

# Adaptive estimation of the conditional intensity of marker-dependent counting processes

F. Comte<sup>a</sup>, S. Gaïffas<sup>b</sup> and A. Guillaux<sup>b</sup>

<sup>a</sup>MAP5, UMR 8145, University Paris Descartes, Paris, France. E-mail: [fabienne.comte@parisdescartes.fr](mailto:fabienne.comte@parisdescartes.fr)

<sup>b</sup>LSTA, University Pierre et Marie Curie, Paris, France. E-mail: [stephane.gaiffas@upmc.fr](mailto:stephane.gaiffas@upmc.fr); [agathe.guillaux@upmc.fr](mailto:agathe.guillaux@upmc.fr)

Received 22 October 2008; revised 12 July 2010; accepted 23 July 2010

**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 nonasymptotic 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.

**Résumé.** Dans ce travail, nous proposons un estimateur original de l'intensité conditionnelle d'un processus de comptage marqué, c'est-à-dire d'un processus de comptage dépendant de covariables. Nous utilisons une méthode de sélection de modèle et nous obtenons pour notre estimateur, une borne non asymptotique du risque quadratique sur un compact. Nous vérifions ensuite que l'estimateur atteint automatiquement une vitesse de convergence sur des classes fonctionnelles de régularité anisotrope fixée mais inconnue. Enfin, nous démontrons une borne inférieure qui garantit l'optimalité de la vitesse obtenue. Une brève illustration de la façon dont fonctionne l'estimateur dans le contexte de l'estimation du taux de risque instantané conditionnel est fournie pour conclure.

MSC: 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 [1] 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  $\Lambda$  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

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

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  $\alpha$  is an unknown deterministic function called intensity.

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

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.) with cumulative distribution function (c.d.f.)  $F_T$ , and  $X$  a vector of covariates in  $\mathbb{R}^d$ . 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

(C):  $T$  and  $C$  are independent conditionally to  $X$ .

In this case, the processes to consider (see, e.g., [1]) are given, for  $i = 1, \dots, n$  and  $z \geq 0$ , by

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

The unknown intensity function  $\alpha$  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  $T$  given  $X$ .

Nonparametric estimation of the hazard rate in presence of covariates was initiated by Beran [5]. Stute [37], Dabrowska [13], McKeague and Utikal [30] and Li and Doss [26] extended his results. Many authors have considered semiparametric estimation of the hazard rate, beginning with [12], see [1] for a review of the enormous literature on semiparametric models. We refer to [20,27] for some recent developments.

Adaptive nonparametric estimation for censored data in presence of covariates has been considered by LeBlanc and Crowley [25] or Castellan and Letué [8] for particular functional Cox models: in these works,  $\alpha(x, z) = \exp(f(x))\alpha_0(z)$ , only  $f$  is estimated. On the other hand, Brunel et al. [7] constructed an optimal adaptive estimator of the conditional density in a general model.

**Example 2 (Cox processes).** Let  $\eta^i$ , for  $i = 1, \dots, n$ , be a Cox process (see [22]) on  $\mathbb{R}_+$  with random mean-measure  $\Lambda^i$  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 Eq. (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 [1]. The nonparametric estimation of the intensity of Poisson processes without covariates has been considered in several papers. We refer to [3,34] 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  $P^1, \dots, P^n$  with finite state space  $\{1, \dots, k\}$  and denote by  $\alpha_{jl}$  the transition intensity from state  $j$  to state  $l$ . For individual  $i$  with covariate  $X_i$ , let  $N_{jl}^i(t)$  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  $N_{jl}^i$  verifies the following Aalen multiplicative intensity model:

$$N_{jl}^i(t) = \int_0^t \alpha_{jl}(X_i, z) Y_j^i(z) dz + M^i(t) \quad \text{for all } t \geq 0,$$

where  $Y_j^i(t) = \mathbb{1}\{P^i(t-) = j\}$  for all  $t \geq 0$ , see [1] or [21]. This setting is discussed in [1], 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 [33] proposed a kernel-type estimator, Grégoire [17] studied cross-validation for these estimators. More recently, Reynaud-Bouret [35] 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 oracle inequalities for our estimator (see Theorems 1 and 2), a resulting upper bound (see Corollary 1) and a lower bound (see Theorem 3), the latter asserts the optimality in the minimax sense. 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 and 6. Some technical proofs are to be found in a longer version of this paper, see [11].

**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  $\alpha$ . Think for instance of the very popular case of proportional hazards Cox model, see [12], it is assumed that  $\alpha(x, z) = \alpha_0(z) \exp(\beta^\top x)$  for some unknown function  $\alpha_0$  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.

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

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

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

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

so that

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

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

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

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

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

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

**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:

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

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

$$d\mu(x, z) = (1 - L_{Z|X}(z, x))F_X(dx) dz, \tag{3}$$

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 [14] or, more recently [18]. Assumption (C) above is weaker than the assumption:  $T$  and  $C$  are independent and  $\mathbb{P}(T \leq C|X, T) = \mathbb{P}(T \leq C|T)$  in [38]. See [16], p. 249, for further discussions on this matter.

## 2.2. Assumptions and notations

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

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

and assume that the following condition holds:

$$(A1) \quad \text{The covariates } X_i \text{ admit a p.d.f. } f_X \text{ such that } \sup_{A_1} |f_X| \leq f_1 < +\infty.$$

Assumption (A1) implies that  $\mu$  admits a density w.r.t. the Lebesgue measure. We denote by  $f$  this density:

$$d\mu(x, z) = f(x, z) dx dz, \quad \text{where } f(x, z) = \mathbb{E}(Y(z)|X = x)f_X(x). \tag{4}$$

We also assume:

$$(A2) \quad \text{There exists } f_0 > 0, \text{ such that } \forall (x, z) \in A_1 \times [0, \tau], f(x, z) \geq f_0.$$

$$(A3) \quad \forall (x, z) \in A_1 \times [0, \tau], \alpha(x, z) \leq \|\alpha\|_{\infty, A} < +\infty.$$

$$(A4) \quad \forall i, \forall t, Y^i(t) \leq C_Y \text{ where } C_Y \text{ is a known fixed constant.}$$

**Remark 3.** Assumption (A2) is fulfilled if  $Y$  is bounded from below in expectation and if  $f_X$  is bounded from below. The requirement that the density of the design is bounded away from zero is standard in regression models, in particular. Assumption (A2) reduces to such a condition in Example 2 (Cox processes), where we have  $f(z, x) = I(z \in [0, \tau])f_X(x)$ . In the general setting of counting processes, a lower bound on the expectation of  $Y$  is classical, see [35], p. 648. In the censored case (Example 1), we can write:

$$\mathbb{E}(Y(z)|X = x) = \mathbb{E}(\mathbb{1}(T \wedge C \geq z)|X = x) = (1 - F_{T|X}(x, z))(1 - G_{C|X}(x, z)).$$

It is a well-known fact (see, e.g., [1], p. 193–194) that the Kaplan–Meier estimator is consistent, for each  $x$  (with no further assumption) only on intervals of the form  $[0, \tau_x]$ , where  $\tau_x < \sup\{s \geq 0, (1 - F_{T|X}(x, s))(1 - G_{C|X}(x, s)) > 0\}$ . We can take  $\tau = \inf_{x \in [0, 1]} \tau_x$ . In view of (3), this justifies our assumption (A2) in this case.

Lastly, in the examples described in Section 1, assumption (A4) is clearly fulfilled with  $C_Y = 1$ . We will set  $C_Y = 1$  in the following for simplicity. This implies together with (A1) that  $\forall(x, z) \in A, |f(x, z)| \leq f_1$ .

### 2.3. Definition of the estimator

We follow the usual model selection paradigm (see, e.g., [29]): 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  $A = [0, 1] \times [0, \tau]$  is defined by

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

where  $F_{m_1}$  and  $H_{m_2}$  are subspaces of  $(L^2 \cap L^\infty)(A_1)$  and  $(L^2 \cap L^\infty)([0, \tau])$ , 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  $\varphi_j^m$  and  $\psi_k^m$  are, respectively, included in  $A_1$  and  $[0, \tau]$ . Here  $j$  and  $k$  are not necessarily integers, they can be pairs of integers, as in the piecewise polynomial or the wavelet cases, see Section 2.4.

**Remark 4.** From a theoretical point of view, we could consider that the covariates  $X$  are in  $\mathbb{R}^d$ . For this end, we would have to consider models of the form  $S_m = F_{m_1} \otimes \dots \otimes F_{m_d} \otimes H_{m_{d+1}}$ . However, this would make the proofs more intricate. Note also that the convergence rate would be slower because of the curse of dimensionality. For the sake of clarity, we restrict ourselves to  $X \in \mathbb{R}$ .

The first step would be to define  $\hat{\alpha}_m = \arg \min_{h \in S_m} \gamma_n(h)$ . 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:

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

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

$$G_m := \left( \frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \varphi_l^m(X_i) \int_0^\tau \psi_k^m(z) \psi_p^m(z) Y^i(z) dz \right)_{(j,k), (l,p) \in J_m \times K_m}$$

and

$$\Upsilon_m := \left( \frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \int_0^\tau \psi_k^m(z) dN^i(z) \right)_{j \in J_m, k \in K_m}.$$

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

$$\hat{\alpha}_m := \begin{cases} \arg \min_{h \in S_m} \gamma_n(h) & \text{on } \hat{\Gamma}_m, \\ 0 & \text{on } \hat{\Gamma}_m^c, \end{cases} \tag{5}$$

where

$$\hat{\Gamma}_m := \{ \min \text{Sp}(G_m) \geq \max(\hat{f}_0/3, n^{-1/2}) \},$$

where  $\text{Sp}(G_m)$  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 (A2)) is required to fulfill the following assumption:

(A5) For any integer  $k \geq 1$ , there are positive constants  $C_0$  and  $n_0$  such that

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_0/n^k \quad \text{for any } n \geq n_0.$$

An estimator satisfying (A5) is defined in Section 3.5, where the constants  $C_0$  and  $n_0$  depend on  $k$  above, on  $f_0, f_1, \tau$  defined in Sections 2.1 and 2.2, and on  $\phi_1, \phi_2$  defined below in Section 2.4. In fact,  $k = 7$  is enough for the proofs. We refer the reader to the proof of Lemma 1, see Section 6, 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:

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} (\gamma_n(\hat{\alpha}_m) + \text{pen}(m)), \tag{6}$$

where  $\text{pen}(m)$  is defined in Theorem 1 below, see Section 3. Our estimator of  $\alpha$  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:

- (M1) For  $i = 1, 2$ ,  $\mathcal{D}_n^{(i)} := \max_{m \in \mathcal{M}_n} D_{m_i} \leq n^{1/4}/\sqrt{\log n}$ . We shall denote by  $\mathcal{F}_n$  (resp.  $\mathcal{H}_n$ ) the space with dimension  $\mathcal{D}_n^{(1)}$  (resp.  $\mathcal{D}_n^{(2)}$ ).
- (M2) There exist  $\phi_1 > 0, \phi_2 > 0$  such that, for all  $u$  in  $F_{m_1}$  and for all  $v$  in  $H_{m_2}$ , we have

$$\sup_{x \in A_1} |u(x)|^2 \leq \phi_1 D_{m_1} \int_{A_1} u^2 \quad \text{and} \quad \sup_{x \in [0, \tau]} |v(x)|^2 \leq \phi_2 D_{m_2} \int_{[0, \tau]} v^2.$$

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

$$\forall h \in S_m \quad \|h\|_{\infty, A} \leq \phi_0 \sqrt{D_{m_1} D_{m_2}} \|h\|_A. \tag{7}$$

- (M3) Nesting condition:

$$D_{m_1} \leq D_{m'_1} \Rightarrow F_{m_1} \subset F_{m'_1} \quad \text{and} \quad D_{m_2} \leq D_{m'_2} \Rightarrow H_{m_2} \subset H_{m'_2}.$$

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

**Remark 5.** We emphasize that  $\phi_2$  depends on  $\tau$  and is in most examples proportional to  $1/\tau$ .

Assumptions (M1)–(M3) 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 [4]):

- [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] Wavelet basis on an interval:  $\text{span}(\Psi_{j,k}: j = l - 1, \dots, m_1, k \in \Lambda(j))$ , where  $l$  and  $m_1$  are integers ( $l$  corresponds to the number of vanishing moments of the basis). The  $\Psi_{j,k}$  are, depending on the localization parameter  $k$ , either translations and dilatations of a pair  $\{\phi, \psi\}$  of scaling function and wavelet with a compact support, or so-called edge scaling functions and wavelets. We give more details in Appendix A.1. By construction, the elements of this basis have their supports included in  $A_1$ , and they have as many vanishing moments as  $\psi$ .
- [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].

Notice that  $\tau^{-1/2}\varphi_1(\cdot/\tau), \dots, \tau^{-1/2}\varphi_D(\cdot/\tau)$  an orthonormal basis in  $L^2([0, \tau])$ , whenever  $\varphi_1, \dots, \varphi_D$  is one in  $L^2([0, 1])$ .

**Remark 6.** The first assumption (M1) prevents the dimension from being too large compared to the number of observations. We can relax considerably this constraint for localized basis: for histogram basis, piecewise polynomial basis and wavelets, (M1) can be relaxed to the weaker condition:  $\mathcal{D}_n^{(i)} \leq \sqrt{n/\log n}$ . Analogously in (M3), we would get  $N_n \leq n/\log n$ . The condition (M2) implies a useful link between the  $L^2$  norm and the infinite norm. The third assumption (M3) implies in particular that  $\forall m, m' \in \mathcal{M}_n, S_m + S_{m'} \subset S_n$ . This condition is useful for the chaining argument used in the proofs, see Section 6.4.

### 3. Main results

#### 3.1. Oracle inequality

We define  $\alpha_m$  as the orthogonal projection of  $\alpha \mathbb{1}(A)$  on  $S_m$ . The estimator  $\hat{\alpha}_{\hat{m}}$ , where  $\hat{\alpha}_m$  and  $\hat{m}$  are given by (5) and (6), respectively, satisfies the following oracle inequality.

**Theorem 1.** Let (A1)–(A5) and (M1)–(M3) hold. Define the following penalty:

$$\text{pen}(m) := K_0(1 + \|\alpha\|_{\infty, A}) \frac{D_{m_1} D_{m_2}}{n}, \tag{8}$$

where  $K_0$  is a numerical constant. We have

$$\mathbb{E}(\|\alpha \mathbb{1}(A) - \hat{\alpha}_{\hat{m}}\|_{\mu}^2) \leq \kappa_0 \inf_{m \in \mathcal{M}_n} \{ \|\alpha \mathbb{1}(A) - \alpha_m\|_{\mu}^2 + \text{pen}(m) \} + \frac{C}{n} \tag{9}$$

for any  $n \geq n_0$ , where  $n_0$  is a constant coming from assumption (A5) (see Section 2.3), where  $\kappa_0$  is a numerical constant and  $C$  is a constant depending on  $\phi_1, \phi_2, \|\alpha\|_{\infty, A}, f_0, f_1$  and  $\tau$ .

The proof of Theorem 1 involves a deviation inequality (see Lemma 5) for the empirical process

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

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 (see Proposition 4 and the detailed proof in [11]).

#### 3.2. Adaptive upper bound

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  $\alpha$  restricted to  $A$  belongs to the anisotropic Besov space  $B_{2, \infty}^{\beta}(A)$  on  $A$  with regularity  $\beta = (\beta_1, \beta_2)$ . Let us recall the definition of  $B_{2, \infty}^{\beta}(A)$ . Let

$\{e_1, e_2\}$  the canonical basis of  $\mathbb{R}^2$  and take  $A_{h,i}^r := \{x \in \mathbb{R}^2; x, x + he_i, \dots, x + rhe_i \in A\}$ , for  $i = 1, 2$ . For  $x \in A_{h,i}^r$ , let

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

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

$$\omega_{r_i,i}(g, t) = \sup_{|h| \leq t} \left( \int_{A_{h,i}^{r_i}} |\Delta_{h,i}^{r_i} g(x)|^2 dx \right)^{1/2}, \quad i = 1, 2.$$

Consider the Besov norm

$$\|\alpha\|_{B_{2,\infty}^\beta(A)} := \|\alpha\|_A + |\alpha|_{B_{2,\infty}^\beta(A)} = \|\alpha\|_A + \sup_{t>0} \sum_{i=1}^2 t^{-\beta_i} \omega_{r_i,i}(g, t). \quad (10)$$

Define the Besov space  $B_{2,\infty}^\beta(A)$  as the set of functions  $g$  such that  $\|g\|_{B_{2,\infty}^\beta(A)} < +\infty$ , and for  $L > 0$ , consider the ball

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

More details concerning Besov spaces can be found in [40]. The next corollary shows that  $\hat{\alpha}_{\hat{m}}$  adapts to the unknown anisotropic smoothness of  $\alpha$ .

**Corollary 1.** *Assume that  $\alpha$  restricted to  $A$  belongs to  $B_{2,\infty}^\beta(A, L)$ , with smoothness  $\beta = (\beta_1, \beta_2)$  such that  $\beta_1 > 1/2$  and  $\beta_2 > 1/2$ . We consider the piecewise polynomial or wavelet spaces described in Section 2.4 (with the regularity of the polynomials and the wavelets larger than  $\beta_i - 1$ ). Then, under the assumptions of Theorem 1, we have*

$$\mathbb{E} \|\alpha - \hat{\alpha}_{\hat{m}}\|_A^2 \leq C n^{-2\bar{\beta}/(2\bar{\beta}+2)},$$

where  $\bar{\beta}$  is the harmonic mean of  $\beta_1$  and  $\beta_2$  (i.e.,  $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$ ) and  $C$  depends on  $L, \tau, \phi_0, f_0, f_1$  and  $\|\alpha\|_{\infty,A}$ .

The rate of convergence achieved by  $\hat{\alpha}_{\hat{m}}$  in Corollary 1 is optimal in the minimax sense as proved in Theorem 3 below. For trigonometric spaces, the result also holds, but for  $\beta_1 > 3/2$  and  $\beta_2 > 3/2$  (because of  $(M1)$ ).

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. Random penalty

It is worth noting that the penalty defined in Eq. (8) involves the unknown quantity  $\|\alpha\|_{\infty,A}$ . This problem occurs occasionally in penalization procedures, see, for instance, [10] or [23]. The solution is to replace it by an estimator:

$$\widehat{\text{pen}}(m) = K_1 \left( 1 + \|\hat{\alpha}_{m^*}\|_{A,\infty} \right) \frac{D_{m_1} D_{m_2}}{n}, \quad (11)$$

where  $K_1$  is a numerical constant and  $\hat{\alpha}_{m^*}$  is a rough estimator of  $\alpha$  computed on an arbitrary space  $S_{m^*}$  with dimension  $D_{m^*} = D_{m_1^*} D_{m_2^*}$ . Let us consider

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} (\gamma_n(\hat{\alpha}_m) + \widehat{\text{pen}}(m)). \quad (12)$$

Then we can prove the following result.

**Theorem 2.** *Let the assumptions of Theorem 1 be satisfied. Consider the estimator  $\hat{\alpha}_{\hat{m}}$  defined by (5)–(12)–(11), where the term  $\hat{\alpha}_{m^*}$  is computed with (5) on a space  $S_{m^*}$  in collection  $[T]$  with dimension  $D_{m^*}$  such that*

$$D_{m_1^*} = D_{m_2^*} = n^{1/4}.$$

*If  $\alpha$  restricted to  $A$  belongs to the anisotropic Besov space  $B_{2,\infty}^\beta(A)$  with regularity  $\beta = (\beta_1, \beta_2)$  such that  $\beta_1 > 2$  and  $\beta_2 > 2$ , then, for  $n$  large enough,*

$$\mathbb{E}(\|\alpha \mathbb{1}(A) - \hat{\alpha}_{\hat{m}}\|_\mu^2) \leq \kappa_1 \inf_{m \in \mathcal{M}_n} \left\{ \|\alpha \mathbb{1}(A) - \alpha_m\|_\mu^2 + (1 + \|\alpha\|_{\infty,A}) \frac{D_{m_1} D_{m_2}}{n} \right\} + \frac{C}{n}, \quad (13)$$

where  $\kappa_1$  is a numerical constant and  $C$  is a constant depending on  $\phi_1, \phi_2, \|\alpha\|_{\infty,A}, f_0, f_1$  and  $\tau$ .

The rate of convergence of  $\hat{\alpha}_{\hat{m}}$  on Besov balls can be deduced, in an obvious manner, from Theorem 2, as Corollary 1 was obtained from Theorem 1.

### 3.4. Lower bound

In the next theorem, we prove that the rate  $n^{-2\bar{\beta}/(2\bar{\beta}+2)}$  is optimal over  $B_{2,\infty}^\beta(A)$  where we recall that  $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$ . The Besov ball  $B_{2,\infty}^\beta(A, L)$  is defined in Section 3.2. Let us denote by  $\mathbb{E}_\alpha$  the integration w.r.t. the joint law  $\mathbb{P}_\alpha$ , when the intensity is  $\alpha$ , of the  $n$ -sample  $(X_i, N^i(z), Y^i(z); z \leq \tau, i = 1, \dots, n)$ .

**Theorem 3.** *Assume that assumption (A1) holds. Then there is a constant  $C > 0$  that depends on  $\beta, L, \tau$  and  $f_1$  such that*

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

for  $n$  large enough, where the infimum is taken among all estimators.

**Remark 7.** *There is a slight difference between the statements of Theorem 3 and Corollary 1: the upper bound in Corollary 1 requires assumption (A2) to be fulfilled (which requires that  $f(x, z) = \mathbb{E}(Y(z)|X = x) f_X(x) \geq f_0$ ) while Theorem 3 does not. However Corollary 1 and Theorem 3 are stated on the same functional sets. This kind of difference between the statements of upper and lower bounds is classical, and can be found in regression models as well, see the discussion in [36], p. 1351, for a regression model.*

### 3.5. Estimation of $f_0$

We recall that  $f$  is the density of  $\mu$ , which is defined in Eq. (4). We define

$$\hat{f}_m = \arg \min_{h \in \mathcal{S}_m} v_n(h), \quad \text{where } v_n(h) = \|h\|^2 - \frac{2}{n} \sum_{i=1}^n \int_0^\tau h(X_i, z) Y^i(z) dz. \quad (14)$$

This estimator admits a simple explicit formulation:

$$\hat{f}_m(x, z) = \sum_{(j,k) \in J_m \times K_m} \hat{b}_{j,k} \varphi_j^m(x) \psi_k^m(y), \quad \text{with } \hat{b}_{j,k} = \frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \int_0^\tau \psi_k^m(z) Y^i(z) dz. \quad (15)$$

As before, we consider estimation of  $f$  over the compact set  $A = [0, 1] \times [0, \tau]$ . 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  $\mathcal{D}_n^{(2)} = \dim(\mathcal{H}_n)$  its dimension (see (M1)) and by  $\ell_n$  its index so that  $H_{\ell_n} = \mathcal{H}_n$ . Consider, for a given  $m_1$ , the estimator

$$\hat{f}_{m_1} := \arg \min_{h \in F_{m_1} \times \mathcal{H}_n} v_n(h)$$

and define an estimator of  $f_0$  by considering any  $\inf_{(x,z) \in A} \hat{f}_{m_1}(x, z)$ . It is worth noticing that an arbitrary choice is sufficient, because only a rough estimation of the lower bound  $f_0$  of  $f$  is useful here. Therefore, the estimator  $\hat{f}_0$  used in (5) for the construction of  $\hat{\alpha}_m$  can be defined, for an arbitrary  $m_1^*$ , see Proposition 1, as

$$\hat{f}_0 := \inf_{(x,z) \in A} \hat{f}_{m_1^*}(x, z) \quad \text{with } D_{m_1^*} = \dim(F_{m_1^*}). \tag{16}$$

**Proposition 1.** Consider  $\hat{f}_0$  defined by (16) in the basis [T] with  $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log n}$ . Assume that  $f \in \mathcal{B}_{2,\infty}^{(\tilde{\beta}_1, \tilde{\beta}_2)}(A)$  with  $\tilde{\beta}_1 > 2, \tilde{\beta}_2 > 2$ . Then, for any  $k \in \mathbb{N}$ , there are positive constants  $n_0$  and  $C_0$  such that

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_0/n^k$$

for any  $n \geq n_0$ , where  $C_0$  and  $n_0$  are constant depending on  $k, \tau, f_0, f_1, \phi_1$  and  $\phi_2$ . This proves that  $\hat{f}_0$  fulfills assumption (A5).

The proof of this result is given in Section 6.

#### 4. Illustration

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  $(X_1, T_1), \dots, (X_n, T_n)$  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  $\varepsilon_i$  with a  $\chi^2(4)$  distribution,  $\sigma = 1/4$  and  $b(x) = 2x + 5$ . Note that in this case, the hazard function to be estimated is

$$\alpha_{\text{NL}}(x, t) = \frac{1}{\sigma} \alpha_\varepsilon \left( \frac{t - b(x)}{\sigma} \right),$$

where  $\alpha_\varepsilon$  denotes the hazard function of  $\varepsilon$ .

*Case (AFT). Accelerated Failure Time model:*

$$\log(T_i) = a + bX_i + \varepsilon_i,$$

where the  $\varepsilon_i$  are standard normal and  $a = 5$  and  $b = 2$ . The hazard function to be estimated is then:

$$\alpha_{\text{AFT}}(x, t) = \frac{\alpha_\varepsilon(\log(t) - (a + bx))}{t}.$$

*Case (PH). Proportional Hazards model (see [8,25]):* in this case, the hazard writes

$$\alpha(x, t) = \exp(b(x))\alpha_0(t).$$

We take  $b(x) = bx$  with  $b = 0.4$  and  $\alpha_0(t) = a\lambda t^{a-1}$ , which is a Weibull hazard function with  $a = 3$  and  $\lambda = 1$ .

We choose to compute and plot our estimators with histogram bases for two reasons: first, it makes the estimator much easier to compute; secondly, it shows very well how the changes are captured, and when an anisotropic choice is performed by the estimation procedure. More sophisticated implementation is beyond the scope of the paper.

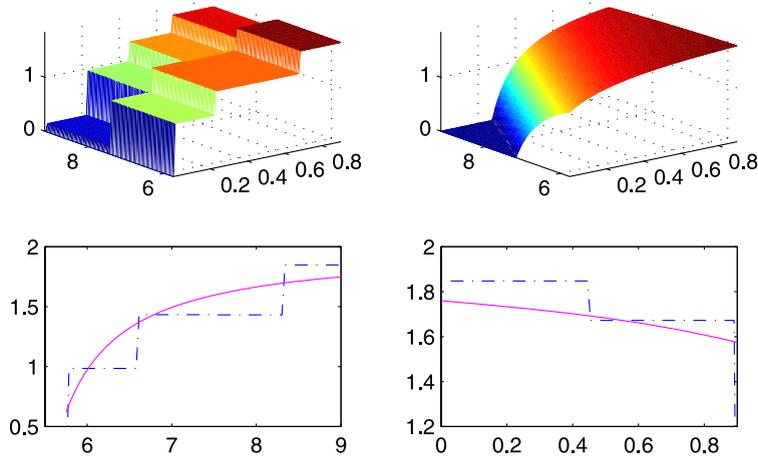


Fig. 1. Case (NL) Estimated (top left) and true (top right) conditional hazard rates and example of cross-sections (bottom) for a fixed value of  $x$  (left) and  $y$  (right).

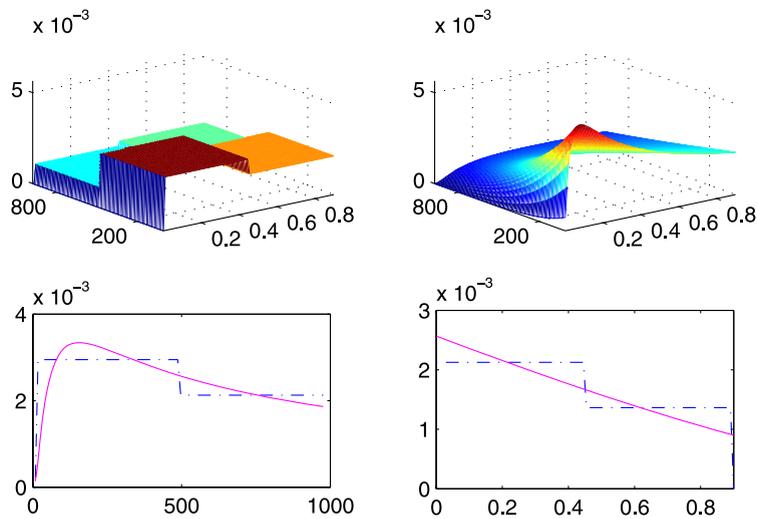


Fig. 2. Case (AFT) Estimated (top left) and true (top right) conditional hazard rates and example of cross-sections (bottom) for a fixed value of  $x$  (left) and  $y$  (right).

The penalty is taken as

$$\widehat{\text{pen}}(m_1, m_2) = \kappa (1 + \|\hat{\alpha}\|_{\infty, A}) \frac{2^{m_1+m_2}}{n},$$

with  $\kappa = 4$ . Note that, for sake of simplicity,  $\|\hat{\alpha}\|_{\infty, A}$  is estimated by  $\max_{j,k} \hat{a}_{j,k}$  (the largest histogram coefficients) instead of the trigonometric basis, which was used for technical reasons in Theorem 2: this is because it makes the procedure faster, since all  $\hat{a}_{j,k}$  are already computed for estimation. These coefficients are computed on the largest space considered (with dimension  $\sqrt{n}$ ).

We can see from Figs 1–3 that the algorithm exploits the opportunity (Figs 1 and 3) of choosing different dimensions in the two directions, and that it gives a good account of the general form of the surfaces.

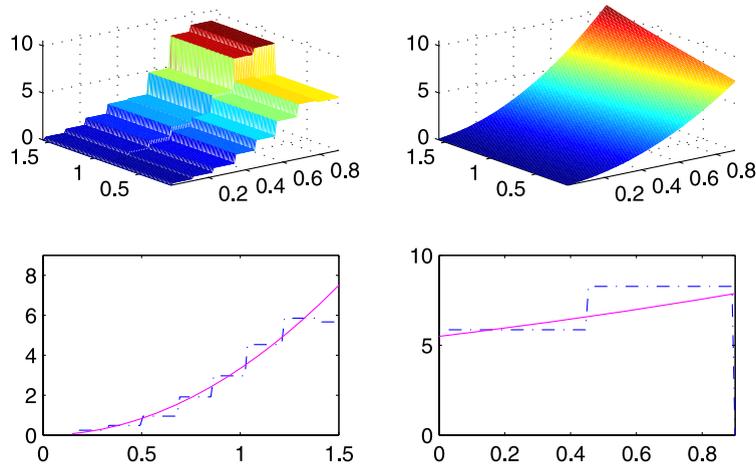


Fig. 3. Case (PH) Estimated (top left) and true (top right) conditional hazard rates and example of cross-sections (bottom) for a fixed value of  $x$  (left) and  $y$  (right).

### 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

$$\langle h_1, h_2 \rangle_n = \frac{1}{n} \sum_{i=1}^n \int_0^\tau h_1(X_i, z) h_2(X_i, z) Y^i(z) dz \mathbb{1}(X_i \in [0, 1]) \tag{17}$$

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

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

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

$$\begin{aligned} \hat{\Gamma}_m &= \{ \min \text{Sp}(G_m) \geq \max(\hat{f}_0/3, n^{-1/2}) \}, & \hat{\Gamma} &:= \bigcap_{m \in \mathcal{M}_n} \hat{\Gamma}_m, \\ \Delta &:= \left\{ \forall h \in \mathcal{S}_n: \left| \frac{\|h\|_n^2}{\|h\|_\mu^2} - 1 \right| \leq \frac{1}{2} \right\} & \text{and } \Omega &:= \left\{ \left| \frac{\hat{f}_0}{f_0} - 1 \right| \leq \frac{1}{2} \right\}. \end{aligned} \tag{18}$$

For  $m \in \mathcal{M}_n$ , we denote by  $\alpha_m$  the orthogonal projection on  $S_m$  of  $\alpha$  restricted to  $A$ . The following decomposition holds

$$\begin{aligned} \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha \mathbb{1}(A)\|_\mu^2) &\leq 2\|\alpha \mathbb{1}(A) - \alpha_m\|_\mu^2 + 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega)) \\ &\quad + 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta \cap \Omega)^c)). \end{aligned} \tag{19}$$

The last term is bounded via the following proposition.

**Proposition 2.** *Under the assumptions of Theorem 1,*

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta \cap \Omega)^c)) \leq C_1/n, \tag{20}$$

where  $C_1$  is a constant depending on  $\tau, \phi_1, \phi_2, \|\alpha\|_{\infty, A}, f_0, f_1$ .

The study of the term  $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega))$  involves two preliminary steps. The first one is the following lemma.

**Lemma 1.** *Under the assumptions of Theorem 1, the following embedding holds: for  $n \geq 4/f_0^2$ , we have*

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

See the proof in Section 6.3. As a consequence, for all  $m \in \mathcal{M}_n$ , the matrices  $G_m$  are invertible on  $\Delta \cap \Omega$ . The second step is the following useful decomposition. Let us define

$$\begin{aligned} v_n(h) &= \frac{1}{n} \sum_{i=1}^n \left( \int_0^\tau h(X_i, z) dN^i(z) - \int_0^\tau h(X_i, z) \alpha(X_i, z) Y^i(z) dz \right) \\ &= \frac{1}{n} \sum_{i=1}^n \int_0^\tau h(X_i, z) dM^i(z), \end{aligned} \tag{21}$$

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

$$\begin{aligned} \gamma_n(h_1) - \gamma_n(h_2) &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 \rangle_n - \frac{2}{n} \sum_{i=1}^n \int_0^\tau (h_1 - h_2)(X_i, z) dN^i(z) \\ &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 - \alpha \rangle_n - 2v_n(h_1 - h_2) \\ &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 - \alpha \mathbb{1}(A) \rangle_n - 2v_n(h_1 - h_2), \end{aligned} \tag{22}$$

where the indicator  $\mathbb{1}(A)$  can be inserted because all other functions in the scalar product are  $A$ -supported. Let us assume that  $n \geq 4/f_0^2$ . Now, on  $\Delta \cap \Omega$ , we have thanks to Lemma 1 and by the definition of  $\hat{m}$ , that

$$\gamma_n(\hat{\alpha}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(\alpha_m) + \text{pen}(m) \quad \forall m \in \mathcal{M}_n.$$

It follows from (22) and from the fact that  $2xy \leq x^2/\theta + \theta y^2$  for any  $x, y, \theta > 0$  that, on  $\Delta \cap \Omega$ ,

$$\begin{aligned} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 &\leq 2\langle \hat{\alpha}_{\hat{m}} - \alpha_m, \alpha \mathbb{1}(A) - \alpha_m \rangle_n + \text{pen}(m) + 2v_n(\hat{\alpha}_{\hat{m}} - \alpha_m) - \text{pen}(\hat{m}) \\ &\leq \frac{1}{4} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 + 4\|\alpha \mathbb{1}(A) - \alpha_m\|_n^2 + \text{pen}(m) \\ &\quad + \frac{1}{4} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4 \sup_{h \in B_{m, \hat{m}}^\mu(0, 1)} v_n^2(h) - \text{pen}(\hat{m}), \end{aligned}$$

where  $B_{m, m'}^\mu(0, 1) := \{h \in S_m + S_{m'} : \|h\|_\mu \leq 1\}$ . Now, we need to introduce a centering factor denoted by  $p(m, m')$ , related to the supremum of the empirical process  $v_n(h)$ .

**Proposition 3.** *Grant the assumptions of Theorem 1. There exists a numerical constant  $\kappa > 0$  such that the following holds. If*

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

then

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E} \left( \sup_{h \in B_{m, m'}^\mu(0, 1)} (v_n^2(h) - p(m, m'))_+ \mathbb{1}(\Delta) \right) \leq \frac{C_2}{n}$$

for  $n$  large enough, where  $C_2$  is a constant depending on  $f_0, \|\alpha\|_{\infty, A}$  and the chosen basis (see Section 2.4).

The proof of Proposition 3 is partly given in Section 6.4 below (see [11] for details). It yields

$$\begin{aligned} \frac{3}{4} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 &\leq 4 \|\alpha \mathbb{1}(A) - \alpha_m\|_n^2 + \text{pen}(m) + \frac{1}{4} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \\ &\quad + 4 \left( \sup_{h \in B_{m, \hat{m}}^\mu(0,1)} v_n^2(h) - p(m, \hat{m}) \right)_+ + 4p(m, \hat{m}) - \text{pen}(\hat{m}). \end{aligned}$$

Now, let fix  $K_0 \geq 4\kappa$ , so that

$$4p(m, m') \leq \text{pen}(m) + \text{pen}(m') \quad \forall m, m',$$

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

$$\begin{aligned} \frac{3}{8} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 &\leq 4 \|\alpha \mathbb{1}(A) - \alpha_m\|_n^2 + 2 \text{pen}(m) \\ &\quad + \frac{1}{4} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4 \sum_{m' \in \mathcal{M}_n} \left( \sup_{h \in B_{m, m'}^\mu(0,1)} v_n^2(h) - p(m, m') \right)_+ \end{aligned} \quad (23)$$

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

$$\frac{1}{8} \|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \leq 4 \|\alpha \mathbb{1}(A) - \alpha_m\|_n^2 + 2 \text{pen}(m) + 4 \sum_{m' \in \mathcal{M}_n} \left( \sup_{h \in B_{m, m'}^\mu(0,1)} v_n^2(h) - p(m, m') \right)_+.$$

Now, Proposition 3 entails

$$\frac{1}{8} \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega)) \leq 4 \|\alpha \mathbb{1}(A) - \alpha_m\|_\mu^2 + 2 \text{pen}(m) + \frac{C_2}{n}. \quad (24)$$

Gathering (19), (20) and (24), we obtain that, for  $n \geq 4/f_0^2$ ,

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha \mathbb{1}(A)\|_\mu^2) \leq 2 \|\alpha_m - \alpha \mathbb{1}(A)\|_\mu^2 + 16 \left( 4 \|\alpha \mathbb{1}(A) - \alpha_m\|_\mu^2 + 2 \text{pen}(m) + \frac{C_2}{n} \right) + \frac{2C_1}{n}$$

for any  $m \in \mathcal{M}_n$ . On the other hand, if  $n \leq 4/f_0^2$ , then  $1/n \geq f_0^2/4$  and it is easy to see that Lemma 3 (see below) entails  $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha \mathbb{1}(A)\|_\mu^2) \leq C/n$  where  $C$  is a constant depending on  $C_B$  from Lemma 3,  $f_0$  and  $\|\alpha \mathbb{1}(A)\|_\mu^2$ . This concludes the proof of Theorem 1.

## 5.2. Proof of Corollary 1

To control the bias term, we use Lemma 6, see Appendix A.2, that gives the approximation result allowing to derive the rate of convergence. If we choose  $S_m$  as one of the finite-dimensional linear span considered in Section A.2, we can apply Lemma 6 to the function  $\alpha_A$ , the restriction of  $\alpha$  to  $A$ . Since  $\alpha_m$  has been defined as the orthogonal projection of  $\alpha_A$  on  $S_m$ , we get using (A1) and (A4):

$$\|\alpha \mathbb{1}(A) - \alpha_m\|_\mu \leq f_1 \|\alpha - \alpha_m\|_A \leq C_3 [D_{m_1}^{-\beta_1} + D_{m_2}^{-\beta_2}],$$

where  $C_3$  depends on the Besov norm of  $\alpha$  and on  $f_1$ . Now, according to Theorem 1 and (A2), we obtain

$$\mathbb{E} \|\hat{\alpha}_{\hat{m}} - \alpha\|_A^2 \leq f_0^{-1} \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\|_\mu^2) \leq C_4 \inf_{m \in \mathcal{M}_n} \left\{ D_{m_1}^{-2\beta_1} + D_{m_2}^{-2\beta_2} + \frac{D_{m_1} D_{m_2}}{n} \right\},$$

where  $C_4$  depends on the Besov norm of  $\alpha$ , and on  $f_0, f_1, \phi_1, \phi_2$  and  $\tau$ . In particular, if  $m^* = (m_1^*, m_2^*)$  is such that

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

then

$$\mathbb{E} \|\hat{\alpha}_{\hat{m}} - \alpha\|_A^2 \leq 2C_4 \left( D_{m_1^*}^{-2\beta_1} + \frac{D_{m_1^*}^{1+\beta_1/\beta_2}}{n} \right) \leq 4n^{-2\beta_1\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} = 4C_4 n^{-2\bar{\beta}/(2\bar{\beta}+2)},$$

where we recall that the harmonic mean of  $\beta_1$  and  $\beta_2$  is  $\bar{\beta} = 2\beta_1\beta_2/(\beta_1 + \beta_2)$ . The condition  $D_{m_1} \leq \sqrt{n}/\log n$  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 satisfied. Both conditions hold if  $\beta_1 > 1/2$  and  $\beta_2 > 1/2$ .

### 5.3. Proof of Theorem 2

The proof follows the line of the proof of Theorem 2.2 in [24], p. 67, so we only give a sketch of proof. Let us define

$$\Lambda = \left\{ \left| \frac{\|\hat{\alpha}_{m^*}\|_\infty}{\|\alpha\|_{\infty,A}} - 1 \right| < \frac{1}{2} \right\},$$

and recall that  $\Delta$  and  $\Omega$  are given by (18). Then we decompose the risk of  $\hat{\alpha}_{\hat{m}}$  as follows:

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2) = \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2 \mathbb{1}(\Lambda \cap \Delta \cap \Omega)) + \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2 \mathbb{1}((\Lambda \cap \Delta \cap \Omega)^c)).$$

The study of the term  $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2 \mathbb{1}(\Lambda \cap \Delta \cap \Omega))$  is very similar to the study of its analogous in the proof of Theorem 1, by using that, on  $\Lambda$ ,

$$\frac{1}{2} \text{pen}(m) \leq \frac{K_0}{K_1} \widehat{\text{pen}}(m) \leq \frac{3}{2} \text{pen}(m). \tag{25}$$

Thus, the algebra starts with  $\widehat{\text{pen}}(m)$  instead of  $\text{pen}(m)$ , and on  $\Lambda$ , it is proportional to  $\text{pen}(m)$  thanks to (25). At the end, only constant multiplicative factors are changed. In other words, taking  $K_1 = 2K_0$ , (24) is simply replaced by

$$\frac{1}{8} \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega \cap \Lambda)) \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 + 4 \text{pen}(m) + \frac{C_2}{n}. \tag{26}$$

The conclusion follows from the following lemma, which is proven in a longer version of the paper, see [11].

**Lemma 2.** *Under the assumptions of Theorem 2,*

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2 \mathbb{1}((\Lambda \cap \Delta \cap \Omega)^c)) \leq C_R/n,$$

where  $C_R$  depends on  $\phi_1, \phi_2, \tau, f_0, f_1$  and  $\|\alpha\|_{\infty,A}$ .

This ends the proof of Theorem 2.

### 5.4. Proof of Theorem 3

In order to prove Theorem 3, we use the following theorem from [41], which is a standard tool for the proof of such a lower bound. We say that  $\partial$  is a *semi-distance* on some set  $\Theta$  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 (see [41]).** *Let  $(\Theta, \partial)$  be a set endowed with a semi-distance  $\partial$ . 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)  $\partial(\theta_j, \theta_k) \geq 2v \forall 0 \leq j < k \leq M,$
- (2)  $P_{\theta_j} \ll P_{\theta_0} \forall 1 \leq j \leq M,$
- (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.

In this proof, we denote by  $\mathbb{P}_{\alpha}$  the distribution of  $(X, N(z), Y(z); z \leq \tau)$  when the intensity of  $N$  is  $\alpha$  and by  $\mathbb{P}_{\alpha}^n$  the distribution of the  $n$ -sample  $(X_i, N^i(z), Y^i(z); z \leq \tau, i = 1, \dots, n).$

We construct a family of functions  $\{\alpha_0, \dots, \alpha_M\}$  that satisfies points (1)–(3). We use the notation  $|A|$  for the area of the rectangle  $A$  (or the length of an interval) and  $\#(R)$  denotes the cardinality of a set  $R$ . Let  $\alpha_0(x, t) = |B|^{-1} \mathbb{1}(t \in B)$  where  $B$  is a compact set such that  $A = [0, 1] \times [0, \tau] \subset B \times B$  and  $|B| \geq 2|A|^{1/2}/L$ . As a consequence, we have  $\alpha_0(x, t) > 0$  for  $(x, t) \in A$  and  $\|\alpha_0\|_{B_{2,\infty}^{\beta}(A)} = \|\alpha_0\|_A + |\alpha_0|_{B_{2,\infty}^{\beta}(A)} \leq L/2$  since  $|\alpha_0|_{B_{2,\infty}^{\beta}(A)} = 0$ , see (10). We shall denote for short  $a_0 = |B|^{-1}$  in the following. Let  $\psi$  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

$$\psi_{j,k}(x, t) = \tau^{-1/2} 2^{(j_1+j_2)/2} \psi(2^{j_1} t/\tau - k_1) \psi(2^{j_2} x - k_2),$$

so that  $\|\psi_{j,k}\|_A = 1$ . Let  $S_{j,k}$  stands for the support of  $\psi_{j,k}$ . We consider the maximal set  $R_j \subset \mathbb{Z}^2$  such that

$$S_{j,k} \subset A \quad \forall k \in R_j \quad \text{and} \quad S_{j,k} \cap S_{j,k'} = \emptyset \quad \forall k, k' \in R_j, k \neq k'. \tag{27}$$

The cardinality of  $R_j$  satisfies  $\#(R_j) = c2^{j_1+j_2}$ , where  $c$  is a positive constant that depends on  $\tau$  and on the support of  $\psi$  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 (27) 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  $\Omega_j$ . Using a result of Varshamov–Gilbert – see [41] – we can find a subset  $\{\omega^{(0)}, \dots, \omega^{(M_j)}\}$  of  $\Omega_j$  such that

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

for any  $0 \leq p < q \leq M_j$ , where  $M_j \geq 2^{\#(R_j)/8}$ . 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  $0 \leq p < q \leq M_j$

$$\|\alpha_p - \alpha_q\|_A \geq \left( \frac{b\#(R_j)}{8n} \right)^{1/2} = 2v_j$$

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

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

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

$$\left(\frac{b2^{j_1+j_2}}{\tau n}\right)^{1/2} \leq \frac{a_0}{3\|\psi\|_{\infty}^2}. \tag{28}$$

Using the Bernstein’s estimate from [19] (see Theorem 3.5, p. 194), we have for  $\psi$  smooth enough that

$$\left\| \sum_{k \in R_j} \omega_k \psi_{j,k} \right\|_{B_{2,\infty}^{\beta}(A)} \leq c_{\tau} (2^{j_1\beta_1} + 2^{j_2\beta_2}) \left\| \sum_{k \in R_j} \omega_k \psi_{j,k} \right\|_A \leq c_{\tau, \psi} (2^{j_1\beta_1} + 2^{j_2\beta_2}) (2^{j_1+j_2})^{1/2},$$

where  $c_{\tau, \psi}$  is a constant that depends on  $\tau$  and  $\psi$ . Note that the Bernstein’s estimate from [19] is stated on the space  $\mathbb{L}^2([0, 1]^2)$  while we consider here  $\mathbb{L}^2([0, 1] \times [0, \tau])$ . An obvious (but tedious) modification of the proof of Hochmuth (it suffices to change the scaling of the moduli of continuity  $\omega_{r_i, i}$  herein) allows to show that the Bernstein’s estimate is the same as for  $\mathbb{L}^2([0, 1]^2)$ , up to a multiplicative constant that depends on  $\tau$ . Hence, if

$$\frac{c_{\tau, \psi} (2^{j_1\beta_1} + 2^{j_2\beta_2}) (2^{j_1+j_2})^{1/2}}{\sqrt{n}} \leq \frac{L}{2\sqrt{b}}, \tag{29}$$

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

Points (2) and (3) are derived using Jacod’s formula (see [1]). Indeed, we can prove that the log-likelihood  $\ell(\alpha, \alpha_0) := \log(d\mathbb{P}_{\alpha}/d\mathbb{P}_{\alpha_0})$  of  $N$  writes

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

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

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

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

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

which gives by integration with respect to  $\mathbb{P}_{\alpha}$

$$K(\mathbb{P}_{\alpha}, \mathbb{P}_{\alpha_0}) \leq \frac{3\|\alpha - \alpha_0\|_{\mu}^2}{2a_0} \leq \frac{3f_1\|\alpha - \alpha_0\|_A^2}{2a_0} \leq \frac{3bf_1\#(R_j)}{2na_0}$$

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

$$\frac{1}{M} \sum_{p=0}^M K(\mathbb{P}_{\alpha_p}^n, \mathbb{P}_{\alpha_0}^n) \leq \frac{3bf_1\#(R_j)}{2a_0} \leq a \log M_j$$

with

$$a := 12bf_1/(a_0 \log 2).$$

So, we take  $b$  small enough, so that  $a < 1/8$  (this is the only constraint on  $b$ ) and point (3) is met in Tsybakov’s [41] theorem. It only remains to choose the levels  $j_1$  and  $j_2$  so that (28) and (29) holds, and to compute the corresponding  $v_j$ . We take  $j = (j_1, j_2)$  such that

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

where  $c_1$  and  $c_2$  are positive constants satisfying  $c_{\tau,\psi}(c_1^{\beta_1} + c_2^{\beta_2})\sqrt{c_1c_2} \leq L/(2\sqrt{b})$ . For this choice,  $2^{j_1+j_2}/n \leq c_1c_2n^{-2\tilde{\beta}/(2\tilde{\beta}+2)}$  so (28) holds for  $n$  large enough and (29) holds and  $v_j \geq c_3n^{-\tilde{\beta}/(2\tilde{\beta}+2)}$  where  $c_3 = c_{\tau,\psi}\sqrt{bc_1c_2}/128$ .

### 6. Proof of the auxiliary results

#### 6.1. Proof of Proposition 1

Let  $\hat{f}_{m_1^*}$  and  $\hat{f}_0$  be defined by (16), with  $m_1^* = (D_{m_1}, \mathcal{D}_n^{(2)})$  and  $D_{m_1} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log n}$ . 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

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

Therefore, we just have to prove that  $\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty, A} > f_0/2) \leq C_0/n^k$ . 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  $\tilde{\beta} > 1$ , the embedding theorem proved in [32], p. 236, implies that  $f$  belongs to  $B_{\infty,\infty}^{(\beta_1^*, \beta_2^*)}(A)$  with  $\beta_1^* = \tilde{\beta}_1(1 - 1/\tilde{\beta})$  and  $\beta_2^* = \tilde{\beta}_2(1 - 1/\tilde{\beta})$ . Moreover, [32] proves that there exists a function  $F_{m^*}$  in the space  $S_{m^*}$  of trigonometric polynomials such that

$$\|F_{m^*} - f\mathbb{1}(A)\| \leq C(D_{m_1^*}^{-\tilde{\beta}_1} + \mathcal{D}_n^{-\tilde{\beta}_2}) \quad \text{and} \quad \|F_{m^*} - f\mathbb{1}(A)\|_{\infty} \leq C(D_{m_1^*}^{-\beta_1^*} + \mathcal{D}_n^{-\beta_2^*}),$$

where  $C$  depends on the Besov norm of  $f$  on  $A$ . Then

$$\begin{aligned} \|f_{m_1^*} - f\mathbb{1}(A)\|_{\infty} &\leq \|f_{m_1^*} - F_{m^*}\|_{\infty} + \|F_{m^*} - f\mathbb{1}(A)\|_{\infty} \\ &\leq \phi_0\sqrt{D_{m_1^*}\mathcal{D}_n^{(2)}}\|f_{m_1^*} - F_{m^*}\| + \|F_{m^*} - f\mathbb{1}(A)\|_{\infty} \\ &\leq \phi_0\sqrt{D_{m_1^*}\mathcal{D}_n^{(2)}}(\|f_{m_1^*} - f\mathbb{1}(A)\| + \|f\mathbb{1}(A) - F_{m^*}\|) + \|F_{m^*} - f\mathbb{1}(A)\|_{\infty} \\ &\leq C'\left[\sqrt{D_{m_1^*}\mathcal{D}_n^{(2)}}(D_{m_1^*}^{-\tilde{\beta}_1} + \mathcal{D}_n^{-\tilde{\beta}_2}) + D_{m_1^*}^{-\beta_1^*} + (\mathcal{D}_n^{(2)})^{-\beta_2^*}\right], \end{aligned}$$

where  $C'$  depends on  $\phi_0$  and the Besov norm of  $f$ . But since  $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\log(n)$ , this proves that  $\|f_{m_1^*} - f\mathbb{1}(A)\|_{\infty} \rightarrow 0$  when  $n \rightarrow +\infty$  as soon as  $\tilde{\beta}_1 > 2$  and  $\tilde{\beta}_2 > 2$ . So, there is  $n_0$  such that for any  $n \geq n_0$ , we have  $\|f_{m_1^*} - f\|_{\infty, A} \leq f_0/4$  and

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

Using (M2), 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

$$\vartheta_n(h) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau (h(X_i, y) Y^i(y) - \mathbb{E}(h(X_i, y) Y^i(y))) dy = \|\sqrt{h}\|_n^2 - \|\sqrt{h}\|_\mu^2. \quad (30)$$

With this notation, and recalling the definition of  $\hat{f}_m$  (see Eq. (15)), we have  $\mathbb{E}(\hat{b}_{j,k}) = b_{j,k}$  and

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

thus

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

Note 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  $U_i^{j,k} = \varphi_j(X_i) \int_0^\tau \psi_k(y) Y^i(y) dy$  are i.i.d. random variables. We apply the Bernstein inequality to the sum of the random variables  $U_i^{j,k}$ . We have

$$|U_i^{j,k}| \leq \|\varphi_j\|_\infty \int_0^\tau |\psi_k(y)| dy \leq \|\varphi_j\|_\infty \left(\tau \int_0^\tau \psi_k^2(y) dy\right)^{1/2} \leq \sqrt{\tau \phi_1 D_{m_1^*}} := c$$

and  $\mathbb{E}[(U_i^{j,k})^2] \leq \tau f_1 =: v^2$ , so the Bernstein inequality gives

$$\mathbb{P}(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \geq x) \leq 2 \exp\left(-\frac{nx^2}{2(v^2 + cx/3)}\right)$$

with  $x = f_0/(4\sqrt{\phi_1\phi_2} D_{m_1^*} \mathcal{D}_n^{(2)})$  and  $v$  and  $c$  defined above. This entails

$$\mathbb{P}\left(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \geq \frac{f_0}{4\sqrt{\phi_1\phi_2} D_{m_1^*} \mathcal{D}_n^{(2)}}\right) \leq 2 \exp\left(-\frac{Cn}{(D_{m_1^*} \mathcal{D}_n^{(2)})^2}\right),$$

where  $C$  is a constant depending on  $f_0, f_1, \tau, \phi_1, \phi_2$ , and since  $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log(n)}$  we obtain

$$\mathbb{P}(\Omega^{\mathbb{G}}) \leq 2\sqrt{n} \exp(-C(\log n)^2) \leq \frac{C_0}{n^k},$$

where  $C_0$  is a constant depending on  $k, f_0, \phi_1, \phi_2, \tau$  and  $f_1$ . This concludes the proof of Proposition 1.

## 6.2. Proof of Proposition 2

### 6.2.1. Proof of Proposition 2

One can write

$$\begin{aligned} \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta \cap \Omega)^{\mathbb{G}})) &\leq \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta^{\mathbb{G}})) + \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Omega^{\mathbb{G}})) \\ &\leq 2f_1 [\mathbb{E}^{1/2}(\|\hat{\alpha}_{\hat{m}}\|^4) (\mathbb{P}^{1/2}(\Delta^{\mathbb{G}}) + \mathbb{P}^{1/2}(\Omega^{\mathbb{G}})) + \|\alpha\|_A^2 (\mathbb{P}(\Omega^{\mathbb{G}}) + \mathbb{P}(\Delta^{\mathbb{G}}))]. \end{aligned}$$

using assumptions (A1) and (A4). Now, assumption (A5) with  $k = 7$  ensures that  $\mathbb{P}(\Omega^{\mathfrak{G}}) \leq C_0/n^7$  for any  $n \geq n_0$ . It remains to bound both  $\mathbb{E}(\|\hat{\alpha}_{\hat{m}}\|^4)$  and  $\mathbb{P}(\Delta^{\mathfrak{G}})$ .

**Lemma 3.** *Under the assumptions of Theorem 1,  $\mathbb{E}(\|\hat{\alpha}_{\hat{m}}\|^4) \leq C_B n^5$ , where  $C_B$  is a constant depending on  $\phi_1, \phi_2, \tau$  and  $\|\alpha\|_{\infty, A}$ .*

**Lemma 4.** *Under the assumptions of Theorem 1, we have  $\mathbb{P}(\Delta^{\mathfrak{G}}) \leq C_k^{(\Delta)}/n^k$  for any  $k \geq 1$ , where  $C_k^{(\Delta)}$  is a constant depending on  $k$ , on the basis, and on  $f_0, f_1$ .*

The proof of Lemma 3 is given below, the proof of Lemma 4 is given in a longer version of the paper, see [11]. Using Lemmas 3 and 4 and assumption (A5), we get

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_{\mu}^2 \mathbb{1}((\Delta \cap \Omega)^{\mathfrak{G}})) \leq C_1/n, \quad (31)$$

where  $C_1$  is a constant depending on  $\tau, \phi_1, \phi_2, \|\alpha\|_{\infty, A}, f_0, f_1$ . This concludes the proof of Proposition 2.

### 6.2.2. Proof of Lemma 3

We present here a short proof, we refer the reader to [11] for more details. Note that  $\hat{\alpha}_{\hat{m}}$  is either 0 or  $\arg \min_{t \in S_{\hat{m}}} \gamma_n(t)$ . In the second case,  $\min \text{Sp}(G_{\hat{m}}) \geq \max(\hat{f}_0/3, n^{-1/2})$ , hence

$$\begin{aligned} \|\hat{\alpha}_{\hat{m}}\|^2 &= \sum_{j,k} (\hat{\alpha}_{j,k}^{\hat{m}})^2 = \|A_{\hat{m}}\|^2 = \|G_{\hat{m}}^{-1} \mathcal{Y}_{\hat{m}}\|^2 \leq (\min \text{Sp}(G_{\hat{m}}))^{-2} \|\mathcal{Y}_{\hat{m}}\|^2 \\ &\leq \min(9/\hat{f}_0^2, n) \frac{1}{n} \sum_{i=1}^n \sum_j \varphi_j^2(X_i) \sum_k \left( \int_0^\tau \psi_k(z) dN^i(z) \right)^2, \end{aligned}$$

where, for short,  $\varphi_j := \varphi_j^{\hat{m}}$  and  $\psi_k := \psi_k^{\hat{m}}$ . So that

$$\|\hat{\alpha}_{\hat{m}}\|^4 \leq n^2 \phi_1^2 (\mathcal{D}_n^{(1)})^2 \mathcal{D}_n^{(2)} \frac{1}{n} \sum_{i=1}^n \sum_k \left( \mathbb{1}_{A_1}(X_i) \int_0^\tau \psi_k(z) dN^i(z) \right)^4, \quad (32)$$

see assumptions (M2) and (A2). In addition, one can write

$$\begin{aligned} &\mathbb{E} \left( \left( \mathbb{1}_{A_1}(X_i) \int_0^\tau \psi_k(z) dN^i(z) \right)^4 \right) \\ &\leq 2^3 \mathbb{E} \left( \left( \mathbb{1}_{A_1}(X_i) \int_0^\tau \psi_k(z) \alpha(X^i, z) Y^i(z) dz \right)^4 \right) + 2^3 \mathbb{E} \left( \left( \int_0^\tau \psi_k(z) dM^i(z) \right)^4 \right). \end{aligned} \quad (33)$$

Using the Burkholder inequality, see, e.g., [28], p. 75, and the fact that the quadratic variation process of each  $M^i$  is  $N^i$  ( $i = 1, \dots, n$ ), we know that there exists a universal constant  $\kappa_b$  such that

$$\sum_k \mathbb{E} \left( \left( \int_0^\tau \psi_k(z) dM^i(z) \right)^4 \right) \leq \kappa_b \mathbb{E} \left( N^i(\tau) \sum_{s: \Delta N^i(s) \neq 0} \sum_k \psi_k^4(s) \right), \quad (34)$$

see [11] for more details. Notice that  $\sum_k \psi_k^4(s) \leq \phi_2^2 (\mathcal{D}_n^{(2)})^2$ , by assumption (M2) and some algebra. It now remains to bound  $\mathbb{E}[(N^i(\tau))^2]$ . By assumptions (A3) and (A4), we have

$$\mathbb{E}[N^1(\tau)]^2 \leq 2(M^1(\tau))^2 + 2 \left( \int_0^\tau \alpha(X_1, z) Y^1(z) dz \right)^2 + 2\tau \|\alpha\|_{\infty, A} + 2(\tau \|\alpha\|_{\infty, A})^2. \quad (35)$$

Combining (33), (34) and (35), we get

$$\begin{aligned} \mathbb{E}\left(\sum_k \left(\int_0^\tau \psi_k(z) dN^i(z)\right)^4\right) &\leq 8\kappa_b \phi_2^2 (\mathcal{D}_n^{(2)})^2 \mathbb{E}[(N^1(\tau))^2] + 8\|\alpha\|_{\infty,A}^4 \tau^2 \sum_k \left(\int_0^\tau \psi_k^2(z) dz\right)^2 \\ &\leq 8\kappa_b \phi_2^2 (\mathcal{D}_n^{(2)})^2 \mathbb{E}[(N^1(\tau))^2] + 8\|\alpha\|_{\infty,A}^4 \tau^2 \mathcal{D}_n^{(2)}. \end{aligned} \quad (36)$$

Then we have, by inserting (36) in (32),

$$\begin{aligned} \mathbb{E}(\|\hat{\alpha}_m\|^4) &\leq (\phi_1 n \mathcal{D}_n^{(1)})^2 \mathcal{D}_n^{(2)} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \sum_k \left(\int_0^\tau \psi_k(z) dN^i(z)\right)^4\right) \\ &\leq C_B n^2 (\mathcal{D}_n^{(1)})^2 (\mathcal{D}_n^{(2)})^3 \leq C_B n^{4.5} \leq C_B n^5, \end{aligned}$$

where  $C_B$  is a constant depending on  $\phi_1$ ,  $\phi_2$ ,  $\tau$  and  $\|\alpha\|_{\infty,A}$ . We use here that  $\mathcal{D}_n^{(i)} \leq \sqrt{n}/\log(n)$  in the case of localized bases [DP], [W], [H]. Note that for basis [T], under (M1), the final order is smaller (namely  $n^{3.25}$  instead of  $n^{4.5}$ ). This concludes the proof of Lemma 3.

### 6.3. Proof of Lemma 1

Let  $m \in \mathcal{M}_n$  be fixed and let  $\ell$  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  $\Delta$  (see (18)):

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

Therefore, on  $\Delta$ , for all  $m \in \mathcal{M}_n$ , we have  $\min \text{Sp}(G_m) \geq f_0/2$ . Moreover, on  $\Omega$ , we have  $f_0 \geq 2\hat{f}_0/3$  and  $\max(\hat{f}_0/3, n^{-1/2}) = \hat{f}_0/3$ , for  $n \geq 4/f_0^2$ .

### 6.4. Proof of Proposition 3

Usually, in model selection (see, e.g., [29]), the penalty is obtained by using the so-called Talagrand's deviation inequality for the maximum of empirical processes. Since the empirical process  $v_n(\cdot)$  (see Eq. (21)) considered here is not bounded, we cannot use directly Talagrand's inequality. Using tools from [42], we prove Bennett and Bernstein type inequalities for  $v_n(\cdot)$ , and using a  $L^2(\mu) - L^\infty$  generic chaining type of technique (see [2,39]), we derive an uniform deviation.

**Lemma 5.** *For any positive  $\delta, \epsilon$  and for any function  $h \in (L^2 \cap L^\infty)(A)$ , we have the following Bennett-type deviation inequality:*

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

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

$$\mathbb{P}(v_n(h) \geq \epsilon, \|h\|_n \leq \delta) \leq \exp\left(-\frac{n\epsilon^2/2}{\|\alpha\|_{A,\infty} \delta^2 + \epsilon \|h\|_{A,\infty}/3}\right) \quad (37)$$

and

$$\mathbb{P}(v_n(h) \geq \delta \sqrt{2\|\alpha\|_{\infty,A} x} + \|h\|_{\infty,A} x, \|h\|_n^2 \leq \delta^2) \leq \exp(-nx). \quad (38)$$

**Proof.** Notice that the process

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

is a locally square integrable martingale with jumps of size less than  $n\|h\|_{\infty, A}$ . As a consequence, Corollary 2.3 of [42] applies almost directly. However to introduce the empirical norm  $\|h\|_n$  in the deviation inequality, we re-derive the majoration of the term

$$S_\tau := \sum_{i=1}^n \sum_{k \geq 2} \frac{a^k}{k!} \int_0^\tau |h(X_i, z)|^k dV_k^i(z),$$

where, for all  $i = 1, \dots, n$ ,  $V_2^i(t) := \langle M^i(t) \rangle$  and, for  $k \geq 3$ , we define  $V_k^i(t)$  as the compensator of the  $k$ -variation process  $\sum_{s \leq t} |\Delta M^i(t)|^k$  of  $M^i(t)$  (see Eq. (A3) on page 1795 in [42]).

In our case, we have,  $n^{-1} \sum_{i=1}^n \int_0^\tau h(X_i, z)^2 dV_2^i(z) \leq \|h\|_n^2 \|\alpha\|_{\infty, A}$ , so that

$$S_\tau \leq \frac{n\delta^2 \|\alpha\|_{\infty, A}}{\|h\|_{\infty, A}^2} (\exp(a\|h\|_{\infty, A}) - 1 - a\|h\|_{\infty, A}),$$

see the proof of Corollary 2.3 of [42]. This majoration together with the proof of Lemma 2.2 in [42] yields the Bennett-type deviation inequality in our lemma. To obtain (37) and (38), we use the fact that  $g(x) \geq 3x^2/(2(x+3))$  for any  $x \geq 0$  and  $g(x) \geq g_2(x)$  for any  $x \geq 0$  where  $g_2(x) := x + 1 - \sqrt{1+2x}$  and  $g_2^{-1}(y) = \sqrt{2y} + y$ , see [6], pp. 366–367.  $\square$

The proof of the next Proposition is given in a longer version of the paper, see [11]. It is obtained from (38) by using a recent  $L^2(\mu) - L^\infty$  generic chaining type of technique (see [2,39]). This method is close to other  $L^2(\mu) - L^\infty$  chaining methods, see among others Theorem 5 in [6], Proposition 7 and Theorems 8 and 9 in [4] or Proposition 4, pp. 282–287 in [10]. We define, for  $\rho > 1$ , the set

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

**Proposition 4.** Let  $\bar{S}$  be a  $D$ -dimensional linear subspace of  $L^2 \cap L^\infty(\mu)$ , and define  $B_\delta$  as the  $L^2(\mu)$  closed ball of  $\bar{S}$  with radius  $\delta$ . The  $L^\infty$ -index of  $\bar{S}$  is defined in the following way:

$$\bar{r} = \frac{1}{\sqrt{D}} \inf_{(\psi_\lambda)} \sup_{\beta \neq 0} \frac{\|\sum_{\lambda \in \Lambda} \beta_\lambda \psi_\lambda\|_{\infty, A}}{|\beta|_\infty}, \tag{39}$$

where the infimum is taken over every orthonormal basis  $(\psi_\lambda)_{\lambda \in \Lambda}$  of  $\bar{S}$ , and where  $|\beta|_\infty$  is the  $\ell_\infty$ -norm of  $\beta \in \mathbb{R}^\Lambda$ . For any  $x > 0$  and  $\delta > 0$ , we have

$$\mathbb{P}_{\Delta_\rho} \left[ \sup_{h \in B_\delta} v_n(h) \geq \kappa_0 \left( \delta_\rho \sqrt{\frac{\|\alpha\|_{\infty, A}(D+x)}{n}} + \delta_\rho \bar{r} \frac{D+x}{n} \right) \right] \leq e^{-x},$$

where  $\kappa_0 = 11.8$ , and where  $\delta_\rho = \delta(2 - 1/\rho)$ , where  $\rho > 1$ .

Now, we can turn to the proof of Proposition 3. We denote by  $D(m, m')$  the dimension of the linear space  $S_m + S'_m$ .

**Proof of Proposition 3.** In Proposition 4, take  $x = D_{m'} + u$ ,  $\delta = 1$ ,  $B_\delta = B_{m, m'}^\mu(0, 1) = \{t \in S_m + S'_m : \|t\|_\mu \leq 1\}$  and  $\rho = 2$  in order to get

$$\mathbb{P}_\Delta \left[ \sup_{h \in B_{m, m'}^\mu(0, 1)} v_n^2(h) \geq \eta^2 \right] \leq 2\mathbb{P}_\Delta \left[ \sup_{h \in B_{m, m'}^\mu(0, 1)} v_n(h) \geq \eta \right] \leq 2e^{-D_{m'} - u},$$

where

$$\begin{aligned} \eta^2 &= \frac{9}{4} \kappa_0^2 \left( \sqrt{\|\alpha\|_{\infty,A} \frac{D(m,m') + D_{m'} + u}{n}} + \bar{r}_{m,m'} \frac{D(m,m') + D_{m'} + u}{n} \right)^2 \\ &\leq \frac{9}{2} \kappa_0^2 \left( \|\alpha\|_{\infty,A} \frac{D(m,m') + D_{m'} + u}{n} + 2\bar{r}_{m,m'}^2 \left( \frac{D(m,m') + D_{m'}}{n} \right)^2 + 2\bar{r}_{m,m'}^2 \frac{u^2}{n^2} \right) \\ &\leq 18\kappa_0^2 \left( (1 + \|\alpha\|_{\infty,A}) \frac{D_m + D_{m'}}{n} + \left( \frac{\|\alpha\|_{\infty,A} u}{n} \vee \bar{r}_{m,m'}^2 \frac{u^2}{n^2} \right) \right), \end{aligned}$$

where we used the fact that

$$\bar{r}_{m,m'}^2 \left( \frac{D(m,m') + D_{m'}}{n} \right)^2 \leq \frac{D(m,m')}{n}$$

for  $n$  large enough (see Appendix B in [11]) and  $D(m,m') \leq D_m + D_{m'}$ . This gives

$$\begin{aligned} \mathbb{P}_\Delta \left[ \sup_{h \in B_{m,m'}^\mu(0,1)} v_n^2(h) \geq \kappa \left( (1 + \|\alpha\|_{\infty,A}) \frac{D_m + D_{m'}}{n} + \left( \frac{\|\alpha\|_{\infty,A} u}{n} \vee \bar{r}_{m,m'}^2 \frac{u^2}{n^2} \right) \right) \right] \\ \leq 2e^{-D_{m'} - u}, \end{aligned} \tag{40}$$

where  $\kappa = 18\kappa_0^2$ . Now, we set  $p(m,m') = \kappa(1 + \|\alpha\|_{\infty,A})(D_m + D_{m'})/n$  with  $\kappa = 18\kappa_0^2$ . This gives

$$\mathbb{P}_\Delta \left[ \sup_{h \in B_{m,m'}^\mu(0,1)} v_n^2(h) \geq p(m,m') + z \right] \leq \begin{cases} 2e^{-D_{m'} - nz/(\kappa\|\alpha\|_{\infty,A})} & \text{if } z \leq \kappa\|\alpha\|_{\infty,A}^2/\bar{r}_{m,m'}^2, \\ 2e^{-D_{m'} - n\sqrt{z}/\sqrt{\kappa\bar{r}_{m,m'}^2}} & \text{if } z > \kappa\|\alpha\|_{\infty,A}^2/\bar{r}_{m,m'}^2, \end{cases}$$

and we obtain that

$$\begin{aligned} &\mathbb{E} \left[ \left( \sup_{h \in B_{m,m'}^\mu(0,1)} v_n^2(h) - p(m,m') \right)_+ \mathbb{1}(\Delta) \right] \\ &\leq \int_0^\infty \mathbb{P}_\Delta \left( \sup_{h \in B_{m,m'}^\mu(0,1)} v_n^2(h) > p(m,m') + z \right) dz \\ &\leq 2e^{-D_{m'}} \left( \int_0^\infty e^{-nz/(\kappa\|\alpha\|_{\infty,A})} dz + \int_0^{+\infty} e^{-n\sqrt{z}/\sqrt{\kappa\bar{r}_{m,m'}^2}} dz \right) \\ &\leq 2e^{-D_{m'}} \frac{\kappa}{n} \left( \|\alpha\|_{\infty,A} \int_0^\infty e^{-v} dv + \frac{\bar{r}_{m,m'}^2}{n} \int_0^\infty e^{-\sqrt{v}} dv \right) \\ &\leq 2e^{-D_{m'}} \frac{\kappa}{n} \left( \|\alpha\|_{\infty,A} + \frac{2\bar{r}_{m,m'}^2}{n} \right) \leq \frac{\kappa_\alpha e^{-D_{m'}}}{n}, \end{aligned}$$

where we used the upper bounds of  $\bar{r}_{m,m'}$  given in Appendix B, see [11], and where  $\kappa_\alpha$  is a constant depending on  $\|\alpha\|_{\infty,A}$ ,  $f_0$  and the basis. It remains to bound from above  $\sum_{m' \in \mathcal{M}_n} e^{-D_{m'}}$ . This term is at most

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

This concludes the proof of Proposition 3 when  $n$  is large enough. The statement of Proposition 3 is obvious for small  $n$ , up to an increased constant  $C_2$ . □

## Appendix A: Some useful tools from wavelet and approximation theory

### A.1. The basis [W]

Consider a pair  $\{\phi, \psi\}$  of scaling function and wavelet, where  $\psi$  has  $K$  vanishing moments. Then  $\phi$  and  $\psi$  have a support width of at least  $2K - 1$ , and there is a pair with minimal support, see [15]. This is the starting point of the construction of an orthonormal wavelet basis of  $\mathbb{L}^2[0, 1]$ , as proposed in [9]. Roughly, the idea is to retain the interior scaling functions (those “far” from the edges 0 and 1), and to add adapted edge scaling functions. This is done in [9], see Section 4 and Theorem 4.4, where the construction allows to keep the orthonormality of the system and the number of vanishing moment unchanged, as well as the number  $2^j$  of scaling function at each resolution  $j$  (which improves a previous construction by [31]). Indeed, if  $l$  is such that  $2^l \geq 2K$ , consider for  $j \geq l - 1$ :

$$\Psi_{j,k} := \begin{cases} \psi_{j,k}^0 & \text{if } j \geq l \text{ and } k = 0, \dots, K - 1, \\ \psi_{j,k} & \text{if } j \geq l \text{ and } k = K, \dots, 2^j - K - 1, \\ \psi_{j,k}^1 & \text{if } j \geq l \text{ and } k = 2^j - K, \dots, 2^j - 1, \\ \phi_{l,k}^0 & \text{if } j = l - 1 \text{ and } k = 0, \dots, K - 1, \\ \phi_{l,k} & \text{if } j = l - 1 \text{ and } k = K, \dots, 2^l - K - 1, \\ \phi_{l,k}^1 & \text{if } j = l - 1 \text{ and } k = 2^l - K, \dots, 2^l - 1, \end{cases}$$

where  $\phi_{j,k} = 2^{j/2}\phi(2^j \cdot -x)$  and  $\psi_{j,k} = 2^{j/2}\psi(2^j \cdot -x)$  are the “interior” dilatations and translations of  $\{\phi, \psi\}$ , and  $\phi_{j,k}^0, \psi_{j,k}^0, \phi_{j,k}^1, \psi_{j,k}^1$  are, at each resolution  $j$ , dilatations of  $2K$  edge scaling functions and wavelets ( $K$  for each edge). We know from [9] that the collection

$$W := \{\Psi_{j,k} : j \geq l - 1, k = 0, \dots, 2^j - 1\}$$

is an orthonormal basis of  $\mathbb{L}^2[0, 1]$ , and the interior and edge wavelets have  $K$  vanishing moments, which ensures that the elements of this collection have the same smoothness as  $\phi$  and  $\psi$ .

### A.2. Some approximation results

An orthonormal basis of  $\mathbb{L}^2[0, 1]^2$  is simply obtained by taking tensor products of two bases [W] for instance. If  $W^{(1)}$  and  $W^{(2)}$  are two basis [W] (we can use two different pairs  $\{\phi^{(1)}, \psi^{(1)}\}$  and  $\{\phi^{(2)}, \psi^{(2)}\}$  with possibly different number of vanishing moments), we can simply consider

$$W^{(1)} \otimes W^{(2)} := \{\Psi_{j_1,k_1}^{(1)} \otimes \Psi_{j_2,k_2}^{(2)} : j_1 \geq l_1 - 1, j_2 \geq l_2 - 1, \\ k_1 = 0, \dots, 2^{j_1} - 1, k_2 = 0, \dots, 2^{j_2} - 1\},$$

where  $\Psi_{j_1,k_1}^{(1)} \otimes \Psi_{j_2,k_2}^{(2)}(x_1, x_2) := \Psi_{j_1,k_1}^{(1)}(x_1)\Psi_{j_2,k_2}^{(2)}(x_2)$ . We can also obtain an orthonormal basis of  $\mathbb{L}^2[0, 1]^2$  by taking tensor products of two collections among the ones considered in Section 2.4. Let us consider  $S_m$  as one of the following:

- A space of piecewise polynomials (see Section 2.4, basis [DP]) of degrees smaller than  $s_i > \beta_i - 1$  ( $i = 1, 2$ ) based on a partition with rectangles of sidelengths  $1/D_{m_1}$  and  $1/D_{m_2}$ .
- A space spanned by tensors products of [W], namely the span of the  $\Psi_{j_1,k_1}^{(1)} \otimes \Psi_{j_2,k_2}^{(2)}$  for  $j_1 \in \{l - 1, \dots, m_1\}$ ,  $j_2 \in \{l - 1, \dots, m_2\}$ ,  $k_1 \in \{0, \dots, 2^{j_1} - 1\}$ ,  $k_2 \in \{0, \dots, 2^{j_2} - 1\}$ , where the  $\Psi_{j,k}^{(1)}$  and  $\Psi_{j,k}^{(2)}$  have 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.

Note that the dimension of each space is  $D_{m_1}D_{m_2}$ . The following result is an easy consequence of results by Hochmuth [19] and Nikol’skii [32] (see [23]).

**Lemma 6.** Let  $s$  belong to  $B_{2,\infty}^{\beta}(A)$  where  $\beta = (\beta_1, \beta_2)$ . We consider that  $S_m$  is one of the spaces above, with dimension  $D_{m_1}D_{m_2}$ . If  $s_m$  is the orthogonal projection of  $s$  on  $S_m$ , then there is a positive constant  $C$  such that

$$\|s - s_m\|_A = \left( \int_A |s - s_m|^2 \right)^{1/2} \leq C [D_{m_1}^{-\beta_1} + D_{m_2}^{-\beta_2}],$$

where  $C$  depends on the Besov norm of  $s$  and on the basis.

## References

- [1] P. K. Andersen, O., Borgan, R. D. Gill and N. Keiding. *Statistical Models Based on Counting Processes*. Springer, New York, 1993. MR1198884
- [2] Y. Baraud. A Bernstein-type inequality for suprema of random processes with an application to statistics. *Bernoulli* (2010). To appear.
- [3] Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields* **149** (2009) 239–284. MR2449129
- [4] A. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** (1999) 301–413. MR1679028
- [5] J. Beran. Nonparametric regression with randomly censored survival data. Technical report, Dept. Statist., Univ. California, Berkeley, 1981.
- [6] L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** (1998) 329–375. MR1653272
- [7] E. Brunel, F. Comte and C. Lacour. Adaptive estimation of the conditional density in presence of censoring. *Sankhyā A* **69** (2007) 734–763. MR2521231
- [8] G. Castellán and F. Letué. Estimation of the Cox regression function via model selection. Chapter of the PhD thesis of F. Letué, Univ. Paris XI-Orsay, 2000.
- [9] A. Cohen, I. Daubechies and P. B. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** (1993) 54–81. MR1256527
- [10] F. Comte. Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli* **7** (2001) 267–298. MR1828506
- [11] F. Comte, S. Gaïffas and A. Guillaou. Adaptive estimation of the conditional intensity of marker-dependent counting processes. Preprint MAP5 2008-16, revised 2010. Available at <http://hal.archives-ouvertes.fr/hal-00333356/fr/>.
- [12] D. R. Cox. Regression models and life-tables (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** (1972) 187–220. MR0341758
- [13] D. M. Dabrowska. Nonparametric regression with censored survival time data. *Scand. J. Statist.* **14** (1987) 181–197. MR0932943
- [14] D. M. Dabrowska. Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Ann. Statist.* **17** (1989) 1157–1167. MR1015143
- [15] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** (1988) 909–996. MR0951745
- [16] M. Delecroix, O. Lopez and V. Patilea. Nonlinear censored regression using synthetic data. *Scand. J. Statist.* **35** (2008) 248–265. MR2418739
- [17] G. Grégoire. Least squares cross-validation for counting processes intensities. *Scand. J. Statist.* **20** (1993) 343–360. MR1276698
- [18] C. Heuchenne and I. Van Keilegom. Location estimation in nonparametric regression with censored data. *J. Multivariate Anal.* **98** (2007) 1558–1582. MR2370107
- [19] R. Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.* **12** (2002) 179–208. MR1884234
- [20] J. Huang. Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** (1999) 1536–1563. MR1742499
- [21] M. Jacobsen. *Statistical Analysis of Counting Processes. Lecture Note in Statistics* **12**. Springer, New York, 1982. MR0676128
- [22] A. F. Karr. *Point Processes and Their Statistical Inference*. Marcel Dekker, New York, 1986. MR0851982
- [23] C. Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.* **43** (2007) 571–597. MR2347097
- [24] C. Lacour. Estimation non paramétrique adaptative pour les chaînes de Markov et les chaînes de Markov cachées. PhD thesis, 2007. Available at <http://www.math.u-psud.fr/~lacour/etudes/>.
- [25] M. LeBlanc and J. Crowley. Adaptive regression splines in the Cox model. *Biometrics* **55** (1999) 204–213.
- [26] G. Li and H. Doss. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23** (1995) 787–823. MR1345201
- [27] O. B. Linton, J. P. Nielsen and S. Van de Geer. Estimating the multiplicative and additive hazard functions by kernel methods. *Ann. Statist.* **31** (2003) 464–492. MR1983538
- [28] R. S. Liptser and A. N. Shirayev. *Theory of Martingales. Mathematics and its Applications (Soviet Series)* **49**. Kluwer Academic, Dordrecht, 1989. MR1022664
- [29] P. Massart. *Concentration Inequalities and Model Selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin, 2007. MR2319879
- [30] I. W. McKeague and K. J. Utikal. Inference for a nonlinear counting process regression model. *Ann. Statist.* **18** (1990) 1172–1187. MR1062704
- [31] Y. Meyer. Ondelettes sur l’intervalle. *Rev. Mat. Iberoamericana* **7** (1991) 115–133. MR1133374
- [32] S. M. Nikol’skii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, New York, 1975. MR0374877
- [33] H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** (1983) 453–466. MR0696058
- [34] P. Reynaud-Bouret. Adaptive estimation of the intensity of nonhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* **126** (2003) 103–153. MR1981635

- [35] P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli* **12** (2006) 633–661. [MR2248231](#)
- [36] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** (1980) 1348–1360. [MR0594650](#)
- [37] W. Stute. Conditional empirical processes. *Ann. Statist.* **14** (1986) 638–647. [MR0840519](#)
- [38] W. Stute. Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.* **23** (1996) 461–471. [MR1439707](#)
- [39] M. Talagrand. *The Generic Chaining*. Springer, Berlin, 2005. [MR2133757](#)
- [40] H. Triebel. *Theory of Function Spaces. III. Monographs in Mathematics* **100**. Birkhäuser, Basel, 2006. [MR2250142](#)
- [41] A. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, Berlin, 2004. [MR2013911](#)
- [42] S. van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** (1995) 1779–1801. [MR1370307](#)