$\mathbb{P}\Big(\nu_n(h) \geq \delta\sqrt{\|\alpha\|_{\infty,A}x} + \|h\|_{\infty,A}x/3, \;\; \|h\|^2_n \leq \delta^2\Big) \leq \exp(-nx). \tag{27}$$

Proof. Remark that $\nu_n(h) = \nu(h, 1)$ where $\nu(h, \cdot)$ is the stochastic process given by

$$n\nu(h,t) := \sum_{i=1}^{n}\int_0^t h(X_i, z)dM^i(z) := n\sum_{i=1}^{n}\nu(h,t)^i.$$

The predictable variation of M^i is given by $\langle M^i(t)\rangle = \int_0^t \alpha(X_i, z)Y^i(z)dz$, so we have

$$\langle n\nu(h,t)^i\rangle = \int_0^t h(X_i, z)^2\alpha(X_i, z)Y^i(z)dz$$

for any $t \in [0,1]$. Moreover, we have $\Delta M^i(t) \in \{0,1\}$ for any $i = 1,\dots,n$ since the counting processes N^i admit intensities. We can write $\nu(h,t)^i = \nu(h,t)^{i,c} + \nu(h,t)^{i,d}$ where $\nu(h,t)^{i,c}$ is a continuous martingale and where $\nu(h,t)^{i,d}$ is a purely discrete martingale (see e.g. Liptser and Shiryayev (1989)). For some $a > 0$ (to be chosen later on) we define $U_a^i(t) := an\nu^i(h,t) - S_a^i(t)$, where $S_a^i(t)$ is the compensator of

$$\frac{1}{2}\langle an\nu(h,t)^{i,c}\rangle + \sum_{s\le t}\Big(\exp(a|\Delta n\nu(h,s)^i|) - 1 - a|\Delta n\nu(h,s)^i|\Big). \tag{28}$$

We know from the proof of Lemma 2.2 and Corollary 2.3 of van de Geer (1995), that $\exp(U_a^i(t))$ is a supermartingale. Using the standard Cramér-Chernoff method (see for instance Massart (2007), Chapter 2), we have, for any $a > 0$:

$$\begin{aligned}
&\mathbb{P}\Big(\nu_n(h) \ge \epsilon, ||h||_n \le \delta\Big)\\
&= \mathbb{P}\Big(\exp(an\nu_n(h)) \ge \exp(na\epsilon), ||h||_n \le \delta\Big)\\
&\le \Big(\mathbb{E}\Big[\exp\Big(an\sum_{i=1}^n \nu(h,1)^i - \sum_{i=1}^n S_a^i(1)\Big)\Big]\Big)^{1/2}\Big(\mathbb{E}\Big[\exp\Big(\sum_{i=1}^n S_a^i(1) - an\epsilon\Big)\mathbb{1}\{||h||_n \le \delta\}\Big]\Big)^{1/2}\\
&\le \Big(\mathbb{E}\Big[\exp\Big(\sum_{i=1}^n S_a^i(1) - an\epsilon\Big)\mathbb{1}\{||h||_n \le \delta\}\Big]\Big)^{1/2}.
\end{aligned}$$

The last inequality holds since $\exp(U_a^i(t)) = \exp(an\nu^i(h,t) - S_a^i(t))$ are independent supermartingales with $U_a^i(0) = 0$, so that $\mathbb{E}[\exp(U_a^i(t))] \le 1$, for $i = 1,\dots,n$.

Let us decompose $M^i = M^{i,c} + M^{i,d}$, with $M^{i,c}$ a continuous martingale and $M^{i,d}$ a purely discrete martingale. The process $V_2^i(t) := \langle M^i(t)\rangle$ is the compensator of the quadratic variation process $[M^i(t)] = \langle M^{i,c}(t)\rangle + \sum_{s\le t}|\Delta M^i(t)|^2$. If $k \ge 3$, we define $V_k^i(t)$ as the compensator of the k-variation process $\sum_{s\le t}|\Delta M^i(t)|^k$ of $M^i(t)$. Since $\Delta M^i(t) \in \{0,1\}$ for all $0 \le t \le 1$, the V_k^i are all equal for $k \ge 3$ and such that $V_k^i(t) \le V_2^i(t)$, for all $k \ge 3$. The process $S_a^i(1)$ has been defined as the compensator of (28). As a consequence, we have:

$$S_a^i(1) = \sum_{k\ge 2}\frac{a^k}{k!}\int_0^1 |h(X_i,z)|^k dV_k^i(z) \le \int_0^1 h(X_i,z)^2 dV_2^i(z) \times \sum_{k\ge 2}\frac{\|h\|_{\infty,A}^{k-2}}{k!}a^k$$

and if $\|h\|_n \le \delta$

$$\sum_{i=1}^n S_a^i(1) \le \bar{S}_a^n := \frac{n\delta^2\|\alpha\|_{\infty,A}}{\|h\|_{\infty,A}^2}\Big(\exp\big(a\|h\|_{\infty,A}\big) - 1 - a\|h\|_{\infty,A}\Big).$$

The minimum of $\bar{S}_a^n - an\epsilon$ for $a > 0$ is achieved by

$$a = \frac{1}{\|h\|_{\infty,A}}\log\Big(\frac{\epsilon\|h\|_{\infty,A}}{\|\alpha\|_{\infty,A}\delta^2} + 1\Big)$$

and is equal to

$$-\frac{n\delta^2\|\alpha\|_{\infty,A}}{\|h\|_{\infty,A}^2}g\Big(\frac{\epsilon\|h\|_{\infty,A}}{\|\alpha\|_{\infty,A}\delta^2}\Big)$$

where we recall that $g(x) = (1+x)\log(1+x) - x$. This concludes the proof of the Bennett inequality. Inequality (26) follows from the fact that $g(x) \geq 3x^2/(2(x+3))$ for any $x \geq 0$. To prove (27), we use the following trick from Birgé and Massart (1998): we have $g(x) \geq g_2(x)$ for any $x \geq 0$ where $g_2(x) := x+1-\sqrt{1+2x}$ and $g_2^{-1}(y) = \sqrt{2y}+y$. $\square$

6.2. **Proof of Proposition 5 (maximal inequality via $L^2 - L^\infty$ chaining).** Using a $L^2 - L^\infty$ chaining method, as in Barron et al. (1999) or Comte (2001), we obtain the following result, which leads to Proposition (5):

Lemma 3. *Let $B^\mu_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_\mu \leq 1\}$. Then*

$$\mathbb{E}\Big(\sup_{h \in B_{m,m'}(0,1)} (\nu_n^2(h) - p(m,m'))_+ \mathbb{1}(\Delta)\Big) \leq C(1+\|\alpha\|_{\infty,A})\frac{e^{-D_{m'}}}{n},$$

where

$$p(m,m') = \kappa(1+\|\alpha\|_{\infty,A})\frac{D_m + D_{m'}}{n}.$$

Proof. The result of Lemma 3 is obtained from Inequality (26) by a $L^2(\mu) - L^\infty$ chaining technique. The method is analogous to the one given in Proposition 4 p. 282-287 in Comte (2001), in Theorem 5 in Birgé and Massart (1998) and in Proposition 7, Theorem 8 and Theorem 9 in Barron et al. (1999). Since the context is different, we give, for the sake of completeness, the details of the proof. It relies on the following lemma (Lemma 9 in Barron *et al.* (1999)):

Lemma (Barron et al. (1999))**.** *Let μ be a positive measure on $[0,1]$. Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be a finite orthonormal system in $L^2 \cap L^\infty(\mu)$ with $|\Lambda| = D$ and $\bar{S}$ be the linear span of $\{\psi_\lambda\}$. Let*

$$\bar{r} = \frac{1}{\sqrt{D}} \sup_{\beta \neq 0} \frac{\|\sum_{\lambda\in\Lambda} \beta_\lambda \psi_\lambda\|_\infty}{|\beta|_\infty}. \tag{29}$$

For any positive δ, one can find a countable set $T \subset \bar{S}$ and a mapping p from $\bar{S}$ to T with the following properties:

- *for any ball $\mathcal{B}$ with radius $\sigma \geq 5\delta$,*
$$|T \cap \mathcal{B}| \leq (B'\sigma/\delta)^D \quad \text{with} \quad B' < 5,$$
- *$\|u - p(u)\|_\mu \leq \delta$ for all u in $\bar{S}$, and*
$$\sup_{u \in p^{-1}(t)} \|u - t\|_\infty \leq \bar{r}\delta, \quad \text{for all } t \text{ in } T.$$

To use this lemma, the main difficulty is often to evaluate $\bar{r}$ in the different contexts. We consider a collection of product models $(S_m)_{m\in\mathcal{M}_n}$ which can be [DP] or [T]. For the sake of place, we omit collection [W] as it right similar to collection [DP]. Recall that $B^\mu_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_\mu \leq 1\}$. We have to compute $\bar{r} = \bar{r}_{m,m'}$ corresponding to $\bar{S} = S_m + S_{m'} \subset \mathcal{S}_n$ on which the norm connection holds. We denote by $D(m,m') = \dim(S_m + S_{m'})$.

- Collection [DP] – As $S_m + S_{m'}$ is a linear space, an orthonormal $L^2(\mu)$-basis $(\psi_\lambda)_{\lambda\in\Lambda_n}$ can be built by orthonormalisation on each sub-rectangle of $(\varphi_\lambda)_{\lambda\in\Lambda_n}$, the orthonormal basis of $\mathcal{S}_n$. Then

$$\begin{aligned}\sup_{\beta\neq 0}\frac{\|\sum_{\lambda\in\Lambda_n}\beta_\lambda\psi_\lambda\|_{\infty,A}}{|\beta|_\infty} &\leq \|\sum_{\lambda\in\Lambda_n}|\psi_\lambda|\|_{\infty,A}\leq (r+1)\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|_{\infty,A}\\ &\leq (r+1)^{3/2}\sqrt{N_n}\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|\\ &\leq (r+1)^{3/2}\sqrt{N_n}\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|_\mu/\sqrt{f_0}\\ &\leq (r+1)^{3/2}\sqrt{N_n/f_0}.\end{aligned}$$

  Thus here $\bar{r}_{m,m'}\leq((r+1)^{3/2}/\sqrt{f_0})\sqrt{N_n/D(m,m')}$.
- Collection [T]– For trigonometric polynomials, we write

$$\begin{aligned}\sup_{\beta\neq 0}\frac{\|\sum_{\lambda\in\Lambda_n}\beta_\lambda\psi_\lambda\|_{\infty,A}}{|\beta|_\infty} &\leq \frac{C\sqrt{N_n}\|\sum_\lambda\beta_\lambda\psi_\lambda\|}{|\beta|_\infty}\leq\frac{C\sqrt{N_n}\|\sum_\lambda\beta_\lambda\psi_\lambda\|_\mu}{\sqrt{f_0}|\beta|_\infty}\\ &\leq \frac{C\sqrt{N_n}\sqrt{\sum_\lambda\beta_\lambda^2}}{\sqrt{f_0}|\beta|_\infty}\leq\frac{C\sqrt{N_nD(m,m')}}{\sqrt{f_0}}.\end{aligned}$$

  Therefore, $\bar{r}_{m,m'}\leq C\sqrt{N_n/f_0}$.

We may now prove Lemma 3. We apply the Lemma from Barron et al. (1999) to the linear space $S_m + S_{m'}$ of dimension $D(m,m')$ and norm connection measured by $\bar{r}_{m,m'}$ bounded above. We consider δ_k-nets $T_k = T_{\delta_k}\cap B^\mu_{m,m'}(0,1)$, with $\delta_k = \delta_0 2^{-k}$ and $\delta_0\leq 1/5$ (to be chosen later). Moreover we set $H_k = \log(|T_k|)\leq D(m,m')\log(5/\delta_k) = D(m,m')[k\log(2)+\log(5/\delta_0)]$. Given some point $h\in B^\mu_{m,m'}(0,1)$, we can find a sequence $\{h_k\}_{k\geq 0}$ with $h_k\in T_k$ such that $\|h-h_k\|^2_\mu\leq\delta_k^2$ and $\|h-h_k\|_{\infty,A}\leq\bar{r}_{m,m'}\delta_k$. Thus we have the following decomposition that holds for any $h\in B^\mu_{m,m'}(0,1)$:

$$h = h_0 + \sum_{k\geq 1}(h_k - h_{k-1}),$$

with $\|h_0\|_\mu\leq 1$, $\|h_0\|_{\infty,A}\leq\bar{r}_{(m,m')}$, and

$$\|h_k - h_{k-1}\|^2_\mu\leq 2(\delta_k^2+\delta_{k-1}^2) = 5\delta_{k-1}^2/2,\quad \|h_k-h_{k-1}\|_{\infty,A}\leq 3\bar{r}_{(m,m')}\delta_{k-1}/2$$

for any $k\geq 1$. In the sequel we denote by $\mathbb{P}_\Delta(\cdot)$ the measure $\mathbb{P}(\cdot\cap\Delta)$, see (16). Let in addition $(\eta_k)_{k\geq 0}$ be a sequence of positive numbers that will be chosen later on and η such that $\eta_0+\sum_{k\geq 1}\eta_k\leq\eta$. We have:

$$\begin{aligned}&\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)}\nu_n(h)>\eta\Big]\\ =\ &\mathbb{P}_\Delta\Big[\exists(h_k)_{k\in\mathbb{N}}\in\prod_{k\in\mathbb{N}}T_k\ /\ \nu_n(h_0)+\sum_{k=1}^{+\infty}\nu_n(h_k-h_{k-1})>\eta_0+\sum_{k\geq 1}\eta_k\Big]\\ \leq\ &\mathbb{P}_1+\mathbb{P}_2\end{aligned}$$

where

$$\mathbb{P}_1 = \sum_{h_0 \in T_0} \mathbb{P}_\Delta(\nu_n(h_0) > \eta_0), \;\; \mathbb{P}_2 = \sum_{k=1}^{\infty} \sum_{\substack{h_{k-1} \in T_{k-1} \\ h_k \in T_k}} \mathbb{P}_\Delta(\nu_n(h_k - h_{k-1}) > \eta_k).$$

Then using Inequality (27), we straightforwardly infer that $\mathbb{P}_1 \leq \exp(H_0 - nx_0)$ and $\mathbb{P}_2 \leq \sum_{k\geq 1} \exp(H_{k-1} + H_k - nx_k)$ if we choose

$$\begin{cases} \eta_0 = \sqrt{3x_0\|\alpha\|_{\infty,A}/2} + \bar{r}_{(m,m')}x_0/3 \\ \eta_k = (1/2)\delta_{k-1}(\sqrt{15\|\alpha\|_{\infty,A}x_k} + \bar{r}_{(m,m')}x_k). \end{cases}$$

Fix $u > 0$ and choose x_0 such that

$$nx_0 = H_0 + D_{m'} + u$$

and for $k \geq 1$, x_k such that

$$nx_k = H_{k-1} + H_k + kD_{m'} + D_{m'} + u.$$

If $D_{m'} \geq 1$, we infer that

$$\mathbb{P}_\Delta\Big(\sup_{h \in B^{\mu}_{m,m'}(0,1)} \nu_n(h) > \eta_0 + \sum_{k\geq 1} \eta_k\Big) \leq e^{-D_{m'}-u}\Big(1 + \sum_{k=1}^{\infty} e^{-kD_{m'}}\Big) \leq 1.6e^{-D_{m'}-u}.$$

Now, it remains to compute $\sum_{k\geq 0} \eta_k$. We note that $\sum_{k=0}^{\infty} \delta_k = \sum_{k=0}^{\infty} k\delta_k = 2\delta_0$. This implies that:

$$\begin{aligned} x_0 &+ \sum_{k=1}^{\infty} \delta_{k-1}x_k \\ &\leq \Big[\log(5/\delta_0) + \delta_0 \sum_{k=1}^{\infty} 2^{-(k-1)}[(2k-1)\log(2) + 2\log(5/\delta_0) + k]\Big]\frac{D(m,m')}{n} \\ &+ \Big(1 + \delta_0 \sum_{k\geq 1} 2^{-(k-1)}\Big)\frac{D_{m'}}{n} + \Big(1 + \delta_0 \sum_{k\geq 1} 2^{-(k-1)}\Big)\frac{u}{n} \\ &\leq \frac{a(\delta_0)D(m,m')}{n} + \frac{1+2\delta_0}{n}(D_{m'} + u), \end{aligned} \tag{30}$$

where $a(\delta_0) = \log(5/\delta_0) + \delta_0(4\log(5/\delta_0) + 6\log(2) + 4)$. This leads to

$$\begin{aligned} \Big(\sum_{k=0}^{\infty} \eta_k\Big)^2 &\leq \frac{1}{4}\Big[\sqrt{2}\Big(\sqrt{3\|\alpha\|_{\infty,A}x_0/2} + \bar{r}_{m,m'}x_0/3\Big) + \sum_{k=1}^{\infty} \delta_{k-1}\Big(\sqrt{15\|\alpha\|_{\infty,A}x_k} + \bar{r}_{m,m'}x_k\Big)\Big]^2 \\ &\leq \frac{1}{4}\Big[\Big(\sqrt{3\|\alpha\|_{\infty,A}x_0} + \sum_{k=1}^{\infty} \delta_{k-1}\sqrt{15\|\alpha\|_{\infty,A}x_k}\Big) + \bar{r}_{m,m'}\Big(\sqrt{2}x_0/3 + \sum_{k=1}^{\infty} \delta_{k-1}x_k\Big)\Big]^2 \\ &\leq \frac{15}{4}\Big[\Big(\sqrt{x_0} + \sum_{k=1}^{\infty} \delta_{k-1}\sqrt{x_k}\Big)^2 \|\alpha\|_{\infty,A} + \bar{r}^2_{m,m'}\Big(x_0 + \sum_{k=0}^{\infty} \delta_{k-1}x_k\Big)^2\Big] \\ &\leq 4\Big[2\Big(x_0 + \sum_{k=1}^{\infty} \delta_{k-1}x_k\Big)\|\alpha\|_{\infty,A} + \bar{r}^2_{m,m'}\Big(x_0 + \sum_{k=1}^{\infty} \delta_{k-1}x_k\Big)^2\Big]. \end{aligned}$$

Now, fix $\delta_0 \leq 1/5$ (say, $\delta_0 = 1/10$) and use the bound (30). The bound for $(\sum_{k=0}^{+\infty} \eta_k)^2$ is less than a quantity proportional to:

$$\Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big)\|\alpha\|_{\infty,A} + \bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big)^2 + \frac{\|\alpha\|_{\infty,A}u}{n} + \bar{r}^2_{m,m'}\frac{u^2}{n^2}.$$

For collection [DP], we use that $\bar{r}^2_{m,m'} \leq (r+1)^3 N_n/(f_0 D(m,m'))$ and $N_n \leq n/\log n$ to obtain the bound:

$$\begin{aligned}\bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big)^2 &\leq c(r+1)^3 \frac{N_n}{f_0 D(m,m')}\frac{D(m,m')^2}{n^2} \\ \leq \quad \frac{c(r+1)^3}{f_0}\frac{N_n D(m,m')}{n^2} &\leq \frac{c(r+1)^3}{f_0}\frac{1}{\log n}\frac{D(m,m')}{n} \leq \frac{D(m,m')}{n}.\end{aligned}$$

For collection [T], we have $\bar{r}_{m,m'} \leq C\sqrt{N_n}$ and $N_n \leq \sqrt{n}/\log n$. We get

$$\bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big)^2 \quad \leq \quad \frac{CN_n D(m,m')^2}{n^2} \leq \frac{C}{\log n}\frac{D(m,m')}{n} \leq \frac{D(m,m')}{n}.$$

Thus, for both the cases, the bound for $(\sum \eta_k)^2$ is proportional to:

$$(1 + \|\alpha\|_{\infty,A})\Big[\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big] + \frac{\|\alpha\|_{\infty,A}u}{n} + \bar{r}^2_{m,m'}\frac{u^2}{n^2}.$$

We obtain, as $D(m,m') \leq D_m + D_{m'}$,

$$\begin{aligned}&\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)}[\nu_n(h)]^2 > \kappa\Big((1+\|\alpha\|_{\infty,A})\frac{D_m + D_{m'}}{n} + (\frac{\|\alpha\|_{\infty,A}u}{n} \vee \bar{r}^2_{m,m'}\frac{u^2}{n^2})\Big)\Big] \\ \leq \quad &\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)}[\nu_n(h)]^2 > \eta^2\Big] \leq 2\,\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)}\nu_n(h) > \eta\Big] \leq 3.2e^{-D_{m'}-u}\end{aligned}$$

so that, if we take $\kappa_\alpha := \kappa(1 + \|\alpha\|_{\infty,A})$,

$$\begin{aligned}&\mathbb{E}\Big[\Big(\sup_{h\in B^\mu_{m,m'}(0,1)}\nu_n^2(h) - p(m,m')\Big)_+ \mathbb{1}(\Delta)\Big] \\ &\qquad \leq \int_0^\infty \mathbb{P}_\Delta\Big(\sup_{h\in B^\mu_{m,m'}(0,1)}\nu_n^2(h) > p(m,m') + u\Big)du \\ &\qquad \leq e^{-D_{m'}}\Big(\int_{2\kappa_\alpha/\bar{r}^2_{m,m'}}^\infty e^{-nu/(2\kappa_\alpha)}du + \int_0^{2\kappa_\alpha/\bar{r}^2_{m,m'}} e^{-n\sqrt{u}/(2\sqrt{\kappa_\alpha}\bar{r}_{m,m'})}du\Big) \\ &\qquad \leq e^{-D_{m'}}\frac{2\kappa_\alpha}{n}\Big(\int_0^\infty e^{-v}dv + \frac{2\bar{r}^2_{m,m'}}{n}\int_0^\infty e^{-\sqrt{v}}dv\Big) \\ &\qquad \leq e^{-D_{m'}}\frac{2\kappa_\alpha}{n}(1 + \frac{4\bar{r}^2_{m,m'}}{n}) \leq \frac{\kappa'_\alpha e^{-D_{m'}}}{n},\end{aligned}$$

where κ'_α is a constant depending on $\|\alpha\|_{\infty,A}$. This ends the proof of Lemma 3.
To conclude the proof of Proposition 5, we just have to bound $\sum_{m'\in\mathcal{M}_n} e^{-D_{m'}}$. This term is at most

$$\sum_{j,k\geq 1} e^{-jk} \quad = \quad \sum_{j=1}^\infty\sum_{k=1}^\infty (e^{-j})^k = \sum_{j=1}^\infty \frac{e^{-j}}{1-e^{-j}} \leq \frac{1}{1-e^{-1}}\sum_{j=1}^\infty e^{-j} = \frac{e^{-1}}{(1-e^{-1})^2}.$$

□

7. Proof of the auxiliary results

7.1. **Proof of Proposition 1.** Let $\hat{f}_{m_1^*}$ and $\hat{f}_0$ be defined by (12), with $m_1^* = (D_{m_1}, \mathcal{D}_n^{(2)})$ with $\log n \leq D_{m_1} \leq n^{1/4}/\sqrt{\log n}$ and $\mathcal{D}_n^{(2)} \leq n^{1/4}/\sqrt{\log n}$, see $(\mathcal{M}_1)$. We remark that, for all $(x,z) \in \mathbb{R}^2$,

$$\hat{f}_{m_1^*}(x,z) = f(x,z) + \hat{f}_{m_1^*}(x,z) - f(x,z) \geq f_0 - \|\hat{f}_{m_1^*} - f\|_{\infty,A}.$$

We deduce that $\|\hat{f}_{m_1^*} - f\|_{\infty,A} \geq f_0 - \hat{f}_0$. In the same manner, $\|\hat{f}_{m_1^*} - f\|_{\infty,A} \geq \hat{f}_0 - f_0$. Thus

$$\mathbb{P}(\Omega^{\complement}) = \mathbb{P}(|f_0 - \hat{f}_0| > f_0/2) \leq \mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2).$$

Therefore, we just have to prove that $\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2) \leq C_k/n^k$.

First remark that $\|\hat{f}_{m_1^*} - f\|_{\infty,A} \leq \|\hat{f}_{m_1^*} - f_{m_1^*}\|_{\infty,A} + \|f_{m_1^*} - f\|_{\infty,A}$. As $f \in B_{2,\infty}^{(\tilde{\beta}_1,\tilde{\beta}_2)}(A)$ with $\bar{\tilde{\beta}} > 1$, the imbedding theorem proved in Nikol'skii (1975) p.236 implies that f belongs to $B_{\infty,\infty}^{(\beta_1^*,\beta_2^*)}(A)$ with $\beta_1^* = \tilde{\beta}_1(1 - 1/\bar{\tilde{\beta}})$ and $\beta_2^* = \tilde{\beta}_2(1 - 1/\bar{\tilde{\beta}})$. Then the approximation lemma of Lacour (2007) recalled in Section 5.2, which is still valid for the trigonometric polynomial spaces with the infinite norm instead of the L^2 norm, yields to

$$\|f_{m_1*} - f\|_{\infty,A} \leq C(D_{m_1*}^{-\beta_1^*} + (\mathcal{D}_n^{(2)})^{-\beta_2^*}).$$

As we assumed that $D_{m_1^*} \geq \log n$, it follows that $\|f_{m_1*} - f\|_{\infty,A}$ tends to zero when $n \to +\infty$. Thus, for n large enough, we have $\|f_{m_1*} - f\|_{\infty,A} \leq f_0/4$ and

$$\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2) \leq \mathbb{P}(\|\hat{f}_{m_1^*} - f_{m_1^*}\|_{\infty,A} > f_0/4).$$

Now, following $(\mathcal{M}2)$, we get

$$\|\hat{f}_{m_1^*} - f_{m_1^*}\|_{\infty,A} \leq \sqrt{\phi_1\phi_2 D_{m_1^*}\mathcal{D}_n^{(2)}}\|\hat{f}_{m_1^*} - f_{m_1^*}\|.$$

Now we define

$$\vartheta_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int \Big(h(X_i,y)Y^i(y) - \mathbb{E}\big(h(X_i,y)Y^i(y)\big)\Big)dy = \|\sqrt{h}\|_n^2 - \|\sqrt{h}\|_\mu^2. \tag{31}$$

With this notation, and reminding of (11) and of the proof of Proposition 2 in Section 3.4, we have

$$\|\hat{f}_{m_1^*} - f_{m_1^*}\|^2 = \sum_{j,k}(\hat{b}_{j,k} - b_{j,k})^2 = \sum_{j,k}\vartheta_n^2(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}).$$

Thus

$$\begin{aligned}
\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2) &\leq \mathbb{P}\Big(\sum_{j,k}\vartheta_n^2(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}) \geq \frac{f_0^2}{16\phi_1\phi_2 D_{m_1^*}\mathcal{D}_n^{(2)}}\Big) \\
&\leq \sum_{j,k}\mathbb{P}\Big(\vartheta_n^2(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}) \geq \frac{f_0^2}{16\phi_1\phi_2(D_{m_1^*}\mathcal{D}_n^{(2)})^2}\Big) \\
&\leq \sum_{j,k}\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \geq \frac{f_0}{4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)}}\Big).
\end{aligned}$$

Notice that $\vartheta_n(\varphi_j^{m_1^*}\otimes\psi_k^{m_1^*})=\frac{1}{n}\sum_1^n(U_i^{j,k}-\mathbb{E}(U_i^{j,k}))$, where $U_i^{j,k}=\varphi_j(X_i)\int\psi_k(y)Y^i(y)dy$ are i.i.d. r.v. We can apply the Bernstein inequality to ϑ_n i.e. to the i.i.d. r.v. $U_i^{j,k}$. Indeed, we have

$$\|U_i^{j,k}\|_\infty\le\|\varphi_j\|_\infty\int|\psi_k(y)|dy\le\|\varphi_j\|_\infty(\int\psi_k^2(y)dy)^{1/2}\le\sqrt{\phi_1 D_{m_1^*}}:=c$$

and $\mathbb{E}[(U_i^{j,k})^2]\le\|f_X\|_{\infty,A}=v^2$. We get

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*}\otimes\psi_k^{m_1^*})|\ge\frac{f_0}{4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)}}\Big)\le 2\exp(-\frac{nx^2/2}{v^2+cx})$$

with $x=f_0/(4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)})$ and v and c are right above. That is:

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*}\otimes\psi_k^{m_1^*})|\ge\frac{f_0}{4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)}}\Big)\le 2\exp\Big(-\frac{Cnf_0^2}{16\phi_1\phi_2(D_{m_1^*}\mathcal{D}_n^{(2)})^2}\Big).$$

As both $D_{m_1^*}$ and $\mathcal{D}_n^{(2)}$ are less than $n^{1/4}/\sqrt{\log(n)}$, we obtain:

$$\mathbb{P}(\Omega^{\complement})\le 2D_{m_1^*}\mathcal{D}_n^{(2)}\exp\Big(-\frac{Cnf_0^2}{16\phi_1\phi_2(D_{m_1^*}\mathcal{D}_n^{(2)})^2}\Big)\le 2\sqrt{n}\exp\Big(-C'(\log n)^2\Big)\le\frac{C'_k}{n^k},$$

for any k arbitrarily large, when n is large enough.

Proof of Proposition 3. Note that $\hat\alpha_{\hat m}$ is either 0 or $\operatorname{argmin}_{t\in S_{\hat m}}\gamma_n(t)$. Let us denote for short $\varphi_j:=\varphi_j^{\hat m}$ and $\psi_k:=\psi_k^{\hat m}$. In the second case, $\min\operatorname{Sp}(G_{\hat m})\ge\max(\hat f_0/3,n^{-1/2})$ and thus

$$\begin{aligned}
\|\hat\alpha_{\hat m}\|^2 &= \sum_{j,k}(\hat a_{j,k}^{\hat m})^2=\|A_{\hat m}\|^2=\|G_{\hat m}^{-1}\Upsilon_{\hat m}\|^2\\
&\le (\min\operatorname{Sp}(G_{\hat m}))^{-2}\|\Upsilon_{\hat m}\|^2\le\min(9/\hat f_0^2,n)\sum_{j,k}\Big(\frac{1}{n}\sum_{i=1}^n\varphi_j(X_i)\int\psi_k(z)dN^i(z)\Big)^2\\
&\le \min(9/\hat f_0^2,n)\frac{1}{n}\sum_{i=1}^n\sum_j\varphi_j^2(X_i)\sum_k\Big(\int\psi_k(z)dN^i(z)\Big)^2\\
&\le \min(9/\hat f_0^2,n)\phi_1\mathcal{D}_n^{(1)}\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\int\psi_k(z)dN^i(z)\Big)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\hat\alpha_{\hat m}\|^4 &\le n^2\phi_1^2(\mathcal{D}_n^{(1)})^2\frac{1}{n}\sum_{i=1}^n\left(\sum_k\Big(\int\psi_k(z)dN^i(z)\Big)^2\right)^2\\
&\le n^2\phi_1^2(\mathcal{D}_n^{(1)})^2\mathcal{D}_n^{(2)}\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\int\psi_k(z)dN^i(z)\Big)^4. \qquad (32)
\end{aligned}$$

Now, we have:

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\Big(\int \psi_k(z)dN^i(z)\Big)^4\Big)\\
\leq\ & 2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)dM^i(z)\Big)^4\Big)+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big).
\end{aligned}$$

Using the Bürkholder Inequality as recalled in Liptser and Shiryayev (1989) p 75, and the fact that the quadratic variation process of each M^i is N^i $(i=1,\dots,n)$, we obtain:

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\Big(\int \psi_k(z)dN^i(z)\Big)^4\Big)\\
\leq\ & 2^3C_b\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k^2(z)dN^i(z)\Big)^2\Big)+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big)\\
\leq\ & 2^3C_b\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\sum_{s:\Delta N^i(s)\neq 0}\psi_k^4(s)\Big)\Big)+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big)\\
\leq\ & 2^3C_b\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\Big(\sum_{s:\Delta N^i(s)\neq 0}\sum_{k}\psi_k^4(s)\Big)\Big)+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big)\\
\leq\ & 2^3C_b\phi_2(\mathcal{D}_n^{(2)})^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big(\Big(\sum_{s:\Delta N^i(s)\neq 0}1\Big)\Big)+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big)\\
\leq\ & 2^3C_b\phi_2(\mathcal{D}_n^{(2)})^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(N^i(1))+2^3\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big)
\end{aligned}$$

This yields, using Assumptions $(\mathcal{A}3)$ and $(\mathcal{A}4)$:

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\Big(\int \psi_k(z)dN^i(z)\Big)^4\Big)\\
\leq\ & C\Big(\phi_2(\mathcal{D}_n^{(2)})^2\mathbb{E}(N^1(1))+\sum_{k}\mathbb{E}\Big(\Big(\int \psi_k(z)\alpha(X,z)Y(z)dz\Big)^4\Big)\Big)\\
\leq\ & C\Big(\phi_2(\mathcal{D}_n^{(2)})^2\mathbb{E}(N^1(1))+\|\alpha\|_{\infty,A}^4\|\sum_{k}\psi_k^2\|_{\infty,A}\sum_{k}\int \psi_k^2(z)dz\Big)\\
\leq\ & C\Big(\phi_2(\mathcal{D}_n^{(2)})^2\mathbb{E}(N^1(1))+\|\alpha\|_{\infty,A}^4\phi_2(\mathcal{D}_n^{(2)})^2\Big). \qquad (33)
\end{aligned}$$

Then we have, by inserting (33) in (32),

$$\begin{aligned}
\mathbb{E}(\|\hat{\alpha}_{\hat{m}}\|^4) &\leq (\phi_1 n\mathcal{D}_n^{(1)})^2\mathcal{D}_n^{(2)}\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\Big(\int \psi_k(z)dN^i(z)\Big)^4\Big)\\
&\leq Cn^2(\mathcal{D}_n^{(1)})^2(\mathcal{D}_n^{(2)})^3\leq C'n^{4.5}\leq C'n^5,
\end{aligned}$$

as we claim that we can reach $\mathcal{D}_n^{(i)} \leq \sqrt{n}/\log(n)$ in the case of localized bases [DP], [W], [H]. Note that for basis [T], under $(\mathcal{M}1)$, the final order is much less (namely $n^{3.25}$ instead of $n^{4.5}$).

Proof of Proposition 4. Define, for $\rho > 1$, the set

$$\Delta_\rho = \{\forall h \in \mathcal{S}_n, \Big| \|h\|_n^2/\|h\|_\mu^2 - 1\Big| \leq 1 - 1/\rho\},$$

where $\mathcal{S}_n$ is the set of maximal dimension of the collection. Remark that $\Delta = \Delta_2$, see (16). First we observe that:

$$\mathbb{P}(\Delta_\rho^{\complement}) \leq \mathbb{P}\Big(\sup_{h \in B^\mu_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| > 1 - 1/\rho\Big)$$

where $\vartheta_n(\cdot)$ is defined by (31) and $B^\mu_{\mathcal{S}_n}(0,1) = \{t \in \mathcal{S}_n, \|t\|_\mu \leq 1\}$. We denote by $(\varphi_j \otimes \psi_k)$ the $\mathbb{L}^2$-orthonormal basis of $\mathcal{S}_n$. If $h(x,y) = \sum_{j,k} a_{j,k}\varphi_j(x)\psi_k(y)$, then

$$\vartheta_n(h^2) = \sum_{j,k,j',k'} a_{j,k}a_{j',k'}\vartheta_n((\varphi_j \otimes \psi_k)(\varphi_{j'} \otimes \psi_{k'})). \tag{34}$$

We obtain

$$\sup_{h \in B^\mu_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| \leq f_0^{-1} \sup_{\sum a_{j,k}^2 \leq 1} \Big| \sum_{j,k,j',k'} a_{j,k}a_{j',k'}\vartheta_n((\varphi_j \otimes \psi_k)(\varphi_{j'} \otimes \psi_{k'}))\Big|. \tag{35}$$

Lemma (Baraud et al. (2001a))**.** *Let $B_{j,j'} = \|\varphi_j\varphi_{j'}\|_{\infty,A}$ and $V_{j,j'} = \|\varphi_j\varphi_{j'}\|_2$. Let, for any symmetric matrix $(A_{j,j'})$*

$$\bar{\rho}(A) := \sup_{\sum b_j^2 \leq 1} \sum_{j,j'} |b_j b_{j'}| A_{j,j'}$$

and $L(\varphi) := \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}$. Then, if $(\mathcal{M}2)$ is satisfied, we have $L(\varphi) \leq \phi_1(\mathcal{D}_n^{(1)})^2$, and $L(\varphi) \leq 5\phi_1^4\mathcal{D}_n^{(1)}$, if the basis is localized (cases [P] or [W]).

Let us define

$$\begin{aligned} x &:= \frac{f_0^2(1-1/\rho)^2}{4\|f_X\|_{\infty,A}(\mathcal{D}_n^{(2)})^2 L(\varphi)} \text{ and} \\ \Theta &:= \Big\{\forall (j,k)\forall(j',k') \quad |\vartheta_n((\varphi_j \otimes \psi_k)(\varphi_{j'} \otimes \psi_{k'}))| \leq 4\Big(B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x}\Big)\Big\}. \end{aligned}$$

Starting from (35), we have, on Θ:

$$\sup_{h \in B^\mu_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| \leq 4f_0^{-1} \sup_{\sum a_{j,k}^2 \leq 1} \sum_{j,j'} (\sum_{k,k'} |a_{j,k}a_{j',k'}|)\Big(B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x}\Big).$$

Thus setting $b_j = \sum_k |a_{j,k}|$, we have $\sum_j b_j^2 \leq \mathcal{D}_n^{(2)}$ and it follows that, on Θ,

$$\begin{aligned}
\sup_{h\in B^{\mu}_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| &\leq f_0^{-1}\mathcal{D}_n^{(2)} \sup_{\sum b_j^2=1} \sum_{j,j'} |b_j b_{j'}| \Big(B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x}\Big) \\
&\leq f_0^{-1}\mathcal{D}_n^{(2)} \Big(\bar\rho(B)x + \bar\rho(V)\sqrt{2\|f_X\|_{\infty,A}x}\Big) \\
&\leq (1-1/\rho)\Big(\frac{f_0(1-1/\rho)}{4\mathcal{D}_n^{(2)}\|f\|_{\infty,A}} \frac{\bar\rho(B)}{L(\varphi)} + \frac{1}{\sqrt{2}}\Big(\frac{\bar\rho^2(V)}{L(\varphi)}\Big)^{1/2}\Big) \\
&\leq (1-1/\rho)\Big(\frac{1}{4}+\frac{1}{\sqrt{2}}\Big) \leq (1-1/\rho).
\end{aligned}$$

Therefore,

$$\mathbb{P}\Big(\sup_{t\in B^{\mu}_{\mathcal{S}_n}(0,1)} |\vartheta_n(t^2)| > 1-\frac{1}{\rho}\Big) \leq \mathbb{P}(\Theta^{\complement}).$$

Let $\phi_\lambda = \varphi_j \otimes \psi_k$ for $\lambda = (j,k)$. To bound $\mathbb{P}(\vartheta_n(\phi_\lambda\phi_{\lambda'}) \geq B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x})$, we will apply the Bernstein inequality given in Birgé and Massart (1998) to the i.i.d. r.v.

$$U_i^{\lambda,\lambda'} = U_i^{(j,k),(j',k')} = \varphi_j(X_i)\varphi_{j'}(X_i)\int \psi_k(y)\psi_{k'}(y)Y^i(y)dy. \tag{36}$$

Under $(\mathcal{A}4)$, the r.v. are bounded

$$|U_i^{\lambda,\lambda'}| \leq \|\varphi_j\varphi_{j'}\|_{\infty,A}\int |\psi_k(y)\psi_{k'}(y)|dy \leq \|\varphi_j\varphi_{j'}\|_{\infty,A} = B_{j,j'}.$$

Moreover, using $(\mathcal{A}4)$ again, we obtain:

$$(U_i^{\lambda,\lambda'})^2 \leq (\varphi_j(X_i)\varphi_{j'}(X_i))^2 \int \psi_k^2(y)dy \int \psi_{k'}^2(y)dy = (\varphi_j(X_i)\varphi_{j'}(X_i))^2$$

and thus

$$\mathbb{E}[(U_i^{\lambda,\lambda'})^2] \leq \mathbb{E}[(\varphi_j(X_i)\varphi_{j'}(X_i))^2] \leq \|f_X\|_{\infty,A}V_{j,j'}^2.$$

We get

$$\mathbb{P}(|\vartheta_n(\phi_\lambda\phi_{\lambda'})| \geq B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x}) \leq 2e^{-nx}.$$

Given that $\mathbb{P}(\Delta_\rho^{\complement}) \leq \mathbb{P}(\Theta^{\complement}) = \sum_{\lambda,\lambda'} \mathbb{P}\Big(|\vartheta_n(\phi_\lambda\phi_{\lambda'})| > B_{j,j'}x + V_{j,j'}\sqrt{2\|f_X\|_{\infty,A}x}\Big)$, we can write:

$$\begin{aligned}
\mathbb{P}(\Delta_\rho^{\complement}) &\leq 2(\mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)})^2 \exp\Big\{ -\frac{nf_0^2(1-1/\rho)^2}{4\|f_X\|_{\infty,A}(\mathcal{D}_n^{(2)})^2 L(\varphi)}\Big\} \\
&\leq 2n^2 \exp\Big\{ -\frac{f_0^2(1-1/\rho)^2}{4\|f_X\|_{\infty,A}} \frac{n}{(\mathcal{D}_n^{(2)})^2 L(\varphi)}\Big\}.
\end{aligned}$$

Following the lemma of Baraud et al. (2001a) above, and using Assumption $(\mathcal{M}_1)$, we have

$$(\mathcal{D}_n^{(2)})^2 L(\varphi) \leq \phi_1(\mathcal{D}_n^{(2)}\mathcal{D}_n^{(1)})^2 \leq \phi_1 n/\log^2(n).$$

And then, we have for any k arbitrarily large, when n is large enough,

$$\mathbb{P}(\Delta_\rho^{\complement}) \leq 2n^2 \exp\Big\{ - \frac{f_0^2(1-1/\rho)^2}{40\|f\|_{\infty,A}\phi_1} \log^2(n) \Big\} \leq \frac{C_k}{n^k}. \tag{37}$$

Now, if the basis is localized, the result is better. In this case, $L(\varphi) \leq 5\phi_1^4 \mathcal{D}_n^{(1)}$. Moreover, take histogram basis in (34), then all terms with $k \neq k'$ vanish and then we can take $b_j = (\sum_k a_{j,k}^2)^{1/2}$ directly. Then, as then $\sum_j b_j^2 \leq 1$, we obtain

$$\mathbb{P}(\Delta_\rho^{\complement}) \quad \leq \quad 2(\mathcal{D}_n^{(1)})^2 \mathcal{D}_n^{(2)} \exp\Big\{ - \frac{n f_0^2(1-1/\rho)^2}{40\|f_X\|_{\infty,A} L(\varphi)} \Big\} \leq 2n^2 \exp\Big\{ - \frac{f_0^2(1-1/\rho)^2}{40\|f_X\|_{\infty,A}} \frac{n}{L(\varphi)} \Big\}.$$

Thus $L(\varphi) \leq 5\phi_1^4 \mathcal{D}_n^{(1)} \leq \phi_1 n/\log^2(n)$ is enough to get (37) again. The proof is easy to extend to any localized basis as [P] or [W], (with $\mathcal{D}_n^{(2)}$ in the bound of $\sum_j b_j^2$ replaced by $r+1$ in case [P] for instance).

Proof of Lemma 1. Let $m \in \mathcal{M}_n$ be fixed and let ℓ be an eigenvalue of G_m. There exists $A_m \neq 0$ with coefficients $(a_\lambda)_\lambda$ such that $G_m A_m = \ell A_m$ and thus $A_m^\top G_m A_m = \ell A_m^\top A_m$. Now, take $h := \sum_\lambda a_\lambda \varphi_\lambda \in S_m$. We have $\|h\|_n^2 = A_m^\top G_m A_m$ and $\|h\|_A^2 = A_m^\top A_m$. Thus, on Δ (see (16)):

$$A_m^\top G_m A_m = \|h\|_n^2 \geq \frac{1}{2}\|h\|_\mu^2 \geq \frac{1}{2} f_0 \|h\|_A^2 = \frac{1}{2} f_0 A_m^\top A_m.$$

Therefore, on Δ, for all $m \in \mathcal{M}_n$, we have $\min \mathrm{Sp}(G_m) \geq f_0/2$. Moreover, on Ω, we have $f_0 \geq 2\hat{f}_0/3$ and $\max(\hat{f}_0/3, n^{-1/2}) = \hat{f}_0$, for $n \geq 36/f_0^2$. □

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York.

Baraud, Y., and Birgé, L. (2008). Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Relat. Fields*, to appear.

Baraud, Y., Comte, F., and Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29(3):839-875.

Baraud, Y., Comte, F., and Viennet, G. (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.* 5, 33-49.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113(3):301-413.

Beran, J. (1981). Nonparametric regression with randomly censored survival data. Technical report, Dept. Statist. Univ. California, Berkeley.

Birgé, L., and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 4(3):329-375.

Brunel, E., Comte, F., and C. Lacour (2007). Adaptive estimation of the conditional density in presence of censoring. *Sankhya* 69(4):734-763.

Comte, F. (2001). Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7(2):267-298.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34, 187-220.

Dabrowska, D. M. (1987). Nonpametric regression with censored survival time data. *Scand. J. Statist.*, 14(3):181-197.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann. Statist. 17(3):1157-1167.

Grégoire, G. (1993). Least squares cross-validation for counting processes intensities. *Scand. J. Statist.*, 20(4):343-360.

Heuchenne, C., and Van Keilegom, I. (2007). Location estimation in nonparametric regression with censored data. *J. Multiv. Anal.*, 98(8):1558-1582.

Hochmuth, R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179-208.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, 27(5):1536-1563.

Jacobsen, M. (1982). *Statistical analysis of counting processes.* Lecture Note in Statistics 12. Springer-Verlag, New York.

Karr, A.F. (1986). *Point processes and their statistical inference.* Probability: Pure and Applied. Marcel Dekker Inc. New York.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics, 129. Springer-Verlag, New York.

Lacour, C. (2007). Adaptive estimation of the transition density of a markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571-597.

Li, G., and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, 23(3):787-823.

Linton, O. B., Nielsen, J. P., and Van de Geer, S. (2003). Estimating the multiplicative and additive hazard fonctions by kernel methods. *Ann. Statist.*, 31(2):464-492.

Liptser, R. S., and Shiryayev, A. N. (1989). *Theory of martingales*, vol. 49 of *Mathematics and its Applications (Soviet Series).* Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze].

Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics.* Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

McKeague, I. W., and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18(3):1172-1187.

Nikol'skii, S. M. (1975). *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11(2):453-466.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of nonhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields*, 126(1):103-153.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4): 633-661.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.*, 14(2): 638-647.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23(4): 461-471.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505-563.

Triebel, H. (2006). *Theory of function spaces. III.* Monographs in Mathematics, 100. Birkhäuser Verlag, Basel, 2006.

Tsybakov, A. (2003a). *Introduction a l'estimation non-paramétrique.* Springer.

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, 23(5):1779–1801.