



# Estimating a density, a hazard rate, and a transition intensity via the $\rho$ -estimation method

Mathieu Sart

## ► To cite this version:

Mathieu Sart. Estimating a density, a hazard rate, and a transition intensity via the  $\rho$ -estimation method. *Annales de l'Institut Henri Poincaré*, 2021, 57 (1), pp.195-249. 10.1214/20-AIHP1076 . hal-01557973v5

**HAL Id: hal-01557973**

**<https://hal.science/hal-01557973v5>**

Submitted on 22 Jul 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ESTIMATING A DENSITY, A HAZARD RATE, AND A TRANSITION INTENSITY VIA THE $\rho$ -ESTIMATION METHOD

MATHIEU SART

**ABSTRACT.** We propose a unified study of three statistical settings by widening the  $\rho$ -estimation method developed in [BBS17]. More specifically, we aim at estimating a density, a hazard rate (from censored data), and a transition intensity of a time inhomogeneous Markov process. We show non-asymptotic risk bounds for an Hellinger-type loss when the models consist, for instance, of piecewise polynomial functions, multimodal functions, or functions whose square root is piecewise convex-concave. Under convex-type assumptions on the models, maximum likelihood estimators coincide with  $\rho$ -estimators, and satisfy therefore our risk bounds. However, our results also apply to some models where the maximum likelihood method does not work. Subsequently, we present an alternative way, based on estimator selection, to define a piecewise polynomial estimator. We control the risk of the estimator and carry out some numerical simulations to compare our approach with a more classical one based on maximum likelihood only.

## 1. INTRODUCTION

**1.1. Statistical settings.** In the present paper, we are interested in estimating a unknown function  $\mathbf{s}_0$  that appears in one of the following frameworks.

**Framework 1** (Density Estimation). *Let  $X$  be a real-valued random variable with density function  $\mathbf{s}_0$  with respect to the Lebesgue measure  $\mu$ . Our aim is to estimate the density  $\mathbf{s}_0$  from the observation of  $n$  independent copies  $X_1, \dots, X_n$  of  $X$ .*

**Framework 2** (Hazard rate estimation for right censored data). *Let  $(T_1, C_1), \dots, (T_n, C_n)$  be  $n$  independent copies of a pair  $(T, C)$  of non-negative random variables. The variable  $C$  may take the value  $+\infty$ . We suppose that  $T$  is independent of  $C$  and that  $T$  admits a density  $f_0$  with respect to the Lebesgue measure  $\mu$ . The target function is the hazard rate  $\mathbf{s}_0$  defined for  $t \geq 0$  by*

$$\mathbf{s}_0(t) = \frac{f_0(t)}{\mathbb{P}(T \geq t)}.$$

*The observations are  $(X_i, D_i)_{1 \leq i \leq n}$  where  $X_i = \min\{T_i, C_i\}$  and  $D_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \\ 0 & \text{otherwise.} \end{cases}$*

**Framework 3** (Estimation of the transition intensity of a Markov process). *We consider a (possibly inhomogeneous) Markov process  $\{X_t, t \geq 0\}$  with the following properties:*

- *The process is cadlag with finite state space, says  $\{0, 1, \dots, m\}$ .*
- *The state 0 is absorbing.*
- *Let, for each interval  $I \subset [0, +\infty)$ ,  $A_I$  be the event: “the process jumps at least two times on  $I$ ”. Then,  $\mathbb{P}(A_I) = o(\mu(I))$  when the length  $\mu(I)$  of  $I$  tends to 0.*

---

2010 *Mathematics Subject Classification.* 62G07, 62G35, 62N02, 62M05.

*Key words and phrases.*  $\rho$ -estimator, maximum likelihood, qualitative assumptions, piecewise polynomial estimation.

- *The transition time*

$$T_{1,0} = \inf \{t > 0, X_{t-} = 1, X_t = 0\},$$

which has values in  $[0, +\infty]$ , is absolutely continuous with respect to the Lebesgue measure  $\mu$  on  $\mathbb{R}$  and satisfies therefore for all Borel set  $A$  of  $\mathbb{R}$ ,

$$\mathbb{P}(T_{1,0} \in A) = \int_A f_0(t) dt,$$

for a suitable non-negative measurable function  $f_0$ .

We consider an observation interval  $I_{obs} \subset [0, +\infty)$  either of the form  $I_{obs} = [0, T]$  with  $T \in (0, +\infty)$  or  $I_{obs} = [0, +\infty)$ . Our aim is to estimate the transition rate  $\mathbf{s}_0$  from state 1 to 0 defined for  $t > 0$  by

$$\mathbf{s}_0(t) = \frac{f_0(t)}{\mathbb{P}(X_{t-} = 1)},$$

from the observation of  $n$  independent copies  $\{X_t^{(i)}, t \in I_{obs}\}$  of  $\{X_t, t \in I_{obs}\}$ .

In all these frameworks, we will always suppose that  $n \geq 3$ . Although numerous estimation strategies can be considered, we will rather focus in this paper on a particular method presented in [BBS17] and named “ $\rho$ -estimation”.

**1.2. About  $\rho$ -estimation in framework 1.** We begin by carrying out the method and some known results in density estimation. The key references are [BBS17, BB16, BB17].

We need a loss in order to measure the quality of an estimator. In  $\rho$ -estimation, we deal with the Hellinger distance  $h$ . It is defined for two non-negative integrable functions  $s$  and  $s'$  by

$$h^2(s, s') = \frac{1}{2} \int_{\mathbb{R}} \left( \sqrt{s(t)} - \sqrt{s'(t)} \right)^2 dt.$$

The aim is then to define an estimator  $\hat{s}$  that minimizes as far as possible the Hellinger distance  $h$  between  $\hat{s}$  and the target  $\mathbf{s}_0$ .

The procedure is based on models  $S$ , that is a collections of densities, which translate, in mathematical terms, the knowledge we have on  $\mathbf{s}_0$ . A model may correspond to different assumptions, such as parametric, regularity, or qualitative ones. This includes in particular models  $S$  for which the maximum likelihood method does not work. Several examples are known in the literature. A very simple one is

$$(1) \quad S = \bigcup_{K \text{ interval of } \mathbb{R}} \{s \mathbb{1}_K, \text{ where } s \text{ is a non-increasing density on } K\},$$

where  $\mathbb{1}_K$  denotes the indicator function of  $K$ . In this model, the log likelihood can be made arbitrarily large, and the maximum likelihood estimator does not exist. By contrast, we may define, and study,  $\rho$ -estimators on  $S$ .

The maximal risk  $R_S(n) = \sup_{\mathbf{s}_0 \in S} \mathbb{E}[h^2(\mathbf{s}_0, \hat{s})]$  of a  $\rho$ -estimator  $\hat{s}$  on  $S$  can be controlled according to different notions that aim at measuring the “complexity” of the model  $S$  (entropy with bracketing, metric dimension, covering numbers...). Interestingly,  $R_S(n)$  achieves the optimal minimax rate of convergence in most cases we know (up to possible logarithmic factors).

This minimax point of view supposes that  $\mathbf{s}_0$  does belong to  $S$ . Such an assumption corresponds to a perfect modelling of the statistical problem, which is scarcely the case in practice. It makes

therefore more sense to study the risk of the estimator  $\hat{s}$  not only when  $\mathbf{s}_0$  lies in  $S$  but more generally when  $\mathbf{s}_0$  is close to the model  $S$ . It turns out that the Hellinger quadratic risk of a  $\rho$ -estimator  $\hat{s}$  can be bounded from above by

$$\mathbb{E}[h^2(\mathbf{s}_0, \hat{s})] \leq C \inf_{s \in S} h^2(\mathbf{s}_0, s) + R_S(n) \quad \text{whatever the density } \mathbf{s}_0,$$

where  $C$  is a universal constant (that is a number). This inequality asserts that a small error in the choice of the model  $S$  induces a small error in the estimation of  $\mathbf{s}_0$ . This is a robustness property. Such a property is not shared in general by the maximum likelihood estimator: it may indeed perform very poorly when  $\mathbf{s}_0 \notin S$  but is close to  $S$  in terms of Hellinger distance.

The rate given by  $R_S(n)$  stands for the worst-case rate over all densities  $\mathbf{s}_0$  of  $S$ . This rate may therefore be very pessimistic in the sense that the estimation may be much faster for some particular densities  $\mathbf{s}_0 \in S$ . The preceding risk bound can be refined to take into account this phenomenon (named superminimaxity in [BB16]). For illustration purposes, consider the model  $S$  defined by (1) and a  $\rho$ -estimator  $\hat{s}$  on  $S$ . Then, the rate of convergence of  $\hat{s}$  is at least  $\sqrt{\ell/n} \log^{3/2}(n/\ell)$  when  $\mathbf{s}_0$  is not only non-increasing on an interval but also piecewise constant on  $\ell$  intervals. In this case, the rate of estimation is much faster than the minimax rate.

There are moreover two additional properties of  $\rho$ -estimators we now briefly mention. First,  $\rho$ -estimators can be related to maximum likelihood ones. Second, it is possible to deal with penalized  $\rho$ -estimators, allowing to cope with model selection.

**1.3. Hazard rate and transition intensity estimation.** These two frameworks have not yet been studied by means of the  $\rho$ -estimation method. They appear in different domains such as reliability or survival analysis. For instance, in medical studies, a variable  $T$  may be used to represent the lifetime of a patient, the hazard rate  $\mathbf{s}_0$  at time  $t$ ,

$$\mathbf{s}_0(t) = \frac{f_0(t)}{\mathbb{P}(T \geq t)} = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq T \leq t+h \mid T \geq t)}{h},$$

then measures the tendency of dying just after  $t$ , given survival to time  $t$ . Since patients may leave the study, the data may be censored. The random variable  $C$  then gives the time of leaving and  $D = \mathbb{1}_{T \leq C}$  indicates whether the patient dies ( $D = 1$ ) or leaves the study ( $D = 0$ ).

In medical trials, a Markov process  $\{X_t, t > 0\}$  may be used to model the evolution of a disease, the state 0 representing (for instance) the death of the patient. The transition rate  $\mathbf{s}_0$  at time  $t$ ,

$$\mathbf{s}_0(t) = \frac{f_0(t)}{\mathbb{P}(X_{t-} = 1)} = \lim_{h \rightarrow 0} \frac{\mathbb{P}(X_{t+h} = 0 \mid X_{t-} = 1)}{h},$$

has similar interpretation than the hazard rate: it measures the risk of dying just after  $t$ , given the disease is in state 1 at time  $t-$ . This framework is actually more general than the one of hazard rate estimation (when the data are uncensored) as  $\mathbf{s}_0$  coincides with the hazard rate of  $T$  when the Markov process is defined by  $X_t = \mathbb{1}_{T \geq t}$ .

In the literature, numerous estimators have been proposed to deal with (at least) one of these two frameworks. We may cite wavelet estimators, Kernel estimators, maximum likelihood estimators, procedures based on  $\mathbb{L}^2$  contrasts. . . However, non-asymptotic studies seem to be rather scarce. We refer to [BC05, RB06, BC08, Pla11, AD10] for results concerning procedures based on (penalized)  $\mathbb{L}^2$  contrasts. We may cite [vdG95, DR02] for a study of non-asymptotic properties of maximum likelihood estimators. We refer to [BB09] for results concerning a selection rule based on pairwise comparisons of histogram type estimators.

**1.4. A generalized procedure.** In this paper, we propose to extend the scope of  $\rho$ -estimation to frameworks 2 and 3. Although it has already been studied in the literature, we do not exclude framework 1 for pedagogical and numerical reasons (see Section 1.5 below).

We measure the risks of our estimators by means of a (possibly random) Hellinger-type distance  $h$  adapted to the framework. In framework 1,  $h$  is the usual Hellinger distance, in framework 2,

$$h^2(s, s') = \frac{1}{2} \int_0^\infty \left( \sqrt{s(t)} - \sqrt{s'(t)} \right)^2 \left( \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \geq t} \right) dt,$$

and in framework 3,

$$h^2(s, s') = \frac{1}{2} \int_{I_{\text{obs}}} \left( \sqrt{s(t)} - \sqrt{s'(t)} \right)^2 \left( \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_{t-}^{(i)}=1} \right) dt.$$

The quality of an estimator  $\hat{s}$  is therefore assessed by  $h^2(\mathbf{s}_0, \hat{s})$ : the smaller  $h^2(\mathbf{s}_0, \hat{s})$ , the better the estimator.

In hazard rate estimation, we also explain how to use the (unknown) loss  $h_E$  defined by

$$h_E^2(s, s') = \frac{1}{2} \int_0^\infty \left( \sqrt{s(t)} - \sqrt{s'(t)} \right)^2 \mathbb{P}(X \geq t) dt,$$

in place of  $h$ .

We develop the  $\rho$ -estimation method in frameworks 2 and 3. We use a new uniform exponential inequality to prove risk bounds and rates of convergence when the target  $\mathbf{s}_0$  is, for instance, a piecewise polynomial function, a multimodal function, or a function whose square root is piecewise convex-concave. These rates correspond to the expected ones, up to possible logarithmic factors. Moreover, they are slightly faster than the ones obtained in [BB16] in density estimation under the same assumptions.

Besides, there is a close connection between maximum likelihood and  $\rho$ -estimation when the models satisfy convexity-type conditions. This allows to deduce results for maximum likelihood estimators from results for  $\rho$ -estimators. Thereby, this paper also includes non-asymptotic risk bounds for maximum likelihood estimators.

**1.5. Estimator selection.** The practical computation of  $\rho$ -estimators seems unfortunately to be numerically out of reach in numerous models. This is the case for instance when  $S = \mathcal{P}_{\ell,r}$  consists of non-negative piecewise polynomial functions of degree  $r$  on  $\ell$  pieces. Although maximum likelihood estimators do not exist on  $\mathcal{P}_{\ell,r}$ ,  $\rho$ -estimators do exist, and we may even control their Hellinger-type risks. However, we do not know how to construct them in practice (in none of the three frameworks).

We then propose an alternative way, based on maximum likelihood and estimator selection to reduce the numerical complexity. More precisely, we carry out a new procedure, inspired from [Sar14], to select among a suitable collection of maximum likelihood estimators. Although the cardinal of this family is large, dynamic programming makes possible the practical implementation of the procedure in favorable situations. We prove an oracle inequality for the selected estimator from which, we deduce, when  $r = 0$ , a risk bound very similar to the one we would obtain for the  $\rho$ -estimator on  $\mathcal{P}_{\ell,0}$ . Besides, we carry out a numerical study in which we compare, in the context of density estimation, our procedure with a selection rule based on maximum likelihood only.

We finally explain how to modify this procedure to select adaptively the number  $\ell$  of pieces from the data. In particular, we show that we can build an estimator that performs well when  $\mathbf{s}_0$  belongs, or is close to, the model  $\mathcal{P}_r = \cup_{\ell=1}^{\infty} \mathcal{P}_{\ell,r}$ . We get a risk bound that almost corresponds to the one we would obtain for the best estimator of the family  $\{\hat{s}_{\ell,r}, \ell \geq 1\}$  where  $\hat{s}_{\ell,r}$  denotes the  $\rho$ -estimator on  $\mathcal{P}_{\ell,r}$ .

**1.6. Organization of the paper.** We carry out in Section 2 the general statistical setting that encompasses the three frameworks and explain the estimation procedure. In Section 3, we present the required assumptions on the models as well as our main result on the theoretical performances of  $\rho$ -estimators. We also show how to use the loss  $h_E$  in hazard rate estimation and relate our estimators to those of maximum likelihood. In Section 4, we deal with estimator selection to define a piecewise polynomial estimator in a more practical way. Section 5 is devoted to numerical simulations. The proofs are deferred to Section 6. The probabilistic tool that enables us to control the risk of  $\rho$ -estimators is quite technical and therefore delayed to Appendix A. We give a simple example to illustrate the robustness of  $\rho$ -estimators in hazard rate estimation in Appendix B. We discuss in more detail the numerical complexity of our procedures in Appendix C.

## 2. THE $\rho$ -ESTIMATION METHOD

**2.1. Statistical setting and notations.** In this paper, the target  $\mathbf{s}_0$  is viewed as the intensity of a random measure ([BB09, AD10, Bar11]). This makes possible a unified treatment of the three frameworks. More precisely, we consider an abstract probability space  $(\Omega, \mathcal{E}, \mathbb{P})$  on which are defined the random variables appearing in the different frameworks. We associate to each framework, and each borel set  $A \in \mathcal{B}(\mathbb{R})$  two random variables  $N(A)$  and  $M(A)$ . We set in density estimation,

$$N(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i), \quad M(A) = \mu(A),$$

and in hazard rate estimation,

$$N(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i) \mathbb{1}_{D_i=1}, \quad M(A) = \frac{1}{n} \sum_{i=1}^n \int_A \mathbb{1}_{X_i \geq t} \mathbb{1}_{[0,+\infty)}(t) dt.$$

In framework 3, we define the jump time of the  $i^{th}$  process

$$T_{1,0}^{(i)} = \inf \left\{ t > 0, X_{t-}^{(i)} = 1, X_t^{(i)} = 0 \right\},$$

and consider

$$N(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_{1,0}^{(i)} \in A} \mathbb{1}_{I_{\text{obs}}}(T_{1,0}^{(i)}), \quad M(A) = \frac{1}{n} \sum_{i=1}^n \int_A \mathbb{1}_{X_{t-}^{(i)}=1} \mathbb{1}_{I_{\text{obs}}}(t) dt.$$

These formulas define two random measures  $N$  and  $M$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  such that

$$\mathbb{E}[N(A)] = \mathbb{E} \left[ \int_A \mathbf{s}_0(t) dM(t) \right] \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

In each of the frameworks, the statistical problem may be reduced to that of estimating  $\mathbf{s}_0$  from the observation of the random measures  $N$  and  $M$ .

As explained in the introduction, we will evaluate the quality of the estimators by using an Hellinger-type loss. This Hellinger-type distance  $h$  can be written simultaneously in the three statistical settings as

$$h^2(s, s') = \frac{1}{2} \int_{\mathbb{R}} \left( \sqrt{s(t)} - \sqrt{s'(t)} \right)^2 dM(t),$$

for all non-negative integrable functions  $s$  and  $s'$  with respect to the measure  $M$ .

We now introduce some notations that will be used all along the paper. We define  $\mathbb{R}_+ = [0, +\infty)$ , and set for  $x, y \in \mathbb{R}$ ,  $x \wedge y = \min(x, y)$ ,  $x \vee y = \max(x, y)$ . The positive part of a real valued function  $f$  is denoted by  $f_+$  and its negative part by  $f_-$ . The distance between a point  $x$  and a set  $A$  in a metric space  $(E, d)$  is denoted by  $d(x, A) = \inf_{y \in A} d(x, y)$ . We denote the cardinal of a set  $A$  by  $|A|$ , and its complement by  $A^c$ . We set  $\log_+ x = \max\{\log x, 1\}$  for all  $x > 0$ . The notations  $c, c', C, C', \dots$  are for the constants. These constants may change from line to line.

**2.2. Heuristics.** Let  $\mathcal{S} = \mathbb{L}_+^1(\mathbb{R}, \mu)$  be the cone of non-negative Lebesgue integrable functions in frameworks 1 and 3, and  $\mathcal{S}$  be the cone of measurable non-negative functions which are locally integrable with respect to  $\mu$  in framework 2. Let now  $S$  be a subset of  $\mathcal{S}$ . Such set will be called a model. Our aim is to define an estimator  $\hat{s}$  with values in  $S$  such that  $h(\mathbf{s}_0, \hat{s})$  is as small as possible.

Consider two arbitrary functions  $s, s'$  of  $\mathcal{S}$ . We begin by defining an approximation  $T_E(s, s')$  of  $h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, s')$ .

Let  $\psi$  be the real-valued function defined for  $x \geq 0$  by  $\psi(x) = \frac{\sqrt{x}-1}{\sqrt{x+1}}$ , and  $\psi(+\infty) = 1$ . For  $s, s' \in \mathcal{S}$ , we set

$$T_E(s, s') = \int_{\mathbb{R}} \psi(s'/s) \mathbf{s}_0 dM - \frac{1}{4} \int_{\mathbb{R}} (s' - s) dM.$$

In this definition, and throughout the paper, we use the conventions  $0/0 = 1$  and  $x/0 = +\infty$  for all  $x > 0$ . Some computations show:

**Lemma 1.** *For all  $s, s' \in \mathcal{S}$ ,*

$$(2) \quad \frac{1}{3} h^2(\mathbf{s}_0, s) - 3h^2(\mathbf{s}_0, s') \leq T_E(s, s') \leq 3h^2(\mathbf{s}_0, s) - \frac{1}{3} h^2(\mathbf{s}_0, s').$$

Let  $S$  be a model and  $s \in S$ . We are interested in evaluating  $h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, S)$ . The smaller this number, the better  $s$ . As  $T_E(s, s')$  is roughly of the order of  $h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, s')$ , it is natural to approximate  $h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, S)$  by  $\gamma_E(s) = \sup_{s' \in S} T_E(s, s')$  and to study the properties of the minimizers of  $\gamma_E(\cdot)$ .

We deduce from the above lemma that for all  $s \in S$ ,

$$\frac{1}{3} h^2(\mathbf{s}_0, s) - 3h^2(\mathbf{s}_0, S) \leq \gamma_E(s) \leq 3h^2(\mathbf{s}_0, s) - \frac{1}{3} h^2(\mathbf{s}_0, S).$$

Minimizing  $\gamma_E(\cdot)$  over  $S$  yields a function  $\bar{s} \in S$  (assuming such a function exists) such that,

$$\frac{1}{3} h^2(\mathbf{s}_0, \bar{s}) - 3h^2(\mathbf{s}_0, S) \leq \gamma_E(\bar{s}) \leq \inf_{s \in S} \gamma_E(s) \leq 3 \inf_{s \in S} h^2(\mathbf{s}_0, s) - \frac{1}{3} h^2(\mathbf{s}_0, S) = \frac{8}{3} h^2(\mathbf{s}_0, S).$$

Therefore,  $h^2(\mathbf{s}_0, \bar{s}) \leq 17h^2(\mathbf{s}_0, S)$ , which means that  $\bar{s}$  is, up to a multiplicative constant, the closest function to  $\mathbf{s}_0$  among the ones of  $S$ .

The approximation  $T_E(s, s')$  of  $h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, s')$  is unknown in practice as it involves  $\mathbf{s}_0$ . This prevents us from minimizing  $\gamma_E(\cdot)$ . It can however be suitably approximated in practice.

**2.3. The procedure.** Let  $T(s, s')$  be the approximation of  $T_E(s, s')$  defined for  $s, s' \in \mathcal{S}$  by

$$T(s, s') = \int_{\mathbb{R}} \psi(s'/s) \, dN - \frac{1}{4} \int_{\mathbb{R}} (s' - s) \, dM.$$

This translates as

$$T(s, s') = \begin{cases} \frac{1}{n} \sum_{i=1}^n \left\{ \psi(s'(X_i)/s(X_i)) - \frac{1}{4} \int_{\mathbb{R}} (s'(t) - s(t)) \, dt \right\} & \text{in framework 1,} \\ \frac{1}{n} \sum_{i=1}^n \left\{ \psi(s'(X_i)/s(X_i)) \mathbb{1}_{D_i=1} - \frac{1}{4} \int_0^\infty (s'(t) - s(t)) \mathbb{1}_{X_i \geq t} \, dt \right\} & \text{in framework 2,} \\ \frac{1}{n} \sum_{i=1}^n \left\{ \psi(s'(T_{1,0}^{(i)})/s(T_{1,0}^{(i)})) \mathbb{1}_{I_{\text{obs}}(T_{1,0}^{(i)})} - \frac{1}{4} \int_{I_{\text{obs}}} (s'(t) - s(t)) \mathbb{1}_{X_{t^-}^{(i)}=1} \, dt \right\} & \text{in framework 3.} \end{cases}$$

Let  $S$  be a model and for  $s \in S$ ,

$$\gamma(s) = \sup_{s' \in S} T(s, s').$$

Any estimator  $\hat{s} \in S$  satisfying

$$(3) \quad \gamma(\hat{s}) \leq \inf_{s \in S} \gamma(s) + 1/n$$

is called a  $\rho$ -estimator.

Remark. Contrary to [BBS17, BB17], we do not assume that  $S$  consists of densities in framework 1 for more flexibility in the choice of models. Likewise, the functions of  $S$  may not be hazard rates in framework 2, or transition intensities in framework 3.

The procedure may also be used to estimate the restriction of  $\mathbf{s}_0$  to an interval  $K$ . Indeed, let  $N'$  be defined by  $N'(A) = N(A \cap K)$  for all  $A \in \mathcal{B}(\mathbb{R})$ . Then,  $\mathbb{E}[N'(A)] = \mathbb{E}[\int_A \mathbf{s}_0 \mathbb{1}_K \, dM]$  and the target function becomes  $\mathbf{s}_0 \mathbb{1}_K$ . Let now  $\mathcal{F}$  be a collection of functions and  $S$  be a model of the form  $S = \{f \mathbb{1}_K, f \in \mathcal{F}\}$ . Since the functions of  $S$  vanish outside  $K$ , we may replace  $N$  in the procedure by  $N'$  without changing the estimator. Thereby, when all functions of  $S$  vanish outside  $K$ , the estimator  $\hat{s}$  actually estimates  $\mathbf{s}_0 \mathbb{1}_K$ .

### 3. RISK BOUNDS FOR $\rho$ -ESTIMATORS

**3.1. Assumptions on models.** We recall that the definition of  $\rho$ -estimators is based on the minimization of a criterion  $\gamma(\cdot)$  on  $S$ . This criterion  $\gamma(\cdot)$  uses the approximation  $T(s, s') \simeq T_E(s, s')$  where  $s, s' \in S$ . Bounding above the risk of the  $\rho$ -estimator requires to bound above the error due to the approximation of  $T_E$  by  $T$ . For more informations on this point, we refer to Appendix A. We here only mention that it is possible to control the error under suitable assumptions on the model  $S$  we now describe.

We consider a non-decreasing collection  $(\mathcal{I}_d)_{d \geq 1}$  of Borel sets. This class may be a collection of unions of at most  $d$  intervals, or more generally, in frameworks 1 and 2, a Vapnik-Chervonenkis class of dimension at most  $2d$ .

**Assumption 1.** *We suppose:*

- in frameworks 1 and 2, that the collection  $\mathcal{I}_d$  is Vapnik-Chervonenkis with dimension at most  $2d$ . Besides, the following technical condition holds: there exists an at most countable set  $\mathcal{I}'_d \subset \mathcal{I}_d$  such that for all  $I \in \mathcal{I}_d$ , there exists a sequence  $(I_m)_{m \geq 0} \in \mathcal{I}'_d{}^{\mathbb{N}}$  satisfying

$$\lim_{m \rightarrow +\infty} \mathbb{1}_{I_m}(t) = \mathbb{1}_I(t)$$

for every  $t \in \mathbb{R}$ .

- in framework 3, that the sets  $I \in \mathcal{I}_d$  are unions of at most  $d$  intervals.

We then consider models  $S$  satisfying the following condition.

**Assumption 2.** *There exist  $\bar{S} \subset S$  and a map  $d_S(\cdot)$  on  $\bar{S}$  such that: for all  $\bar{s} \in \bar{S}$ ,  $s \in S$ ,  $u > 0$ , the set  $\{t \in \mathbb{R}, s(t) > u\bar{s}(t)\}$  belongs to  $\mathcal{I}_{d_S(\bar{s})}$ .*

This assumption applies for several models of interest, including some which are well suited for estimating functions under smooth or shape constraints. We carry out below three examples.

Let  $\ell \geq 1$  and  $\mathcal{M}_\ell$  be the family that gathers all the collections  $m$  of size  $\ell$  of the form

$$(4) \quad m = \{[x_1, x_2], (x_2, x_3], (x_3, x_4], \dots, (x_\ell, x_{\ell+1}]\},$$

where  $x_1 < x_2 < \dots < x_{\ell+1}$  are  $\ell + 1$  real numbers (with the convention that  $m = \{[x_1, x_2]\}$  when  $\ell = 1$ ).

We define for  $r \geq 0$  the model  $\mathcal{P}_{\ell,r}$  by

$$(5) \quad \mathcal{P}_{\ell,r} = \left\{ \sum_{K \in m} s_K \mathbb{1}_K, \text{ where } m \in \mathcal{M}_\ell \text{ and } s_K \text{ is a polynomial function of degree at most } r \right. \\ \left. \text{that is non-negative on } K \right\}.$$

**Proposition 1.** *Let for  $d \geq 1$ ,  $\mathcal{I}_d$  be the collection of unions of at most  $d$  intervals. Then, Assumption 2 is fulfilled with  $S = \bar{S} = \mathcal{P}_{\ell,r}$ , and for all  $\bar{s} \in \bar{S}$ ,  $d_S(\bar{s}) = (\ell + 2)(r + 2)$ .*

We may also consider the model composed of piecewise monotone functions and the one composed of functions whose square roots are piecewise convex-concave. They are defined for  $k \geq 1$  by

$$(6) \quad \mathcal{F}_k = \mathcal{S} \cap \left\{ \sum_{K \in m} s_K \mathbb{1}_K, \text{ where } m \in \mathcal{M}_k \text{ and } s_K \text{ is monotone on } K \right\}. \\ \mathcal{G}_k = \mathcal{S} \cap \left\{ \sum_{K \in m} s_K^2 \mathbb{1}_K, \text{ where } m \in \mathcal{M}_k \text{ and } s_K \text{ is a non-negative function that is either} \right. \\ \left. \text{convex or concave on } K \right\}.$$

We recall here that  $\mathcal{S}$  is the cone of non-negative integrable functions in frameworks 1 and 3, and the cone of non-negative locally integrable functions in framework 2.

When  $S = \mathcal{F}_k$ , we may define  $\bar{S}$  as the set of functions of  $\mathcal{F}_k$  that are piecewise constant on a finite number of pieces. When  $S = \mathcal{G}_k$ , we may define  $\bar{S}$  as the set of functions of  $\mathcal{G}_k$  whose square root is piecewise affine (on a finite number of pieces). More precisely, the proposition below is given by [BB16]:

**Proposition 2.** *Let for  $d \geq 1$ ,  $\mathcal{I}_d$  be the collection of unions of at most  $d$  intervals. Assumption 2 is fulfilled with:*

- $S = \mathcal{F}_k$ ,  $\bar{S} = S \cap (\cup_{\ell=1}^{\infty} \mathcal{P}_{\ell,0}) \subset S$  and for all  $\bar{s} \in S \cap \mathcal{P}_{\ell,0}$ ,  $d_S(\bar{s}) = (3/2)(k + \ell + 5)$ .
- $S = \mathcal{G}_k$ ,  $\bar{S} = S \cap (\cup_{\ell=1}^{\infty} \mathcal{P}_{\ell,1,sq.root}) \subset S$ , where  $\mathcal{P}_{\ell,1,sq.root} = \{s \in \mathcal{S}, \sqrt{s} \in \mathcal{P}_{\ell,1}\}$ , and for all  $\bar{s} \in S \cap \mathcal{P}_{\ell,1,sq.root}$ ,  $d_S(\bar{s}) = 3(k + \ell + 5)$ .

### 3.2. A uniform risk bound.

**Theorem 3.** *Let  $(\mathcal{I}_d)_{d \geq 1}$  be a non-decreasing collection of Borel sets that fulfils Assumption 1. For all  $\xi > 0$ , there exists an event of probability lower bounded by  $1 - e^{-n\xi}$  on which: for all model  $S$  satisfying Assumption 2 and all  $\rho$ -estimator  $\hat{s}$  on  $S$ ,*

$$(7) \quad h^2(\mathbf{s}_0, \hat{s}) \leq C \inf_{\bar{s} \in \bar{S}} \left\{ h^2(\mathbf{s}_0, \bar{s}) + \frac{d_S(\bar{s})}{n} \log_+^2 \left( \frac{n}{d_S(\bar{s})} \right) + \xi \log_+(1/\xi) \right\}.$$

In particular,

$$(8) \quad \mathbb{E} [h^2(\mathbf{s}_0, \hat{s})] \leq C' \inf_{\bar{s} \in \bar{S}} \left\{ \mathbb{E} [h^2(\mathbf{s}_0, \bar{s})] + \frac{d_S(\bar{s})}{n} \log_+^2 \left( \frac{n}{d_S(\bar{s})} \right) \right\}.$$

In the above inequalities  $C$  and  $C'$  are universal positive constants.

Define  $R_S(\mathbf{s}_0, n)$  by

$$(9) \quad R_S(\mathbf{s}_0, n) = \inf_{\bar{s} \in \bar{S}} \left\{ \mathbb{E} [h^2(\mathbf{s}_0, \bar{s})] + \frac{d_S(\bar{s})}{n} \log_+^2 \left( \frac{n}{d_S(\bar{s})} \right) \right\}.$$

It follows from (8) that  $R_S(\mathbf{s}_0, n)$  is – up to a universal multiplicative constant – an upper-bound of the Hellinger quadratic risk  $\mathbb{E} [h^2(\mathbf{s}_0, \hat{s})]$  of a  $\rho$ -estimator  $\hat{s}$  on  $S$ . It then remains to compute  $R_S(\mathbf{s}_0, n)$  to deduce (an upper bound of) the rate of convergence of the  $\rho$ -estimator when  $\mathbf{s}_0 \in S$ . Let us now discuss what is new here.

First, in density estimation, our risk bound slightly improves the one of [BB16] in the sense that our variance term involves a smaller exponent on the logarithm. The logarithmic term cannot be avoided in general under our assumptions (see Section 3.3). It is an open question to know whether the power 2 can be replaced by 1. A careful comparison between the two results show that our assumptions on models are more stringent. But, the conclusion is also stronger: the event on which (7) holds true depends on  $S$  through  $(\mathcal{I}_d)_{d \geq 1}$  only. It remains the same when the model  $S$  changes but not the collection  $(\mathcal{I}_d)_{d \geq 1}$ . This property is the keystone of Section 4.1 (see (17)) and at the heart of the developments of subsequent sections.

Second, our study is not restricted to density estimation but encompasses three frameworks. It is noteworthy that the risk bound is of the same form in the three frameworks: only the Hellinger loss depends on the framework. Thereby, we can effortlessly transfer results in density estimation to frameworks 2 and 3. In particular, the properties of “robustness” or “superminimaxity” described in the introduction remain valid in hazard rate and transition intensity estimation. For the sake of illustration, we make the risk bounds explicit when  $S = \mathcal{P}_{\ell,r}$ ,  $S = \mathcal{F}_k$  and  $S = \mathcal{G}_k$  in Sections 3.3 and 3.4.

**3.3. Risks of  $\rho$ -estimators on  $\mathcal{P}_{\ell,r}$ .** When  $S = \mathcal{P}_{\ell,r}$ , Theorem 3 asserts

$$\begin{aligned} R_{\mathcal{P}_{\ell,r}}(\mathbf{s}_0, n) &= \mathbb{E} [h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r})] + \frac{(\ell + 2)(r + 2) \log_+^2(n/((\ell + 2)(r + 2)))}{n} \\ (10) \quad &\leq \mathbb{E} [h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r})] + \frac{6\ell(r + 1) \log_+^2(n/(\ell(r + 1)))}{n}. \end{aligned}$$

When  $\mathbf{s}_0 \notin \mathcal{P}_{\ell,r}$ , but is close to  $\mathcal{P}_{\ell,r}$ ,  $\mathbb{E} [h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r})]$  is an approximation term that may be interpreted as a robustness term. It is, for instance, small when  $\sqrt{\mathbf{s}_0}$  is smooth with bounded support, see [DY90].

When  $\mathbf{s}_0$  does belong to  $\mathcal{P}_{\ell,r}$ ,  $R_{\mathcal{P}_{\ell,r}}(\mathbf{s}_0, n)$  becomes

$$R_{\mathcal{P}_{\ell,r}}(\mathbf{s}_0, n) \leq \frac{6\ell(r + 1) \log_+^2(n/\ell(r + 1))}{n}.$$

This bound is valid in the three frameworks and is almost optimal. Only the power 2 on the logarithm is sub-optimal in general (the optimal risk bound involves a power 1 instead of 2 in framework 1 when  $r = 0$ , see [BM98, BB17]).

**3.4. Risks of  $\rho$ -estimators on  $\mathcal{F}_k$  and  $\mathcal{G}_k$ .** Bounding  $R_S(\mathbf{s}_0, n)$  from above is no more difficult in frameworks 2 and 3 than in density estimation. Thereby, we may easily get upper-bounds in frameworks 2 and 3 from results obtained in the literature in density estimation. More precisely, two bounds on  $R_S(\mathbf{s}_0, n)$  can be deduced from the results of [BB16]: when  $S = \mathcal{F}_k$  and when  $S = \mathcal{G}_k$ . They are given below.

**Corollary 1.** *There exist universal constants  $C, C'$  and a map  $V(\cdot)$  on  $\mathcal{F}_k$  such that for all  $\mathbf{s}_0 \in \mathcal{F}_k$ ,*

$$(11) \quad R_{\mathcal{F}_k}(\mathbf{s}_0, n) \leq C \inf_{\substack{\ell \geq 1 \\ \bar{\mathbf{s}} \in \mathcal{P}_{\ell,0} \cap \mathcal{F}_k}} \left\{ \mathbb{E} [h^2(\mathbf{s}_0, \bar{\mathbf{s}})] + \frac{k + \ell}{n} \log_+^2 \left( \frac{n}{k + \ell} \right) \right\}$$

$$(12) \quad \leq C' \left\{ V(\mathbf{s}_0) \left( \frac{\log^2 n}{n} \right)^{2/3} + \frac{k \log^2 n}{n} \right\}.$$

Moreover, there exist universal constants  $C'', C'''$  and a map  $W(\cdot)$  on  $\mathcal{G}_k$  such that for all  $\mathbf{s}_0 \in \mathcal{G}_k$ ,

$$\begin{aligned} R_{\mathcal{G}_k}(\mathbf{s}_0, n) &\leq C'' \inf_{\substack{\ell \geq 1 \\ \bar{\mathbf{s}} \in \mathcal{P}_{\ell,1, \text{sq.root}} \cap \mathcal{G}_k}} \left\{ \mathbb{E} [h^2(\mathbf{s}_0, \bar{\mathbf{s}})] + \frac{k + \ell}{n} \log_+^2 \left( \frac{n}{k + \ell} \right) \right\} \\ (13) \quad &\leq C''' \left\{ W(\mathbf{s}_0) \left( \frac{\log^2 n}{n} \right)^{4/5} + \frac{k \log^2 n}{n} \right\}. \end{aligned}$$

This corollary gives therefore (an upper-bound of) the rates of convergence of  $\rho$ -estimators on  $\mathcal{F}_k$  and  $\mathcal{G}_k$  in the three frameworks, and in particular, in the two new ones. Up to logarithmic factors, we recover the expected rate  $n^{-1/3}$  for a multimodal function  $\mathbf{s}_0$  (using the loss  $h$ ). Moreover, this rate becomes parametric when  $\mathbf{s}_0 \in \mathcal{F}_k$  is also piecewise constant on  $\ell$  pieces. More generally, the convergence rate we get is faster than  $n^{-1/3}$  if  $\ell$  is allowed to depend on  $n$ , but in such a way that  $\ell n^{-1/3}$  tends to 0 (up to logarithmic factors). The number  $\ell$  as well as the pieces on which  $\mathbf{s}_0$  is constant are possibly unknown to the statistician. This reasoning applies in a similar way to the collection  $\mathcal{G}_k$ .

Like-minded results have been obtained for least squares estimators in the regression setting. We refer to [Zha02, Cha14, BT15, CGS15, GS15, Bel18, CL19]. In isotonic regression, the risk of the

least squares estimator is, in general, bounded from above by a quantity of the order of  $n^{-1/3}$ . This bound becomes of the order of  $\sqrt{(\ell/n) \log(n/\ell)}$  when the regression function is piecewise constant on  $\ell$  pieces. Likewise, in convex regression, the result is better when the target is piecewise affine. This is the same phenomenon as described above for our estimators, although our results involve additional logarithmic factors.

In (12) and (13), the terms  $V(\mathbf{s}_0)$ ,  $W(\mathbf{s}_0)$  measure, in some sense, the “variations” of  $\sqrt{\mathbf{s}_0}$ . To reduce the size of this paper, we propose to make explicit  $V(\cdot)$  only.

Let for  $K \subset \mathbb{R}$ ,

$$M_E(K) = \begin{cases} \mu(K) & \text{in framework 1,} \\ \int_K \mathbb{P}(X \geq t) \mathbb{1}_{[0, +\infty)}(t) dt & \text{in framework 2,} \\ \int_K \mathbb{P}(X_{t-} = 1) \mathbb{1}_{I_{\text{obs}}}(t) dt & \text{in framework 3.} \end{cases}$$

The map  $V(\cdot)$  is then defined for  $\mathbf{s}_0 \in \mathcal{F}_k$  by

$$V(\mathbf{s}_0) = \inf_m \sum_{K \in m} \left[ M_E(K) \left( \sqrt{\sup_{x \in K} \mathbf{s}_0(x)} - \sqrt{\inf_{x \in K} \mathbf{s}_0(x)} \right)^2 \right]^{1/3},$$

where the infimum runs over all collections  $m \in \mathcal{M}_k$  for which  $\mathbf{s}_0 = \sum_{K \in m} s_K \mathbb{1}_K$  where  $s_K$  is monotone on  $K$ .

It can be bounded from above as follows when  $k = 2$  and  $\mathbf{s}_0 \in \mathcal{F}_2$ :

- in framework 1,

$$V(\mathbf{s}_0) \leq 2L_{\text{supp}}^{1/3} (\sup_{x \in \mathbb{R}} \mathbf{s}_0(x))^{1/3},$$

whenever the support of  $\mathbf{s}_0$  is of finite length  $L_{\text{supp}}$ .

- in framework 2,

$$V(\mathbf{s}_0) \leq 2(\mathbb{E}(X))^{1/3} \left( \sup_{x \geq 0} \mathbf{s}_0(x) \right)^{1/3},$$

whenever  $X$  has finite expectation.

- in framework 3,

$$V(\mathbf{s}_0) \leq 2(\mathbb{E}(T_1))^{1/3} \left( \sup_{x \in I_{\text{obs}}} \mathbf{s}_0(x) \right)^{1/3},$$

whenever  $T_1 = \int_{I_{\text{obs}}} \mathbb{1}_{X_{t-}=1} dt$  has finite expectation.

Contrary to density estimation, the size  $L_{\text{supp}}$  of the support of  $\mathbf{s}_0$  is not involved in these upper-bounds in hazard rate and transition intensity estimation.

**3.5. About the risk bounds in hazard rate estimation.** The empirical distance  $h$  has been scarcely used in hazard rate estimation. The only papers we are aware of that deal with this loss are [vdG95, BB09]. But other losses may be considered, and in particular deterministic losses. Since  $\mathbf{s}_0$  is not integrable on  $(0, +\infty)$ , it is not possible to use a loss of the form

$$h_{\text{unif}}^2(\mathbf{s}_0, \hat{s}) = \frac{1}{2} \int_0^\alpha \left( \sqrt{\mathbf{s}_0(t)} - \sqrt{\hat{s}(t)} \right)^2 dt,$$

when  $\alpha = +\infty$ . Setting  $\alpha < +\infty$  amounts to measuring the quality of the estimation on an interval of finite length only. Moreover, the rate of convergence of  $h_{\text{unif}}^2(\mathbf{s}_0, \hat{s})$  depends generally on  $\alpha$  when  $\alpha$  goes to infinity with  $n$ .

In the present paper, we dealt with  $h$ , but we may also consider

$$h_E^2(\mathbf{s}_0, \hat{s}) = \frac{1}{2} \int_0^\infty \left( \sqrt{\mathbf{s}_0(t)} - \sqrt{\hat{s}(t)} \right)^2 \mathbb{P}(X \geq t) dt.$$

The difference between  $h_E$  and  $h$  lies in the fact that  $h_E$  involves the (unknown) survival function  $G(t) = \mathbb{P}(X \geq t)$  of  $X$  whereas  $h$  involves its empirical counterpart  $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \geq t}$ . Note that the quality of the estimation is not measured uniformly on  $(0, +\infty)$  but according to the difficulty of the problem. In particular, the larger  $t$ , the farther  $\hat{s}(t)$  may be from  $\mathbf{s}_0(t)$ .

Let us mention that we may always relate  $h_{unif}$  to  $h_E$  since

$$h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \alpha]}, \hat{s} \mathbb{1}_{[0, \alpha]}) \leq h_{unif}^2(\mathbf{s}_0, \hat{s}) \leq (G(\alpha))^{-1} h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \alpha]}, \hat{s} \mathbb{1}_{[0, \alpha]}),$$

when  $G(\alpha) > 0$ . Likewise, we may relate  $h_E$  to  $h$  as shown below.

**Proposition 4.** *Consider framework 2, suppose  $n \geq 1043$ , that  $G$  is continuous,  $\mathbb{E}[X] < \infty$  and that the density  $f_0$  of  $T$  is square-integrable:  $\int_0^\infty f_0^2 d\mu < \infty$ . Let  $\hat{\alpha}$  be a positive random variable such that*

$$(14) \quad 150 \frac{\log n}{n} \leq G_n(\hat{\alpha}) \leq 151 \frac{\log n}{n}.$$

*There exists a universal constant  $C$  such that, for all estimator  $\hat{s} \in \mathcal{S}$ , the truncated estimator  $\tilde{s}$  defined by*

$$\tilde{s} = \min\{\hat{s}, n^3\} \mathbb{1}_{[0, \hat{\alpha}]}$$

*satisfies*

$$\mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s})] \leq C \left\{ \mathbb{E} [h^2(\mathbf{s}_0, \hat{s})] + \frac{\log n}{n} + \frac{\mathbb{E}[X] + \int_0^\infty f_0^2 d\mu}{n^2} \right\}.$$

Thereby, a truncation argument allows to deduce risk bounds for the deterministic loss  $h_E$  from the ones established with  $h$ . For instance, let  $\hat{s}$  be a  $\rho$ -estimator satisfying the assumptions of Theorem 3. Then, the preceding proposition asserts:

$$\begin{aligned} \mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s})] &\leq C \inf_{\bar{s} \in \bar{S}} \left\{ h_E^2(\mathbf{s}_0, \bar{s}) + \frac{d_S(\bar{s})}{n} \log_+^2 \left( \frac{n}{d_S(\bar{s})} \right) + \frac{\mathbb{E}[X] + \int_0^\infty f_0^2 d\mu}{n^2} \right\} \\ &\leq C \left\{ R_S(\mathbf{s}_0, n) + \frac{\mathbb{E}[X] + \int_0^\infty f_0^2 d\mu}{n^2} \right\}. \end{aligned}$$

Up to a remaining term, and a modification of the multiplicative constant, this bound corresponds to the one we would get for  $\mathbb{E} [h^2(\mathbf{s}_0, \hat{s})]$  and that is written in (8). In particular, the rates of convergence we got for  $S = \mathcal{F}_k$  and  $S = \mathcal{G}_k$  are valid for  $\tilde{s}$  and the Hellinger loss  $h_E$ .

**3.6. Connection with maximum likelihood estimation.** The  $\rho$ -estimation procedure differs from that of maximum likelihood. Nevertheless, the two approaches are very close in some situations, see [BBS17, BB17] for results in density estimation. This point is crucial to reduce the numerical complexity when  $S = \mathcal{P}_{\ell, r}$  (see Section 4).

We define

$$(15) \quad \mathcal{L}(s) = \int_{\mathbb{R}} \log s dN - \int_{\mathbb{R}} s dM \quad \text{for all } s \in \mathcal{S},$$

and call maximum likelihood estimator any estimator maximizing  $\mathcal{L}(\cdot)$  on  $S$ . In the above formula, and throughout the paper, the convention  $\log 0 = -\infty$  is used.

In framework 1, the term  $\int_{\mathbb{R}} s \, dM = \int_{\mathbb{R}} s \, d\mu$  plays the role of a Lagrange term as  $s \in S$  may not be a density. In framework 2,  $\mathcal{L}(s)$  is the usual log likelihood when  $s$  is a hazard rate, up to some terms constant in  $s$ . The same is true for framework 3, see the literature of counting processes *e.g.* equation (3.2) of [Ant89] (using that  $s$  is an Aalen's multiplicative intensity).

We may write

$$T(s, s') = \int_{\mathbb{R}} \tanh\left(\frac{\log s' - \log s}{4}\right) \, dN - \frac{1}{4} \int_{\mathbb{R}} (s' - s) \, dM \quad \text{for all } s, s' \in \mathcal{S}.$$

As  $\tanh(x) \simeq x$  when  $x \simeq 0$ , we deduce that if  $\tilde{s}$  maximizes  $\mathcal{L}(\cdot)$  and  $s' \simeq \tilde{s}$ ,

$$\begin{aligned} T(\tilde{s}, s') &\simeq \frac{1}{4} \left( \int_{\mathbb{R}} \log s' \, dN - \int_{\mathbb{R}} \log \tilde{s} \, dN \right) - \frac{1}{4} \left( \int_{\mathbb{R}} s' \, dM - \int_{\mathbb{R}} \tilde{s} \, dM \right) \\ &\simeq \frac{1}{4} (\mathcal{L}(s') - \mathcal{L}(\tilde{s})). \end{aligned}$$

Thereby,  $T(\tilde{s}, s')$  is likely non-positive. Under suitable properties of  $S$ , this result does not only occur when  $s' \simeq \tilde{s}$ , but also for all  $s' \in S$ , which implies that  $\gamma(\tilde{s}) = 0$ . In particular,  $\tilde{s}$  is a  $\rho$ -estimator.

**Theorem 5.** *Suppose that  $S$  is a convex subset of  $\mathcal{S}$ . Let  $K$  be a subset of  $\mathbb{R}$  such that  $\{x \in \mathbb{R}, s(x) \neq 0\} \subset K$  for all  $s \in S$ . Define*

$$\mathcal{L}_K(s) = \int_K \log s \, dN - \int_K s \, dM \quad \text{for all } s \in S,$$

*and suppose that  $\sup_{s \in S} \mathcal{L}_K(s) \notin \{-\infty, +\infty\}$ .*

*If there exists an estimator  $\tilde{s} \in S$  such that  $\mathcal{L}_K(s) \leq \mathcal{L}_K(\tilde{s})$  for all  $s \in S$ , then  $\gamma(\tilde{s}) = 0$  and  $\tilde{s}$  is a  $\rho$ -estimator. Conversely, assume that there exists a  $\rho$ -estimator  $\hat{s} \in S$  such that  $\gamma(\hat{s}) = 0$ . Then, for all  $s \in S$ ,  $\mathcal{L}_K(s) \leq \mathcal{L}_K(\hat{s})$ , and  $\hat{s}$  maximizes  $\mathcal{L}_K(\cdot)$  over  $S$ .*

When  $K = \mathbb{R}$ ,  $\mathcal{L}_K(\cdot) = \mathcal{L}(\cdot)$ , which means that results on maximum likelihood estimators may be derived from that of  $\rho$ -estimators and vice versa. A similar result for convex sets of densities was obtained by Su Weijie in the context of framework 1 and was recently included in [BB17]. Using sets  $K$  not equal to  $\mathbb{R}$  may be of interest to remove some observations that would make the log likelihood identically equal to  $-\infty$ . In that case, we rather estimate the restriction of  $\mathbf{s}_0$  to  $K$  as illustrated below.

We consider the convex model  $S$  in framework 1 defined by

$$(16) \quad S = \{s \mathbb{1}_{(0, +\infty)}, s \text{ is a non-increasing function of } \mathcal{S} \text{ on } \mathbb{R}\}.$$

When the random variables  $X_i$  are positive, which in particular holds true almost surely if  $\mathbf{s}_0$  does belong to  $S$ , the maximum likelihood estimator exists on  $S$  and is known as the Grenander estimator, see [Gre56]. We deduce from the above theorem with  $K = \mathbb{R}$  that this estimator is, in this case, a  $\rho$ -estimator. When some of the random variables  $X_i$  are non-positive,  $\mathcal{L}(s) = -\infty$  for all  $s \in S$ , and we cannot maximize  $\mathcal{L}(\cdot)$  over  $S$  to design an estimator. However, the  $\rho$ -estimation approach works and still coincides with the maximum likelihood one, up to minor modifications. Indeed, in this case, the preceding theorem can be used with  $K = (0, +\infty)$ . Then,  $\mathcal{L}_K(s)$  takes the form

$$\mathcal{L}_K(s) = \frac{1}{n} \sum_{\substack{i \in \{1, \dots, n\} \\ X_i > 0}} \log s(X_i) - \int_0^\infty s(t) \, dt \quad \text{for all } s \in \mathcal{S}.$$

Let  $\tilde{s}$  be the Grenander estimator based on the random variables  $X_1, \dots, X_n$  that are positive. This estimator is a density and maximizes the map

$$s \mapsto \frac{1}{n_0} \sum_{\substack{i \in \{1, \dots, n\} \\ X_i > 0}} \log s(X_i)$$

over the densities  $s$  of  $S$ , where  $n_0$  is the number of positive random variables among  $X_1, \dots, X_n$ . One can verify that the estimator that maximizes  $\mathcal{L}_K(\cdot)$  over  $S$ , and which is thus a  $\rho$ -estimator on  $S$ , is  $\hat{s} = (n_0/n)\tilde{s}$ . Note that  $\int_{\mathbb{R}} \hat{s} d\mu = n_0/n$ , which means that  $\hat{s}$  is not a density unless all the observations  $X_i$  are positive. This is due to the fact that  $\hat{s}$  here estimates the restriction of  $\mathbf{s}_0$  to  $(0, +\infty)$  (which cannot be a density when some observations  $X_i$  are negative).

Let us mention that a maximum likelihood estimator may not be rate optimal. We refer to Theorem 3 of [BM93] for an example of convex set of densities where this phenomenon occurs. As  $S$  is convex in that example, the maximum likelihood estimator is also a  $\rho$ -estimator. This means that there are unfortunately  $\rho$ -estimators that do not reach the optimal minimax rate of convergence.

It is sometimes convenient to consider models  $S$  of the form  $S = \{f^2, f \in \mathcal{F}\}$  where  $\mathcal{F}$  consists of non-negative functions. The set  $\mathcal{F}$  can then be interpreted as a translation of the knowledge one has on  $\sqrt{\mathbf{s}_0}$ . For instance, if  $\mathcal{F}$  denotes the set of non-negative concave functions on  $[0, +\infty)$  vanishing on  $(-\infty, 0)$ , the assumption  $\mathbf{s}_0 \in S$  means that  $\sqrt{\mathbf{s}_0}$  is concave on  $[0, +\infty)$  with support in  $[0, +\infty)$ . It turns out that the connection between  $\rho$ - and maximum likelihood estimators remains valid when the convexity assumption is put on  $\mathcal{F}$  instead of  $S$ .

**Theorem 6.** *Let  $\mathcal{F}$  be a convex set of non-negative functions such that  $S = \{f^2, f \in \mathcal{F}\}$  is included in  $\mathcal{S}$ . Let  $K$  be a subset of  $\mathbb{R}$  such that  $\{x \in \mathbb{R}, f(x) \neq 0\} \subset K$  for all  $f \in \mathcal{F}$ . Then, if  $\sup_{s \in S} \mathcal{L}_K(s) \notin \{-\infty, +\infty\}$ , the conclusions of Theorem 5 apply to  $S$ : any maximizer  $\tilde{s} \in S$  of  $\mathcal{L}_K(\cdot)$  on  $S$  vanishes  $\gamma(\cdot)$ , and any  $\hat{s} \in S$  vanishing  $\gamma(\cdot)$  maximizes  $\mathcal{L}_K(\cdot)$  over  $S$ .*

#### 4. FROM THEORY TO PRACTICE: ESTIMATOR SELECTION

It is often difficult in practice to find a global minimum of  $\gamma(\cdot)$  and thus to build  $\rho$ -estimators. In particular, we do not know how to construct a  $\rho$ -estimator on the model  $S = \mathcal{P}_{\ell, r}$ . In this section, our goal is to propose an alternative way, more numerically friendly, to define an estimator with similar statistical properties on this model. We then explain how to make the estimator adaptive with respect to  $\ell$ .

**4.1. Maximum likelihood estimation.** We start by considering a much simpler model than  $\mathcal{P}_{\ell, r}$ . More precisely, we consider a partition  $m \in \mathcal{M}_\ell$  and define the collection  $\mathcal{P}_r(m) \subset \mathcal{P}_{\ell, r}$  of piecewise polynomial functions on  $m$ :

$$\mathcal{P}_r(m) = \left\{ \sum_{K \in m} s_K \mathbb{1}_K, \text{ for all } K \in m, s_K \text{ is a polynomial function of degree at most } r, \right. \\ \left. \text{non-negative on } K \right\}.$$

As  $\mathcal{P}_r(m)$  is convex, our procedure is essentially reduced to that of maximum likelihood, which is easier to implement in practice. Elementary computations show:

**Lemma 2.** *Let  $r \geq 0$ ,  $\ell \geq 1$ ,  $m \in \mathcal{M}_\ell$ , and for  $K \in m$ ,*

$$\mathcal{P}_r(K) = \{s \mathbb{1}_K, s \text{ is a polynomial function of degree at most } r \text{ and non-negative on } K\}.$$

*Then,  $\sup_{s \in \mathcal{P}_r(K)} \mathcal{L}_K(s)$  is finite and achieved at a point  $\hat{s}_K$ .*

*Moreover,  $\hat{s}_m = \sum_{K \in m} \hat{s}_K$  is a  $\rho$ -estimator on the model  $S = \mathcal{P}_r(m)$  that vanishes  $\gamma(\cdot)$ . If  $N(\cup_{K \in m} K) = N(\mathbb{R})$ ,  $\hat{s}_m$  maximizes  $\mathcal{L}(\cdot)$  and is therefore also a maximum likelihood estimator.*

It follows from Theorem 3 that there is an event that does not depend on  $m$ , that holds true with probability larger than  $1 - e^{-n\xi}$ , and on which:

$$(17) \quad h^2(\mathbf{s}_0, \hat{s}_m) \leq C \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + \frac{\ell(r+1)}{n} \log_+^2 \left( \frac{n}{\ell(r+1)} \right) + \xi \log_+(1/\xi) \right\}.$$

Our objective in the subsequent sections is to define a good partition, that is a partition  $m$  for which the bias term  $h^2(\mathbf{s}_0, \mathcal{P}_r(m))$  is small.

Remark: the preceding inequality shows that  $\hat{s}_m$  is robust with respect to model misspecification measured in terms of  $h$ . Let us mention that this property is not always true for maximum likelihood estimators, see Section 2.3 of [Bir06] for an example in density estimation. This is not specific to this framework, see Appendix B for an example in hazard rate estimation.

**4.2. Estimator selection.** Let  $(m_\lambda)_{\lambda \in \Lambda}$  be a family of (possibly random) partitions of  $\mathcal{M}_\ell$ . The preceding section explains how to compute a  $\rho$ -estimator  $\hat{s}_{m_\lambda}$  on  $\mathcal{P}_r(m_\lambda)$ . We deduce from Proposition 1 that the (random) model  $S = \{\hat{s}_{m_\lambda}, \lambda \in \Lambda\}$  fulfils Assumption 2 with  $\bar{S} = S$ ,  $d_S(\hat{s}_{m_\lambda}) = (\ell+2)(r+2)$ . Applying our procedure to  $S$  leads to an estimator of the form  $\hat{s} = \hat{s}_{m_{\hat{\lambda}}}$  that satisfies on an event of probability lower bounded by  $1 - e^{-n\xi}$ ,

$$(18) \quad \begin{aligned} h^2(\mathbf{s}_0, \hat{s}_{m_{\hat{\lambda}}}) &\leq C \left\{ \inf_{\lambda \in \Lambda} h^2(\mathbf{s}_0, \hat{s}_{m_\lambda}) + \frac{(r+1)\ell \log_+^2(n/(\ell(r+1)))}{n} + \xi \log_+(1/\xi) \right\}, \\ &\leq C' \left\{ \inf_{\lambda \in \Lambda} h^2(\mathbf{s}_0, \mathcal{P}_r(m_\lambda)) + \frac{(r+1)\ell \log_+^2(n/(\ell(r+1)))}{n} + \xi \log_+(1/\xi) \right\}, \end{aligned}$$

using (17). Here,  $C, C'$  are universal constants.

Instead of dealing with  $\mathcal{P}_{\ell,r}$  — which seems to be very tricky in practice — we may restrict our procedure to the set  $S = \{\hat{s}_{m_\lambda}, \lambda \in \Lambda\}$ . Minimizing our criterion then requires to compute  $T(\hat{s}_{m_\lambda}, \hat{s}_{m_{\lambda'}})$ , for every pair  $(\hat{s}_{m_\lambda}, \hat{s}_{m_{\lambda'}}) \in S^2$ . This is numerically feasible when  $S$  is finite and not too large.

As we see in (18), we should take  $\{m_\lambda, \lambda \in \Lambda\}$  as large as possible to improve on the theoretical performances of the selected estimator. Ideally,  $\{m_\lambda, \lambda \in \Lambda\} = \mathcal{M}_\ell$  to recover the risk bound of a  $\rho$ -estimator  $\hat{s}$  on  $\mathcal{P}_{\ell,r}$ :

$$(19) \quad h^2(\mathbf{s}_0, \hat{s}_{m_{\hat{\lambda}}}) \leq C' \left\{ h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r}) + \frac{(r+1)\ell \log_+^2(n/(\ell(r+1)))}{n} + \xi \log_+(1/\xi) \right\}.$$

There is therefore a trade-off between the theoretical and computational properties of  $\hat{s}_{m_{\hat{\lambda}}}$ . The larger the collection, the better the theoretical properties, but the longer it takes to compute the estimator.

**4.3. Selecting among a special collection of piecewise polynomial estimators.** In this section, we propose to deal with a special but possibly very large collection  $\{\hat{s}_{m_\lambda}, \lambda \in \Lambda\}$  of piecewise polynomial  $\rho$ -estimators. This collection being very rich, we hope to recover a theoretical risk bound akin to (19). Moreover, we propose a criterion  $\gamma_2(\cdot)$  that uses the particular structure of this collection to reduce the numerical complexity.

We consider a (possibly random) collection of distinct random variables  $\{Y_i, i \in \hat{I}\}$  where  $\hat{I}$  is a (possibly random) set such that  $\hat{n} = |\hat{I}| \geq 2$ . Since the random variables  $(Y_i)_{i \in \hat{I}}$  are distinct almost surely, we may order them:  $Y_{(1)} < Y_{(2)} < \dots < Y_{(\hat{n})}$ . We define the collection  $\widehat{\mathcal{M}}$  that gathers all the partitions  $m$  of  $[Y_{(1)}, Y_{(\hat{n})}]$  of the form

$$m = \{[Y_{(1)}, Y_{(n_1)}], (Y_{(n_1)}, Y_{(n_2)}], (Y_{(n_2)}, Y_{(n_3)}], \dots, (Y_{(n_k)}, Y_{(\hat{n})})\},$$

where  $k \geq 0$  and  $1 < n_1 < n_2 < \dots < n_k < \hat{n}$  with the convention that  $m = \{[Y_{(1)}, Y_{(\hat{n})}]\}$  when  $k = 0$ . We set for  $\ell \in \{1, \dots, \hat{n} - 1\}$ ,

$$\widehat{\mathcal{M}}_\ell = \{m \in \widehat{\mathcal{M}}, |m| = \ell\}.$$

Note that  $\widehat{\mathcal{M}}_\ell \subset \mathcal{M}_\ell$ , but  $\widehat{\mathcal{M}}_\ell \neq \mathcal{M}_\ell$ .

We now consider a random variable  $\hat{\ell}$  with values in  $\{1, \dots, \hat{n} - 1\}$ . For each  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ , we define the piecewise polynomial  $\rho$ -estimator  $\hat{s}_m$  on  $\mathcal{P}_r(m)$  as explained in Lemma 2. This construction implies that there exist  $\hat{\ell}$  intervals on which  $\hat{s}_m$  is a polynomial function. Moreover,  $\hat{s}_m$  may have spikes on  $\{Y_{(1)}, Y_{(2)}, \dots, Y_{(\hat{n})}\}$  only. We aim at selecting an estimator among the family  $\{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ .

We define for  $m \in \widehat{\mathcal{M}}$ ,  $K \in m$  and  $m' \in \widehat{\mathcal{M}}$ , the partition  $m' \vee K$  of  $K$  by

$$(20) \quad m' \vee K = \{K' \cap K, K' \in m', K' \cap K \neq \emptyset\}.$$

We consider a positive number  $L$  and define the criterion  $\gamma_2(\cdot)$  for  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$  by

$$(21) \quad \gamma_2(\hat{s}_m) = \sum_{K \in m} \sup_{m' \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ T(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K) - L|m' \vee K| \frac{(r+1) \log_+^2(n/(r+1))}{n} \right\}.$$

The selected estimator is then any estimator  $\hat{s}_{\hat{m}}$  of the collection  $\{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$  minimizing  $\gamma_2(\cdot)$ :

$$(22) \quad \gamma_2(\hat{s}_{\hat{m}}) = \min_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \gamma_2(\hat{s}_m).$$

**Theorem 7.** *There exists a universal constant  $L_0$  such that if  $L \geq L_0$ , any estimator  $\hat{s}_{\hat{m}}$  minimizing (22) satisfies for all  $\xi > 0$ , and probability larger than  $1 - e^{-n\xi}$ ,*

$$(23) \quad h^2(\mathbf{s}_0, \hat{s}_{\hat{m}}) \leq C \left\{ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}.$$

In particular,

$$(24) \quad \mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] \leq C' \mathbb{E} \left[ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} \right].$$

In the above inequalities,  $C$  and  $C'$  are universal constants. Moreover, the event on which (23) holds may be chosen independently of  $(Y_i)_{i \in \hat{I}}$ ,  $\hat{\ell}$ ,  $\hat{n}$ ,  $r$  and the value of  $L$  (when  $L \geq L_0$ ).

This risk bound is very similar to the one (18) obtained when the first criterion  $\gamma(\cdot)$  is minimized on  $S = \{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$  (we only slightly lose on the variance term). The main interest of this procedure compared to the first one lies in its numerical complexity. To avoid a digression, we defer the discussion on numerical aspects to Section 5 and Appendix C.

Let us mention that our procedure is in line with [Sar14] where a numerical complexity reduction work was carried out. More precisely, it is explained there how to select a partition  $m$  from the collection of dyadic partitions of [DY90]. This is very effective to estimate a function under smoothness assumptions (that is, under the assumption that the target belongs to a Besov space). Our aim in this paper is different. Instead, we want to design an estimator with similar statistical properties to that of a  $\rho$ -estimator on  $\mathcal{P}_{\ell,r}$ . This is why we use a much richer collection than in [Sar14].

We now consider framework 1. When  $Y_{(1)} \leq \min_{1 \leq i \leq n} X_i \leq \max_{1 \leq i \leq n} X_i \leq Y_{(\hat{n})}$ ,  $\hat{s}_m$  maximizes  $\mathcal{L}(\cdot)$  and is therefore a maximum likelihood estimator. It is then natural to compare our estimator  $\hat{s}_{\hat{m}}$  to the one  $\hat{s}_{\tilde{m}}$  that maximizes the log likelihood  $\mathcal{L}(\hat{s}_m)$  over  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ . We refer to Section 5 for numerical simulations when  $\{Y_i, i \in \hat{I}\} = \{X_i, i \in \{1, \dots, n\}\}$  and  $r = 0$ . We do not know theoretical results for  $\hat{s}_{\tilde{m}}$ . However, when  $\{Y_i, i \in \hat{I}\}$  is not random, then results concerning  $\hat{s}_{\tilde{m}}$  may be found in the literature. We refer to Theorem 3.2 of [Cas99] (when  $r = 0$ ) and Theorem 2 of [BBM99] (when  $r \geq 0$ ) for upper-bounds of the Hellinger risk in density estimation. Note that they put restrictions either on  $\mathbf{s}_0$ , or on the minimal length of the intervals  $K$  of the partitions  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ . Besides, contrary to ours, their upper-bounds involve the Kullback Leibler divergence.

The bias term in (23) depends on the collection  $\widehat{\mathcal{M}}_{\hat{\ell}}$  and thus on the choice of  $\{Y_i, i \in \hat{I}\}$ . In general, this bias term may be larger than the one we would obtain for a  $\rho$ -estimator on  $\mathcal{P}_{\ell,r}$ . As the  $Y_i$  are allowed to be random, we may choose them according to the data, and in such a way that the bias term becomes comparable to the one we would obtain for the  $\rho$ -estimator, at least when  $r = 0$ .

Suppose that  $\{Y_i, i \in \hat{I}\}$  is rich enough to satisfy

$$(25) \quad N(A) = N(\{Y_i, i \in \hat{I}\} \cap A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

For instance, we may define  $\{Y_i, i \in \hat{I}\}$  as follows:

- in framework 1, we may set  $\hat{I} = \{1, \dots, n\}$ , and for all  $i \in \hat{I}$ ,  $Y_i = X_i$ ,
- in framework 2, suppose that the random variables  $X_i$  are distinct almost surely. Then, we may consider a set  $\hat{I} \subset \{1, \dots, n\}$ , such that  $|\hat{I}| \geq 2$ ,  $\hat{I} \supset \{i \in \{1, \dots, n\}, D_i = 1\}$  and define for all  $i \in \hat{I}$ ,  $Y_i = X_i$ ,
- in framework 3, we may consider a set  $\hat{I} \subset \{1, \dots, n\}$ , such that  $|\hat{I}| \geq 2$  and  $\hat{I} \supset \{i \in \{1, \dots, n\}, T_{1,0}^{(i)} \in I_{\text{obs}}\}$ , and define for all  $i \in \hat{I}$ ,  $Y_i = T_{1,0}^{(i)}$ .

We may show:

**Corollary 2.** *Suppose (25) and that  $\mathbf{s}_0$  vanishes outside the interval  $[0, L_{\text{supp}}]$ . Let  $\mathcal{P}_{\ell,0,[0,L_{\text{supp}}]}$  be the collection of step functions based on partitions  $m$  of  $[0, L_{\text{supp}}]$  belonging to  $\mathcal{M}_{\hat{\ell}}$ .*

There exists a universal constant  $L_0$  such that if  $L \geq L_0$ ,  $r = 0$ , any estimator  $\hat{s}_{\hat{m}}$  minimizing (22) satisfies for all  $\xi > 0$ , and probability larger than  $1 - e^{-n\xi}$ ,

$$(26) \quad h^2(\mathbf{s}_0, \hat{s}_{\hat{m}}) \leq C \left\{ h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0, [0, L_{\text{supp}}]}) + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}.$$

In particular,

$$\mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] \leq C' \mathbb{E} \left[ h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0, [0, L_{\text{supp}}]}) + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} \right].$$

In the above inequalities,  $C$  and  $C'$  are universal constants.

Note that the procedure does not require the knowledge of the support  $[0, L_{\text{supp}}]$  of  $\mathbf{s}_0$ . Moreover, the only difference between  $h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0, [0, L_{\text{supp}}]})$  and  $h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0})$  lies in the fact that the functions  $s \in \mathcal{P}_{\hat{\ell}, 0}$  may be based on partitions of an interval  $K$  different from  $[0, L_{\text{supp}}]$ . In particular,

$$h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}+2, 0, [0, L_{\text{supp}}]}) \leq h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0}) \leq h^2(\mathbf{s}_0, \mathcal{P}_{\hat{\ell}, 0, [0, L_{\text{supp}}]}).$$

We recover, up to slight modifications, the risk bound we would obtain for a  $\rho$ -estimator on  $\mathcal{P}_{\hat{\ell}, 0}$ . The estimator  $\hat{s}_{\hat{m}}$  has however the advantage of being computable in practice when  $n$  and  $\hat{\ell}$  are small enough, see Appendix C. We refer to Section 4.4 below for more general results (in particular to deal with  $r \geq 1$ ).

**4.4. Adaptive piecewise polynomial estimation.** The preceding section explains how to define a piecewise polynomial estimator on  $\hat{\ell}$  pieces. We show in this section that the criterion may be modified to choose  $\hat{\ell}$  from the data, and get an estimator adaptive with respect to  $\hat{\ell}$ .

We define for  $\ell_{\max} \in \{1, \dots, \hat{n} - 1\}$  the collection  $\widehat{\mathcal{M}}_{\leq \ell_{\max}}$  of partitions  $m \in \widehat{\mathcal{M}}$  whose cardinal is at most  $\ell_{\max}$ ,

$$\widehat{\mathcal{M}}_{\leq \ell_{\max}} = \left\{ m \in \widehat{\mathcal{M}}, |m| \leq \ell_{\max} \right\} = \bigcup_{\ell=1}^{\ell_{\max}} \widehat{\mathcal{M}}_{\ell}.$$

We consider a random variable  $\hat{\ell}_{\max}$  with values in  $\{1, \dots, \hat{n} - 1\}$  and aim at selecting an estimator among  $\{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}\}$ .

We consider  $L > 0$  and set for  $m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}$ ,

$$\gamma_3(\hat{s}_m) = \sum_{K \in m} \sup_{m' \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}} \left\{ T(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K) - L|m' \vee K| \frac{(r+1) \log_+^2(n/(r+1))}{n} \right\}.$$

The selected estimator  $\hat{s}_{\hat{m}}$  is any estimator of the family satisfying

$$(27) \quad \gamma_3(\hat{s}_{\hat{m}}) + 2L \frac{(r+1)|\hat{m}| \log_+^2(n/(r+1))}{n} \\ = \inf_{m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}} \left\{ \gamma_3(\hat{s}_m) + 2L(r+1) \frac{|m| \log_+^2(n/(r+1))}{n} \right\}.$$

The numerical complexity of this procedure is similar to that of Section 4.3. Moreover, we prove:

**Theorem 8.** *There exists a universal constant  $L_0$  such that if  $L \geq L_0$ , any estimator  $\hat{s}_{\hat{m}}$  satisfying (27) satisfies for all  $\xi > 0$ , and probability larger than  $1 - e^{-n\xi}$ ,*

$$(28) \quad h^2(\mathbf{s}_0, \hat{s}_{\hat{m}}) \leq C \inf_{m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}} \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + L \frac{(r+1)|m| \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}.$$

In particular,

$$(29) \quad \mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] \leq C' \mathbb{E} \left[ \inf_{m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}} \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + L \frac{(r+1)|m| \log_+^2(n/(r+1))}{n} \right\} \right].$$

In the above inequalities,  $C$  and  $C'$  are universal constants. Moreover, the event on which (28) holds may be chosen independently of  $(Y_i)_{i \in \hat{I}}$ ,  $\hat{\ell}_{\max}$ ,  $\hat{n}$ ,  $r$  and the value of  $L$  (when  $L \geq L_0$ ).

This risk bound improves when  $\hat{\ell}_{\max}$  grows. Moreover, (29) implies

$$\mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] \leq C' \mathbb{E} \left[ \inf_{1 \leq \hat{\ell} \leq \hat{\ell}_{\max}} \left\{ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \{h^2(\mathbf{s}_0, \mathcal{P}_r(m))\} + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} \right\} \right].$$

The right-hand side of this inequality corresponds to the bound (24) achieved by the estimator of Section 4.3 when the choice of  $\hat{\ell}$  is the best possible among  $\{1, \dots, \hat{\ell}_{\max}\}$ .

The quality and the construction of the estimator  $\hat{s}_{\hat{m}}$  still depends on  $\{Y_i, i \in \hat{I}\}$ . However, when  $\hat{\ell}_{\max} = \hat{n} - 1$ , and when  $\{Y_i, i \in \hat{I}\}$  is rich enough, the infimum in (29) can be taken over the infinite collection  $\mathcal{M} = \bigcup_{\ell \geq 1} \mathcal{M}_{\ell}$  (up to a modification of  $C'$ ), as shown below.

**Lemma 3.** *Suppose that  $\{Y_i, i \in \hat{I}\}$  is chosen in such a way that  $N$  satisfies (25). There exists a universal constant  $C$  such that for all  $\xi > 0$  and probability larger than  $1 - e^{-n\xi}$ : for all  $\ell \geq 1$ ,*

$$\inf_{m \in \widehat{\mathcal{M}}_{\leq 2\ell+3}} h^2(\mathbf{s}_0, \hat{s}_m) \leq C \left\{ h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r}) + \frac{\ell(r+1) \log_+^2(n/(\ell(r+1)))}{n} + \xi \log_+(1/\xi) \right\}.$$

Suppose now that  $\hat{\ell}_{\max} = \hat{n} - 1$ , that (25) is satisfied, and that  $L \geq \max\{1, L_0\}$ . We then deduce from  $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}$  and (29),

$$\mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] \leq C' \mathbb{E} \left[ \inf_{m \in \widehat{\mathcal{M}}} \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + L \frac{|m|(r+1) \log_+^2(n/(r+1))}{n} \right\} \right],$$

where  $C'$  is a universal constant. Lemma 3 implies

$$\begin{aligned} \mathbb{E} [h^2(\mathbf{s}_0, \hat{s}_{\hat{m}})] &\leq C'' \mathbb{E} \left[ \inf_{\ell \geq 1} \left\{ h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r}) + L \frac{\ell(r+1) \log_+^2(n/(\ell(r+1)))}{n} \right\} \right], \\ &\leq C'' \inf_{\ell \geq 1} \mathcal{R}(\ell), \end{aligned}$$

where

$$\mathcal{R}(\ell) = \mathbb{E} [h^2(\mathbf{s}_0, \mathcal{P}_{\ell,r})] + L \frac{\ell(r+1) \log_+^2(n/(\ell(r+1)))}{n}.$$

This term  $\mathcal{R}(\ell)$  can be interpreted as an upper-bound of the risk of a  $\rho$ -estimator on  $\mathcal{P}_{\ell,r}$  (up to constants), barely worse than the one given by Theorem 3 and that is written in (19).

## 5. NUMERICAL SIMULATIONS

**5.1. About  $L$ .** We propose in this section a simple way to tune the parameter  $L$  in the procedure of Section 4.3. In theory, any value of  $L$  larger than  $L_0$  is suitable (where  $L_0$  stands for a universal constant). Unfortunately, the value of  $L_0$  is too large to be used in practice, and we do not know the smallest value of  $L_0$  that would make the risk bound valid.

A simple solution to solve the choice of this calibration parameter  $L$  is to consider a collection  $\mathbb{L}$  of such parameters and to select among them. More precisely, we minimize  $\gamma_2(\cdot)$  for each  $L \in \mathbb{L}$  and denote the resulting estimator by  $\hat{s}_{\hat{m}_L}$  to emphasize that it depends on  $L$ . We then pick out an estimator among  $\{\hat{s}_{\hat{m}_L}, L \in \mathbb{L}\}$  by the first selection rule, see Section 4.2.

A  $\rho$ -estimator on the (random) model  $\{\hat{s}_{\hat{m}_L}, L \in \mathbb{L}\}$  is of the form  $\hat{s} = \hat{s}_{\hat{m}_L}$  and satisfies for all  $\xi > 0$  and probability larger than  $1 - e^{-n\xi}$ ,

$$h^2(\mathbf{s}_0, \hat{s}) \leq C \left[ \inf_{L \in \mathbb{L}} \{h^2(\mathbf{s}_0, \hat{s}_{\hat{m}_L})\} + \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right],$$

where  $C$  is a universal constant. If  $\mathbb{L}$  contains at least one number  $L$  larger than  $L_0$ , we derive from (23),

$$h^2(\mathbf{s}_0, \hat{s}) \leq C' \left[ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \{h^2(\mathbf{s}_0, \mathcal{P}_r(m))\} + \min \left( 1, \inf_{\substack{L \in \mathbb{L}, \\ L \geq L_0}} L \right) \frac{\hat{\ell}(r+1) \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right],$$

where  $C'$  is a universal constant.

Thereby, the estimator  $\hat{s}$  no longer depends on the particular choice of  $L$  but rather on the collection  $\mathbb{L}$ . The larger  $\mathbb{L}$ , the better the risk bound. However, the numerical complexity of the whole procedure increases with the size of  $\mathbb{L}$ , and the constant  $C'$  above may be larger than in (23).

**5.2. Results.** We consider framework 1,  $r = 0$ ,  $\ell \in \{1, \dots, n\}$ ,  $\{Y_i, i \in \hat{I}\} = \{X_1, \dots, X_n\}$  and the (random) collection  $\widehat{\mathcal{M}}_{\ell}$  consisting of partitions of  $[X_{(1)}, X_{(n)}]$  of size  $\ell$  defined in Section 4.3. For each  $m \in \widehat{\mathcal{M}}_{\ell}$ , we consider the  $\rho$ - and maximum likelihood estimator  $\hat{s}_m$  on  $\mathcal{P}_0(m)$  defined by

$$\hat{s}_m = \sum_{K \in m} \frac{N(K)}{\mu(K)} \quad \text{with } N(K) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_K(X_i).$$

We carry out in this section a numerical study to compare two selection rules described in Sections 4.3 and 5.1.

- The first procedure is based on the likelihood. We select the partition  $\hat{m}^{(1,\ell)} \in \widehat{\mathcal{M}}_{\ell}$  by maximizing the map

$$m \mapsto \mathcal{L}(\hat{s}_m) = \frac{1}{n} \sum_{i=1}^n \log \hat{s}_m(X_i) \quad \text{over } m \in \widehat{\mathcal{M}}_{\ell}.$$

- The second procedure is based on the  $\rho$ -estimation method. We consider a set  $A$  consisting of 300 equally spaced points over  $[0, 3]$ , and define

$$\mathbb{L} = \left\{ \frac{a}{\log^2 n}, a \in A \right\}.$$

For each  $L \in \mathbb{L}$ , we use the procedure of Section 4.3 specified in (21) and (22) to get a partition  $\hat{m}_L \in \widehat{\mathcal{M}}_\ell$ . We then use the procedure of Section 4.2 to pick out an estimator among  $\{\hat{s}_{\hat{m}_L}, L \in \mathbb{L}\}$  as explained in Section 5.1. This leads to a selected partition of the form  $\hat{m}_{\hat{L}} \in \widehat{\mathcal{M}}_\ell$  that will be denoted in the sequel by  $\hat{m}^{(2,\ell)}$ .

We consider four densities  $\mathbf{s}_0$ :

**Example 1.**  $\mathbf{s}_0$  is the density of a Normal distribution

$$\mathbf{s}_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for all } x \in \mathbb{R}.$$

**Example 2.**  $\mathbf{s}_0$  is the density of a log Normal distribution

$$\mathbf{s}_0(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2} \log^2 x} \mathbb{1}_{(0,+\infty)}(x) \quad \text{for all } x \in \mathbb{R}.$$

**Example 3.**  $\mathbf{s}_0$  is the density of an exponential distribution

$$\mathbf{s}_0(x) = e^{-x} \mathbb{1}_{[0,+\infty)}(x) \quad \text{for all } x \in \mathbb{R}.$$

**Example 4.**  $\mathbf{s}_0$  is the density of a mixture of uniform distributions

$$\mathbf{s}_0(x) = \frac{1}{2} \times 3 \mathbb{1}_{[0,1/3]}(x) + \frac{1}{8} \times 3 \mathbb{1}_{[1/3,2/3]}(x) + \frac{3}{8} \times 3 \mathbb{1}_{[2/3,1]}(x) \quad \text{for all } x \in \mathbb{R}.$$

We simulate  $N_{\text{rep}}$  samples  $(X_1, \dots, X_n)$  according to a density  $\mathbf{s}_0$  defined above, and compute, in each of these samples the two selected estimators. Let, for  $k \in \{1, 2\}$  and  $i \in \{1, \dots, N_{\text{rep}}\}$ ,  $\hat{s}_{\hat{m}^{(k,\ell,i)}}$  be the value of the estimator corresponding to the  $k^{\text{th}}$  procedure and the  $i^{\text{th}}$  sample. We evaluate the quality of the estimators by

$$\widehat{R}(k, \ell) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} h^2(\mathbf{s}_0, \hat{s}_{\hat{m}^{(k,\ell,i)}}).$$

We estimate the probability that the two estimators coincide by

$$\widehat{P}_{\text{equal}}(\ell) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} \mathbb{1}_{\hat{m}^{(2,\ell,i)} = \hat{m}^{(1,\ell,i)}}$$

Results are summarized in Figures 1 (when  $n = 50$ ) and 2 (when  $n = 100$ ).

Numerically, we observe in these examples that the two estimators  $\hat{s}_{\hat{m}^{(1,\ell)}}$  and  $\hat{s}_{\hat{m}^{(2,\ell)}}$  perform similarly. Their risks are close and the estimators may even coincide. In Example 4,  $\mathbf{s}_0$  does belong to  $\mathcal{P}_{3,0}$  and the fractions  $\widehat{R}(2, \ell)/\widehat{R}(1, \ell)$  are very close to 1. In the other examples,  $\mathbf{s}_0$  is not piecewise constant, and the robustness properties of the second procedure may be useful. The fractions  $\widehat{R}(2, \ell)/\widehat{R}(1, \ell)$  suggest indeed that the second procedure improves the risk of the first one by a few percent, at least when the size  $\ell$  of the partitions is well adapted to the underlying density, that is when  $\ell$  corresponds to the smallest values of  $\widehat{R}(1, \ell)$  and  $\widehat{R}(2, \ell)$ .

**Remark.** The fractions  $\widehat{R}(2, \ell)/\widehat{R}(1, \ell)$  are computed with all significant digits and are then rounded.

|   | Ex 1  | Ex 2  | Ex 3  | Ex 4  |   | Ex 1  | Ex 2  | Ex 3  | Ex 4  |
|---|-------|-------|-------|-------|---|-------|-------|-------|-------|
| $\widehat{R}(1, 2)$                           | 0.057 | 0.078 | 0.064 | 0.052 | $\widehat{R}(1, 5)$                           | 0.062 | 0.063 | 0.061 | 0.060 |
| $\widehat{R}(2, 2)$                           | 0.057 | 0.080 | 0.065 | 0.051 | $\widehat{R}(2, 5)$                           | 0.059 | 0.062 | 0.059 | 0.060 |
| $\frac{\widehat{R}(2, 2)}{\widehat{R}(1, 2)}$ | 1.00  | 1.02  | 1.02  | 0.99  | $\frac{\widehat{R}(2, 5)}{\widehat{R}(1, 5)}$ | 0.95  | 0.98  | 0.98  | 1.00  |
| $\widehat{P}_{equal}(2)$                      | 0.76  | 0.75  | 0.80  | 0.78  | $\widehat{P}_{equal}(5)$                      | 0.27  | 0.33  | 0.32  | 0.39  |
| $\widehat{R}(1, 3)$                           | 0.052 | 0.056 | 0.053 | 0.048 | $\widehat{R}(1, 6)$                           | 0.067 | 0.068 | 0.066 | 0.065 |
| $\widehat{R}(2, 3)$                           | 0.047 | 0.055 | 0.052 | 0.047 | $\widehat{R}(2, 6)$                           | 0.065 | 0.067 | 0.065 | 0.065 |
| $\frac{\widehat{R}(2, 3)}{\widehat{R}(1, 3)}$ | 0.91  | 0.98  | 0.97  | 0.99  | $\frac{\widehat{R}(2, 6)}{\widehat{R}(1, 6)}$ | 0.97  | 0.99  | 0.99  | 1.00  |
| $\widehat{P}_{equal}(3)$                      | 0.63  | 0.64  | 0.66  | 0.57  | $\widehat{P}_{equal}(6)$                      | 0.28  | 0.33  | 0.33  | 0.37  |
| $\widehat{R}(1, 4)$                           | 0.057 | 0.058 | 0.056 | 0.054 | $\widehat{R}(1, 7)$                           | 0.071 | 0.072 | 0.071 | 0.070 |
| $\widehat{R}(2, 4)$                           | 0.052 | 0.055 | 0.053 | 0.053 | $\widehat{R}(2, 7)$                           | 0.070 | 0.072 | 0.070 | 0.071 |
| $\frac{\widehat{R}(2, 4)}{\widehat{R}(1, 4)}$ | 0.92  | 0.94  | 0.95  | 0.98  | $\frac{\widehat{R}(2, 7)}{\widehat{R}(1, 7)}$ | 0.99  | 1.00  | 1.00  | 1.00  |
| $\widehat{P}_{equal}(4)$                      | 0.32  | 0.40  | 0.40  | 0.43  | $\widehat{P}_{equal}(7)$                      | 0.32  | 0.36  | 0.35  | 0.41  |

FIGURE 1. Results for simulated data with  $n = 50$ ,  $N_{\text{rep}} = 10000$ .

|   | Ex 1  | Ex 2  | Ex 3  | Ex 4  |   | Ex 1  | Ex 2  | Ex 3  | Ex 4  |
|---|-------|-------|-------|-------|---|-------|-------|-------|-------|
| $\widehat{R}(1, 2)$                           | 0.055 | 0.074 | 0.056 | 0.035 | $\widehat{R}(1, 5)$                           | 0.038 | 0.038 | 0.037 | 0.033 |
| $\widehat{R}(2, 2)$                           | 0.056 | 0.076 | 0.057 | 0.034 | $\widehat{R}(2, 5)$                           | 0.035 | 0.034 | 0.035 | 0.033 |
| $\frac{\widehat{R}(2, 2)}{\widehat{R}(1, 2)}$ | 1.03  | 1.02  | 1.02  | 0.98  | $\frac{\widehat{R}(2, 5)}{\widehat{R}(1, 5)}$ | 0.92  | 0.94  | 0.95  | 1.00  |
| $\widehat{P}_{equal}(2)$                      | 0.63  | 0.60  | 0.70  | 0.80  | $\widehat{P}_{equal}(5)$                      | 0.15  | 0.18  | 0.17  | 0.23  |
| $\widehat{R}(1, 3)$                           | 0.034 | 0.042 | 0.037 | 0.023 | $\widehat{R}(1, 6)$                           | 0.041 | 0.040 | 0.039 | 0.037 |
| $\widehat{R}(2, 3)$                           | 0.033 | 0.042 | 0.036 | 0.024 | $\widehat{R}(2, 6)$                           | 0.039 | 0.040 | 0.038 | 0.037 |
| $\frac{\widehat{R}(2, 3)}{\widehat{R}(1, 3)}$ | 0.96  | 1.00  | 0.98  | 1.01  | $\frac{\widehat{R}(2, 6)}{\widehat{R}(1, 6)}$ | 0.95  | 0.97  | 0.98  | 1.00  |
| $\widehat{P}_{equal}(3)$                      | 0.71  | 0.63  | 0.63  | 0.57  | $\widehat{P}_{equal}(6)$                      | 0.10  | 0.15  | 0.11  | 0.19  |
| $\widehat{R}(1, 4)$                           | 0.036 | 0.035 | 0.034 | 0.028 | $\widehat{R}(1, 7)$                           | 0.044 | 0.043 | 0.043 | 0.40  |
| $\widehat{R}(2, 4)$                           | 0.032 | 0.034 | 0.032 | 0.028 | $\widehat{R}(2, 7)$                           | 0.043 | 0.043 | 0.042 | 0.40  |
| $\frac{\widehat{R}(2, 4)}{\widehat{R}(1, 4)}$ | 0.90  | 0.96  | 0.94  | 0.98  | $\frac{\widehat{R}(2, 7)}{\widehat{R}(1, 7)}$ | 0.96  | 0.99  | 0.98  | 1.00  |
| $\widehat{P}_{equal}(4)$                      | 0.29  | 0.39  | 0.35  | 0.33  | $\widehat{P}_{equal}(7)$                      | 0.09  | 0.11  | 0.11  | 0.16  |

FIGURE 2. Results for simulated data with  $n = 100$ ,  $N_{\text{rep}} = 1000$ .

## 6. PROOFS

6.1. **Proof of Lemma 1.** Let  $\sqrt{q} = (\sqrt{s} + \sqrt{s'})/2$  and

$$K = \{t \in \mathbb{R}, s(t) \neq 0 \text{ or } s'(t) \neq 0\}.$$

Then,

$$\frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} (\sqrt{s_0} - \sqrt{q})^2 dM = \frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} s_0 dM + \frac{1}{2} \int_K (\sqrt{s'} - \sqrt{s}) \sqrt{q} dM$$

$$- \int_K \left( \sqrt{s'} - \sqrt{s} \right) \sqrt{\mathbf{s}_0} \, dM.$$

Note that

$$h^2(\mathbf{s}_0, s') - h^2(\mathbf{s}_0, s) = \frac{1}{2} \int_K (s' - s) \, dM + \int_K \sqrt{\mathbf{s}_0} \left( \sqrt{s} - \sqrt{s'} \right) \, dM.$$

Therefore,

$$\begin{aligned} \frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} (\sqrt{\mathbf{s}_0} - \sqrt{q})^2 \, dM &= \frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} \mathbf{s}_0 \, dM + \frac{1}{2} \int_K \left( \sqrt{s'} - \sqrt{s} \right) \sqrt{q} \, dM \\ &\quad - \frac{1}{2} \int_K (s' - s) \, dM + h^2(\mathbf{s}_0, s') - h^2(\mathbf{s}_0, s) \\ (30) \qquad \qquad \qquad &= T_E(s, s') + h^2(\mathbf{s}_0, s') - h^2(\mathbf{s}_0, s). \end{aligned}$$

Now,

$$\begin{aligned} \frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} (\sqrt{\mathbf{s}_0} - \sqrt{q})^2 \, dM &= \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{s} + \sqrt{s'}} \left( \sqrt{\mathbf{s}_0} - \frac{\sqrt{s} + \sqrt{s'}}{2} \right)^2 \, dM \\ &\leq \int_K \left( \sqrt{\mathbf{s}_0} - \frac{\sqrt{s} + \sqrt{s'}}{2} \right)^2 \, dM \\ &\leq \frac{1}{4} \int_K \left( (\sqrt{\mathbf{s}_0} - \sqrt{s}) + (\sqrt{\mathbf{s}_0} - \sqrt{s'}) \right)^2 \, dM. \end{aligned}$$

By using the inequality  $(x + y)^2 \leq (1 + \alpha)x^2 + (1 + \alpha^{-1})y^2$ ,

$$\begin{aligned} \frac{1}{2} \int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} (\sqrt{\mathbf{s}_0} - \sqrt{q})^2 \, dM &\leq \frac{1 + \alpha}{4} \int_K (\sqrt{\mathbf{s}_0} - \sqrt{s})^2 \, dM + \frac{1 + \alpha^{-1}}{4} \int_K (\sqrt{\mathbf{s}_0} - \sqrt{s'})^2 \, dM \\ &\leq \frac{1 + \alpha}{2} h^2(\mathbf{s}_0, s) + \frac{1 + \alpha^{-1}}{2} h^2(\mathbf{s}_0, s'). \end{aligned}$$

We now put this inequality into (30) to get

$$T_E(s, s') \leq \frac{3 + \alpha}{2} h^2(\mathbf{s}_0, s) - \frac{1 - \alpha^{-1}}{2} h^2(\mathbf{s}_0, s').$$

The right-hand side of (2) follows from this inequality with  $\alpha = 3$ . As to the left-hand side, note that we also have (setting  $\alpha = 3$ , and exchanging the role of  $s$  and  $s'$ ),

$$T_E(s', s) \leq 3h^2(\mathbf{s}_0, s') - \frac{1}{3}h^2(\mathbf{s}_0, s).$$

Yet,  $T_E(s, s') = -T_E(s', s)$  and hence  $T_E(s, s') \geq \frac{1}{3}h^2(\mathbf{s}_0, s) - 3h^2(\mathbf{s}_0, s')$  as wished.  $\square$

**6.2. Proof of Proposition 1 for  $S = \mathcal{P}_{\ell, r}$ .** Let  $s, \bar{s} \in \mathcal{P}_{\ell, r}$ . There exist two partitions  $m_1, m_2$  of  $\mathbb{R}$  into intervals such that  $|m_1| = \ell + 2$  and  $|m_2| = \ell + 2$  and such that  $s$  (respectively  $\bar{s}$ ) is polynomial on each element  $K_1 \in m_1$  (respectively  $K_2 \in m_2$ ). Let

$$m = \{K_1 \cap K_2, (K_1, K_2) \in m_1 \times m_2, K_1 \cap K_2 \neq \emptyset\}.$$

Then,  $m$  is a partition of  $\mathbb{R}$  into intervals such that  $|m| \leq |m_1| + |m_2| \leq 2\ell + 4$ . Moreover, we may write  $s$  and  $\bar{s}$  as

$$s = \sum_{K \in m} s_K \mathbb{1}_K \quad \text{and} \quad \bar{s} = \sum_{K \in m} \bar{s}_K \mathbb{1}_K,$$

where  $s_K$  and  $\bar{s}_K$  are non-negative polynomial functions on  $K$  of degree at most  $r$ . Let  $P_K = s_K - u\bar{s}_K$ . Now,

$$\{t \in \mathbb{R}, s(t) > u\bar{s}(t)\} = \bigcup_{K \in m} \{t \in K, P_K(t) > 0\}.$$

Let  $\mathcal{Z}$  be the set gathering the zeros of  $P_K$ . If  $\mathcal{Z} = \emptyset$ , then  $P_K$  is either positive, or negative on  $\mathbb{R}$  and the set  $\{t \in K, P_K(t) > 0\}$  is either empty or the interval  $K$ . If  $\mathcal{Z} = \mathbb{R}$ , then  $P_K = 0$  and  $\{t \in K, P_K(t) > 0\} = \emptyset$ . Suppose now that  $\mathcal{Z} \neq \emptyset$  and  $\mathcal{Z} \neq \mathbb{R}$ . We may write  $\mathcal{Z} = \{b_1, \dots, b_k\}$  with  $b_1 < b_2 < \dots < b_k$  and  $k \leq r$ . We set  $b_0 = -\infty$  and  $b_{k+1} = +\infty$ . For all  $j \in \{0, \dots, k\}$ ,  $P_K$  is either positive or negative on  $(b_j, b_{j+1})$ , and its sign changes with  $j$ . Therefore, the set  $\{t \in K, P_K(t) > 0\}$  is a union of at most  $k/2 + 1$  intervals.

Finally, for all  $K \in m$ ,  $\{t \in K, P_K(t) > 0\}$  is always a union of at most  $r/2 + 1$  intervals, which implies that  $\{t \in \mathbb{R}, s(t) > u\bar{s}(t)\}$  is a union of at most  $(r/2 + 1)(2\ell + 4)$  intervals.  $\square$

**6.3. Proof of Theorem 3.** Let for  $d \geq 1$ ,

$$(31) \quad \vartheta(d) = \frac{d}{n} \log_+^2 \left( \frac{n}{d} \right).$$

We need to prove that there exist a universal constant  $C$  and an event  $\Omega_\xi$  such that  $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$  and on which any  $\rho$ -estimator  $\hat{s}$  on  $S$  satisfies

$$(32) \quad h^2(\mathbf{s}_0, \hat{s}) \leq C \inf_{\bar{s} \in \bar{S}} \{h^2(\mathbf{s}_0, \bar{s}) + \vartheta(d_S(\bar{s})) + \xi \log_+(1/\xi)\}.$$

We introduce the following notations. We define for all odd integer  $d \geq 3$ ,

$$\mathcal{J}_d = \left\{ \bigcap_{r=1}^{\infty} A_r, (A_r)_{r \geq 1} \text{ is a non-increasing sequence of } \mathcal{I}_{(d-1)/2} \right\},$$

$$\bar{\mathcal{J}}_d = \mathcal{J}_d \cup \{\mathbb{R} \setminus A, A \in \mathcal{J}_d\}.$$

Let  $s, s' \in \mathcal{S}$ . Suppose that there exists  $d \geq 1$  such that for all  $u > 0$ , the set  $\{t \in \mathbb{R}, s'(t) > us(t)\}$  belongs to  $\mathcal{I}_d$ . Then,  $d_{s'}(s)$  stands for any number  $d$  such that

$$\{t \in \mathbb{R}, s'(t) > us(t)\}$$

belongs to  $\mathcal{I}_d$  (for all  $u > 0$ ). If the preceding assumption does not hold, we set  $d_{s'}(s) = +\infty$ . We define for all odd integers  $d \geq 3$ ,

$$\mathcal{G}_d = \{\psi(s'/s), s, s' \in \mathcal{S}, d_{s'}(s) = (d-1)/2\}.$$

We will apply Theorem 11 (in Appendix A) to the class  $\mathcal{F} = \mathcal{G}_d$ . We begin with the following elementary claim:

**Claim 1.** *The functions  $f \in \mathcal{G}_d$  satisfy  $|f| \leq 1$ . Moreover, the collection*

$$(33) \quad \mathcal{A} = \{\{t \in \mathbb{R}, f_+(t) > u\}, f \in \mathcal{G}_d, u \in (0, 1)\} \cup \{\{t \in \mathbb{R}, f_-(t) > u\}, f \in \mathcal{G}_d, u \in (0, 1)\}$$

*is included in  $\bar{\mathcal{J}}_d$ .*

*Proof.* Let  $f \in \mathcal{G}_d$  written as  $f = \psi(s'/s)$ . Then,

$$\begin{aligned} \{t \in \mathbb{R}, f_+(t) > u\} &= \{t \in \mathbb{R}, \psi_+(s'(t)/s(t)) > u\} \\ &= \{t \in \mathbb{R}, s(t) \neq 0, \psi_+(s'(t)/s(t)) > u\} \cup \{t \in \mathbb{R}, s(t) = 0, s'(t) > 0\} \\ &= \{t \in \mathbb{R}, s(t) \neq 0, s'(t) > vs(t)\} \cup \{t \in \mathbb{R}, s(t) = 0, s'(t) > 0\}, \end{aligned}$$

where  $v = \psi^{-1}(u) \in (0, +\infty)$ . Therefore,

$$\{t \in \mathbb{R}, f_+(t) > u\} = \{t \in \mathbb{R}, s'(t) > vs(t)\},$$

belongs to  $\mathcal{I}_{(d-1)/2} \subset \bar{\mathcal{J}}_d$ .

Now, note that  $\psi_-(x) = \psi_+(1/x)$ . Hence,

$$\{t \in \mathbb{R}, f_-(t) > u\} = \{t \in \mathbb{R}, \psi_+(s(t)/s'(t)) > u\}.$$

By exchanging the role of  $s$  and  $s'$  in the above computations, we derive

$$\begin{aligned} \{t \in \mathbb{R}, f_-(t) > u\} &= \{t \in \mathbb{R}, s(t) > vs'(t)\} \\ &= \{t \in \mathbb{R}, s'(t) < (1/v)s(t)\}. \end{aligned}$$

Now, for all  $r \geq 1$ ,

$$\{t \in \mathbb{R}, s'(t) > (1 - 1/(2r))(1/v)s(t)\} \in \mathcal{I}_{(d-1)/2}.$$

Therefore,

$$\{t \in \mathbb{R}, s'(t) \geq (1/v)s(t)\} = \bigcap_{r=1}^{\infty} \{t \in \mathbb{R}, s'(t) > (1 - 1/(2r))(1/v)s(t)\}$$

belongs to  $\mathcal{J}_d$  and  $\{t \in \mathbb{R}, f_-(t) > u\} = \mathbb{R} \setminus \{t \in \mathbb{R}, s'(t) \geq (1/v)s(t)\}$  belongs to  $\bar{\mathcal{J}}_d$ .  $\square$

**Claim 2.** *The collection  $\bar{\mathcal{J}}_d$  is Vapnik-Chervonenkis with dimension at most  $2d - 1 \leq 2d$ . Moreover, in framework 3, each set  $A \in \bar{\mathcal{J}}_d$  is a union of at most  $(d + 1)/2 \leq d$  intervals.*

*Proof.* Let  $t_1, \dots, t_{2n} \in \mathbb{R}$  and a non-increasing sequence  $(A_r)_{r \geq 1}$ . Then,

$$\left| \bigcap_{r=1}^{\infty} (\{t_1, \dots, t_{2n}\} \cap A_r) \right| = \lim_{r \rightarrow +\infty} |\{t_1, \dots, t_{2n}\} \cap A_r|.$$

The non-increasing sequence  $(|\{t_1, \dots, t_{2n}\} \cap A_r|)_{r \geq 1}$  consists of integers. Therefore, there exists  $r_0$  such that  $|\{t_1, \dots, t_{2n}\} \cap A_r| = |\{t_1, \dots, t_{2n}\} \cap A_{r_0}|$  for all  $r \geq r_0$ . Hence,

$$\bigcap_{r=1}^{\infty} (\{t_1, \dots, t_{2n}\} \cap A_r) = \{t_1, \dots, t_{2n}\} \cap A_{r_0}.$$

This implies that

$$\{\{t_1, \dots, t_{2n}\} \cap A, A \in \mathcal{J}_d\} = \{\{t_1, \dots, t_{2n}\} \cap A, A \in \mathcal{I}_{(d-1)/2}\}$$

and  $S_{\mathcal{J}_d}(2n) = S_{\mathcal{I}_{(d-1)/2}}(2n)$ . Therefore,  $\mathcal{J}_d$  is Vapnik-Chervonenkis with dimension at most  $d - 1$ . We deduce that  $\bar{\mathcal{J}}_d$  is Vapnik-Chervonenkis with dimension at most  $2(d - 1) + 1 \leq 2d$ .

The two following elementary results show that each set  $A \in \bar{\mathcal{J}}_d$  is a union of at most  $(d + 1)/2$  intervals in framework 3:

- For all union  $A$  of at most  $(d - 1)/2$  intervals,  $\mathbb{R} \setminus A$  is a union of at most  $(d + 1)/2$  intervals.

- For all non-increasing sequence  $(A_r)_{r \geq 1}$  consisting of unions of at most  $(d-1)/2$  intervals,  $\bigcap_{r=1}^{\infty} A_r$  is a union of at most  $(d-1)/2$  intervals.

□

**Claim 3.** For all  $s, s' \in \mathcal{S}$ ,

$$\int_{\mathbb{R}} \psi^2(s'/s) \mathbf{s}_0 \, dM \leq 4 (h^2(\mathbf{s}_0, s) + h^2(\mathbf{s}_0, s')).$$

*Proof of Claim 3.* The proof of this lemma follows from some computations as in Section 8.4 of [Bar11] (see also Proposition 3 of [BB17]). Let  $\sqrt{q} = (\sqrt{s} + \sqrt{s'})/2$  and

$$K = \{t \in \mathbb{R}, s(t) \neq 0 \text{ or } s'(t) \neq 0\}.$$

Then,

$$\begin{aligned} \int_{\mathbb{R}} \psi^2\left(\frac{s'}{s}\right) \mathbf{s}_0 \, dM &= \int_K \psi^2\left(\frac{s'}{s}\right) \mathbf{s}_0 \, dM \\ &= \frac{1}{4} \int_K \left(\sqrt{s'} - \sqrt{s}\right)^2 \frac{\mathbf{s}_0}{q} \, dM \\ &= \frac{1}{4} \int_K \left(\sqrt{s'} - \sqrt{s}\right)^2 \left(\sqrt{\frac{\mathbf{s}_0}{q}} - 1 + 1\right)^2 \, dM \\ &\leq \frac{1}{2} \int_K \left(\sqrt{s'} - \sqrt{s}\right)^2 \left(\sqrt{\frac{\mathbf{s}_0}{q}} - 1\right)^2 \, dM + \frac{1}{2} \int_K \left(\sqrt{s'} - \sqrt{s}\right)^2 \, dM \\ &\leq \frac{1}{2} \int_K \frac{\left(\sqrt{s'} - \sqrt{s}\right)^2}{q} (\sqrt{\mathbf{s}_0} - \sqrt{q})^2 \, dM + h^2(s, s') \\ &\leq 2 \int_K (\sqrt{\mathbf{s}_0} - \sqrt{q})^2 \, dM + h^2(s, s') \\ &\leq \frac{1}{2} \int_K \left((\sqrt{\mathbf{s}_0} - \sqrt{s}) + (\sqrt{\mathbf{s}_0} - \sqrt{s'})\right)^2 \, dM + h^2(s, s') \\ &\leq \int_K (\sqrt{\mathbf{s}_0} - \sqrt{s})^2 \, dM + \int_K (\sqrt{\mathbf{s}_0} - \sqrt{s'})^2 \, dM + h^2(s, s') \\ &\leq 2h^2(\mathbf{s}_0, s) + 2h^2(\mathbf{s}_0, s') + h^2(s, s'). \end{aligned}$$

We complete the proof by using  $h^2(s, s') \leq 2h^2(\mathbf{s}_0, s) + 2h^2(\mathbf{s}_0, s')$ . □

The lemma below is at the core of the proof of Theorem 3.

**Lemma 4.** For all  $\xi > 0$ , there exists an event  $\Omega_\xi$  such that  $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$  and on which: for all  $\varepsilon \in (0, 1/12)$ ,  $s, s' \in \mathcal{S}$ ,

$$(34) \quad T(s, s') \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s') + c_1 \min\{\vartheta(d_{s'}(s)), \vartheta(d_s(s'))\} + c_2 \xi \log_+(1/\xi).$$

In the above inequality,  $c_1$  and  $c_2$  only depend on  $\varepsilon$  and the convention  $\vartheta(+\infty) = +\infty$  is used.

*Proof.* Let  $d \geq 3$  be an odd integer. Claims 1 and 2 show that the assumptions of Theorem 11 in Appendix A are satisfied with  $\mathcal{F} = \mathcal{G}_d$ . Therefore, there exists for all  $\xi > 0$  an event  $\Omega_\xi(d)$  such that

$\mathbb{P} [\Omega_\xi(d)] \geq 1 - e^{-n\xi}$  and on which: for all  $\varepsilon > 0$ ,  $f \in \mathcal{G}_d$  of the form  $f = \psi(s'/s)$ , with  $s, s' \in \mathcal{S}$ ,

$$|Z(f)| \leq \varepsilon v(f) + c [\vartheta(2d_{s'}(s) + 1) + \xi \log_+(1/\xi)].$$

In this inequality,  $c$  only depends on  $\varepsilon$ . Let  $\Omega_\xi = \bigcap_{\substack{d \text{ odd} \\ d \geq 3}} \Omega_{\xi+(2 \log(1+d))/n}(d)$ . Then,

$$\mathbb{P} [(\Omega_\xi)^c] \leq \sum_{\substack{d \text{ odd} \\ d \geq 3}} \mathbb{P} [(\Omega_{\xi+(2 \log(1+d))/n}(d))^c] \leq \sum_{d=1}^{\infty} \frac{e^{-n\xi}}{(1+d)^2} \leq e^{-n\xi}.$$

Moreover, on  $\Omega_\xi$ : for all  $s, s' \in \mathcal{S}$ ,  $f = \psi(s'/s)$  such that  $d_{s'}(s) < \infty$ ,

$$\begin{aligned} |Z(f)| &\leq \varepsilon v(f) + c \vartheta(2d_{s'}(s) + 1) + c \left[ \left( \xi + \frac{2 \log(2 + 2d_{s'}(s))}{n} \right) \log_+ \left( \frac{1}{\xi + \frac{2 \log(2 + 2d_{s'}(s))}{n}} \right) \right] \\ &\leq \varepsilon v(f) + c \vartheta(2d_{s'}(s) + 1) + \frac{2c \log(2 + 2d_{s'}(s))}{n} \log_+ \left( \frac{n}{2 \log(2 + 2d_{s'}(s))} \right) + c \xi \log_+(1/\xi) \\ &\leq \varepsilon v(f) + c' \vartheta(d_{s'}(s)) + c \xi \log_+(1/\xi), \end{aligned}$$

where  $c'$  only depends on  $\varepsilon$ . This last inequality remains true when  $d_{s'}(s) = \infty$  using the convention  $\vartheta(+\infty) = +\infty$ . Moreover, as  $|Z(-f)| = |Z(f)|$ ,  $v(-f) = v(f)$ ,  $\psi(s/s') = -\psi(s'/s)$ , we may exchange the role of  $s$  and  $s'$  in the preceding inequality to get on  $\Omega_\xi$ :

$$(35) \quad |Z(f)| \leq \varepsilon v(f) + c' \min \{ \vartheta(d_s(s')), \vartheta(d_{s'}(s)) \} + c \xi \log_+(1/\xi).$$

Now, it follows from (2) that for all  $s, s' \in \mathcal{S}$ ,

$$(36) \quad T(s, s') \leq 3h^2(\mathbf{s}_0, s) - \frac{1}{3}h^2(\mathbf{s}_0, s') + Z(\psi(s'/s)).$$

Therefore, we deduce from Claim 3 and from (35) that on  $\Omega_\xi$ : for all  $s, s' \in \mathcal{S}$ ,

$$T(s, s') \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s') + c' \min \{ \vartheta(d_s(s')), \vartheta(d_{s'}(s)) \} + c \xi \log_+(1/\xi),$$

which proves (34) with  $c_1 = c'$  and  $c_2 = c$ .  $\square$

We now finish the proof of Theorem 3. Lemma 4 implies that on  $\Omega_\xi$ : for all  $s, s' \in S$ ,

$$(37) \quad T(s, s') \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s') + c_1 \vartheta(d_{s'}(s)) + c_2 \xi \log_+(1/\xi).$$

Thus, for all  $s \in S$ ,

$$(38) \quad \gamma(s) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, S) + c_1 \sup_{s' \in S} \vartheta(d_{s'}(s)) + c_2 \xi \log_+(1/\xi).$$

By using  $T(s, s') = -T(s', s)$ , we deduce from (37) that for all  $s, s' \in S$ ,

$$\frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s') - (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - c_1 \vartheta(d_{s'}(s)) - c_2 \xi \log_+(1/\xi) \leq T(s', s).$$

Any  $\rho$ -estimator  $\hat{s}$  satisfies on  $\Omega_\xi$ : for all  $s \in S$ ,

$$\begin{aligned} (39) \quad \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, \hat{s}) - (3 + 4\varepsilon)h^2(\mathbf{s}_0, s') - c_1 \vartheta(d_{\hat{s}}(s)) - c_2 \xi \log_+(1/\xi) &\leq T(\hat{s}, s) \\ &\leq \gamma(\hat{s}) \\ &\leq \gamma(s) + 1/n. \end{aligned}$$

Using now (38) and  $1/n \leq \sup_{s' \in S} \vartheta(d_{s'}(s))$ , we deduce when  $\varepsilon \in (0, 1/12)$ ,

$$(40) \quad h^2(\mathbf{s}_0, \hat{s}) \leq \inf_{s \in \bar{S}} \left\{ c_{1,\varepsilon} h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, S) + c_{2,\varepsilon} \sup_{s' \in S} \vartheta(d_{s'}(s)) + c_{2,\varepsilon} \xi \log_+(1/\xi) \right\},$$

with  $c_{1,\varepsilon} = 24(3 + 4\varepsilon)/(4 - 3\varepsilon)$ , and with  $c_{2,\varepsilon}$  depending only on  $\varepsilon$ .

When  $s \in \bar{S}$  and  $s' \in S$ , Assumption 2 says that  $d_{s'}(s)$  may be defined by  $d_{s'}(s) = d_S(s)$ . Therefore, (40) becomes

$$h^2(\mathbf{s}_0, \hat{s}) \leq \inf_{s \in \bar{S}} \left\{ c_{1,\varepsilon} h^2(\mathbf{s}_0, s) - h^2(\mathbf{s}_0, S) + c_{2,\varepsilon} \vartheta(d_S(s)) + c_{2,\varepsilon} \xi \log_+(1/\xi) \right\},$$

and it remains to choose  $\varepsilon$  arbitrarily in  $(0, 1/12)$  to prove the theorem.  $\square$

**6.4. Proof of Proposition 4.** We begin by proving the following lemma.

**Lemma 5.** *Consider framework 2, and suppose that  $G$  is continuous. Let  $\xi > 0$  and suppose that there exists a positive random variable  $\hat{\alpha}_\xi$  satisfying*

$$(41) \quad (70 + 16\xi) \frac{\log n}{n} \leq G_n(\hat{\alpha}_\xi) \leq (71 + 16\xi) \frac{\log n}{n}.$$

*There exists an event which holds true with probability larger than  $1 - n^{-\xi}$  and on which: for all  $t \in [0, \hat{\alpha}_\xi]$ ,  $G_n(t) \leq 4G(t)$  and  $G(t) \leq (9/4)G_n(t)$ . Moreover, for all estimator  $\hat{s} \in \mathcal{S}$ ,*

$$(42) \quad h_E^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) \leq (9/4)h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + (80 + 18\xi) \frac{\log n}{n}$$

$$(43) \quad h^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) \leq 4h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + (321 + 73\xi) \frac{\log n}{n}.$$

*Proof of Lemma 5.* We need the celebrated Vapnick-Chervonenkis inequalities for relative deviations (Theorem 14 in Section D.3). We apply this result with  $\mathcal{B} = \{[t, +\infty), t \geq 0\}$ ,  $S_{\mathcal{B}}(2n) = 2n + 1$  and  $x^2 = 4\xi(\log n)/n + 4(\log 2 + \log(16n + 8))/n$ . There exists therefore an event  $\Omega_{1,\xi}$  such that  $\mathbb{P}(\Omega_{1,\xi}) \geq 1 - (1/2)n^{-\xi}$  and on which: for all  $t \geq 0$ ,

$$\left| \sqrt{G_n(t)} - \sqrt{G(t)} \right| \leq x.$$

The assumption on  $\hat{\alpha}_\xi$  ensures that  $4x^2 \leq G_n(\hat{\alpha}_\xi)$  and hence: for all  $t \in [0, \hat{\alpha}_\xi]$ ,

$$\left| \sqrt{G_n(t)} - \sqrt{G(t)} \right| \leq (1/2)\sqrt{G_n(\hat{\alpha}_\xi)} \leq (1/2)\sqrt{G_n(t)}.$$

We deduce: for all  $t \in [0, \hat{\alpha}_\xi]$ ,  $G_n(t) \leq 4G(t)$  and  $G(t) \leq (9/4)G_n(t)$ .

Now,

$$\begin{aligned} h_E^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) &\leq h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + h_E^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0) \\ &\leq (9/4)h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + h_E^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0). \end{aligned}$$

If  $G_C$  (resp  $G_T$ ) denotes the survival function of  $C$  (resp  $T$ ),  $G = G_C G_T$  and thus

$$h_E^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0) = \frac{1}{2} \int_{\hat{\alpha}_\xi}^{\infty} \mathbf{s}_0 G \, d\mu = \frac{1}{2} \int_{\hat{\alpha}_\xi}^{\infty} \mathbf{s}_0 G_C G_T \, d\mu = \frac{1}{2} \int_{\hat{\alpha}_\xi}^{\infty} f_0 G_C \, d\mu.$$

Therefore,

$$h_E^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0) \leq \left( \frac{1}{2} \int_{\hat{\alpha}_\xi}^{\infty} f_0 d\mu \right) \times G_C(\hat{\alpha}_\xi) \leq \frac{1}{2} G_T(\hat{\alpha}_\xi) G_C(\hat{\alpha}_\xi) \leq \frac{1}{2} G(\hat{\alpha}_\xi).$$

It remains to use  $G(\hat{\alpha}_\xi) \leq (9/4)G_n(\hat{\alpha}_\xi)$ ,  $G_n(\hat{\alpha}_\xi) \leq (71 + 16\xi)(\log n)/n$  to get (42) on  $\Omega_{1,\xi}$ .

As to (43), note that

$$(44) \quad \begin{aligned} h^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) &\leq h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + h^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0) \\ &\leq 4h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + h^2(\mathbf{s}_0 \mathbb{1}_{[\hat{\alpha}_\xi, +\infty)}, 0). \end{aligned}$$

If  $(9/2)(71 + 16\xi)(\log n)/n \geq 1$ , we set  $\alpha = 0$ . If not, there exists a number  $\alpha \geq 0$  such that

$$G(\alpha) = \frac{9}{2}(71 + 16\xi) \frac{\log n}{n}.$$

On  $\Omega_{\xi,1}$ ,

$$G(\hat{\alpha}_\xi) \leq \frac{9}{4}G_n(\hat{\alpha}_\xi) \leq \frac{9}{4}(71 + 16\xi) \frac{\log n}{n}.$$

Therefore  $\alpha \leq \hat{\alpha}_\xi$  and (44) leads to

$$(45) \quad h^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) \leq 4h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + h^2(\mathbf{s}_0 \mathbb{1}_{[\alpha, +\infty)}, 0).$$

It follows from Bernstein's deviation inequality, and more precisely from (86), that there exists an event  $\Omega_{\xi,2}$  such that  $\mathbb{P}[\Omega_{\xi,2}] \geq 1 - (1/2)n^{-\xi}$  and on which

$$\int_{\alpha}^{\infty} \mathbf{s}_0 G_n \leq 2 \int_{\alpha}^{\infty} \mathbf{s}_0 G d\mu + 2 \frac{\xi \log n + \log 2}{n}.$$

Since  $\int_{\alpha}^{\infty} \mathbf{s}_0 G d\mu \leq G(\alpha)$  (see the preceding computations),

$$\int_{\alpha}^{\infty} \mathbf{s}_0 G_n \leq 2G(\alpha) + 2 \frac{\xi \log n + \log 2}{n}.$$

We deduce,

$$h^2(\mathbf{s}_0 \mathbb{1}_{[\alpha, +\infty)}, 0) \leq G(\alpha) + \frac{\xi \log n + \log 2}{n}.$$

By using (45), we get on  $\Omega_{\xi,1} \cap \Omega_{\xi,2}$ ,

$$h^2(\mathbf{s}_0, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) \leq 4h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]}) + G(\alpha) + \frac{\xi \log n + \log 2}{n}.$$

No matter if  $(9/2)(71 + 16\xi)(\log n)/n$  is smaller or larger than 1,

$$G(\alpha) \leq \frac{9}{2}(71 + 16\xi) \frac{\log n}{n},$$

which shows (43). □

**Lemma 6.** *Consider framework 2, and suppose that  $G$  is continuous. Let  $\xi > 0$  and suppose that there exists a positive random variable  $\hat{\alpha}_\xi$  satisfying (41). There exists an event which holds true with probability larger than  $1 - n^{-\xi}$  and on which: for all  $\kappa_n > 0$ , all estimator  $\hat{s} \in \mathcal{S}$ , and  $\tilde{s} = \min\{\hat{s}, \kappa_n\} \mathbb{1}_{[0, \hat{\alpha}_\xi]}$ ,*

$$(46) \quad h_E^2(\mathbf{s}_0, \tilde{s}) \leq (9/2)h^2(\mathbf{s}_0, \hat{s}) + 9\mathbb{P}\left[f_0(T) > 70\kappa_n \frac{\log n}{n}\right] + (803 + 183\xi) \frac{\log n}{n}.$$

*Proof of Lemma 6.* We derive from (42) that with probability larger than  $1 - n^{-\xi}$ ,

$$h_E^2(\mathbf{s}_0, \tilde{s}) \leq (9/4)h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \tilde{s}) + (80 + 18\xi) \frac{\log n}{n}.$$

Now,

$$\begin{aligned} h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]}, \tilde{s}) &= h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} \leq \kappa_n}, \hat{s} \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} \leq \kappa_n}) + h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}) \\ &\leq h^2(\mathbf{s}_0, \hat{s}) + h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}). \end{aligned}$$

We need to bound above the second term of this inequality. We have,

$$\begin{aligned} h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n}) &\leq h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 \leq \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 \leq \kappa_n}) \\ &\quad + h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 > \kappa_n}). \end{aligned}$$

Yet,

$$\begin{aligned} h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 \leq \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 \leq \kappa_n}) &\leq h^2(\mathbf{s}_0, \hat{s}), \\ h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\hat{s} > \kappa_n, \mathbf{s}_0 > \kappa_n}) &\leq h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}). \end{aligned}$$

By putting all these inequalities together,

$$(47) \quad h_E^2(\mathbf{s}_0, \tilde{s}) \leq (9/2)h^2(\mathbf{s}_0, \hat{s}) + (9/4)h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}) + (80 + 18\xi) \frac{\log n}{n}.$$

We now derive from (43),

$$\begin{aligned} h^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}) &\leq 4h_E^2(\mathbf{s}_0 \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}, \kappa_n \mathbb{1}_{[0, \hat{\alpha}_\xi]} \mathbb{1}_{\mathbf{s}_0 > \kappa_n}) + (321 + 73\xi) \frac{\log n}{n} \\ &\leq 2 \int_0^{\hat{\alpha}_\xi} (\mathbf{s}_0 + \kappa_n) \mathbb{1}_{\mathbf{s}_0 > \kappa_n} G \, d\mu + (321 + 73\xi) \frac{\log n}{n} \\ &\leq 4 \int_0^{\hat{\alpha}_\xi} \mathbf{s}_0 \mathbb{1}_{\mathbf{s}_0 > \kappa_n} G \, d\mu + (321 + 73\xi) \frac{\log n}{n} \\ &\leq 4 \int_0^{\hat{\alpha}_\xi} f_0 \mathbb{1}_{\mathbf{s}_0 > \kappa_n} \, d\mu + (321 + 73\xi) \frac{\log n}{n}. \end{aligned}$$

As  $G$  is a non-increasing function, we deduce that for all  $t \in [0, \hat{\alpha}_\xi]$ ,

$$G(t) \geq G(\hat{\alpha}_\xi) \geq (1/4)G_n(\hat{\alpha}_\xi) \geq (70 + 16\xi) \frac{\log n}{n} \geq 70 \frac{\log n}{n}.$$

Moreover,  $\mathbf{s}_0(t) > \kappa_n$  implies when  $t \in [0, \hat{\alpha}_\xi]$ ,

$$f_0(t) > \kappa_n G(t) > 70 \kappa_n \frac{\log n}{n}.$$

Therefore,

$$\int_0^{\hat{\alpha}_\xi} f_0 \mathbb{1}_{\mathbf{s}_0 > \kappa_n} \, d\mu \leq \mathbb{P} \left[ f_0(T) > 70 \kappa_n \frac{\log n}{n} \right].$$

□

We are now in position to state:

**Proposition 9.** *Consider framework 2, and suppose that  $G$  is continuous. Let  $\xi > 0$  and suppose that there exists a positive random variable  $\hat{\alpha}_\xi$  satisfying (41). Then, for all estimator  $\hat{s} \in \mathcal{S}$ , and  $\tilde{s} = \min\{\hat{s}, \kappa_n\} \mathbb{1}_{[0, \hat{\alpha}_\xi]}$ ,*

$$\mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s})] \leq (9/2) \mathbb{E} [h^2(\mathbf{s}_0, \hat{s})] + (803 + 183\xi) \frac{\log n}{n} + 9 \mathbb{P} \left[ f_0(T) > 70\kappa_n \frac{\log n}{n} \right] + \frac{1 + \kappa_n \mathbb{E}[X]}{2n^\xi}.$$

*Proof of Proposition 9.* Let  $\mathcal{A}_\xi$  be the event given by Lemma 6. Then  $\mathbb{P}(\mathcal{A}_\xi) \geq 1 - n^{-\xi}$ . Moreover,

$$\mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s})] \leq \mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s}) \mathbb{1}_{\mathcal{A}_\xi}] + \mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s}) \mathbb{1}_{\mathcal{A}_\xi^c}].$$

We have,

$$h_E^2(\mathbf{s}_0, \tilde{s}) \leq \frac{1}{2} \int_0^\infty (\mathbf{s}_0 + \tilde{s}) G \, d\mu \leq \frac{1}{2} \left( 1 + \int_0^\infty \kappa_n G \, d\mu \right) \leq \frac{1 + \kappa_n \mathbb{E}[X]}{2},$$

and hence,

$$\mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s})] \leq \mathbb{E} [h_E^2(\mathbf{s}_0, \tilde{s}) \mathbb{1}_{\mathcal{A}_\xi}] + \frac{1 + \kappa_n \mathbb{E}[X]}{2n^\xi}.$$

It remains to use (46) to finish the proof.  $\square$

We may use this result with  $\xi = 5$ ,  $\kappa_n = n^3$  and apply Markov's inequality

$$\mathbb{P} \left[ f_0(T) > 70\kappa_n \frac{\log n}{n} \right] \leq \frac{n}{70\kappa_n \log n} \int_0^\infty f_0^2 \, d\mu$$

to get Proposition 4. Note that there exists a random variable  $\hat{\alpha}$  satisfying (14) if  $150(\log n)/n < 1$  that is  $n \geq 1043$ .  $\square$

**6.5. Proof of Theorem 5.** In each framework, the measure  $N$  can be put in the form  $N(A) = n^{-1} \sum_{i \in \hat{I}} \mathbb{1}_A(Y_i)$  where  $\hat{I} \subset \{1, \dots, n\}$ , and where the  $Y_i$  are suitable real-valued random variables. For instance, in framework 1,  $\hat{I} = \{1, \dots, n\}$ ,  $Y_i = X_i$ , in framework 2,  $\hat{I} = \{i \in \{1, \dots, n\}, D_i = 1\}$ ,  $Y_i = X_i$ , and in framework 3,  $\hat{I} = \{i \in \{1, \dots, n\}, T_{1,0}^{(i)} \in I_{\text{obs}}\}$ ,  $Y_i = T_{1,0}^{(i)}$ .

Set  $\hat{J} = \{i \in \hat{I}, Y_i \in K\}$ . Then, for  $s, s' \in S$ ,  $T(s, s')$  and  $\mathcal{L}_K(s)$  take the form

$$T(s, s') = \frac{1}{n} \sum_{j \in \hat{J}} \psi \left( \frac{s'(Y_j)}{s(Y_j)} \right) - \frac{1}{4} \int_K (s' - s) \, dM$$

$$\mathcal{L}_K(s) = \frac{1}{n} \sum_{j \in \hat{J}} \log s(Y_j) - \int_K s \, dM.$$

The proof is straightforward if  $\hat{J} = \emptyset$  since then  $4T(s, s') = \mathcal{L}_K(s') - \mathcal{L}_K(s)$  and  $4\gamma(s) = \sup_{s' \in S} \mathcal{L}_K(s') - \mathcal{L}_K(s)$ . We suppose from now on that  $\hat{J} \neq \emptyset$ .

**Claim 4.** *Let  $S' = \{s \in S, \mathcal{L}_K(s) \neq -\infty\}$  and  $s \in S'$ . Then,  $\sup_{s' \in S'} T(s, s') = 0$  if and only if  $\sup_{s' \in S'} \mathcal{L}_K(s') - \mathcal{L}_K(s) = 0$ .*

*Proof.* Suppose that  $\sup_{s' \in S'} \mathcal{L}_K(s') - \mathcal{L}_K(s) = 0$ .

Let  $S'_1 = \{s \in S', s = s', N \text{ a.s}\}$  and  $S'_2 = S' \setminus S'_1$ . When  $s \in S'_1$ ,

$$\begin{aligned} T(s, s') &= -\frac{1}{4} \int_K (s' - s) \, dM \\ &= \frac{1}{4} (\mathcal{L}_K(s') - \mathcal{L}_K(s)). \end{aligned}$$

Therefore,  $T(s, s') \leq 0$ .

Let now  $s \in S'_2$ ,  $u \in [0, 1]$  and  $\zeta = s' - s$ . Note that  $s + u\zeta = (1 - u)s + us' \in S'$  and thus  $\mathcal{L}_K(s + u\zeta) - \mathcal{L}_K(s) \leq 0$ . We introduce the real-valued map  $\wp_1$  for  $u \in [0, 1]$  defined by

$$\begin{aligned} \wp_1(u) &= \mathcal{L}_K(s + u\zeta) - \mathcal{L}_K(s) \\ &= \frac{1}{n} \sum_{j \in \hat{J}} \log \left( \frac{s(Y_j) + u\zeta(Y_j)}{s(Y_j)} \right) - u \int_K \zeta \, dM. \end{aligned}$$

We now define  $\wp_2$  for  $u \in [0, 1]$  by

$$\begin{aligned} \wp_2(u) &= 4T(s, s + u\zeta) \\ &= \frac{4}{n} \sum_{j \in \hat{J}} \psi \left( \frac{s(Y_j) + u\zeta(Y_j)}{s(Y_j)} \right) - u \int_K \zeta \, dM. \end{aligned}$$

Some computations show that  $\wp_1$  and  $\wp_2$  are twice differentiable on  $[0, 1]$  and

$$\begin{aligned} \wp_1(0) &= \wp_2(0) = 0 \\ \wp_1'(0) &= \wp_2'(0) = \frac{1}{n} \sum_{j \in \hat{J}} \frac{\zeta(Y_j)}{s(Y_j)} - \int_K \zeta \, dM \\ \wp_1''(0) &= \wp_2''(0) = -\frac{1}{n} \sum_{j \in \hat{J}} \left( \frac{\zeta(Y_j)}{s(Y_j)} \right)^2. \end{aligned}$$

Therefore,  $\wp_1''(0)$  and  $\wp_2''(0)$  are always negative.

Since  $\wp_1(u)$  is non-positive for all  $u \in [0, 1]$ ,  $\wp_1'(0) \leq 0$ . The above computations show the existence of  $u_1 \in (0, 1]$  such that  $\wp_2(u) \leq 0$  for all  $u \in [0, u_1]$ . Now,  $\wp_2$  is concave, and hence non-positive on  $[0, 1]$ . In particular,  $\wp_2(1) = T(s, s') \leq 0$ .

Likewise,  $\sup_{s' \in S'} T(s, s') = 0$  implies  $\sup_{s' \in S'} \mathcal{L}_K(s') - \mathcal{L}_K(s) = 0$ .

□

Let  $\tilde{s} \in S$  such that  $\mathcal{L}_K(\tilde{s}) \geq \mathcal{L}_K(s)$  for all  $s \in S$  and  $\mathcal{L}_K(\tilde{s}) \neq -\infty$ . The above claim then shows that  $T(\tilde{s}, s) \leq 0$  for all  $s \in S$  such that  $\mathcal{L}_K(s) \neq -\infty$ . Choose now  $s \in S$  such that  $\mathcal{L}_K(s) = -\infty$ . Define for  $u \in [0, 1]$ ,  $s_u = (1 - u)\tilde{s} + us \in S$  and note that  $s_1 = s$ . If  $u \in [0, 1)$ ,  $\mathcal{L}_K(s_u) \neq -\infty$  and thus  $T(\tilde{s}, s_u) \leq 0$ . The continuity of the map  $u \in [0, 1] \mapsto T(\tilde{s}, s_u)$  ensures that  $T(\tilde{s}, s) \leq 0$ . Finally,  $\gamma(\tilde{s}) = 0$ .

Conversely, let  $\hat{s}$  be a  $\rho$ -estimator satisfying  $\gamma(\hat{s}) = 0$ . We begin by proving that  $\mathcal{L}_K(\hat{s}) \neq -\infty$ . Consider  $s \in S$  such that  $\mathcal{L}_K(s) \neq -\infty$  and define for  $u \in [0, 1]$ ,  $s_u = (1 - u)\hat{s} + us \in S$ ,

$$\begin{aligned} \wp_3(u) &= T(\hat{s}, s_u) \\ &= \frac{1}{n} \sum_{j \in \hat{J}} \psi \left( \frac{(1 - u)\hat{s}(Y_j) + us(Y_j)}{\hat{s}(Y_j)} \right) - \frac{1}{4} \int_K (s_u - \hat{s}) \, dM. \end{aligned}$$

When  $j \in \hat{J}$ ,  $s(Y_j) > 0$ . Therefore, if  $\hat{J}' = \{j \in \hat{J}, \hat{s}(Y_j) = 0\}$  and  $u \in (0, 1]$ ,

$$\wp_3(u) = \frac{|\hat{J}'|}{n} + \frac{1}{n} \sum_{j \in \hat{J} \setminus \hat{J}'} \psi \left( \frac{(1 - u)\hat{s}(Y_j) + us(Y_j)}{\hat{s}(Y_j)} \right) - \frac{1}{4} \int_K (s_u - \hat{s}) \, dM.$$

Therefore, if  $\hat{J}' \neq \emptyset$  choosing  $u > 0$  small enough leads to  $\wp_3(u) \geq |\hat{J}'|/(2n) > 0$ , which is impossible as  $\gamma(\hat{s}) = 0$ . Therefore,  $\hat{J}' = \emptyset$  and  $\mathcal{L}_K(\hat{s}) \neq -\infty$ . The claim then asserts that for all  $s \in S$  such that  $\mathcal{L}_K(s) \neq -\infty$ ,  $\mathcal{L}_K(s) \leq \mathcal{L}_K(\hat{s})$ . This inequality being true if  $\mathcal{L}_K(s) = -\infty$ , the proof is complete.  $\square$

**6.6. Sketch of the proof of Theorem 6.** We define the elements  $Y_i$ ,  $\hat{I}$ ,  $\hat{J}$  as in the proof of Theorem 5. Let for  $x \in [0, +\infty]$ ,  $\psi_2(x) = \psi(x^2)$  and for  $f, f' \in \mathcal{F}$ ,

$$\begin{aligned} T_2(f, f') &= T(f^2, f'^2) = \frac{1}{n} \sum_{j \in \hat{J}} \psi_2 \left( \frac{f'(Y_j)}{f(Y_j)} \right) - \frac{1}{4} \int_K (f'^2 - f^2) \, dM, \\ \mathcal{L}_{K,2}(f) &= \mathcal{L}_K(f^2) = \frac{2}{n} \sum_{j \in \hat{J}} \log f(Y_j) - \int_K f^2 \, dM. \end{aligned}$$

The proof is very similar to the one of Theorem 5. The main change lies in the replacement of the symbols  $S$ ,  $T$ ,  $\mathcal{L}_K$  by  $\mathcal{F}$ ,  $T_2$ ,  $\mathcal{L}_{K,2}$ . We will only give some insight into why Claim 4 remains valid under these modifications.

As in the proof of Theorem 5, we may suppose that  $\hat{J} \neq \emptyset$ .

**Claim 5.** *Let  $\mathcal{F}' = \{f \in \mathcal{F}, \mathcal{L}_{K,2}(f) \neq -\infty\}$  and  $f \in \mathcal{F}'$ . Then,  $\sup_{f' \in \mathcal{F}'} T_2(f, f') = 0$  if and only if  $\sup_{f' \in \mathcal{F}'} \mathcal{L}_{K,2}(f') - \mathcal{L}_{K,2}(f) = 0$ .*

*Sketch of the proof.* We prove that  $\sup_{f' \in \mathcal{F}'} \mathcal{L}_{K,2}(f') - \mathcal{L}_{K,2}(f) = 0$  implies  $\sup_{f' \in \mathcal{F}'} T_2(f, f') = 0$ . The proof of the converse is similar.

Let  $\mathcal{F}'_1 = \{f' \in \mathcal{F}', f' = f, N \text{ a.s.}\}$  and  $\mathcal{F}'_2 = \mathcal{F}' \setminus \mathcal{F}'_1$ . As in the proof of Claim 4,  $T_2(f, f') = (\mathcal{L}_{K,2}(f') - \mathcal{L}_{K,2}(f))/4$  when  $f' \in \mathcal{F}'_1$  and is thus non-positive. Let now  $f' \in \mathcal{F}'_2$ ,  $u \in [0, 1]$  and  $\zeta = f' - f$ . Note that  $f + u\zeta = (1 - u)f + uf' \in \mathcal{F}'$  and thus  $\mathcal{L}_{K,2}(f + u\zeta) - \mathcal{L}_{K,2}(f) \leq 0$ .

We introduce the real-valued map  $\wp_1$  for  $u \in [0, 1]$  defined by

$$\begin{aligned} \wp_1(u) &= \mathcal{L}_{K,2}(f + u\zeta) - \mathcal{L}_{K,2}(f) \\ &= \frac{2}{n} \sum_{j \in \hat{J}} \log \left( \frac{f(Y_j) + u\zeta(Y_j)}{f(Y_j)} \right) - u^2 \int_K \zeta^2 \, dM - 2u \int_K \zeta f \, dM. \end{aligned}$$

We now define  $\wp_2$  for  $u \in [0, 1]$  by

$$\begin{aligned}\wp_2(u) &= 4T_2(f, f + u\zeta) \\ &= \frac{4}{n} \sum_{j \in J} \psi_2 \left( \frac{f(Y_j) + u\zeta(Y_j)}{f(Y_j)} \right) - u^2 \int_K \zeta^2 dM - 2u \int_K \zeta f dM.\end{aligned}$$

Some computations show that  $\wp_1(0) = \wp_2(0) = 0$ ,  $\wp_1'(0) = \wp_2'(0)$ ,  $\wp_1''(0) = \wp_2''(0) < 0$ .

As  $\wp_1(u)$  is non-positive for all  $u \in [0, 1]$ ,  $\wp_1'(0) \leq 0$ . There exists therefore  $u_1 \in (0, 1]$  such that  $\wp_2(u) \leq 0$  for all  $u \in [0, u_1]$ . Since  $\psi_2$  is concave,  $\wp_2$  is also concave, and  $\wp_2$  is non-positive on  $[0, 1]$ . In particular,  $\wp_2(1) = T_2(f, f') \leq 0$ .

□

**6.7. Proof of Lemma 2.** As in the proof of Theorem 5, the measure  $N$  can be put in the form  $N(A) = n^{-1} \sum_{i \in \hat{I}} \mathbb{1}_A(Y_i)$  where  $\hat{I} \subset \{1, \dots, n\}$ , and where the  $Y_i$  are suitable real-valued random variables.

Note that if  $K \cap \{Y_{(1)}, \dots, Y_{(\hat{n})}\} = \emptyset$  then,

$$\mathcal{L}_K(s) = - \int_K s(t) dM(t),$$

and the supremum  $\sup_{s \in \mathcal{P}_r(K)} \mathcal{L}_K(s)$  is achieved at  $\hat{s}_K = 0$  and equals 0. We now suppose that  $K \cap \{Y_{(1)}, \dots, Y_{(\hat{n})}\} \neq \emptyset$ .

Let  $G_n$  be the Radon–Nikodym derivative of  $M$  with respect to the Lebesgue measure  $\mu$ . Then,  $G_n = 1$  in framework 1,  $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \geq t} \mathbb{1}_{[0, +\infty)}(t)$  in framework 2 and

$$G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_{t-}^{(i)} = 1} \mathbb{1}_{I_{\text{obs}}}(t)$$

in framework 3. Let  $k$  be the largest integer of  $\{1, \dots, \hat{n}\}$  such that  $Y_{(k)}$  belongs to  $K$  and  $K' = K \cap (-\infty, Y_{(k)}]$ . There exists some  $\alpha > 0$  such that  $(Y_{(k)} - \alpha, Y_{(k)}) \subset K'$ . Moreover, we can choose  $\alpha$  small enough to get  $G_n(t) \geq 1/n$  for all  $t \in (Y_{(k)} - \alpha, Y_{(k)})$ .

Let now  $s \in \mathcal{P}_r(K)$ . Then,  $\mathcal{L}_K(s)$  takes the form

$$\mathcal{L}_K(s) = \frac{1}{n} \sum_{i \in \hat{I}} (\log s(Y_i)) \mathbb{1}_K(Y_i) - \int_K s(t) G_n(t) dt,$$

and is bounded from above by

$$\mathcal{L}_K(s) \leq \log_+ \left( \sup_{t \in K'} s(t) \right) - \frac{1}{n} \int_{Y_{(k)} - \alpha}^{Y_{(k)}} s(t) dt.$$

We endow the linear space consisting of polynomial functions of degree at most  $r$  with the two following norms:

$$\|s\|_1 = \int_{Y_{(k)} - \alpha}^{Y_{(k)}} |s(t)| dt, \quad \|s\|_\infty = \sup_{t \in K'} |s(t)|.$$

This linear space being of finite dimension, there exists  $C$  such that  $\|s\|_\infty \leq C\|s\|_1$  for all  $s \in \mathcal{P}_r(K)$ . Now,

$$\mathcal{L}_K(s) \leq \log_+ (C\|s\|_1) - \frac{\|s\|_1}{n}.$$

The continuous map  $\mathcal{L}_K(\cdot)$  tends therefore to  $-\infty$  when  $\|s\|_1 \rightarrow +\infty$ . As there exists at least a function  $s \in \mathcal{P}_r(K)$  such that  $\mathcal{L}_K(s) \neq -\infty$ ,  $\hat{s}_K$  does exist.

For the second part of the lemma, we use Theorem 5 to deduce that  $T(\hat{s}_K, s_K) \leq 0$  for all  $s_K \in \mathcal{P}_r(K)$ . If  $s \in \mathcal{P}_r(m)$  is of the form  $s = \sum_{K \in m} s_K$ ,

$$T(\hat{s}_m, s) = \sum_{K \in m} T(\hat{s}_K, s_K) \leq 0.$$

Thus,  $\gamma(\hat{s}_m) = 0$  and  $\hat{s}_m$  is a  $\rho$ -estimator on  $\mathcal{P}_r(m)$ .

Finally,  $\hat{s}_m$  maximizes

$$\mathcal{L}_{\cup_{K \in m}}(s) = \int_{\cup_{K \in m}} \log s \, dN - \int_{\cup_{K \in m}} s \, dM$$

over  $s \in \mathcal{P}_r(m)$ . As  $s$  vanishes outside  $\cup_{K \in m}$ ,  $\mathcal{L}_{\cup_{K \in m}}(s) = \mathcal{L}(s)$  if  $N(\cup_{K \in m}) = N(\mathbb{R})$ .  $\square$

**6.8. Proofs of Theorems 7 and 8.** It is convenient for the demonstration to encompass the two procedures into a more general selection rule we now describe. Theorems 7 and 8 follow from Theorem 10 below. Their proofs are given in Sections 6.8.2 and 6.8.3.

We consider an arbitrary (possibly random) set  $\hat{\Lambda}$ . For each  $\lambda \in \hat{\Lambda}$ , we consider an estimator  $\hat{s}_\lambda$  with values in  $\mathcal{S}$ . Our aim is to select an estimator among the collection  $\{\hat{s}_\lambda, \lambda \in \hat{\Lambda}\}$ .

We consider for each  $\lambda \in \hat{\Lambda}$  a (possibly random) model  $\hat{S}_\lambda \subset \mathcal{S}$ . We associate to each  $\lambda \in \hat{\Lambda}$ ,  $s \in \hat{S}_\lambda$ , two penalty terms  $\text{pen}_{1,\lambda}(s)$  and  $\text{pen}_2(\lambda)$ . We finally define the criterion  $\gamma_4$  by

$$\gamma_4(\hat{s}_\lambda) = \sup_{s \in \hat{S}_\lambda} [T(\hat{s}_\lambda, s) - \text{pen}_{1,\lambda}(s)].$$

The selected estimator  $\hat{s}_{\hat{\lambda}}$  is then any estimator among  $\{\hat{s}_\lambda, \lambda \in \hat{\Lambda}\}$  satisfying

$$\gamma_4(\hat{s}_{\hat{\lambda}}) + 2\text{pen}_2(\hat{\lambda}) \leq \inf_{\lambda \in \hat{\Lambda}} \{\gamma_4(\hat{s}_\lambda) + 2\text{pen}_2(\lambda)\} + 1/n.$$

The risk of this estimator is bounded from above as follows.

**Theorem 10.** *Let  $(\mathcal{I}_d)_{d \geq 1}$  be a non-decreasing collection of Borel sets that fulfils Assumption 1, and  $d_{s'}(s)$  be the notation introduced in Section 6.3. Let for  $\xi > 0$ ,  $\Omega_\xi$  be the event given by Lemma 4. It satisfies  $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ . We suppose that there exist two real valued maps,  $\Delta(\cdot) \geq 0$  on  $\hat{\Lambda}$ , and  $d(\cdot) \geq 1$  on  $\mathcal{S}$  such that*

$$(48) \quad d_{\hat{s}_\lambda}(s) \leq d(s) + \Delta(\lambda) \quad \text{for all } \lambda \in \hat{\Lambda}, s \in \hat{S}_\lambda.$$

*We suppose that there exist a (possibly random) model  $\hat{S} \subset \bigcap_{\lambda \in \hat{\Lambda}} \hat{S}_\lambda$  and a map  $\text{pen}_1(\cdot)$  on  $\hat{S}$  such that*

$$(49) \quad \text{pen}_{1,\lambda}(s) \leq \text{pen}_1(s) + \text{pen}_2(\lambda) \quad \text{for all } s \in \hat{S}, \lambda \in \hat{\Lambda}.$$

There exists a universal constant  $L_1$  such that if for all  $\lambda \in \hat{\Lambda}$ ,  $s \in \hat{S}_\lambda$ ,

$$(50) \quad \begin{aligned} \text{pen}_{1,\lambda}(s) &\geq L_1 \frac{d(s)}{n} \log_+^2 \left( \frac{n}{d(s)} \right) \\ \text{pen}_2(\lambda) &\geq L_1 \frac{\Delta(\lambda)}{n} \log_+^2 \left( \frac{n}{\Delta(\lambda)} \right), \end{aligned}$$

and if for all  $s \in \hat{S}$ ,

$$\text{pen}_1(s) \geq L_1 \frac{d(s)}{n} \log_+^2 \left( \frac{n}{d(s)} \right),$$

then, on  $\Omega_\xi$ :

$$h^2(\mathbf{s}_0, \hat{s}_\lambda) \leq c \left( \inf_{\lambda \in \hat{\Lambda}} \{h^2(\mathbf{s}_0, \hat{s}_\lambda) + \text{pen}_2(\lambda)\} + \inf_{s \in \hat{S}} \{h^2(\mathbf{s}_0, s) + \text{pen}_1(s)\} + \xi \log_+(1/\xi) \right).$$

In the above inequality,  $c$  is a universal constant and the convention  $0 \times \log_+^2(n/0) = 0$  is used when  $\Delta(\lambda) = 0$ .

6.8.1. *Proof of Theorem 10.* Let  $\varepsilon \in (0, 1/12)$ . Lemma 4 asserts the following on  $\Omega_\xi$ : for all  $s, s' \in \mathcal{S}$

$$(51) \quad T(s, s') \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s') + c_1 \min \{\vartheta(d_{s'}(s)), \vartheta(d_s(s'))\} + c_2 \xi \log_+(1/\xi)$$

where  $c_1, c_2$  only depend on  $\varepsilon$  and where  $\vartheta(\cdot)$  is given by (31).

Let  $\lambda \in \hat{\Lambda}$  and  $s \in \hat{S}_\lambda$ . Then,

$$T(\hat{s}_\lambda, s) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, s) + c_1 \vartheta(d_{\hat{s}_\lambda}(s)) + c_2 \xi \log_+(1/\xi).$$

Note that  $\vartheta(d_1) \leq 1.48\vartheta(d_2)$  for all  $d_1 \leq d_2$ . Therefore,

$$\begin{aligned} T(\hat{s}_\lambda, s) &\leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) + 1.48c_1 \vartheta(d(s) + \Delta(\lambda)) + c_2 \xi \log_+(1/\xi) \\ &\leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) + 1.48c_1 \vartheta(d(s)) + 1.48c_1 \vartheta(\Delta(\lambda)) + c_2 \xi \log_+(1/\xi). \end{aligned}$$

If  $L_1$  is large enough,

$$T(\hat{s}_\lambda, s) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) + \text{pen}_{1,\lambda}(s) + \text{pen}_2(\lambda) + c_2 \xi \log_+(1/\xi),$$

and hence

$$(52) \quad \gamma_4(\hat{s}_\lambda) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) + \text{pen}_2(\lambda) + c_2 \xi \log_+(1/\xi).$$

We now derive from (51) that for all  $s \in \hat{S}_\lambda$ ,

$$T(s, \hat{s}_\lambda) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, \hat{s}_\lambda) + c_1 \vartheta(d_{\hat{s}_\lambda}(s)) + c_2 \xi \log_+(1/\xi).$$

Using moreover that  $T(\hat{s}_\lambda, s) = -T(s, \hat{s}_\lambda)$  we deduce,

$$\begin{aligned} \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, \hat{s}_\lambda) &\leq T(\hat{s}_\lambda, s) + (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) + c_1 \vartheta(d_{\hat{s}_\lambda}(s)) + c_2 \xi \log_+(1/\xi) \\ &\leq T(\hat{s}_\lambda, s) + (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) + 1.48c_1 \vartheta(d(s) + \Delta(\hat{\lambda})) + c_2 \xi \log_+(1/\xi) \\ &\leq T(\hat{s}_\lambda, s) + (3 + 4\varepsilon)h^2(\mathbf{s}_0, s) + 1.48c_1 \vartheta(d(s)) + 1.48c_1 \vartheta(\Delta(\hat{\lambda})) + c_2 \xi \log_+(1/\xi). \end{aligned}$$

If  $L_1$  is large enough,

$$\begin{aligned} \frac{4-3\varepsilon}{12} h^2(\mathbf{s}_0, \hat{s}_\lambda) &\leq T(\hat{s}_\lambda, s) + (3+4\varepsilon)h^2(\mathbf{s}_0, s) + \frac{1}{2}\text{pen}_{1,\hat{\lambda}}(s) + \frac{1}{2}\text{pen}_2(\hat{\lambda}) \\ &\quad + c_2\xi \log_+(1/\xi) - 1/n \\ &\leq \left[ T(\hat{s}_\lambda, s) - \text{pen}_{1,\hat{\lambda}}(s) \right] + \frac{1}{2}\text{pen}_2(\hat{\lambda}) + \left[ (3+4\varepsilon)h^2(\mathbf{s}_0, s) + \frac{3}{2}\text{pen}_{1,\hat{\lambda}}(s) \right] \\ &\quad + c_2\xi \log_+(1/\xi) - 1/n. \end{aligned}$$

Since this inequality is valid for all  $s \in \hat{S}_\lambda$  and  $\hat{S} \subset \hat{S}_\lambda$ ,

$$\begin{aligned} \frac{4-3\varepsilon}{12} h^2(\mathbf{s}_0, \hat{s}_\lambda) &\leq \gamma_4(\hat{s}_\lambda) + \frac{1}{2}\text{pen}_2(\hat{\lambda}) + \inf_{s \in \hat{S}} \left\{ 3(1+\varepsilon)h^2(\mathbf{s}_0, s) + \frac{3}{2}\text{pen}_{1,\hat{\lambda}}(s) \right\} \\ &\quad + c_2\xi \log_+(1/\xi) - 1/n. \end{aligned}$$

We deduce from (49),

$$\begin{aligned} \frac{4-3\varepsilon}{12} h^2(\mathbf{s}_0, \hat{s}_\lambda) &\leq \gamma_4(\hat{s}_\lambda) + 2\text{pen}_2(\hat{\lambda}) + \inf_{s \in \hat{S}} \left\{ 3(1+\varepsilon)h^2(\mathbf{s}_0, s) + \frac{3}{2}\text{pen}_1(s) \right\} \\ &\quad + c_2\xi \log_+(1/\xi) - 1/n. \end{aligned}$$

By using the definition of  $\hat{\lambda}$  and (52), we get for all  $\lambda \in \hat{\Lambda}$ ,

$$\begin{aligned} \frac{4-3\varepsilon}{12} h^2(\mathbf{s}_0, \hat{s}_\lambda) &\leq \gamma_4(\hat{s}_\lambda) + 2\text{pen}_2(\lambda) + \inf_{s \in \hat{S}} \left\{ 3(1+\varepsilon)h^2(\mathbf{s}_0, s) + \frac{3}{2}\text{pen}_1(s) \right\} + 2c_2\xi \log_+(1/\xi) \\ &\leq (3+4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_\lambda) + 3\text{pen}_2(\lambda) + \inf_{s \in \hat{S}} \left\{ 3(1+\varepsilon)h^2(\mathbf{s}_0, s) + \frac{3}{2}\text{pen}_1(s) \right\} \\ &\quad + 2c_2\xi \log_+(1/\xi). \end{aligned}$$

It remains to take the infimum over  $\lambda \in \hat{\Lambda}$  to finish the proof.  $\square$

**6.8.2. Proof of Theorem 7.** We will apply the selection rule developed in Section 6.8 to pick out an estimator among  $\{\hat{s}_\lambda, \lambda \in \hat{\Lambda}\} = \{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ . For this purpose, we need to explain the values of the different parameters involved in the procedure. We define  $\mathcal{I}_d$  as the collection of unions of at most  $d$  intervals. We set  $\hat{S} = \{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ , and for  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ ,

$$\hat{S}_m = \left\{ \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K, m_K \in \widehat{\mathcal{M}}_{\hat{\ell}} \right\}.$$

Note that the assumption  $\hat{S} \subset \bigcap_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \hat{S}_m$  of Theorem 10 is fulfilled. We define for  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ ,  $K \in m$  and  $m_K \in \widehat{\mathcal{M}}_{\hat{\ell}}$ , the partition  $m_K \vee K$  of  $K$  by (20). A function  $s \in \hat{S}_m$  of the form  $s = \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K$  is piecewise polynomial. In the sequel,  $m(s)$  designs a partition of  $\widehat{\mathcal{M}}$  of the form

$$m(s) = \bigcup_{K \in m} m_K \vee K,$$

with minimal length that is such that

$$|m(s)| = \inf \left\{ \sum_{K \in m} |m_K \vee K|, s = \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K \right\}.$$

Let  $\bar{S} = \bigcup_{k=1}^{\infty} \mathcal{P}_{k,r}$  and note that  $\hat{S}_m \subset \bar{S}$  for all  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ . Let  $s \in \bar{S}$  and  $k \geq 1$  be the smallest integer for which  $s \in \mathcal{P}_{k,r}$ . It follows from Proposition 1 that Assumption 2 is satisfied with  $S = \mathcal{P}_{\hat{\ell} \vee k, r}$  and  $d_{\mathcal{P}_{\hat{\ell} \vee k, r}}(s) = ((\hat{\ell} \vee k) + 2)(r + 2)$ . In particular, for all  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$  and  $s \in \bar{S}$ , we may set since  $\hat{s}_m \in \mathcal{P}_{\hat{\ell}, r}$ ,

$$d_{\hat{s}_m}(s) = \inf_{\substack{k \geq 1 \\ \mathcal{P}_{k,r} \ni s}} ((\hat{\ell} \vee k) + 2)(r + 2).$$

We now define  $d(\cdot)$  for  $s \in \bar{S}$  and  $\Delta(\cdot)$  for  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$  by

$$d(s) = \inf_{\substack{k \geq 1 \\ \mathcal{P}_{k,r} \ni s}} (k + 2)(r + 2), \quad \Delta(m) = \hat{\ell}(r + 2).$$

We define  $d(\cdot)$  arbitrarily when  $s \notin \bar{S}$ . Note that (48) is satisfied. We now define  $L_0 = 6L_1$  and the penalties for  $L \geq L_0$ ,  $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$  and  $s \in \hat{S}_m$  by

$$\text{pen}_{1,m}(s) = L \frac{(r+1)|m(s)| \log_+^2(n/(r+1))}{n}, \quad \text{pen}_2(m) = L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n}.$$

The first penalty satisfies the lower bound (50) since

$$d(s) \leq (r+2)(|m(s)| + 2) \leq 6(r+1)|m(s)| \quad \text{for all } s \in \hat{S}_m.$$

It remains to define  $\text{pen}_1(s)$  for  $s \in \hat{S} = \{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ .

**Claim 6.** For all  $m, m' \in \widehat{\mathcal{M}}$ ,  $|m(\hat{s}_{m'})| \leq |m| + |m'|$ .

*Proof of Claim 6.* We have,

$$\begin{aligned} |m(\hat{s}_{m'})| &\leq \sum_{K \in m} |m'_K \vee K| \\ &\leq \sum_{K \in m} |\{K \cap K', K' \in m, K \cap K' \neq \emptyset\}| \\ &\leq |\{K \cap K', (K, K') \in m \times m', K \cap K' \neq \emptyset\}|. \end{aligned}$$

Since  $m$  and  $m'$  are collections of intervals,  $|m(\hat{s}_{m'})| \leq |m| + |m'|$ . □

It then follows that for all  $m, m' \in \widehat{\mathcal{M}}_{\hat{\ell}}$ ,

$$\text{pen}_{1,m}(\hat{s}_{m'}) \leq L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \text{pen}_2(m).$$

The penalty defined by

$$\text{pen}_1(\hat{s}_{m'}) = L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n}$$

satisfies therefore (49).

Note now that the selection rules described in Sections 6.8 and 4.3 coincide. Theorem 10 controls the risk of the selected estimator: for all  $\xi > 0$ , there exists an event  $\Omega_\xi$  of probability larger than  $1 - e^{-n\xi}$ , and on which:

$$h^2(\mathbf{s}_0, \hat{s}_{\hat{m}}) \leq C \left( \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \{h^2(\mathbf{s}_0, \hat{s}_m) + \text{pen}_2(m)\} + \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \{h^2(\mathbf{s}_0, \hat{s}_m) + \text{pen}_1(\hat{s}_m)\} + \xi \log_+(1/\xi) \right),$$

where  $C$  is a universal constant. By using the definition of the penalty terms,

$$h^2(\mathbf{s}_0, \hat{s}_{\hat{m}}) \leq C' \left\{ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \{h^2(\mathbf{s}_0, \hat{s}_m)\} + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\},$$

where  $C'$  is a universal constant. It then remains to use the fact that  $\hat{s}_m$  is a  $\rho$ -estimator on  $\mathcal{P}_r(m)$  to get a bound on  $h^2(\mathbf{s}_0, \hat{s}_m)$  on the same event  $\Omega_\xi$  (the event that appears in Theorem 3 to control the risk of a  $\rho$ -estimator is the same that the one that appears in Theorem 10. It is, in each case, defined by Lemma 4).  $\square$

**6.8.3. Proof of Theorem 8.** The proof is almost the same as in the case of Theorem 7. The modifications are very mild, and this is the reason why we only specify the values of the different parameters involved in the procedure of Section 6.8:

$$\begin{aligned} \hat{S} &= \{\hat{s}_m, m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}\} \\ \hat{S}_m &= \left\{ \sum_{K \in m} \hat{s}_{m_K} \mathbb{1}_K, m_K \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}} \right\} \quad \text{for all } m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}} \\ \text{pen}_1(\hat{s}_m) &= \text{pen}_2(m) = L \frac{(r+1)|m| \log_+^2(n/(r+1))}{n} \quad \text{for all } m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}. \end{aligned}$$

$\square$

**6.9. Proof of Corollary 2.** This corollary follows from (23) and the lemma below whose proof is given in the next section.

**Lemma 7.** *Suppose that condition (25) is met and  $r = 0$ . Then, for all  $\xi > 0$ , the following holds with probability larger than  $1 - e^{-n\xi}$ : for all  $\ell \in \{1, \dots, \hat{n} - 1\}$ ,  $m \in \mathcal{M}_\ell$  written as in (4) and such that  $Y_{(1)}$  and  $Y_{(\hat{n})}$  belong to  $[x_1, x_{\ell+1}]$ , there exists  $m' \in \widehat{\mathcal{M}}_\ell$  such that the  $\rho$ -estimator  $\hat{s}_{m'}$  satisfies*

$$(53) \quad h^2(\mathbf{s}_0, \hat{s}_{m'}) \leq C \left\{ h^2(\mathbf{s}_0, \mathcal{P}_0(m)) + \frac{\ell \log_+^2(n/\ell)}{n} + \xi \log_+(1/\xi) \right\}.$$

$\square$

**6.10. Proofs of Lemmas 7 and 3.** We set  $\mathcal{M} = \bigcup_{\ell \geq 1} \mathcal{M}_\ell$ . The following claim will be useful in the sequel.

**Claim 7.** *Let  $\xi > 0$  and  $\Omega_\xi$  be the event given by Lemma 4 when  $\mathcal{I}_d$  is the collection of unions of at most  $d$  intervals. Then,  $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ . Let  $\eta \geq 0$ ,  $r \geq 0$ , and  $m, m' \in \mathcal{M}$ . The following holds on  $\Omega_\xi$ : for all piecewise polynomial estimators  $\hat{s}_m \in \mathcal{P}_r(m)$ ,  $\hat{s}_{m'} \in \mathcal{P}_r(m')$  such that  $T(\hat{s}_m, \hat{s}_{m'}) \geq -\eta$ ,*

$$h^2(\mathbf{s}_0, \hat{s}_{m'}) \leq C \left\{ h^2(\mathbf{s}_0, \hat{s}_m) + \frac{(r+1)(|m| + |m'|)}{n} \log_+^2 \left( \frac{n}{(r+1)(|m| + |m'|)} \right) + \xi \log_+(1/\xi) + \eta \right\}.$$

Moreover, if  $\hat{s}_m$  is a  $\rho$ -estimator on  $\mathcal{P}_r(m)$ ,

$$h^2(\mathbf{s}_0, \hat{s}_m) \leq C \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + \frac{(r+1)|m|}{n} \log_+^2 \left( \frac{n}{(r+1)|m|} \right) + \xi \log_+(1/\xi) \right\}.$$

In the above inequalities,  $C$  is universal.

*Proof.* Let  $\varepsilon = 1/24$ . On  $\Omega_\xi$ :

$$(54) \quad T(\hat{s}_m, \hat{s}_{m'}) \leq (3 + 4\varepsilon)h^2(\mathbf{s}_0, \hat{s}_m) - \frac{4 - 3\varepsilon}{12}h^2(\mathbf{s}_0, \hat{s}_{m'}) + c_1\vartheta(d_{\hat{s}_{m'}}(\hat{s}_m)) + c_2\xi \log_+(1/\xi),$$

where  $c_1$  and  $c_2$  are universal constants. Now,  $\hat{s}_m$  and  $\hat{s}_{m'}$  belong to  $\mathcal{P}_r(m'')$  where

$$m'' = \{K \cap K', (K, K') \in m \times m', K \cap K' \neq \emptyset\}.$$

Yet,  $|m''| \leq |m| + |m'|$ . Thereby,  $\hat{s}_m$  and  $\hat{s}_{m'}$  belong to  $\mathcal{P}_{|m|+|m'|,r}$  and it follows from Proposition 1 that we may set

$$d_{\hat{s}_m'}(\hat{s}_m) = (|m| + |m'| + 2)(r + 2).$$

We now bound  $\vartheta(d_{\hat{s}_m'}(\hat{s}_m))$  from above in (54), and then use  $T(\hat{s}_m, \hat{s}_{m'}) \geq -\eta$  to prove the first inequality of the claim. The second one follows from Theorem 3 and Proposition 1.  $\square$

*Proof of Lemma 7.* Let  $m \in \mathcal{M}'_\ell$  be a collection written as

$$m = \{[x_1, x_2], (x_2, x_3], (x_3, x_4], \dots, (x_\ell, x_{\ell+1}]\}$$

and such that  $x_1 \leq Y_{(1)}$ , and  $Y_{(\hat{n})} \leq x_{\ell+1}$ . We may define a partition  $\bar{m} \in \mathcal{M}'_\ell$  of the form

$$\bar{m} = \{[\bar{x}_1, \bar{x}_2], (\bar{x}_2, \bar{x}_3], (\bar{x}_3, \bar{x}_4], \dots, (\bar{x}_\ell, \bar{x}_{\ell+1}]\}$$

where  $\bar{x}_1 = Y_{(1)}$  and  $\bar{x}_{\ell+1} = Y_{(\hat{n})}$  and whose intervals are included into the ones of  $m$ .

Let  $\hat{s}_m$  and  $\hat{s}_{\bar{m}}$  be  $\rho$ -estimators on  $\mathcal{P}_0(m)$  and  $\mathcal{P}_0(\bar{m})$  respectively. We order the intervals of  $\bar{m}$  as follows. We define  $\ell$  intervals  $I_1, \dots, I_\ell$  such that  $\bar{m} = \{I_1, \dots, I_\ell\}$  and such that the value  $\hat{s}_{\bar{m}}$  on  $I_j$ , denoted by  $\hat{s}_{I_j}$ , is non-decreasing when  $j$  grows up. We denote the endpoints of  $I_j$  by  $a_j < b_j$ . We now define  $j_1$  as the largest integer of  $\{1, \dots, \hat{n}\}$  such that  $Y_{(j_1)} \leq a_j$  and  $j_2$  as the smallest integer such that  $Y_{(j_2)} \geq b_j$ . When  $j_1 = 1$ , we set  $K_j = [Y_{(j_1)}, Y_{(j_2)}]$  and when  $j_1 \neq 1$ , we set  $K_j = (Y_{(j_1)}, Y_{(j_2)}]$ . Note that  $K_j$  is the smallest interval containing  $I_j$  that is either of the form  $[Y_{(j_1)}, Y_{(j_2)}]$  or  $(Y_{(j_1)}, Y_{(j_2)}]$ .

Define  $J_1 = K_1$  and for  $j \in \{2, \dots, \ell\}$ ,  $J_j = K_j \setminus \bigcup_{i=1}^{j-1} K_i$ . Since  $K_i \not\subset K_j$  when  $i \neq j$ ,  $K_j \setminus K_i$  is an interval. Therefore,  $J_j = \bigcap_{i=1}^j (K_j \setminus K_i)$  is also an interval. When it is not empty, it is either of the form  $[Y_{(1)}, Y_{(i)}]$  with  $i > 1$  or  $(Y_{(i_1)}, Y_{(i_2)}]$  with  $i_1 < i_2$ . The collection  $\bar{m}' = \{J_j, j \in \{1, \dots, \ell\}\}$  defines therefore a partition of  $[Y_{(1)}, Y_{(\hat{n})}]$  that belongs to  $\widehat{\mathcal{M}}_{\ell'}$  with  $\ell' \leq \ell$  (we must remove the empty sets). Let  $s$  be the step function of  $\mathcal{P}_0(\bar{m}')$  defined by

$$s = \sum_{j=1}^{\ell} \hat{s}_{I_j} \mathbb{1}_{J_j}.$$

We now prove that  $s \leq \hat{s}_{\bar{m}}$ . When  $t \notin [Y_{(1)}, Y_{(\hat{n})}]$ ,  $s(t) = \hat{s}_{\bar{m}}(t) = 0$ . When  $t \in [Y_{(1)}, Y_{(\hat{n})}]$ , there exist  $j \in \{1, \dots, \ell\}$  such that  $t \in I_j$  and  $j' \leq j$  such that  $t \in J_{j'}$ . Therefore,  $s(t) = \hat{s}_{I_{j'}}$ . By using that  $\hat{s}_{I_{j'}} \leq \hat{s}_{I_j}$ , we finally deduce that  $s(t) \leq \hat{s}_{\bar{m}}(t)$ .

Consider an interval  $I_j \in \bar{m}$  and let us denote the cardinal of  $\{Y_{(i)}, Y_{(i)} \in I_j, i \in \{1, \dots, \hat{n}\}\}$  by  $k_j$ . When  $k_j \geq 3$ , there exists at least  $k_j - 2$  random variables  $Y_{(i)}$  that belong to  $I_j$  but not to

$\cup_{j' \in \{1, \dots, \hat{n}\}, j' \neq j} K_{j'}$ . Such  $Y_{(i)}$  belong therefore to  $J_j$  and satisfy  $s(Y_{(i)}) = \hat{s}_{\bar{m}}(Y_{(i)})$ . Therefore,

$$\begin{aligned} \left| \{Y_{(i)}, s(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}), i \in \{1, \dots, \hat{n}\}\} \right| &= \sum_{j=1}^{\ell} \left| \{Y_{(i)}, s(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}), Y_{(i)} \in I_j, i \in \{1, \dots, \hat{n}\}\} \right| \\ (55) \qquad \qquad \qquad &\leq 2\ell. \end{aligned}$$

It follows from  $s \leq \hat{s}_{\bar{m}}$ , (55) and (25) that  $T(s, \hat{s}_{\bar{m}}) \leq 2\ell/n$ . We now use Claim 7 to get on  $\Omega_{\xi}$

$$(56) \qquad h^2(\mathbf{s}_0, s) \leq C \left\{ h^2(\mathbf{s}_0, \hat{s}_{\bar{m}}) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\},$$

where  $C$  is universal.

We may refine the partition  $\bar{m}' \in \widehat{\mathcal{M}}_{\ell'}$  to get  $m' \in \widehat{\mathcal{M}}_{\ell}$  such that  $\mathcal{P}_0(\bar{m}') \subset \mathcal{P}_0(m')$ . Let  $\hat{s}_{m'}$  and  $\hat{s}_{\bar{m}'}$  be  $\rho$ -estimators on  $\mathcal{P}_0(m')$  and  $\mathcal{P}_0(\bar{m}')$  respectively. There exists a universal constant  $C'$  such that on  $\Omega_{\xi}$ :

$$h^2(\mathbf{s}_0, \hat{s}_{m'}) \leq C' \left\{ h^2(\mathbf{s}_0, \mathcal{P}_0(m')) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}.$$

By using that  $s \in \mathcal{P}_0(\bar{m}') \subset \mathcal{P}_0(m')$  and (56),

$$\begin{aligned} h^2(\mathbf{s}_0, \hat{s}_{m'}) &\leq C' \left\{ h^2(\mathbf{s}_0, s) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}, \\ (57) \qquad \qquad \qquad &\leq C'' \left\{ h^2(\mathbf{s}_0, \hat{s}_{\bar{m}}) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}, \end{aligned}$$

where  $C''$  is universal. Note now that  $\hat{s}_{\bar{m}} \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]} \in \mathcal{P}_0(\bar{m})$  and thus  $T(\hat{s}_{\bar{m}}, \hat{s}_{\bar{m}} \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]}) \leq 0$  as  $\hat{s}_{\bar{m}}$  is a  $\rho$ -estimator on the convex model  $\mathcal{P}_0(\bar{m})$  (see Theorem 5 and Lemma 2). Now,

$$\begin{aligned} T(\hat{s}_{\bar{m}}, \hat{s}_m) &= T(\hat{s}_{\bar{m}}, \hat{s}_{\bar{m}} \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]}) + \frac{1}{4} \left( \int_{\mathbb{R}} \hat{s}_m \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]} dM - \int_{\mathbb{R}} \hat{s}_{\bar{m}} dM \right) \\ &\leq 0. \end{aligned}$$

Therefore, Claim 7 asserts that

$$(58) \qquad h^2(\mathbf{s}_0, \hat{s}_{\bar{m}}) \leq C''' \left\{ h^2(\mathbf{s}_0, \hat{s}_m) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\},$$

where  $C'''$  is universal. By using that  $\hat{s}_m$  is a  $\rho$ -estimator,

$$(59) \qquad h^2(\mathbf{s}_0, \hat{s}_m) \leq C'''' \left\{ h^2(\mathbf{s}_0, \mathcal{P}_0(m)) + \frac{\ell}{n} \log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}.$$

It remains to put inequalities (57), (58) and (59) together to finish the proof.  $\square$

*Proof of Lemma 3.* Note that we may always suppose that

$$\{K \cap \{Y_{(1)}, \dots, Y_{(\hat{n})}\}, K \in m\}$$

contains  $Y_{(1)}$  and  $Y_{(\hat{n})}$  (up to an increase of  $|m|$  by 2). Let

$$m_1 = \{K \in m, \{Y_{(1)}, \dots, Y_{(\hat{n})}\} \cap K \neq \emptyset\}.$$

Then,  $m_1 \neq \emptyset$  and we may write  $m_1 = \{K_j, j \in \{1, \dots, k\}\}$  where  $1 \leq k \leq \ell$  and where  $K_j$  is an interval with endpoints  $a_j, b_j$  satisfying  $a_1 < b_1 \leq a_2 < b_2 < \dots$ .

Let us recall that the  $\rho$ -estimator  $\hat{s}_m$  is of the form

$$\hat{s}_m = \sum_{K \in m} \hat{s}_K \quad \text{where } \hat{s}_K \text{ maximizes } \mathcal{L}_K(\cdot) \text{ over } \mathcal{P}_r(K).$$

When  $K \in m$  does not belong to  $m_1$ ,  $\hat{s}_K = 0$  and hence

$$\hat{s}_m = \sum_{j=1}^k \hat{s}_{K_j}.$$

For each  $j \in \{1, \dots, k\}$ , we set  $\alpha_j = \min \{Y_{(i)}, Y_{(i)} \in K_j\}$ ,  $\beta_j = \max \{Y_{(i)}, Y_{(i)} \in K_j\}$ . We define for  $j \in \{2, \dots, k-1\}$ ,  $J_{2j} = (\beta_j, \alpha_{j+1}]$  and for  $j \in \{2, \dots, k\}$ ,  $J_{2j-1} = (\alpha_j, \beta_j]$ . If  $\beta_1 = Y_{(1)}$ , we set  $J_1 = \emptyset$ ,  $J_2 = [\beta_1, \alpha_2]$  and if  $\beta_1 > Y_{(1)}$ ,  $J_1 = [Y_{(1)}, \beta_1]$ ,  $J_2 = (\beta_1, \alpha_2]$ . Note that  $J_{2j-1} \subset K_j$  for all  $j \in \{1, \dots, k\}$ . The collection  $m' = \{J_j, j \in \{1, \dots, 2k-1\}\}$  defines a partition of  $\widehat{\mathcal{M}}$  such that  $|m'| \leq 2k-1$ . We define the  $\rho$ -estimator

$$\hat{s}_{m'} = \sum_{j=1}^k \hat{s}_{J_{2j-1}} + \sum_{j=1}^{k-1} \hat{s}_{J_{2j}},$$

where  $\hat{s}_A$  maximizes  $\mathcal{L}_A(\cdot)$  over  $\mathcal{P}_r(A)$  for all non-empty interval  $A$  with the convention that  $\hat{s}_\emptyset = 0$  when  $A = \emptyset$ . We now consider

$$\tilde{s}_{m'} = \sum_{j=1}^k \hat{s}_{J_{2j-1}}.$$

Note that  $\tilde{s}_{m'}$  also belongs to the random model  $\mathcal{P}_r(m')$  and hence  $T(\hat{s}_{m'}, \tilde{s}_{m'}) \leq 0$ . We deduce from Claim 7 and from  $|m'| \leq 2k-1 \leq 2\ell-1$ , that on  $\Omega_\xi$

$$(60) \quad h^2(\mathbf{s}_0, \hat{s}_{m'}) \leq C \left\{ h^2(\mathbf{s}_0, \tilde{s}_{m'}) + \frac{(r+1)\ell}{n} \log_+^2 \left( \frac{n}{(r+1)\ell} \right) + \xi \log_+(1/\xi) \right\},$$

where  $C$  is universal.

Now, for all  $j \in \{1, \dots, k\}$ , such that  $J_{2j-1} \neq \emptyset$ ,

$$(61) \quad T(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}) \leq 0,$$

since  $\hat{s}_{J_{2j-1}}$  maximizes  $\mathcal{L}_{J_{2j-1}}(\cdot)$  over  $\mathcal{P}_r(J_{2j-1})$  and that  $\hat{s}_{K_j} \mathbb{1}_{J_{2j-1}} \in \mathcal{P}_r(J_{2j-1})$ . When  $J_{2j-1} = \emptyset$ ,  $T(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}) = 0$ , and thus (61) also holds.

We define

$$A = \bigcup_{j=1}^k J_{2j-1}.$$

We deduce from (61) that  $T(\tilde{s}_{m'} \mathbb{1}_A, \hat{s}_m \mathbb{1}_A) \leq 0$ . Therefore,

$$\begin{aligned} T(\tilde{s}_{m'}, \hat{s}_m) &= T(\tilde{s}_{m'} \mathbb{1}_A, \hat{s}_m \mathbb{1}_A) + T(0, \hat{s}_m \mathbb{1}_{A^c}) \\ &\leq 0 + T(0, \hat{s}_m \mathbb{1}_{A^c}) \\ &\leq \int_{A^c} \psi(\hat{s}_m/0) \, dN, \end{aligned}$$

where we recall the conventions  $\psi(0/0) = \psi(1) = 0$ ,  $\psi(x/0) = \psi(\infty) = 1$  for all  $x > 0$ . Let  $B = \bigcup_{j=1}^k K_j$ . Note that  $\hat{s}_m$  vanishes outside  $B$  and thus, as  $|\psi| \leq 1$ ,

$$(62) \quad T(\tilde{s}_{m'}, \hat{s}_m) \leq \int_{B \cap A^c} \psi(\hat{s}_m/0) \, dN \leq N(B \cap A^c).$$

Now,

$$N(B \cap A^c) = \sum_{j=1}^k \{N(K_j) - N(J_{2j-1})\}.$$

Since  $\alpha_j, \beta_j \in \{Y_i, i \in \hat{I}\}$ , we deduce from (25) that  $N(K_j) - N(J_{2j-1}) = N(\{\alpha_j\})$ . In each of the frameworks,  $N(\{\alpha_j\}) \leq 1/n$  and thus  $N(B \cap A^c) \leq k/n$ . By using (62), we get  $T(\tilde{s}_{m'}, \hat{s}_m) \leq k/n$ . Claim 7 with  $\eta = k/n \leq \ell/n$  ensures that on  $\Omega_\xi$ :

$$h^2(\mathbf{s}_0, \tilde{s}_{m'}) \leq C' \left\{ h^2(\mathbf{s}_0, \hat{s}_m) + \frac{(r+1)\ell}{n} \log_+^2 \left( \frac{n}{(r+1)\ell} \right) + \xi \log_+(1/\xi) \right\},$$

where  $C''$  is universal. Since  $\hat{s}_m$  is a  $\rho$ -estimator on  $\mathcal{P}_r(m)$ , we deduce that on the same event  $\Omega_\xi$ :

$$h^2(\mathbf{s}_0, \hat{s}_m) \leq C'' \left\{ h^2(\mathbf{s}_0, \mathcal{P}_r(m)) + \frac{(r+1)\ell}{n} \log_+^2 \left( \frac{n}{(r+1)\ell} \right) + \xi \log_+(1/\xi) \right\},$$

where  $C''$  is universal. It then remains to combine the two last inequalities with (60) to finish the proof.  $\square$

## REFERENCES

- [AD10] Nathalie Akakpo and Cécile Durot. Histogram selection for possibly censored data. *Mathematical Methods of Statistics*, 19(3):189–218, 2010.
- [Ant89] Anestis Antoniadis. A penalty method for nonparametric estimation of the intensity function of a counting process. *Annals of the Institute of Statistical Mathematics*, 41(4):781–807, 1989.
- [Bar11] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401, 2011.
- [Bar16] Yannick Baraud. Bounding the expectation of the supremum of an empirical process over a (weak) vc-major class. *Electronic journal of statistics*, 10(2):1709–1728, 2016.
- [BB09] Yannick Baraud and Lucien Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284, 2009.
- [BB16] Yannick Baraud and Lucien Birgé.  $\rho$ -estimators for shape restricted density estimation. *Stochastic Processes and their Applications*, 126(12):3888–3912, 2016.
- [BB17] Yannick Baraud and Lucien Birgé.  $\rho$ -estimators revisited: general theory and applications. *arXiv preprint*, 2017.
- [BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.
- [BBS17] Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection:  $\rho$ -estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.
- [BC05] Elodie Brunel and Fabienne Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā: The Indian Journal of Statistics*, pages 441–475, 2005.
- [BC08] Elodie Brunel and Fabienne Comte. Adaptive estimation of hazard rate with censored data. *Communications in Statistics—Theory and Methods*, 37(8):1284–1305, 2008.

- [Bel18] Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.
- [Bir06] Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l’Institut Henri Poincaré. Probabilités et Statistique*, 42(3):273–325, 2006.
- [BM93] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.
- [BM98] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [BT15] Pierre C Bellec and Alexandre B Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16(56):1879–1892, 2015.
- [Cas99] Gwénaëlle Castellan. Modified akaike’s criterion for histogram density estimation. *Technical report*, 1999.
- [CGS15] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.
- [Cha14] Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.
- [CL19] Sabyasachi Chatterjee and John Lafferty. Adaptive risk bounds in unimodal regression. *Bernoulli*, 25(1):1–25, 2019.
- [CR04] Fabienne Comte and Yves Rozenholc. A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, 56(3):449–473, 2004.
- [DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.
- [DR02] Sebastian Dohler and Ludger Ruschendorf. Adaptive estimation of hazard functions. *Probability and mathematical statistics - Wroclaw University*, 22(2):355–379, 2002.
- [DY90] R. DeVore and X. Yu. Degree of adaptive approximation. *Mathematics of Computation*, 55:625–635, 1990.
- [GK06] Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.
- [GN15] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.
- [Gre56] Ulf Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(1):70–96, 1956.
- [GS15] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1-2):379–411, 2015.
- [Kan92] Yuichiro Kanazawa. An optimal variable cell histogram based on the sample spacings. *The Annals of Statistics*, 20(1):291–304, 1992.
- [Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2007. École d’été de Probabilités de Saint-Flour.
- [Oka59] Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10(1):29–35, 1959.

- [Pla11] Sandra Plancade. Non parametric estimation of hazard rate in presence of censoring. *Metrika*, 74(3):313–347, 2011.
- [RB06] Patricia Reynaud-Bouret. Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.
- [Sar14] Mathieu Sart. Estimation of the transition density of a Markov chain. *Annales de l’Institut Henri Poincaré. Probabilités et Statistique*, 50(3):1028–1068, 2014.
- [Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.
- [vdG95] Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, 23(5):1779–1801, 1995.
- [Zha02] Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.

## APPENDIX A. A UNIFORM EXPONENTIAL INEQUALITY.

In this section, we present the key tool to control the error due to the approximation  $T(s, s') \simeq T_E(s, s')$ . It is used in the proof of Theorem 3. Its proof is given in Appendix D.

We introduce for any bounded function  $f \in \mathcal{S}$ , the centered random variable

$$Z(f) = \int_{\mathbb{R}} f \, dN - \int_{\mathbb{R}} f \mathbf{s}_0 \, dM.$$

It can also be written as

$$Z(f) = \begin{cases} \frac{1}{n} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) & \text{in framework 1,} \\ \frac{1}{n} \sum_{i=1}^n \left( f(X_i) \mathbb{1}_{D_i=1} - \int_0^\infty f(t) \mathbf{s}_0(t) \mathbb{1}_{X_i \geq t} \, dt \right) & \text{in framework 2,} \\ \frac{1}{n} \sum_{i=1}^n \left( f(T_{1,0}^{(i)}) \mathbb{1}_{I_{\text{obs}}}(T_{1,0}^{(i)}) - \int_{I_{\text{obs}}} f(t) \mathbf{s}_0(t) \mathbb{1}_{X_{t-}^{(i)}=1} \, dt \right) & \text{in framework 3.} \end{cases}$$

This variable  $Z(f)$  measures the approximation error of  $T_E(s, s')$  by  $T(s, s')$  when  $f = \psi(s'/s)$ . The theorem below allows to control the deviations of  $Z(f)$ .

**Theorem 11.** *Let  $\mathcal{F} \subset \mathcal{S}$  be a set of functions  $f$  such that  $|f(t)| \leq 1$  for all  $f \in \mathcal{F}$ ,  $t \in \mathbb{R}$ . Let  $\mathcal{A}$  be the collection of sets defined by*

$$\mathcal{A} = \{ \{t \in \mathbb{R}, f_+(t) > u\}, f \in \mathcal{F}, u \in (0, 1) \} \cup \{ \{t \in \mathbb{R}, f_-(t) > u\}, f \in \mathcal{F}, u \in (0, 1) \}.$$

*Suppose:*

- *in frameworks 1 and 2 that  $\mathcal{A}$  is a Vapnik-Chervonenkis class of dimension at most  $2d$ . Moreover, there exists an at most countable set  $\mathcal{A}' \subset \mathcal{A}$  satisfying the following technical assertion: for all  $A \in \mathcal{A}$ , there exists a sequence  $(A_m)_{m \geq 0} \in \mathcal{A}'^{\mathbb{N}}$  such that  $\lim_{m \rightarrow +\infty} \mathbb{1}_{A_m}(t) = \mathbb{1}_A(t)$  for every  $t \in \mathbb{R}$ .*
- *in framework 3, that each set  $A \in \mathcal{A}$  is a union of at most  $d$  intervals.*

*Let, for  $f \in \mathcal{F}$ ,*

$$v(f) = \int_{\mathbb{R}} f^2 \mathbf{s}_0 \, dM = \begin{cases} \mathbb{E}[f^2(X)] & \text{in framework 1,} \\ n^{-1} \sum_{i=1}^n \int_0^\infty f^2(t) \mathbf{s}_0(t) \mathbb{1}_{X_i \geq t} \, dt & \text{in framework 2,} \\ n^{-1} \sum_{i=1}^n \int_{I_{\text{obs}}} f^2(t) \mathbf{s}_0(t) \mathbb{1}_{X_{t-}^{(i)}=1} \, dt & \text{in framework 3.} \end{cases}$$

*Then, there exists for all  $\xi > 0$  an event which holds true with probability larger than  $1 - e^{-n\xi}$  and on which: for all  $f \in \mathcal{F}$ , and  $\varepsilon > 0$ ,*

$$(63) \quad |Z(f)| \leq \varepsilon v(f) + C_\varepsilon \left\{ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right\}.$$

*In the above inequality,  $C_\varepsilon$  only depends on  $\varepsilon$ .*

Theorem 11 is well tailored for bounding the risk of a  $\rho$ -estimator. Indeed, when  $f = \psi(s'/s)$ , the random variable  $v(f)$  can be bounded from above as follows: for all  $s, s' \in \mathcal{S}$ ,

$$v(\psi(s'/s)) \leq 4(h^2(\mathbf{s}_0, s) + h^2(\mathbf{s}_0, s')).$$

For the proof of this inequality, we refer to Claim 3 page 26. Under suitable assumptions on the collection  $\mathcal{F} = \{\psi(s'/s), s, s' \in S\}$ , Inequality (63) roughly says that with high probability (and  $\varepsilon = 1/24$ ):

$$(64) \quad |T(s, s') - T_E(s, s')| \leq \frac{1}{6} (h^2(\mathbf{s}_0, s) + h^2(\mathbf{s}_0, s')) + R_S(n) \quad \text{for all } s, s' \in S.$$

The term  $R_S(n)$  depends on the probability of the event on which (64) holds true and the complexity of  $S$ . The approximation  $T(s, s') \simeq T_E(s, s')$  is then accurate enough to control the risk of a  $\rho$ -estimator  $\hat{s}$ . Indeed, we deduce from (2), that for all  $s, s' \in S$ ,

$$\frac{1}{6}h^2(\mathbf{s}_0, s) - \frac{19}{6}h^2(\mathbf{s}_0, s') - R_S(n) \leq T(s, s') \leq \frac{19}{6}h^2(\mathbf{s}_0, s) - \frac{1}{6}h^2(\mathbf{s}_0, s') + R_S(n).$$

We may then replace  $T_E(s, s')$  by  $T(s, s')$  and  $\gamma_E(\cdot)$  by  $\gamma(\cdot)$  in the computations of Section 2.2 to bound the risk of a  $\rho$ -estimator  $\hat{s}$ . This leads to

$$h^2(\mathbf{s}_0, \hat{s}) \leq 37h^2(\mathbf{s}_0, S) + 12R_S(n) + 6/n.$$

For a rigorous result, we refer to Theorem 3.

As a by-product of the proof of Theorem 11 we get the following proposition (to be proved in Section D.6) which may be of independent interest:

**Proposition 12.** *Consider framework 1 and an at most countable set  $\mathcal{F} \subset \mathcal{S}$  of functions  $f$  such that  $|f(t)| \leq 1$  for all  $t \in \mathbb{R}$ ,  $f \in \mathcal{F}$ . Let for  $u \in (0, 1)$ ,  $\mathcal{A}_u$  be the collection of sets defined by*

$$\mathcal{A}_u = \{\{t \in \mathbb{R}, f_+(t) > u\}, f \in \mathcal{F}\} \cup \{\{t \in \mathbb{R}, f_-(t) > u\}, f \in \mathcal{F}\},$$

*and  $S_{\mathcal{A}_u}(2n)$  be the Vapnik-Chervonenkis shatter coefficient*

$$S_{\mathcal{A}_u}(2n) = \max_{t_1, \dots, t_{2n} \in \mathbb{R}} |\{\{t_1, \dots, t_{2n}\} \cap A, A \in \mathcal{A}_u\}|.$$

*Let  $\sigma^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X)]$  and  $r_n = \sup_{u \in (0, 1)} \log_+ |S_{\mathcal{A}_u}(2n)|$ . Then, there exist universal constants  $C, C'$  such that*

$$(65) \quad \begin{aligned} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} |Z(f)| \right] &\leq C \inf_{\eta \in (0, 1)} \left\{ \sigma \sqrt{\log(1/\eta)} + \int_0^\eta \sup_{f \in \mathcal{F}} \sqrt{\mathbb{P}(|f(X)| > u)} du \right\} \sqrt{\frac{r_n}{n}} + C \frac{r_n}{n} \\ &\leq C' \left[ \sigma \sqrt{\frac{r_n \log_+(1/\sigma)}{n}} + \frac{r_n}{n} \right]. \end{aligned}$$

This proposition gives a bound on  $\mathbb{E} [\sup_{f \in \mathcal{F}} |Z(f)|]$  that involves the Vapnik-Chervonenkis shatter coefficients  $S_{\mathcal{A}_u}(2n)$  of  $\mathcal{A}_u$ . This result may not be as sharp as the bounds based on covering numbers (see Theorem 3.1 of [GK06]). It is, however, rather convenient in the situations where the shatter coefficients are easier to control than the covering numbers.

Our proposition may be viewed as a refined version of a result of [Bar16] when the random variables  $X_i$  are identically distributed and when  $\mathcal{A}_u$  is Vapnik-Chervonenkis with dimension  $d$  (apart from constants). Such an assumption corresponds to a notion of (weak) VC-major class. In that case, Sauer's lemma [Sau72] implies

$$(66) \quad \mathbb{E} \left[ \sup_{f \in \mathcal{F}} |Z(f)| \right] \leq C'' \left[ \sigma \sqrt{\frac{d \log_+(n/d) \log_+(1/\sigma)}{n}} + \frac{d \log_+(n/d)}{n} \right],$$

where  $C''$  is a number. If we put aside the constant  $C''$ , the main difference between this bound and Inequality (2.8) of [Bar16] lies in the position of the logarithmic term  $\log_+(1/\sigma)$ : it is here involved inside the square root while it is outside in [Bar16].

APPENDIX B. ROBUSTNESS OF  $\rho$ - AND MAXIMUM LIKELIHOOD ESTIMATORS

We show here that the behaviour of a  $\rho$ -estimator can be very different from that of a maximum likelihood estimator in framework 2.

**Proposition 13.** *Let  $\alpha > 0$ ,  $s_\alpha(t) = t^{-1} \mathbb{1}_{t \geq \alpha}$  be the hazard rate of a Pareto distribution, and the model  $S = \{s_\alpha, \alpha > 0\}$ . Suppose that the data are not censored, that is  $C = +\infty$  almost surely, and that the target hazard rate is  $\mathbf{s}_0(t) = f_0(t)/\mathbb{P}(T \geq t)$  where  $f_0$  is a mixture of two Pareto distributions defined for  $\eta \in (0, 1/2)$  by*

$$f_0(t) = \frac{\eta}{n} \frac{\mathbb{1}_{t \geq \eta}}{t^2} + \left(1 - \frac{1}{n}\right) \frac{\mathbb{1}_{t \geq 1}}{t^2}.$$

*Then, there exists a universal positive constant  $c$  such that the risk of a maximum likelihood estimator  $\tilde{s}$  is bounded from below by*

$$\mathbb{E}[h^2(\mathbf{s}_0, \tilde{s})] > c \log(1/(2\eta)).$$

*By contrast, the risk of a  $\rho$ -estimator can be bounded from above by*

$$\sup_{\eta \in (0, 1/2)} \mathbb{E}[h^2(\mathbf{s}_0, \hat{s})] \leq C \frac{\log^2 n}{n},$$

*where  $C$  is a universal constant.*

The risk bound  $\log^2 n/n$  of the  $\rho$ -estimator is likely not optimal. However, we observe that it does not increase when  $\eta$  runs from  $1/2$  to  $0$ . This is different for the maximum likelihood estimator: its risk may be made arbitrarily large by playing with  $\eta$ .

*Proof of Proposition 13.* The survival function  $G$  and the hazard rate  $\mathbf{s}_0$  of  $T$  are given by

$$G(t) = \mathbb{1}_{t \in [0, \eta)} + \left(1 - \frac{1}{n} + \frac{\eta}{nt}\right) \mathbb{1}_{t \in [\eta, 1)} + \left(1 - \frac{1}{n} + \frac{\eta}{n}\right) \frac{\mathbb{1}_{t \geq 1}}{t},$$

$$\mathbf{s}_0(t) = \frac{\eta}{(n-1)t + \eta} \frac{\mathbb{1}_{t \in [\eta, 1)}}{t} + \frac{\mathbb{1}_{t \geq 1}}{t}.$$

The maximum likelihood estimator on  $S = \{s_\alpha, \alpha > 0\}$  is  $\tilde{s} = s_{\tilde{\alpha}}$  where  $\tilde{\alpha} = \min_{1 \leq i \leq n} X_i$ .

Let  $\mathcal{A}_n$  be the event on which there exists at least one observation  $X_i$  smaller than  $2\eta$ . Then,

$$\mathbb{P}(\mathcal{A}_n) = 1 - (1 - 1/(2n))^n > 1 - e^{-1/2} > 0.39.$$

Let now  $t_0$  be defined by the relation  $G(t_0) = 2/3$  and  $\mathcal{B}_n$  be the event on which  $n/2$  observations  $X_i$  are not smaller than  $t_0$ . Then,  $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(Y \geq n/2)$  where  $Y$  is binomially distributed with parameters  $(n, 2/3)$ . Therefore,  $\mathbb{P}(\mathcal{B}_n) \geq 0.74$  as  $n \geq 3$ , and

$$(67) \quad \mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n) \geq 1 - (\mathbb{P}(\mathcal{A}_n^c) + \mathbb{P}(\mathcal{B}_n^c)) \geq 0.13.$$

Define the empirical survival function  $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \geq t}$ . Then,  $G_n(t_0) \geq 1/2$  on  $\mathcal{B}_n$ . Since  $G(1) \geq 1 - 1/n \geq 2/3$  when  $n \geq 3$ ,  $t_0$  is not smaller than  $1$ . We deduce on  $\mathcal{B}_n$  that  $G_n(t) \geq 1/2$  for all  $t \leq 1$ . Therefore,

$$h^2(\mathbf{s}_0, \tilde{s}) \geq \frac{1}{4} \int_{2\eta}^1 \left( \sqrt{\tilde{s}(t)} - \sqrt{\mathbf{s}_0(t)} \right)^2 dt.$$

Now, on  $\mathcal{A}_n$ ,  $\tilde{s}(t) = 1/t$  when  $t \geq 2\eta$  and hence using the definition of  $\mathbf{s}_0$ :

$$h^2(\mathbf{s}_0, \tilde{s}) \geq \frac{1}{4} \int_{2\eta}^1 \frac{1}{t} \left( 1 - \sqrt{\frac{\eta}{(n-1)t + \eta}} \right)^2 dt.$$

Now,

$$\inf_{t \in [2\eta, 1]} \left( 1 - \sqrt{\frac{\eta}{(n-1)t + \eta}} \right)^2 \geq \left( 1 - \sqrt{\frac{\eta}{2(n-1)\eta + \eta}} \right)^2 > 0.17.$$

Therefore,  $h^2(\mathbf{s}_0, \tilde{s}) > 0.04 \log(1/(2\eta))$  and

$$\mathbb{E}[h^2(\mathbf{s}_0, \tilde{s})] > 0.04 \log(1/(2\eta)) \mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n).$$

It remains to use (67) to conclude.

We now turn to  $\rho$ -estimation. The model  $S$  fulfils Assumption 2 with  $d_S(\cdot) = 1$ ,  $\bar{S} = S$  and where  $\mathcal{I}_1$  is the collection of intervals. A  $\rho$ -estimator  $\hat{s}$  on  $S$  satisfies therefore

$$\mathbb{E}[h^2(\mathbf{s}_0, \hat{s})] \leq C \left\{ \mathbb{E}[h^2(\mathbf{s}_0, S)] + \frac{\log^2 n}{n} \right\}.$$

Now, the bias term can be bounded above by  $\mathbb{E}[h^2(\mathbf{s}_0, S)] \leq \mathbb{E}[h^2(\mathbf{s}_0, s_1)]$  with  $s_1(t) = t^{-1} \mathbb{1}_{t \geq 1}$ . Since  $\mathbf{s}_0$  differs from  $s_1$  only on  $[\eta, 1)$ ,

$$\mathbb{E}[h^2(\mathbf{s}_0, S)] \leq \frac{1}{2} \int_{\eta}^1 \left( \sqrt{\mathbf{s}_0(t)} - 0 \right)^2 G(t) dt = \frac{\mathbb{P}(T \in [\eta, 1])}{2} = \frac{G(\eta) - G(1)}{2} = \frac{1 - \eta}{n} \leq \frac{1}{n}.$$

This concludes the proof.  $\square$

## APPENDIX C. NUMERICAL COMPLEXITY

We discuss here the numerical complexity of the procedure defined in Section 4.3.

Optimizing a criterion  $\text{crit}(\cdot)$  on  $\widehat{\mathcal{M}}_{\hat{\ell}}$  is not an easy task as  $\widehat{\mathcal{M}}_{\hat{\ell}}$  is often very large: computing all the  $\text{crit}(m)$  is numerically prohibitive. Luckily, dynamic programming allows to solve more efficiently the optimization problem when the criterion takes the form

$$(68) \quad \text{crit}(m) = \sum_{K \in m} F(K).$$

Let  $\mathcal{K}$  be the collection of intervals with endpoints in  $\{Y_{(i)}, i \in \{1, \dots, \hat{n}\}\}$ . Let  $\kappa$  be the number of computations needed to compute all the  $F(K)$  when  $K$  varies among  $\mathcal{K}$ . Then, we may minimize or maximize  $\text{crit}(\cdot)$  on  $\widehat{\mathcal{M}}_{\hat{\ell}}$  by a dynamic programming algorithm in at most  $\kappa + \mathcal{O}(\hat{n}^2 \hat{\ell})$  operations. A description of this algorithm may be found in [Kan92] and in Section 4.2 of [CR04].

Here,  $\gamma_2(\hat{s}_m)$  is equal to  $\gamma_2(\hat{s}_m) = \sum_{K \in m} F(K)$  where

$$F(K) = \sup_{m' \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ \sum_{K' \in m'} [T(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'}) - L(r+1) \log_+^2(n/(r+1)) n^{-1} \mathbb{1}_{K \cap K'}] \right\},$$

and where  $\hat{s}_K, \hat{s}_{K'}$  are defined in Lemma 2. Algorithms may therefore be used to compute each  $F(K)$  and to minimize  $\gamma_2(\cdot)$ .

Suppose first that  $r = 0$ . We begin by storing in memory all the  $\hat{s}_K, N(K), M(K)$  with  $K \in \mathcal{K}$ . This requires at most  $\mathcal{O}(c \cdot n \cdot \hat{n}^2)$  operations where  $c = 1$  in frameworks 1, 2 and where  $c$  depends on the Markov processes in framework 3. Let  $K \in \mathcal{K}$ . Since  $K \cap K' \in \mathcal{K}$ , we may compute all the

$$T(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'}) = \psi(\hat{s}_{K'}/\hat{s}_K) N(K \cap K') - \frac{1}{4} (\hat{s}_{K'} - \hat{s}_K) M(K \cap K'),$$

when  $K'$  runs among  $\mathcal{K}$  in at most  $\mathcal{O}(\hat{n}^2)$  operations. Computing all the  $F(K)$  with  $\mathcal{O}(\hat{n}^2)$  dynamic programming algorithms requires at most  $\mathcal{O}(\hat{n}^4 \hat{\ell})$  additional operations. It then remains to minimize  $\gamma_2(\cdot)$ . Finally, the number of operations needed to define  $\hat{s}_m$  is at most  $\mathcal{O}(\hat{n}^2(c\hat{n} + \hat{n}^2 \hat{\ell}))$ .

When  $r \geq 1$ , the procedure is similar, although slower. We define for  $K, K' \in \mathcal{K}$ ,  $k_1$  (respectively  $k_2$ ) as the maximum number of operations needed to know  $\hat{s}_K$  (respectively  $\int_{K \cap K'} \hat{s}_K dM$ ). We begin by storing in memory all the  $\hat{s}_K$ , all the atoms of  $N$  on  $K$ , and all the  $\int_{K \cap K'} \hat{s}_K dM$ . This requires at most  $\mathcal{O}((k_1 + n)\hat{n}^2 + k_2 \hat{n}^4)$  operations. Since  $K \cap K' \in \mathcal{K}$ , we may compute each  $T(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'})$  in at most  $\mathcal{O}(nrN(K \cap K') + 1)$  additional operations. We store all the  $T(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'})$ , and use  $\mathcal{O}(\hat{n}^2)$  dynamic programming algorithms to know all the  $F(K)$ . A last algorithm is used to minimize  $\gamma_2(\cdot)$ . Finally, the minimization of  $\gamma_2(\cdot)$  may be performed in at most  $\mathcal{O}(\hat{n}^4(\hat{\ell} + k_2) + (k_1 + n)\hat{n}^2 + rn)$  operations.

The estimator  $\hat{s}_{\hat{m}}$  may therefore be computed in practice in favorable situations. Numerical simulations are also possible (see Section 5). Unfortunately, the preceding bounds may be quite large, especially when  $\hat{n}$  is large. They are, however, much smaller than the number of computations we would need to perform to minimize the first criterion  $\gamma(\cdot)$  by a naive algorithm that would require to know every  $T(\hat{s}_m, \hat{s}_{m'})$ .

Remark. We may also optimize a criterion of the form (68) when  $\widehat{\mathcal{M}}_{\hat{\ell}}$  is replaced by  $\widehat{\mathcal{M}}_{\leq \hat{\ell}} = \{m \in \widehat{\mathcal{M}}, |m| \leq \hat{\ell}\}$ . This only increases the computational cost of the algorithm by  $\mathcal{O}(\hat{\ell})$ .

## APPENDIX D. PROOFS OF THE RESULTS OF APPENDIX A

**D.1. Proof of Theorem 11.** Let  $M_{\mathbf{s}_0}$  be the (possibly random) measure defined by

$$M_{\mathbf{s}_0}(A) = \int_A \mathbf{s}_0 \, dM \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

Let, for each  $A \in \mathcal{B}(\mathbb{R})$ ,  $Q(A)$  be a random variable such that  $Q(\cdot)$  defines a measure on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Let  $f$  be a bounded function and

$$Z_Q(f) = \int_{\mathbb{R}} f \, dQ - \mathbb{E} \left[ \int_{\mathbb{R}} f \, dQ \right].$$

Since  $\mathbb{E}[N(A)] = \mathbb{E}[M_{\mathbf{s}_0}(A)]$  for all  $A \in \mathcal{B}(\mathbb{R})$ ,  $Z(f)$  can be written as  $Z(f) = Z_N(f) - Z_{M_{\mathbf{s}_0}}(f)$ .

We now aim at controlling the deviations of  $Z_Q$ . We begin by showing that this issue boils down to a suitable control of the deviations of  $Q(A) - \mathbb{E}[Q(A)]$ .

**Lemma 8.** *Let  $\mathcal{F}$  be a collection of functions of  $\mathcal{S}$  such that  $|f| \leq 1$  for all  $f \in \mathcal{F}$ . Consider a collection  $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$  such that*

$$\mathcal{A} \supset \{ \{t \in \mathbb{R}, f_+(t) > u\}, f \in \mathcal{F}, u \in (0, 1) \} \cup \{ \{x \in \mathbb{R}, f_-(t) > u\}, f \in \mathcal{F}, u \in (0, 1) \}.$$

*Suppose that  $\sup_{A \in \mathcal{A}} \mathbb{E}[Q(A)] \leq 1$ , that there exist non-negative numbers  $\alpha, \beta$ , and an event on which: for all  $A \in \mathcal{A}$ ,*

$$(69) \quad |Q(A) - \mathbb{E}[Q(A)]| \leq \sqrt{\alpha} \left( \sqrt{Q(A)} + \sqrt{\mathbb{E}[Q(A)]} \right) + \beta.$$

*Then, on this event, for all  $f \in \mathcal{F}$ ,*

$$(70) \quad |Z_Q(f)| \leq C \left\{ \sqrt{\alpha v_Q(f) \log_+(1/v_Q(f))} + \alpha + \beta \right\},$$

*where*

$$v_Q(f) = \min \left\{ \int_{\mathbb{R}} f^2 \, dQ, \mathbb{E} \left[ \int_{\mathbb{R}} f^2 \, dQ \right] \right\}.$$

*Moreover, for all  $\varepsilon > 0$ ,*

$$(71) \quad |Z_Q(f)| \leq \varepsilon v_Q(f) + C_\varepsilon \alpha \log_+(1/\alpha) + C\beta.$$

*The constant  $C$  is universal whereas  $C_\varepsilon$  only depends on  $\varepsilon$ .*

The proof of this result is delayed to Section D.2 below. As  $N$  is an empirical measure, we can prove that (69) is valid with  $Q = N$  on an event of high probability by using, for example, the Vapnik-Chervonenkis inequalities for relative deviations. More precisely:

**Lemma 9.** *Let  $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$  be a collection of Borel sets and  $S_{\mathcal{A}}(2n)$  be the Vapnik-Chervonenkis shatter coefficient of  $\mathcal{A}$  defined by*

$$S_{\mathcal{A}}(2n) = \max_{t_1, \dots, t_{2n} \in \mathbb{R}} |\{ \{t_1, \dots, t_{2n}\} \cap A, A \in \mathcal{A} \}|.$$

*Suppose that there exists an at most countable set  $\mathcal{A}' \subset \mathcal{A}$  such that: for all  $A \in \mathcal{A}$ , there exists a sequence  $(A_m)_{m \geq 0} \in \mathcal{A}'^{\mathbb{N}}$  satisfying  $\lim_{m \rightarrow +\infty} \mathbb{1}_{A_m}(t) = \mathbb{1}_A(t)$  for every  $t \in \mathbb{R}$ .*

*Let now  $\xi > 0$ . Then, there exist a universal constant  $c_1$  and an event  $\Omega_{\xi, 1}$  such that  $\mathbb{P}[\Omega_{\xi, 1}] \geq 1 - e^{-n\xi}$  and on which (69) holds for  $Q = N$  and for all  $A \in \mathcal{A}$  with  $\alpha = c_1[\log_+ |S_{\mathcal{A}}(2n)| + n\xi]/n$  and  $\beta = 0$ .*

We now need to study the deviations of  $Z_{M_{\mathbf{s}_0}}$ . There is nothing to prove in framework 1 since then  $Z_{M_{\mathbf{s}_0}} = 0$ . The lemma below allows to control the deviations of  $Z_{M_{\mathbf{s}_0}}(f)$  for all bounded function  $f$ . It is proved in Section D.4.

**Lemma 10.** *Define  $G(t) = \mathbb{P}(X \geq t)$  in framework 2 and  $G(t) = \mathbb{P}(X_{t-} = 1)\mathbb{1}_{I_{obs}}(t)$  in framework 3. For all  $\xi > 0$ , there exists an event  $\Omega_{\xi,2}$  such that  $\mathbb{P}[\Omega_{\xi,2}] \geq 1 - e^{-n\xi}$  and on which: for all  $\varepsilon > 0$ , and measurable function  $f$  such that  $|f(t)| \leq 1$  for all  $t \in \mathbb{R}$ ,*

$$(72) \quad |Z_{M_{\mathbf{s}_0}}(f)| \leq \varepsilon v_{M_{\mathbf{s}_0}}(f) + C'_\varepsilon \inf_{\alpha \geq 0} \left\{ [\xi + 1/n] \int_0^\alpha \mathbf{s}_0 d\mu + \int_\alpha^\infty \mathbf{s}_0 G d\mu + \xi + 1/n \right\}.$$

Moreover, in framework 2,

$$(73) \quad |Z_{M_{\mathbf{s}_0}}(f)| \leq \varepsilon v_{M_{\mathbf{s}_0}}(f) + C''_\varepsilon (\xi + 1/n) \log_+ [1/(\xi + 1/n)].$$

In these inequalities,  $C'_\varepsilon$  and  $C''_\varepsilon$  only depend on  $\varepsilon$ .

The computation of the infimum in (72) may be avoided in framework 3 when one restricts the collection among which the functions  $f$  may vary. More precisely, a bound on  $Z_{M_{\mathbf{s}_0}}(f)$  may be deduced from the following result to be proved in Section D.5 and from Lemma 9.

**Lemma 11.** *Consider framework 3. Let for  $d \geq 1$ ,  $\mathcal{I}_{2d}$  be the class of unions of at most  $2d$  intervals. Then, there exist a universal constant  $c_2$  and an event  $\Omega_{\xi,2}$  such that  $\mathbb{P}[\Omega_{\xi,2}] \geq 1 - e^{-n\xi}$  and on which (69) holds true for  $Q = M_{\mathbf{s}_0}$  for all  $A \in \mathcal{I}_{2d}$  with  $\alpha = \beta = c_2[d \log_+(n/d) + n\xi]/n$ .*

We are now in position to prove (63). In framework 1, we use Lemmas 8 and 9 to deduce on  $\Omega_{\xi,1}$ :

$$|Z(f)| \leq \varepsilon v(f) + C_\varepsilon \frac{c_1[\log_+ |S_{\mathcal{A}}(2n)| + n\xi]}{n} \log_+ \left( \frac{n}{c_1[\log_+ |S_{\mathcal{A}}(2n)| + n\xi]} \right).$$

Sauer's lemma shows that there exists  $c'_1$  such that  $\log_+ |S_{\mathcal{A}}(2n)| \leq c'_1 d \log_+(n/d)$ . Inequality (63) then follows from elementary computations.

We now turn to frameworks 2 and 3. As in framework 1, Lemmas 8 and 9 imply on  $\Omega_{\xi+(\log 2)/n,1}$ :

$$|Z_N(f)| \leq \varepsilon v_N(f) + C_\varepsilon \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right],$$

where  $C_\varepsilon$  only depends on  $\varepsilon$ . It follows from Lemmas 8, 10 and 11 that on  $\Omega_{\xi+(\log 2)/n,2}$ :

$$|Z_{M_{\mathbf{s}_0}}(f)| \leq \varepsilon v_{M_{\mathbf{s}_0}}(f) + C'_\varepsilon \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right],$$

where  $C'_\varepsilon$  only depends on  $\varepsilon$ . Since  $|Z(f)| \leq |Z_N(f)| + |Z_{M_{\mathbf{s}_0}}(f)|$ , we get on  $\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2}$ :

$$(74) \quad |Z(f)| \leq \varepsilon (v_N(f) + v_{M_{\mathbf{s}_0}}(f)) + (C_\varepsilon + C'_\varepsilon) \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right].$$

Now,

$$(75) \quad v_N(f) \leq \mathbb{E} \left[ \int_{\mathbb{R}} f^2 dN \right] = \mathbb{E} \left[ \int_{\mathbb{R}} f^2 dM_{\mathbf{s}_0} \right].$$

In framework 2, we deduce from Lemma 10 with  $\varepsilon = 1/2$  that on  $\Omega_{\xi+(\log 2)/n,2}$ :

$$(76) \quad \begin{aligned} \left| \int_{\mathbb{R}} f^2 dM_{\mathbf{s}_0} - \mathbb{E} \left[ \int_{\mathbb{R}} f^2 dM_{\mathbf{s}_0} \right] \right| &\leq \frac{1}{2} \mathbb{E} \left[ \int_{\mathbb{R}} f^4 dM_{\mathbf{s}_0} \right] + C(\xi + 1/n) \log_+ [1/(\xi + 1/n)] \\ &\leq \frac{1}{2} \mathbb{E} \left[ \int_{\mathbb{R}} f^4 dM_{\mathbf{s}_0} \right] + C \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right], \end{aligned}$$

where  $C$  is universal. In framework 3,

$$\{t \in \mathbb{R}, f^2(t) > u\} = \{t \in \mathbb{R}, f_+(t) > \sqrt{u}\} \cup \{t \in \mathbb{R}, f_-(t) > \sqrt{u}\}$$

is a union of at most  $2d$  intervals. Lemma 11 with Lemma 9 show that (76) remains valid on  $\Omega_{\xi+(\log 2)/n,2}$ . Moreover, as  $|f| \leq 1$ ,

$$\mathbb{E} \left[ \int_{\mathbb{R}} f^2 dM_{\mathbf{s}_0} \right] \leq 2 \int_{\mathbb{R}} f^2 dM_{\mathbf{s}_0} + 2C \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right].$$

By using (75),

$$v_N(f) \leq 2v_{M_{\mathbf{s}_0}}(f) + 2C \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right].$$

We put this inequality in (74) and use  $v_{M_{\mathbf{s}_0}}(f) \leq v(f)$  to show (63) on  $\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2}$ . Moreover,

$$\mathbb{P} \left[ (\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2})^c \right] \leq e^{-n\xi}.$$

□

**D.2. Proof of Lemma 8.** For convenience, and to make the proof more readable, we introduce a new notation. Given expressions  $x, y \in \mathbb{R}$ , the assertion: there exists a universal constant  $C$  such that  $x \leq Cy$  is written in the sequel as  $x \preccurlyeq y$ . The claim below follows from elementary computations.

**Claim 8.** *When (69) holds true,*

$$(77) \quad |Q(A) - \mathbb{E}[Q(A)]| \preccurlyeq \sqrt{\alpha \min\{Q(A), \mathbb{E}[Q(A)]\}} + \alpha + \beta.$$

*Proof of Claim 8.* For reasons of symmetry, we may suppose that  $Q(A) \geq \mathbb{E}[Q(A)]$  to prove (77). We derive from (69),

$$\begin{aligned} |Q(A) - \mathbb{E}[Q(A)]| &\preccurlyeq \sqrt{\alpha Q(A)} + \alpha + \beta \\ &\preccurlyeq \sqrt{\alpha (Q(A) - \mathbb{E}[Q(A)])} + \sqrt{\alpha \mathbb{E}[Q(A)]} + \alpha + \beta. \end{aligned}$$

For all  $\varepsilon > 0$ , we deduce from the inequality  $2\sqrt{xy} \leq \varepsilon x + \varepsilon^{-1}y$ , that

$$|Q(A) - \mathbb{E}[Q(A)]| \leq \frac{1}{2} |Q(A) - \mathbb{E}[Q(A)]| + C \left[ \sqrt{\alpha \mathbb{E}[Q(A)]} + \alpha + \beta \right],$$

where  $C$  is universal. This shows (77). □

Without loss of generality, we prove the lemma when the functions  $f$  of  $\mathcal{F}$  are non-negative. We suppose moreover that we are on an event on which (69) holds true. Let for  $u \in (0, 1)$ ,  $A_{f,u} = \{t \in \mathbb{R}, f(t) > u\}$ . As in [Bar16], the notion of integral is a great help: for all  $t \in \mathbb{R}$ ,

$$f(t) = \int_0^1 \mathbb{1}_{A_{f,u}}(t) du.$$

Let  $\varepsilon > 0$  and  $\eta \in (0, 1]$  to be specified later. Since

$$f^2(t) = 2 \int_0^1 u \mathbb{1}_{A_{f,u}}(t) du,$$

we get

$$\begin{aligned} |Z(f)| - \varepsilon v_Q(f) &= \left| \int_0^1 (Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]) du \right| - 2\varepsilon \min \left\{ \int_0^1 u Q(A_{f,u}) du, \int_0^1 u \mathbb{E}[Q(A_{f,u})] du \right\} \\ &\leq \int_0^1 \{|Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| - 2\varepsilon u \min \{Q(A_{f,u}), \mathbb{E}[Q(A_{f,u})]\}\} du \\ &\leq \int_0^\eta |Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| du \\ (78) \quad &+ \int_\eta^1 \{|Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| - 2\varepsilon u \min \{Q(A_{f,u}), \mathbb{E}[Q(A_{f,u})]\}\} du. \end{aligned}$$

It follows from (77) and the inequality  $2\sqrt{xy} \leq \varepsilon x + \varepsilon^{-1}y$ ,

$$|Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| - 2\varepsilon u \min \{Q(A_{f,u}), \mathbb{E}[Q(A_{f,u})]\} \leq \alpha/(\varepsilon u) + \alpha + \beta.$$

We deduce,

$$(79) \quad |Z(f)| - \varepsilon v_Q(f) \leq \sqrt{\alpha} \int_0^\eta \sqrt{\mathbb{E}[Q(A_{f,u})]} du + (\alpha/\varepsilon) \log(1/\eta) + \alpha + \beta.$$

We now optimize this result with respect to  $\varepsilon$  and  $\eta$ :

$$|Z(f)| \leq \sqrt{\alpha} \inf_{\eta \in (0,1]} \left\{ \sqrt{v_Q(f) \log(1/\eta)} + \int_0^\eta \sqrt{\mathbb{E}[Q(A_{f,u})]} du \right\} + \alpha + \beta.$$

It remains to use  $\mathbb{E}[Q(A_{f,u})] \leq 1$  to prove (70).

Elementary computations then show (71). As  $x \mapsto x \log_+(1/x)$  is non-decreasing, for all  $x \leq y$ ,  $xy \log_+(1/x) \leq y^2 \log_+(1/y)$ . Moreover, when  $x \geq y$ ,  $\log_+(1/x) \leq \log_+(1/y)$  and hence  $xy \log_+(1/x) \leq xy \log_+(1/y)$ . Therefore, for all  $x, y > 0$ ,

$$\begin{aligned} xy \log_+(1/x) &\leq \max\{x, y\} y \log_+(1/y) \\ &\leq (x + y) y \log_+(1/y). \end{aligned}$$

We thus obtain for all  $\varepsilon > 0$ ,

$$\begin{aligned} 2\sqrt{xy \log_+(1/x)} &\leq \varepsilon(x + y) + \varepsilon^{-1}y \log_+(1/y) \\ &\leq \varepsilon x + C_\varepsilon y \log_+(1/y), \end{aligned}$$

where  $C_\varepsilon = \varepsilon + \varepsilon^{-1}$ . We use this inequality with  $x = v_Q(f)$  and  $y = \alpha$  to get

$$2\sqrt{\alpha v_Q(f) \log_+(1/v_Q(f))} \leq \varepsilon v_Q(f) + C_\varepsilon \alpha \log_+(1/\alpha).$$

This proves (71). □

**D.3. Proof of Lemma 9.** Let  $A \in \mathcal{A}$  and  $(A_m)_{m \geq 1}$  be the sequence given by the theorem. Then,  $N(A_m)$  converges to  $N(A)$ . Moreover, it follows from the dominated convergence theorem that  $\mathbb{E} \left[ \int_{\mathbb{R}} |\mathbb{1}_{A_m} - \mathbb{1}_A| dN \right]$  converges to 0. In particular,  $\mathbb{E}[N(A_m)]$  converges to  $\mathbb{E}[N(A)]$ . We deduce that

$$\sup_{A \in \mathcal{A}} |\sqrt{N(A)} - \sqrt{\mathbb{E}[N(A)]}| = \sup_{A \in \mathcal{A}'} |\sqrt{N(A)} - \sqrt{\mathbb{E}[N(A)]}|.$$

From now on, we may therefore consider, without loss of generality, the collection  $\mathcal{A}$  as at most countable. We then use the celebrated Vapnik-Chervonenkis inequalities for relative deviation recalled below (see for instance page 24 of [DL12]):

**Theorem 14** (Vapnik-Chervonenkis inequalities for relative deviation). *Let  $Z_1, \dots, Z_n$  be  $n$  independent and identically distributed random variables with values in a space  $\mathcal{X}$ . Let  $\mathcal{B}$  be an at most countable collection of measurable sets. Define the empirical measure  $\nu_n(B) = n^{-1} \sum_{i=1}^n \mathbb{1}_B(Z_i)$ ,  $\nu(B) = \mathbb{E}[\nu_n(B)]$  and the Vapnik-Chervonenkis shatter coefficient*

$$S_{\mathcal{B}}(2n) = \max_{z_1, \dots, z_{2n} \in \mathcal{X}} |\{\{z_1, \dots, z_{2n}\} \cap B, B \in \mathcal{B}\}|.$$

Then, for all  $x > 0$ ,

$$\begin{aligned} \mathbb{P} \left( \sup_{B \in \mathcal{B}} \frac{\nu(B) - \nu_n(B)}{\sqrt{\nu(B)}} \geq x \right) &\leq 4S_{\mathcal{B}}(2n)e^{-nx^2/4} \\ \mathbb{P} \left( \sup_{B \in \mathcal{B}} \frac{\nu_n(B) - \nu(B)}{\sqrt{\nu_n(B)}} \geq x \right) &\leq 4S_{\mathcal{B}}(2n)e^{-nx^2/4}. \end{aligned}$$

In particular,

$$(80) \quad \mathbb{P} \left( \sup_{B \in \mathcal{B}} \left| \sqrt{\nu_n(B)} - \sqrt{\nu(B)} \right| \geq x \right) \leq 8S_{\mathcal{B}}(2n)e^{-nx^2/4}.$$

Assume that we are within framework 1. Then, the random measure  $N$  is the empirical measure of  $X_1, \dots, X_n$ . Now (80) with  $\mathcal{B} = \mathcal{A}$ ,

$$x^2 = \frac{4}{n} (\log 8 + \log_+ |S_{\mathcal{A}}(2n)| + n\xi)$$

shows that (69) holds true with probability larger than  $1 - e^{-n\xi}$ ,  $\alpha = x^2$ ,  $\beta = 0$ .

The proof in frameworks 2 and 3 is very similar since  $N$  is an empirical measure for suitable random variables with values in  $\mathcal{X} = \mathbb{R} \times \{0, 1\}$ :  $Z_i = (X_i, \mathbb{1}_{D_i=1})$  in framework 2 and  $Z_i = (T_{1,0}^{(i)} \mathbb{1}_{T_{1,0}^{(i)} \in I_{\text{obs}}}, \mathbb{1}_{T_{1,0}^{(i)} \in I_{\text{obs}}})$  in framework 3. We apply (80) with  $\mathcal{B} = \{A \times \{1\}, A \in \mathcal{A}\}$ . Moreover,

$$\begin{aligned} |S_{\mathcal{B}}(2n)| &\leq \max_{x_1, \dots, x_{2n} \in \mathbb{R}} |\{\{x_1, \dots, x_{2n}\} \cap A, A \in \mathcal{A}\}| \\ &\leq |S_{\mathcal{A}}(2n)|. \end{aligned}$$

We end the proof as in framework 1. □

#### D.4. Proof of Lemma 10.

**Claim 9.** *Let for  $t \geq 0$ ,  $V(t) = \mathbb{1}_{X \geq t}$  in framework 2 and  $V(t) = \mathbb{1}_{X_{t-}=1} \mathbb{1}_{I_{obs}}(t)$  in framework 3. Then, for any Borel set  $A \subset [0, +\infty)$ , and  $k \geq 1$ ,*

$$\mathbb{E} \left[ \left( \int_A \mathbf{s}_0(t) V(t) dt \right)^k \right] \leq k! \int_A \mathbf{s}_0(t) \mathbb{E}[V(t)] dt.$$

The proof of this lemma is deferred to Sections D.4.1 and D.4.2 below. We define  $\alpha > 0$ ,

$$R_\alpha = \int_0^\alpha \mathbf{s}_0 \left( \sqrt{G_n} - \sqrt{G} \right)^2 d\mu,$$

and we prove:

**Claim 10.** *For all  $\xi > 0$  and probability larger than  $1 - 2e^{-n\xi}$ ,*

$$R_\alpha \leq 3\xi \int_0^\alpha \mathbf{s}_0 d\mu.$$

*Proof.* Let  $u = n/3$  and  $\lambda = u / \int_0^\alpha \mathbf{s}_0 d\mu$ . It follows from Jensen's inequality that

$$\exp \left( u \frac{R_\alpha}{\int_0^\alpha \mathbf{s}_0 d\mu} \right) \leq \frac{1}{\int_0^\alpha \mathbf{s}_0 d\mu} \int_0^\alpha \mathbf{s}_0 \exp \left( u (\sqrt{G_n} - \sqrt{G})^2 \right) d\mu.$$

We deduce,

$$(81) \quad \mathbb{E} [\exp (\lambda R_\alpha)] \leq \frac{1}{\int_0^\alpha \mathbf{s}_0 d\mu} \int_0^\alpha \mathbf{s}_0(t) \mathbb{E} \left[ \exp \left( u (\sqrt{G_n(t)} - \sqrt{G(t)})^2 \right) \right] dt.$$

Now,

$$\begin{aligned} \mathbb{E} \left[ \exp \left( u (\sqrt{G_n(t)} - \sqrt{G(t)})^2 \right) \right] &= 1 + \int_1^\infty \mathbb{P} \left[ \exp \left( u (\sqrt{G_n(t)} - \sqrt{G(t)})^2 \right) \geq x \right] dx \\ &= 1 + \int_1^\infty \mathbb{P} \left[ \left| \sqrt{G_n(t)} - \sqrt{G(t)} \right| \geq \sqrt{\frac{\log x}{u}} \right] dx. \end{aligned}$$

The random variable  $nG_n(t)$  is binomially distributed with parameters  $(n, G(t))$ . Therefore, we derive from Theorems 3 and 4 of [Oka59] that

$$\mathbb{P} \left[ \left| \sqrt{G_n(t)} - \sqrt{G(t)} \right| \geq \sqrt{\frac{\log x}{u}} \right] \leq \frac{2}{x^{n/u}}.$$

Hence,

$$\begin{aligned} \mathbb{E} \left[ \exp \left( u (\sqrt{G_n(t)} - \sqrt{G(t)})^2 \right) \right] &\leq 1 + \int_1^\infty \frac{2}{x^{n/u}} dx \\ &\leq 1 + \int_1^\infty \frac{2}{x^3} dx \\ &\leq 2. \end{aligned}$$

By (81),  $\mathbb{E} [\exp (\lambda R_\alpha)] \leq 2$  and by Markov's inequality,

$$\mathbb{P} [R_\alpha \geq n\xi/\lambda] = \mathbb{P} \left[ e^{\lambda R_\alpha} \geq e^{n\xi} \right] \leq 2e^{-n\xi}.$$

□

We now prove Lemma 10. We have,

$$\begin{aligned}
|Z_{M_{\mathbf{s}_0}}(f)| &\leq \int_0^\infty \mathbf{s}_0 |f(G_n - G)| \, d\mu \\
&\leq \int_{G_n \leq G} \mathbf{s}_0 \left| f\left(\sqrt{G_n} - \sqrt{G}\right) \left(\sqrt{G} - \sqrt{G_n} + 2\sqrt{G_n}\right) \right| \, d\mu \\
&\quad + \int_{G_n > G} \mathbf{s}_0 \left| f\left(\sqrt{G_n} - \sqrt{G}\right) \left(\sqrt{G_n} - \sqrt{G} + 2\sqrt{G}\right) \right| \, d\mu \\
&\leq \int_0^\infty \mathbf{s}_0 |f| \left(\sqrt{G_n} - \sqrt{G}\right)^2 \, d\mu + 2 \int_0^\infty \mathbf{s}_0 |f| \left|\sqrt{G_n} - \sqrt{G}\right| \sqrt{\min\{G_n, G\}} \, d\mu.
\end{aligned}$$

Define

$$R = R_\infty = \int_0^\infty \mathbf{s}_0 \left(\sqrt{G_n} - \sqrt{G}\right)^2 \, d\mu.$$

By using  $|f| \leq 1$  and Cauchy-Schwarz inequality,

$$\begin{aligned}
|Z_{M_{\mathbf{s}_0}}(f)| &\leq R + 2\sqrt{R} \sqrt{\int_0^\infty \mathbf{s}_0 f^2 \min\{G_n, G\} \, d\mu} \\
&\leq R + 2\sqrt{R} \sqrt{v_{M_{\mathbf{s}_0}}(f)}.
\end{aligned}$$

Therefore, for all  $\varepsilon > 0$ , using that  $2\sqrt{xy} \leq \varepsilon^{-1}x + \varepsilon y$ ,

$$(82) \quad |Z_{M_{\mathbf{s}_0}}(f)| \leq \varepsilon v_{M_{\mathbf{s}_0}}(f) + (1 + \varepsilon^{-1})R.$$

It remains to bound  $R$  from above. We consider  $\alpha \geq 0$  that minimizes

$$[\xi + 1/n] \int_0^\alpha \mathbf{s}_0 \, d\mu + \int_\alpha^\infty \mathbf{s}_0 G \, d\mu + \xi + 1/n.$$

Now,

$$(83) \quad R \leq R_\alpha + \int_\alpha^\infty \mathbf{s}_0 (G_n + G) \, d\mu.$$

With probability larger than  $1 - (1/2)e^{-n\xi}$ , Claim 10 ensures that

$$(84) \quad R_\alpha \leq 3 \left( \xi + \frac{\log 4}{n} \right) \int_0^\alpha \mathbf{s}_0 \, d\mu.$$

Now,

$$\int_\alpha^\infty \mathbf{s}_0 G_n \, d\mu = \frac{1}{n} \sum_{i=1}^n \int_\alpha^\infty \mathbf{s}_0(t) V_i(t) \, dt,$$

where  $V_i(t) = \mathbb{1}_{X_i \geq t}$  in framework 2 and  $V_i(t) = \mathbb{1}_{X_{t-1}^{(i)} = 1} \mathbb{1}_{I_{\text{obs}}}(t)$  in framework 3. Claim 9 shows that the assumptions of Bernstein's deviation inequality are satisfied (Proposition 2.9 of [Mas07]). Therefore, with probability larger than  $1 - (1/2)e^{-n\xi}$ ,

$$(85) \quad \int_\alpha^\infty \mathbf{s}_0 [G_n - G] \, d\mu \leq 2 \sqrt{\left( \xi + \frac{\log 2}{n} \right) \int_\alpha^\infty \mathbf{s}_0 G \, d\mu + \xi + \frac{\log 2}{n}}.$$

Since  $2\sqrt{xy} \leq x + y$ ,

$$(86) \quad \int_\alpha^\infty \mathbf{s}_0 G_n \, d\mu \leq 2 \int_\alpha^\infty \mathbf{s}_0 G \, d\mu + 2 \left( \xi + \frac{\log 2}{n} \right).$$

By putting (84) and (86) into (83), we get with probability larger than  $1 - e^{-n\xi}$ ,

$$R \leq 3 \left[ \frac{\log 4}{n} + \xi \right] \int_0^\alpha \mathbf{s}_0 d\mu + 3 \int_\alpha^\infty \mathbf{s}_0 G d\mu + 2\xi + \frac{2 \log 2}{n}.$$

It then remains to use the definition of  $\alpha$  and (82) to prove (72). Moreover, in framework 2,

$$\int_\alpha^\infty \mathbf{s}_0 G d\mu = \int_\alpha^\infty \frac{f_0(t)}{\mathbb{P}(T \geq t)} \mathbb{P}(T \geq t) \mathbb{P}(C \geq t) dt \leq \int_\alpha^\infty f_0 d\mu,$$

and

$$\int_0^\alpha \mathbf{s}_0 d\mu = \int_0^\alpha \frac{f_0(t)}{\mathbb{P}(T \geq t)} dt = -\log \left( \int_\alpha^\infty f_0 d\mu \right),$$

which gives (73).  $\square$

D.4.1. *Proof of Claim 9 in framework 2.* We define for  $k \geq 1$ ,

$$J_k = \int_{\substack{t_1, \dots, t_k \in A \\ t_1 < t_2 < \dots < t_k}} \left( \prod_{j=1}^k \mathbf{s}_0(t_j) \right) \mathbb{P}(X \geq t_k) dt_1 dt_2 \dots dt_k.$$

We have,

$$\begin{aligned} \mathbb{E} \left[ \left( \int_A \mathbf{s}_0(t) \mathbb{1}_{X \geq t} dt \right)^k \right] &= \mathbb{E} \left[ \int_{A^k} \prod_{j=1}^k \mathbf{s}_0(t_j) \mathbb{1}_{X \geq t_j} dt_1 dt_2 \dots dt_k \right] \\ &= \int_{A^k} \left( \prod_{j=1}^k \mathbf{s}_0(t_j) \right) \mathbb{P}(X \geq \max\{t_1, \dots, t_k\}) dt_1 dt_2 \dots dt_k \\ &= k! J_k. \end{aligned}$$

Now,

$$J_k \leq \int_{\substack{t_1, \dots, t_{k-1} \in A \\ t_1 < t_2 < \dots < t_{k-1}}} \left( \prod_{j=1}^{k-1} \mathbf{s}_0(t_j) \right) \left( \int_{t_{k-1}}^\infty \mathbf{s}_0(t_k) \mathbb{P}(X \geq t_k) dt_k \right) dt_1 dt_2 \dots dt_{k-1},$$

and

$$\begin{aligned} \int_{t_{k-1}}^\infty \mathbf{s}_0(t_k) \mathbb{P}(X \geq t_k) dt_k &= \int_{t_{k-1}}^\infty f_0(t_k) \mathbb{P}(C \geq t_k) dt_k \\ &\leq \left( \int_{t_{k-1}}^\infty f_0(t_k) dt_k \right) \mathbb{P}(C \geq t_{k-1}) \\ &\leq \mathbb{P}(T \geq t_{k-1}) \mathbb{P}(C \geq t_{k-1}) \\ &\leq \mathbb{P}(X \geq t_{k-1}). \end{aligned}$$

Therefore,

$$\begin{aligned} J_k &\leq \int_{\substack{t_1, \dots, t_{k-1} \in A \\ t_1 < t_2 < \dots < t_{k-1}}} \left( \prod_{j=1}^{k-1} \mathbf{s}_0(t_j) \right) \mathbb{P}(X \geq t_{k-1}) dt_1 dt_2 \dots dt_{k-1} \\ &\leq J_{k-1}. \end{aligned}$$

By induction,  $J_k \leq J_1$ .  $\square$

D.4.2. *Proof of Claim 9 in framework 3.*

**Claim 11.** *Let  $t > 0$ ,  $\mathcal{F}_t = \sigma(X_v, v \leq t)$  be the  $\sigma$ -algebra generated by the family of random variables  $X_v$ ,  $v \in [0, t]$ . Let  $B$  be an event  $\mathcal{F}_t$ -measurable. Let  $\mu_B$  be the measure defined for all  $A \in \mathcal{B}(\mathbb{R})$  by*

$$\mu_B(A) = \mathbb{P}(B \text{ and } T_{1,0} \in A).$$

*Then, for  $\mu$ -almost all  $u > t$ ,*

$$(87) \quad \frac{d\mu_B}{du}(u) = \mathbb{P}(B \text{ and } X_{u-} = 1) \mathbf{s}_0(u).$$

*Proof.* First of all,  $\mu_B$  is absolutely continuous with respect to the Lebesgue measure  $\mu$  and admits therefore a Radon-Nikodym derivative. We now aim to show that this derivative is given by (87) for almost all  $u > t$ .

Let  $Z_h(u)$  be the random variable giving the number of jumps of the Markov process in  $[u-h, u+h]$ . Then,  $\mathbb{P}(Z_h(u) \geq 2) = o(h)$  when  $h \rightarrow 0$ . We deduce,

$$\mu_B([u, u+h]) = \mathbb{P}(B, Z_h(u) = 1, T_{1,0} \in [u, u+h]) + o(h).$$

When  $Z_h(u) = 1$ ,  $T_{1,0} \in [u, u+h]$  is equivalent to  $X_{u-} = 1$  and  $X_{u+h} = 0$ . This yields

$$\begin{aligned} \mu_B([u, u+h]) &= \mathbb{P}(B, Z_h(u) = 1, X_{u-} = 1, X_{u+h} = 0) + o(h) \\ &= \mathbb{P}(B, X_{u-} = 1, X_{u+h} = 0) + o(h) \\ &= \mathbb{P}(B, X_{u-} = 1) \mathbb{P}(X_{u+h} = 0 \mid B, X_{u-} = 1) + o(h). \end{aligned}$$

As  $B$  is  $\mathcal{F}_t$ -measurable and  $u > t$ ,

$$(88) \quad \begin{aligned} \mu_B([u, u+h]) &= \mathbb{P}(B, X_{u-} = 1) \mathbb{P}(X_{u+h} = 0 \mid X_{u-} = 1) + o(h) \\ &= \mathbb{P}(B, X_{u-} = 1) \frac{\mathbb{P}(X_{u-} = 1, X_{u+h} = 0)}{\mathbb{P}(X_{u-} = 1)} + o(h). \end{aligned}$$

Now,

$$\begin{aligned} \mathbb{P}(X_{u-} = 1, X_{u+h} = 0) &= \mathbb{P}(X_{u-} = 1, X_{u+h} = 0, Z_h(u) = 1) + o(h) \\ &= \mathbb{P}(T_{1,0} \in [u, u+h], Z_h(u) = 1) + o(h) \\ &= \mathbb{P}(T_{1,0} \in [u, u+h]) + o(h). \end{aligned}$$

Finally, by putting this inequality into (88),

$$\begin{aligned} \mu_B([u, u+h]) &= \mathbb{P}(B, X_{u-} = 1) \frac{\mathbb{P}(T_{1,0} \in [u, u+h])}{\mathbb{P}(X_{u-} = 1)} + o(h) \\ &= \mathbb{P}(B, X_{u-} = 1) \frac{hf_0(u)}{\mathbb{P}(X_{u-} = 1)} + o(h) \\ &= h\mathbb{P}(B, X_{u-} = 1) \mathbf{s}_0(u) + o(h), \end{aligned}$$

which proves (87). □

We now return to the proof of Claim 9. Without loss of generality, we suppose that  $A \subset I_{\text{obs}}$ . Define for  $k \geq 1$ ,

$$J_k = \int_{\substack{t_1, \dots, t_k \in A \\ t_1 < t_2 < \dots < t_k}} \left( \prod_{j=1}^k \mathbf{s}_0(t_j) \right) \mathbb{P}(X_{t_1-} = 1, \dots, X_{t_k-} = 1) dt_1 dt_2 \dots dt_k.$$

We have,

$$\begin{aligned}
\mathbb{E} \left[ \left( \int_A \mathbf{s}_0(t) \mathbb{1}_{X_{t-}=1} dt \right)^k \right] &= \mathbb{E} \left[ \int_{A^k} \prod_{j=1}^k \mathbf{s}_0(t_j) \mathbb{1}_{X_{t_j-}=1} dt_1 dt_2 \dots dt_k \right] \\
&= \int_{A^k} \left( \prod_{j=1}^k \mathbf{s}_0(t_j) \right) \mathbb{P}(X_{t_1-}=1, \dots, X_{t_k-}=1) dt_1 dt_2 \dots dt_k \\
&= k! J_k.
\end{aligned}$$

Yet,

$$J_k \leq \int_{\substack{t_1, \dots, t_{k-1} \in A \\ t_1 < t_2 < \dots < t_{k-1}}} \left( \prod_{j=1}^{k-1} \mathbf{s}_0(t_j) \right) \left( \int_{t_{k-1}}^{\infty} \mathbf{s}_0(t_k) \mathbb{P}(X_{t_1-}=1, \dots, X_{t_k-}=1) dt_k \right) dt_1 dt_2 \dots dt_{k-1}.$$

Let  $B = [X_{t_1-}=1, \dots, X_{t_{k-1}-}=1] \in \mathcal{F}_{t_{k-1}}$ . Then,

$$\begin{aligned}
\int_{t_{k-1}}^{\infty} \mathbf{s}_0(t_k) \mathbb{P}(X_{t_1-}=1, \dots, X_{t_k-}=1) dt_k &= \int_{t_{k-1}}^{\infty} \frac{d\mu_B}{dt_k}(t_k) dt_k \\
&= \mu_B([t_{k-1}, +\infty)) \\
&= \mathbb{P}(X_{t_1-}=1, \dots, X_{t_{k-1}-}=1 \text{ and } T_{1,0} \geq t_{k-1}) \\
&\leq \mathbb{P}(X_{t_1-}=1, \dots, X_{t_{k-1}-}=1).
\end{aligned}$$

Therefore,  $J_k \leq J_{k-1}$  and by induction  $J_k \leq J_1$ .  $\square$

**D.5. Proof of Lemma 11.** Let  $\overline{\mathcal{I}}_{2d}$  be the class of unions of at most  $2d$  intervals with endpoints in  $(\mathbb{Q} \cap [0, +\infty)) \cup \{+\infty\}$ . Then,  $\overline{\mathcal{I}}_{2d}$  is at most countable, and we only need to prove (69) when  $A \in \overline{\mathcal{I}}_{2d}$ .

The random measure  $M_{\mathbf{s}_0}$  is of the form

$$M_{\mathbf{s}_0}(A) = \frac{1}{n} \sum_{i=1}^n \int_A \mathbf{s}_0(t) \mathbb{1}_{X_{t-}^{(i)}=1} \mathbb{1}_{I_{\text{obs}}}(t) dt \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

There exist independent random variables  $Z_1, \dots, Z_n$  such that

$$M_{\mathbf{s}_0}(A) = \frac{1}{n} \sum_{i=1}^n f_A(Z_i) \quad \text{with } f_A(Z_i) = \int_A \mathbf{s}_0(t) \mathbb{1}_{X_{t-}^{(i)}=1} \mathbb{1}_{I_{\text{obs}}}(t) dt.$$

We measure the complexity of the family  $\{f_A, A \in \overline{\mathcal{I}}_{2d}\}$  by means of the notion of entropy with bracketing:

**Claim 12.** *For all  $\delta > 0$ , there exists a collection  $\mathcal{C}_\delta$  of functions of the form  $f_A$  with  $A \in \overline{\mathcal{I}}_{2d}$ . The cardinal of this set can be bounded by  $\log |\mathcal{C}_\delta| \leq cd \log_+(1/\delta^2)$ , where  $c$  is a universal constant. Moreover, for all  $A \in \overline{\mathcal{I}}_{2d}$ , there exist  $f_{A_1}, f_{A_2} \in \mathcal{C}_\delta$  such that  $f_{A_1} \leq f_A \leq f_{A_2}$  and such that for all  $k \geq 1$ ,*

$$\mathbb{E} \left[ (f_{A_2}(Z_1) - f_{A_1}(Z_1))^k \right] \leq \frac{k!}{2} \delta^2.$$

*Proof.* First of all, we only need to prove the claim when  $\delta$  is smaller than 1, what we will do in the sequel.

We endow  $\overline{\mathcal{I}}_{2d}$  with the distance  $dist$  defined for  $A_1, A_2 \in \overline{\mathcal{I}}_{2d}$  by

$$dist(A_1, A_2) = \mathbb{E}[M_{\mathbf{s}_0}(A_1 \Delta A_2)] \quad \text{where } A_1 \Delta A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1).$$

We may write  $dist(A_1, A_2)$  as

$$dist(A_1, A_2) = \int_{I_{\text{obs}}} |\mathbb{1}_{A_1}(t) - \mathbb{1}_{A_2}(t)| f_0(t) dt.$$

We introduce the real valued function  $F$  defined for  $x \geq 0$  by

$$F(x) = \int_0^x f_0(t) dt.$$

Since  $F$  is a continuous non-decreasing function such that  $F([0, +\infty)) \subset [0, 1]$ , there exist an even integer  $\ell \in [2, 8d/\delta^2 + 2]$ , and  $\ell$  non-negative numbers  $(x_1, x_2, \dots, x_{\ell-1}, x_\ell) \in \{0\} \times \mathbb{Q}^{\ell-2} \times \{+\infty\}$  such that

$$\max_{1 \leq i \leq \ell-1} \{F(x_{i+1}) - F(x_i)\} \leq \delta^2/(8d).$$

We may suppose that  $\ell \geq 2d$ . Let  $\mathcal{X} = \{x_1, x_2, \dots, x_\ell\}$ , and  $\overline{\mathcal{I}}_{dis}$  be the collection of unions of at most  $2d$  closed intervals whose endpoints belong to  $\mathcal{X}$ .

When  $k \leq \ell/2$ , choosing  $k$  disjoint closed intervals whose endpoints belong to  $\mathcal{X}$  amounts to choosing  $2k$  numbers among  $\mathcal{X}$ . When  $k > \ell/2$ , we cannot find  $k$  disjoint closed intervals with endpoints in  $\mathcal{X}$ . The cardinality of  $\overline{\mathcal{I}}_{dis}$  is therefore bounded by

$$|\overline{\mathcal{I}}_{dis}| \leq \sum_{k=0}^{2d} C_\ell^{2k}.$$

Standard combinatorial arguments (see, for instance, Lemma 6 of [BBM99]) show that  $|\overline{\mathcal{I}}_{dis}| \leq (\ell e/(2d))^{2d}$ . Using now that  $\ell \leq 8d/\delta^2 + 2$ , we derive that

$$\log |\overline{\mathcal{I}}_{dis}| \leq cd \log_+(1/\delta^2)$$

for a suitable universal constant  $c$ .

For each set  $A \in \overline{\mathcal{I}}_{2d}$ , we now show that there exist  $A_1, A_2 \in \overline{\mathcal{I}}_{dis}$  such that  $f_{A_1} \leq f_A \leq f_{A_2}$  and  $dist(A_1, A_2) \leq \delta^2/2$ . Let  $A \in \overline{\mathcal{I}}_{2d}$  be written as  $A = \bigcup_{k=1}^{2d} A_k$  where  $A_k$  is an interval whose endpoints are  $a_k \leq b_k$ . For each  $k \in \{1, \dots, 2d\}$ , there exist  $a_k^{(1)} \leq a_k^{(2)} \leq b_k^{(1)} \leq b_k^{(2)} \in \mathcal{X}$  such that

$$a_k^{(1)} \leq a_k \leq a_k^{(2)}, \quad b_k^{(1)} \leq b_k \leq b_k^{(2)},$$

and

$$F(a_k^{(2)}) - F(a_k^{(1)}) \leq \delta^2/(8d), \quad F(b_k^{(2)}) - F(b_k^{(1)}) \leq \delta^2/(8d).$$

Define the closed intervals

$$A_k^{(1)} = \{x \in \mathbb{R}, a_k^{(2)} \leq x \leq b_k^{(1)}\}, \quad A_k^{(2)} = \{x \in \mathbb{R}, a_k^{(1)} \leq x \leq b_k^{(2)}\}.$$

Then,  $A_1 = \bigcup_{k=1}^{2d} A_k^{(1)}$  and  $A_2 = \bigcup_{k=1}^{2d} A_k^{(2)}$  belong to  $\overline{\mathcal{I}}_{dis}$  and satisfy  $f_{A_1} \leq f_A \leq f_{A_2}$ . Moreover,

$$A_2 \Delta A_1 \subset \bigcup_{k=1}^{2d} [a_k^{(1)}, a_k^{(2)}) \cup (b_k^{(1)}, b_k^{(2)}],$$

and hence,

$$\begin{aligned}
\text{dist}(A_1, A_2) &\leq \sum_{k=1}^{2d} \int_{[a_k^{(1)}, a_k^{(2)}] \cup (b_k^{(1)}, b_k^{(2)}]} f_0(t) dt \\
&\leq \sum_{k=1}^{2d} \left( F(a_k^{(2)}) - F(a_k^{(1)}) + F(b_k^{(2)}) - F(b_k^{(1)}) \right) \\
&\leq \sum_{k=1}^{2d} (\delta^2/(8d) + \delta^2/(8d)) \\
&\leq \delta^2/2.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E} \left[ (f_{A_2}(Z_1) - f_{A_1}(Z_1))^k \right] &= \mathbb{E} \left[ (f_{A_2 \setminus A_1}(Z_1))^k \right] \\
&\leq k! \mathbb{E} [M_{\mathbf{s}_0}(A_2 \setminus A_1)] \quad \text{thanks to Claim 9} \\
&\leq k! \text{dist}(A_1, A_2) \\
&\leq k! \delta^2/2,
\end{aligned}$$

which completes the proof with  $\mathcal{C}_\delta = \{f_A, A \in \overline{\mathcal{I}}_{dis}\}$ .  $\square$

We will use several times an exponential inequality of [Mas07] to prove Lemma 11. We keep the notation  $\preccurlyeq$  introduced at the beginning of Section D.2.

Set for  $\delta > 0$ ,  $\mathcal{B}_\delta = \mathcal{C}_\delta \cup \{-f, f \in \mathcal{C}_\delta\}$ . Note that

$$\log |\mathcal{B}_\delta| \leq \log 2 + \log |\mathcal{C}_\delta| \leq c_1 d \log_+(1/\delta^2),$$

where  $c_1$  is a universal constant. We set  $H(\delta) = c_1 d \log_+(1/\delta^2)$  and for  $\sigma \in (0, 1]$ ,

$$E = \sqrt{n} \int_0^\sigma \sqrt{H(u) \wedge n} du + 2(1 + \sigma)H(\sigma).$$

Simple arguments allow to bound  $E$  from above, see for instance page 190 of [GN15]: the fundamental theorem of calculus shows

$$\begin{aligned}
\sigma \sqrt{\log(e/\sigma)} &= \int_0^\sigma \left( \sqrt{\log(e/u)} - \frac{1}{2\sqrt{\log(e/u)}} \right) du \\
&\geq \int_0^\sigma \sqrt{\log_+(1/u)} du - \sigma/2.
\end{aligned}$$

Consequently,

$$(89) \quad E \preccurlyeq \sigma \sqrt{nd \log_+(1/\sigma^2)} + d \log_+(1/\sigma^2).$$

Consider  $\xi > 0$  and define  $J$  as the (possibly empty) set of non-negative integers  $j$  such that  $2^{-j} \geq d/(2n)$ . Let, for  $j \in J$ ,  $x_j = 2 \log(j+1) + 1 + n\xi$ ,  $\overline{\mathcal{A}}_j = \{A \in \overline{\mathcal{I}}_{2d}, 2^{-j-1} \leq \mathbb{E}[M_{\mathbf{s}_0}(A)] \leq 2^{-j}\}$ . Claims 9 and 12 show that assumptions of Corollary 6.9 of [Mas07] are satisfied with  $\mathcal{F} = \{f_A, -f_A, A \in \overline{\mathcal{A}}_j\}$ ,  $\sigma^2 = 2^{-j+1}$ ,  $b = 1$ , and  $H(\delta) = c_1 d \log_+(1/\delta^2)$ . There exists therefore an event  $\Omega_j$  such that  $\mathbb{P}(\Omega_j) \geq 1 - e^{-x_j}$  and on which: for all  $A \in \overline{\mathcal{A}}_j$ ,

$$n |M_{\mathbf{s}_0}(A) - \mathbb{E}[M_{\mathbf{s}_0}(A)]| \preccurlyeq E + \sigma \sqrt{nx_j} + x_j.$$

Hence,

$$|M_{\mathbf{s}_0}(A) - \mathbb{E}[M_{\mathbf{s}_0}(A)]| \preceq \sigma \sqrt{\frac{d \log_+(1/\sigma^2) + x_j}{n}} + \frac{d \log_+(1/\sigma^2) + x_j}{n}.$$

As  $\sigma^2 \leq 4\mathbb{E}[M_{\mathbf{s}_0}(A)]$ ,  $\sigma^2 \geq d/n$ , and  $x_j \preceq \log_+(n/d) + n\xi$ ,

$$|M_{\mathbf{s}_0}(A) - \mathbb{E}[M_{\mathbf{s}_0}(A)]| \preceq \sqrt{\mathbb{E}[M_{\mathbf{s}_0}(A)]} \sqrt{\frac{d \log_+(n/d) + n\xi}{n}} + \frac{d \log_+(n/d) + n\xi}{n}.$$

Let now  $\overline{\mathcal{A}} = \{A \in \overline{\mathcal{I}}_{2d}, \mathbb{E}[M_{\mathbf{s}_0}(A)] \leq d/(2n)\}$ . We apply Corollary 6.9 of [Mas07] with  $\mathcal{F} = \{f_A, -f_A, A \in \overline{\mathcal{A}}\}$ ,  $b = 1$ ,  $\sigma^2 = \min\{d/n, 2\}$ . We deduce that there exists an event  $\Omega'$  such that  $\mathbb{P}(\Omega') \geq 1 - (1/2)e^{-n\xi}$  and on which: for all  $A \in \overline{\mathcal{A}}$ ,

$$|M_{\mathbf{s}_0}(A) - \mathbb{E}[M_{\mathbf{s}_0}(A)]| \preceq \sigma \sqrt{\frac{d \log_+(1/\sigma^2) + n\xi + \log 2}{n}} + \frac{d \log_+(1/\sigma^2) + n\xi + \log 2}{n}.$$

Since  $\sigma \leq \sqrt{d/n} \leq \sqrt{(d \log_+(n/d) + n\xi)/n}$ ,

$$|M_{\mathbf{s}_0}(A) - \mathbb{E}[M_{\mathbf{s}_0}(A)]| \preceq \frac{d \log_+(n/d) + n\xi}{n}.$$

We deduce from these computations that (69) holds true with  $Q = M_{\mathbf{s}_0}$  on the event  $\Omega' \cap (\cap_{j \in J} \Omega_j)$  for all  $A \in \bigcup_{j \in J} \overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$  with  $\alpha, \beta$  of the form  $c_2(d \log_+(n/d) + n\xi)/n$ . Now,  $\overline{\mathcal{I}}_{2d} = \bigcup_{j \in J} \overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$ , and

$$\begin{aligned} \mathbb{P} \left[ \left( \Omega' \cap \left( \bigcap_{j \in J} \Omega_j \right) \right)^c \right] &\leq \mathbb{P}[\Omega'^c] + \sum_{j \in J} \mathbb{P}[\Omega_j^c] \\ &\leq \frac{e^{-n\xi}}{2} + \sum_{j=1}^{\infty} \frac{e^{-n\xi}}{j^2 e} \\ &\leq e^{-n\xi}. \end{aligned}$$

□

**D.6. Proof of Proposition 12.** Suppose without loss of generality that the functions  $f$  are non-negative. Consider  $\varepsilon > 0$  and  $\eta \in (0, 1)$ . We derive from the beginning of the proof of Theorem 11 that for all  $f \in \mathcal{F}$ , and  $u \in (0, 1)$ , there exists  $A_u \in \mathcal{A}_u$  (we omit the subscript  $f$ ) such that

$$\begin{aligned} |Z(f)| &\leq \varepsilon \sigma^2 + \int_0^\eta |N(A_u) - \mathbb{E}[N(A_u)]| \, du \\ &\quad + \int_\eta^1 \{|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u \mathbb{E}[N(A_u)]\} \, du. \end{aligned}$$

Therefore,

$$\begin{aligned} (90) \quad \mathbb{E} \left[ \sup_{f \in \mathcal{F}} |Z(f)| \right] &\leq \varepsilon \sigma^2 + \int_0^\eta \mathbb{E} \left[ \sup_{A_u \in \mathcal{A}_u} |N(A_u) - \mathbb{E}[N(A_u)]| \right] \, du \\ &\quad + \int_\eta^1 \mathbb{E} \left[ \sup_{A_u \in \mathcal{A}_u} \{|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u \mathbb{E}[N(A_u)]\} \right] \, du. \end{aligned}$$

Let now  $\xi > 0$ . As (69) holds true for all  $A \in \mathcal{A}_u$ , on an event  $\Omega_{\xi,u}$  such that  $\mathbb{P}[\Omega_{\xi,u}] \geq 1 - e^{-n\xi}$ , with  $\alpha = c[\log_+ |S_{\mathcal{A}_u}(2n)| + n\xi]/n$ ,  $\beta = 0$ , we deduce from Claim 8 that on this event: for all  $A_u \in \mathcal{A}_u$ ,

$$|N(A_u) - \mathbb{E}[N(A_u)]| \leq C \left[ \sqrt{\frac{\log_+ |S_{\mathcal{A}_u}(2n)| + n\xi}{n}} \sqrt{\mathbb{E}[N(A_u)]} + \frac{\log_+ |S_{\mathcal{A}_u}(2n)| + n\xi}{n} \right],$$

and using that  $\sqrt{xy} \leq C/(8\varepsilon u)x + (2\varepsilon u/C)y$ ,

$$|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u \mathbb{E}[N(A_u)] \leq C' \left\{ \frac{\log_+ |S_{\mathcal{A}_u}(2n)| + n\xi}{n} + \frac{\log_+ |S_{\mathcal{A}_u}(2n)| + n\xi}{n\varepsilon u} \right\}.$$

In these two inequalities,  $C$  and  $C'$  are universal constants.

We integrate these inequalities with respect to  $\xi$  to get

$$\mathbb{E} \left[ \sup_{A_u \in \mathcal{A}_u} |N(A_u) - \mathbb{E}[N(A_u)]| \right] \leq C'' \left[ \sqrt{\frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{n}} \sup_{A_u \in \mathcal{A}_u} \sqrt{\mathbb{E}[N(A_u)]} + \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{n} \right],$$

and

$$\mathbb{E} \left[ \sup_{A_u \in \mathcal{A}_u} \{|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u \mathbb{E}[N(A_u)]\} \right] \leq C'' \left[ \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{n} + \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{n\varepsilon u} \right]$$

where  $C''$  is universal. We now deduce from (90),

$$\begin{aligned} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} |Z(f)| \right] &\leq \varepsilon \sigma^2 + \frac{C''}{n\varepsilon} \int_{\eta}^1 \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{u} du \\ &\quad + \frac{C''}{\sqrt{n}} \int_0^{\eta} \sqrt{\left( \sup_{A_u \in \mathcal{A}_u} \mathbb{E}[N(A_u)] \right) (\log_+ |S_{\mathcal{A}_u}(2n)|)} du + \frac{C''}{n} \int_0^1 \log_+ |S_{\mathcal{A}_u}(2n)| du. \end{aligned}$$

As  $\varepsilon > 0$  and  $\eta \in (0, 1)$  are arbitrary,

$$\begin{aligned} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} |Z(f)| \right] &\leq \frac{C'''}{\sqrt{n}} \inf_{\eta \in (0,1)} \left\{ \sigma \sqrt{\int_{\eta}^1 \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{u} du} + \int_0^{\eta} \sqrt{\left( \sup_{A_u \in \mathcal{A}_u} \mathbb{E}[N(A_u)] \right) (\log_+ |S_{\mathcal{A}_u}(2n)|)} du \right\} \\ &\quad + \frac{C'''}{n} \int_0^1 \log_+ |S_{\mathcal{A}_u}(2n)| du, \end{aligned}$$

where  $C'''$  is a universal constant. It then remains to bound from above these integrals.  $\square$

UNIV LYON, UJM-SAINT-ÉTIENNE, CNRS UMR 5208, INSTITUT CAMILLE JORDAN, 10 RUE TRÉFILERIE, CS 82301, F-42023 SAINT-ETIENNE CEDEX 2, FRANCE

*E-mail address:* mathieu.sart@univ-st-etienne.fr