Estimating a density, a hazard rate, and a transition intensity via the ρ-estimation method

Mathieu Sart

ESTIMATING A DENSITY, A HAZARD RATE, AND A TRANSITION INTENSITY VIA THE ρ-ESTIMATION METHOD

MATHIEU SART

Abstract. We propose a unified study of three statistical settings by widening the ρ-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 relate the performance of ρ-estimators to deviations of an empirical process. We deduce 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 ρ-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 interesting 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 μ. 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 μ. 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 $\{0, 1, \dots, m\}$.*

Date: November, 2018.

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

Key words and phrases. ρ-estimator, maximum likelihood, qualitative assumptions, piecewise polynomial estimation.

- *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*.*
- *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* μ *on* $\mathbb{R}$ *and satisfies therefore for all Borel set* A *of* $\mathbb{R}$*,*

$$\mathbb{P}\left(T_{1,0} \in A\right) = \int_A f_0(t)\,\mathrm{d}t,$$

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*

$$\mathbf{s_0}(t) = \frac{f_0(t)}{\mathbb{P}\left(X_{t-} = 1\right)},$$

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 "ρ-estimation".

1.2. **About ρ-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 ρ-estimation, we deal with the Hellinger distance h. It is defined for two non-negative integrable functions s and s' by

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

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

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

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, ρ-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 ρ-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 ρ-estimator $\hat{s}$ can be bounded 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 ρ-estimator $\hat{s}$ on S. Then, the rate of convergence of $\hat{s}$ is at least $\sqrt{d/n}\log^{3/2}(n/d)$ when $\mathbf{s_0}$ is not only non-increasing on an interval but also piecewise constant on d intervals. In this case, the rate of estimation is much faster than the minimax rate.

There are moreover two additional properties of ρ-estimators we now briefly mention. First, ρ-estimators can be related to maximum likelihood ones. Second, it is possible to deal with penalized ρ-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 ρ-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,

$$\mathbf{s_0}(t) = \frac{f_0(t)}{\mathbb{P}(T \geq t)} = \lim_{h \to 0} \frac{\mathbb{P}(t \leq T \leq t+h \mid T \geq t)}{h},$$

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,

$$\mathbf{s_0}(t) = \frac{f_0(t)}{\mathbb{P}(X_{t-} = 1)} = \lim_{h \to 0} \frac{\mathbb{P}(X_{t+h} = 0 \mid X_{t-} = 1)}{h},$$

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 be rather scarce. We refer to [BC05, RB06, BC08, Pla09, 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 ρ-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,

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

and in framework 3,

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

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

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

in place of h.

We develop the ρ-estimation method in frameworks 2 and 3. We explain how to relate the Hellinger-type risk $h(\mathbf{s_0}, \hat{s})$ of a ρ-estimator $\hat{s}$ to the deviations of an empirical process. We then carry out a uniform exponential inequality, and use it to deduce 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 logarithm factors. Moreover, they are slightly faster that the ones obtained in [BB16] in density estimation under same assumptions.

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

1.5. **Estimator selection.** The practical computation of ρ-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 ℓ pieces. Although maximum likelihood estimators do not exist on $\mathcal{P}_{\ell,r}$, ρ-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 large cardinal of this family, dynamic programming makes it possible the practical implementation of the procedure in favourable 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 ρ-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 ℓ 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 ρ-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. We then explain the estimation procedure and relate it to the maximum likelihood one. In Section 3, we present the probabilistic tool that enables us to control the risk of ρ-estimators. We then present the required assumptions on the models as well as our main result on the theoretical performances of ρ-estimators. In Section 4, we deal with estimator selection to define a piecewise polynomial estimator. Section 5 is devoted to numerical simulations. The proofs are deferred to Section 6.

2. The ρ-estimation method

2.1. **Statistical setting and notations.** In this paper, the target $\mathbf{s_0}$ is viewed as an intensity of a random measure ([BB09, AD10, Bar11]). This makes it 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,

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

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

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

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

$$\mathbb{E}[N(A)] = \mathbb{E}\left[\int_A \mathbf{s_0}(t) \, \mathrm{d}M(t) \right] \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

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

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

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 $\boldsymbol{\mathcal{S}} = \mathbb{L}^1_+(\mathbb{R},\mu)$ be the cone of non-negative Lebesgue integrable functions in frameworks 1 and 3, and $\boldsymbol{\mathcal{S}}$ be the cone of measurable non-negative functions which are locally integrable with respect to μ in framework 2. Let now S be a subset of $\boldsymbol{\mathcal{S}}$. Such set will be named 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 $\boldsymbol{\mathcal{S}}$. We begin by defining an approximation $T_E(s,s')$ of $h^2(\mathbf{s_0},s) - h^2(\mathbf{s_0},s')$.

Let ψ 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 $s, s' \in \boldsymbol{\mathcal{S}}$, we set

$$T_E(s,s') = \int_{\mathbb{R}} \psi\left(s'/s\right)\mathbf{s_0}\,\mathrm{d}M - \frac{1}{4}\int_{\mathbb{R}}\left(s'-s\right)\mathrm{d}M.$$

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 $s, s' \in \boldsymbol{\mathcal{S}}$,*

$$\frac{1}{3}h^2(\mathbf{s_0},s) - 3h^2(\mathbf{s_0},s') \leq T_E(s,s') \leq 3h^2(\mathbf{s_0},s) - \frac{1}{3}h^2(\mathbf{s_0},s'). \tag{2}$$

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 of $\mathbf{s_0}$ 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 $s, s' \in \boldsymbol{\mathcal{S}}$ by

$$T(s, s') = \int_{\mathbb{R}} \psi\left(s'/s\right) \, \mathrm{d}N - \frac{1}{4} \int_{\mathbb{R}} \left(s' - s\right) \, \mathrm{d}M.$$

This is the shorten formula of

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

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

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

is called ρ-estimator.

Remark 1. 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, $\mathbb{E}[N'(A)] = \mathbb{E}[\int_A \mathbf{s_0} \mathbb{1}_K \, \mathrm{d}M]$ 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 $\mathbf{s_0}\mathbb{1}_K$.

Although the frameworks are different, the procedures in frameworks 2 and 3 may be related to that of framework 1 (when the data are not censored in framework 2). Consider for instance framework 2 and suppose that $C = +\infty$. Let S be a model and define $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \geq t} \mathbb{1}_{[0,+\infty)}(t)$. The collection $S' = \{sG_n, \, s \in S\}$ is random and does not consist of densities. However, $G_n(t)$ estimates the probability $\mathbb{P}(X \geq t)$ when $t \geq 0$. It may then be natural to use the procedure in framework 1 with S' to estimate the density of X. In that case, an estimator is a ρ-estimator on S' in density estimation if and only if it is of the form $\hat{s}G_n$ where $\hat{s}$ is a ρ-estimator on S in hazard rate estimation. A similar reasoning applies to framework 3 (replace $G_n(t)$ by $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_{t-}^{(i)}=1} \mathbb{1}_{I_{\mathrm{obs}}}(t)$).

Remark 2. Two ingredients are required to define the procedure. First, we need to approximate $h^2(\mathbf{s_0}, s) - h^2(\mathbf{s_0}, s')$ by a quantity $T_E(s, s')$ that satisfies an inequality akin to (2). Second, we need a random variable $T(s, s')$ that can be computed in practice and that is close enough to $T_E(s, s')$ (for more details about the meaning of "close enough", we refer to Section 3.1). When s and s' are supposed to be densities in framework 1, $T_E(s, s')$ becomes $\int_{\mathbb{R}} \psi(s'/s)\mathbf{s_0}\,\mathrm{d}\mu$. We then recover an approximation of $h^2(\mathbf{s_0}, s) - h^2(\mathbf{s_0}, s')$ that appears in [BB17] (with a slight improvement of the numerical ratio a_0/a_1 in their Assumption 1).

2.4. **Connection with maximum likelihood estimation.** The ρ-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.

We define

$$\mathcal{L}(s) = \int_{\mathbb{R}} \log s \,\mathrm{d}N - \int_{\mathbb{R}} s \,\mathrm{d}M \quad \text{for all } s \in \boldsymbol{\mathcal{S}}, \tag{4}$$

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 $\int_{\mathbb{R}} s\,\mathrm{d}M = \int_{\mathbb{R}} s\,\mathrm{d}\mu$ 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

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

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

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

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 ρ-estimator.

Theorem 1. *Suppose that S is a convex subset of $\boldsymbol{\mathcal{S}}$. 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*

$$\mathcal{L}_K(s) = \int_K \log s \,\mathrm{d}N - \int_K s \,\mathrm{d}M \quad \textit{for all } s \in S,$$

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 ρ-estimator. Conversely, assume that there exists a ρ-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 ρ-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

$$S = \left\{ s\mathbb{1}_{(0,+\infty)},\ s \text{ is a non-increasing function of } \boldsymbol{\mathcal{S}} \text{ on } \mathbb{R} \right\}. \tag{5}$$

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 ρ-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 ρ-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

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

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 ρ-estimator on S, is $\hat{s} = (n_0/n)\tilde{s}$. Note that $\int_{\mathbb{R}} \hat{s}\,\mathrm{d}\mu = n_0/n$, 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 ρ-estimator. This means that there are unfortunately ρ-estimators that do not reach the optimal minimax rate of convergence.

It is sometimes convenient to consider models S of the form $S = \left\{ f^2,\ f \in \mathcal{F} \right\}$ 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 ρ- and maximum likelihood estimators remains valid when the convexity assumption is put on $\mathcal{F}$ instead of S.

Theorem 2. *Let $\mathcal{F}$ be a convex set of non-negative functions such that $S = \left\{ f^2,\ f \in \mathcal{F} \right\}$ is included in $\boldsymbol{\mathcal{S}}$. 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 1 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.*

3. Risk bounds of ρ-estimators

3.1. **A uniform exponential inequality.** We recall that the definition of ρ-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 ρ-estimator requires to bound above the error due to the approximation of T_E by T.

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

$$Z(f) = \int_{\mathbb{R}} f \,\mathrm{d}N - \int_{\mathbb{R}} f\, \mathbf{s_0} \,\mathrm{d}M.$$

It can also be written as

$$Z(f) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right) & \text{in framework 1,} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(f(X_i)\mathbb{1}_{D_i=1} - \int_0^\infty f(t)\mathbf{s_0}(t)\mathbb{1}_{X_i \geq t}\,\mathrm{d}t \right) & \text{in framework 2,} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(f(T_{1,0}^{(i)})\mathbb{1}_{I_{\mathrm{obs}}}(T_{1,0}^{(i)}) - \int_{I_{\mathrm{obs}}} f(t)\mathbf{s_0}(t)\mathbb{1}_{X_{t-}^{(i)}=1}\,\mathrm{d}t \right) & \text{in framework 3.} \end{cases}$$

Note that $Z(f)$ 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 3. *Let $\mathcal{F} \subset \boldsymbol{\mathcal{S}}$ 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*

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

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 $\lim_{m\to+\infty} \mathbb{1}_{A_m}(t) = \mathbb{1}_A(t)$ 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 $f \in \mathcal{F}$,

$$\upsilon(f) = \int_{\mathbb{R}} f^2 \mathbf{s_0} \,\mathrm{d}M = \begin{cases} \mathbb{E}[f^2(X)] & \textit{in framework 1,} \\ n^{-1}\sum_{i=1}^n \int_0^\infty f^2(t)\mathbf{s_0}(t)\mathbb{1}_{X_i\geq t}\,\mathrm{d}t & \textit{in framework 2,} \\ n^{-1}\sum_{i=1}^n \int_{I_{obs}} f^2(t)\mathbf{s_0}(t)\mathbb{1}_{X_{t-}^{(i)}=1}\,\mathrm{d}t & \textit{in framework 3.} \end{cases}$$

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 $\varepsilon > 0$,

$$|Z(f)| \leq \varepsilon \upsilon(f) + C_\varepsilon \left\{ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right\}. \tag{6}$$

In the above inequality, C_ε only depends on ε.

A class consisting of unions of at most d intervals is Vapnik-Chervonenkis with dimension at most $2d$. Therefore, the condition in framework 3 is stronger than in the two first frameworks. It remains however general enough to control the risks of ρ-estimators in several models S of interest (see the next section). As a by-product of the proof of the theorem, we get the following proposition which may be of independent interest:

Proposition 4. *Consider framework 1 and an at most countable set $\mathcal{F} \subset \boldsymbol{\mathcal{S}}$ 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

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z(f)|\right] \leq C \inf_{\eta\in(0,1)} \left\{\sigma\sqrt{\log(1/\eta)} + \int_0^\eta \sup_{f\in\mathcal{F}} \sqrt{\mathbb{P}\left(|f(X)| > u\right)}\ \mathrm{d}u\right\} \sqrt{\frac{r_n}{n}} + C\frac{r_n}{n}$$

$$\leq C' \left[\sigma\sqrt{\frac{r_n \log_+(1/\sigma)}{n}} + \frac{r_n}{n}\right]. \tag{7}$$

This proposition gives a bound on $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z(f)|\right]$ 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

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z(f)|\right] \leq C'' \left[\sigma\sqrt{\frac{d\log_+(n/d)\log_+(1/\sigma)}{n}} + \frac{d\log_+(n/d)}{n}\right], \tag{8}$$

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].

Theorem 3 is well tailored for bounding the risk of a ρ-estimator. Indeed, when $f = \psi(s'/s)$, the random variable $v(f)$ can be bounded above as follows.

Lemma 2. *For all $s, s' \in \boldsymbol{\mathcal{S}}$,*

$$\int_{\mathbb{R}} \psi^2(s'/s)\, \mathbf{s_0}\, \mathrm{d}M \leq 4\left(h^2(\mathbf{s_0}, s) + h^2(\mathbf{s_0}, s')\right).$$

Now, under suitable assumptions on the collection $\mathcal{F} = \{\psi(s'/s), s, s' \in S\}$, Inequality (6) roughly says that with high probability (and $\varepsilon = 1/24$):

$$\left|T(s,s') - T_E(s,s')\right| \leq \frac{1}{6}\left(h^2(\mathbf{s_0}, s) + h^2(\mathbf{s_0}, s')\right) + R_S(n) \quad \text{for all } s, s' \in S. \tag{9}$$

The term $R_S(n)$ depends on the probability of the event on which (9) 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 ρ-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 ρ-estimator $\hat{s}$. This would lead to

$$h^2(\mathbf{s_0},\hat{s}) \leq 37h^2(\mathbf{s_0},S) + 12R_S(n) + 6/n.$$

We need, however, to explain the assumptions to put on the model S to make inequality (9) more precise and rigorous.

3.2. **Assumptions on models.** We consider a non-decreasing collection $(\mathcal{I}_d)_{d\geq 1}$ of Borel sets. In framework 3, $\mathcal{I}_d$ is the collection of unions of at most d intervals. In frameworks 1 and 2, $\mathcal{I}_d$ may be more general. More precisely:

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*
$$\lim_{m\to+\infty} \mathbb{1}_{I_m}(t) = \mathbb{1}_I(t)$$
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 ℓ of the form

$$m = \{[x_1,x_2],(x_2,x_3],(x_3,x_4],\dots,(x_\ell,x_{\ell+1}]\}, \tag{10}$$

where $x_1 < x_2 < \cdots < x_{\ell+1}$ 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

$$\mathcal{P}_{\ell,r} = \Bigg\{ \sum_{K\in m} s_K \mathbb{1}_K, \text{ where } m \in \mathcal{M}_\ell \text{ and } s_K \text{ is a polynomial function of degree at most } r \text{ that is non-negative on } K \Bigg\}. \tag{11}$$

Proposition 5. *Let for $d \geq 1$, $\mathcal{I}_d$ be the collection of unions of at most d intervals. Then, Assumption 2 is fulfilled with $S = \mathcal{P}_{\ell,r}$, $\bar{S} \subset \mathcal{P}_{\ell,r}$ and for all $\bar{s} \in \bar{S}$, $d_S(\bar{s}) = (\ell+2)(2r+1)$.*

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

$$\mathcal{F}_k = \boldsymbol{\mathcal{S}} \cap \left\{ \sum_{K \in m} s_K \mathbb{1}_K, \text{where } m \in \mathcal{M}_k \text{ and } s_K \text{ is monotone on the interior of } K \right\}. \tag{12}$$

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

The proposition below is given by [BB16]:

Proposition 6. *Let for $d \geq 1$, $\mathcal{I}_d$ be the collection of unions of at most d intervals. Assumption 2 is fulfilled with:*

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

3.3. **A uniform risk bound.**

Theorem 7. *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 which holds true with probability larger than $1 - e^{-n\xi}$ and on which: for all model S satisfying Assumption 2 and all ρ-estimator $\hat{s}$ on S,*

$$h^2(\mathbf{s_0}, \hat{s}) \leq C \inf_{\bar{s} \in \bar{S}} \left\{ h^2(\mathbf{s_0}, \bar{s}) + \frac{d_S(\bar{s})}{n} \log_+^2 \left(\frac{n}{d_S(\bar{s})} \right) + \xi \log_+(1/\xi) \right\}. \tag{13}$$

In particular,

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s})\right] \leq C' \inf_{\bar{s} \in \bar{S}} \left\{ \mathbb{E}\left[h^2(\mathbf{s_0}, \bar{s})\right] + \frac{d_S(\bar{s})}{n} \log_+^2 \left(\frac{n}{d_S(\bar{s})} \right) \right\}. \tag{14}$$

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

Define $R_S(\mathbf{s_0}, n)$ by

$$R_S(\mathbf{s_0}, n) = \inf_{\bar{s} \in \bar{S}} \left\{ \mathbb{E}\left[h^2(\mathbf{s_0}, \bar{s})\right] + \frac{d_S(\bar{s})}{n} \log_+^2 \left(\frac{n}{d_S(\bar{s})} \right) \right\}. \tag{15}$$

It follows from (14) that $R_S(\mathbf{s_0}, n)$ is – up to a universal multiplicative constant – an upper-bound of the Hellinger quadratic risk $\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s})\right]$ of a ρ-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 ρ-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 logarithm term cannot be avoided in general under our assumptions (see Section 3.5). 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 (13) holds true depends on S through $(I_d)_{d \geq 1}$ only. It remains the same when the model S changes but not the collection $(I_d)_{d \geq 1}$.

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 3.5 and 3.7.

3.4. **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

$$h^2_{unif}(\mathbf{s_0},\hat{s}) = \frac{1}{2}\int_0^\alpha \left(\sqrt{\mathbf{s_0}(t)} - \sqrt{\hat{s}(t)}\right)^2 \mathrm{d}t,$$

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 $h^2_{unif}(\mathbf{s_0},\hat{s})$ depends generally on α when α goes to infinity with n.

In the present paper, we prefer losses that measure the quality of the estimation of $\mathbf{s_0}$ on the whole half-line $(0,+\infty)$. We dealt with h, but we may also consider

$$h^2_E(\mathbf{s_0},\hat{s}) = \frac{1}{2}\int_0^\infty \left(\sqrt{\mathbf{s_0}(t)} - \sqrt{\hat{s}(t)}\right)^2 \mathbb{P}\left(X \geq t\right) \mathrm{d}t.$$

The difference between h_E and h lies in the fact that h_E involves the (unknown) survival function $G(t) = \mathbb{P}\left(X \geq t\right)$ of X whereas h involves its empirical part $G_n(t) = n^{-1}\sum_{i=1}^n \mathbb{1}_{X_i \geq t}$. Note that the quality of the estimation is not measured uniformly on $(0,+\infty)$ but accordingly 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

$$h^2_E(\mathbf{s_0}\mathbb{1}_{[0,\alpha]},\hat{s}\mathbb{1}_{[0,\alpha]}) \leq h^2_{unif}(\mathbf{s_0},\hat{s}) \leq \left(G(\alpha)\right)^{-1} h^2_E(\mathbf{s_0}\mathbb{1}_{[0,\alpha]},\hat{s}\mathbb{1}_{[0,\alpha]}),$$

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

Proposition 8. *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: $\int_0^\infty f_0^2 \,\mathrm{d}\mu < \infty$. Let $\hat{\alpha}$ be a positive random variable such that*

$$150\frac{\log n}{n} \leq G_n(\hat{\alpha}) \leq 151\frac{\log n}{n}. \tag{16}$$

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

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

satisfies

$$\mathbb{E}\left[h^2_E(\mathbf{s_0},\tilde{s})\right] \leq C\left\{\mathbb{E}\left[h^2(\mathbf{s_0},\hat{s})\right] + \frac{\log n}{n} + \frac{\mathbb{E}[X] + \int_0^\infty f_0^2 \,\mathrm{d}\mu}{n^2}\right\}.$$

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 ρ-estimator satisfying the assumptions of Theorem 7. Then, the preceding proposition asserts:

$$\begin{aligned}\mathbb{E}\left[h_E^2(\mathbf{s_0},\tilde{s})\right] &\leq C \inf_{\bar{s}\in\bar{S}}\left\{h_E^2(\mathbf{s_0},\bar{s}) + \frac{d_S(\bar{s})}{n}\log_+^2\left(\frac{n}{d_S(\bar{s})}\right) + \frac{\mathbb{E}[X]+\int_0^\infty f_0^2\,\mathrm{d}\mu}{n^2}\right\}\\ &\leq C\left\{R_S(\mathbf{s_0},n) + \frac{\mathbb{E}[X]+\int_0^\infty f_0^2\,\mathrm{d}\mu}{n^2}\right\}.\end{aligned}$$

Up to a remaining term, and a modification of the multiplicative constant, this bound corresponds to the one we would get for $\mathbb{E}\left[h^2(\mathbf{s_0},\hat{s})\right]$ and that is written in (14).

3.5. **Risks of ρ-estimators on $\mathcal{P}_{\ell,r}$.** We now make explicit the risk bound given by Theorem 7 when $S=\mathcal{P}_{\ell,r}$ is defined by (11). We have,

$$\begin{aligned}R_{\mathcal{P}_{\ell,r}}(\mathbf{s_0},n) &= \mathbb{E}\left[h^2(\mathbf{s_0},\mathcal{P}_{\ell,r})\right] + \frac{(\ell+2)(2r+1)\log_+^2\left(n/(\ell+2)(2r+1)\right)}{n}\\ &\leq \mathbb{E}\left[h^2(\mathbf{s_0},\mathcal{P}_{\ell,r})\right] + \frac{9\ell(r\vee 1)\log_+^2\left(n/\ell(r\vee 1)\right)}{n}. \qquad (17)\end{aligned}$$

When $\mathbf{s_0}\notin\mathcal{P}_{\ell,r}$, but is close to $\mathcal{P}_{\ell,r}$, $\mathbb{E}\left[h^2(\mathbf{s_0},\mathcal{P}_{\ell,r})\right]$ 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

$$R_{\mathcal{P}_{\ell,r}}(\mathbf{s_0},n) \leq \frac{9\ell(r\vee 1)\log_+^2\left(n/\ell(r\vee 1)\right)}{n}.$$

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]).

3.6. **About piecewise polynomial maximum likelihood estimation.** The risk bound given by Theorem 7 also applies to maximum likelihood estimators when the assumptions of Theorem 1 or 2 are met with $K=\mathbb{R}$.

For instance, consider $m\in\mathcal{M}_\ell$ and define $\mathcal{P}_r(m)$ as the convex subset of $\mathcal{P}_{\ell,r}$ whose functions are polynomial on each interval K of m:

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

We may prove:

Lemma 3. *Let $r\geq 0$, $\ell\geq 1$, $m\in\mathcal{M}_\ell$, and for $K\in m$,*

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

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 ρ-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 7 that there exists an event that does not depend on m, that holds true with probability larger than $1-e^{-n\xi}$, and on which:

$$h^2(\mathbf{s_0},\hat{s}_m)\leq C\left\{h^2(\mathbf{s_0},\mathcal{P}_r(m))+9\frac{|m|(r\vee 1)}{n}\log_+^2\left(\frac{n}{|m|(r\vee 1)}\right)+\xi\log_+(1/\xi)\right\}.$$

In particular, the estimator $\hat{s}_m$ 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. We carry out below another elementary example in framework 2.

Proposition 9. *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 below by

$$\mathbb{E}[h^2(\mathbf{s_0},\tilde{s})]>c\log(1/(2\eta)).$$

By contrast, the risk of a ρ-estimator can be bounded 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 ρ-estimator is likely not optimal. However, we observe that it does not increase when η runs from $1/2$ to 0. This is different for the maximum likelihood estimator: its risk may be made arbitrarily large by playing with η.

3.7. **Risks of ρ-estimators on $\mathcal{F}_k$ and $\mathscr{G}_k$.** 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 $S=\mathscr{G}_k$. 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 $\mathbf{s_0}\in\mathcal{F}_k$,*

$$\begin{aligned}R_{\mathcal{F}_k}(\mathbf{s_0},n)&\leq C\inf_{\substack{\ell\geq 1\\ \bar{s}\in\mathcal{P}_{\ell,0}\cap\mathcal{F}_k}}\left\{\mathbb{E}\left[h^2(\mathbf{s_0},\bar{s})\right]+\frac{k+\ell}{n}\log_+^2\left(\frac{n}{k+\ell}\right)\right\}\\ &\leq C'\left\{V(\mathbf{s_0})\left(\frac{\log^2 n}{n}\right)^{2/3}+\frac{k\log^2 n}{n}\right\}.\end{aligned}\tag{18}$$

Moreover, there exist universal constants C'', C''' and a map $W(\cdot)$ on $\mathscr{G}_k$ such that for all $\mathbf{s_0} \in \mathscr{G}_k$,

$$R_{\mathscr{G}_k}(\mathbf{s_0}, n) \leq C'' \inf_{\substack{\ell \geq 1 \\ \bar{s} \in \mathcal{P}_{\ell,1,sq.root} \cap \mathscr{G}_k}} \left\{ \mathbb{E}\left[h^2(\mathbf{s_0}, \bar{s})\right] + \frac{k+\ell}{n} \log_+^2\left(\frac{n}{k+\ell}\right) \right\}$$
$$\leq C''' \left\{ W(\mathbf{s_0}) \left(\frac{\log^2 n}{n}\right)^{4/5} + \frac{k \log^2 n}{n} \right\}.$$

This corollary gives therefore (an upper-bound of) the rates of convergence of ρ-estimators on $\mathcal{F}_k$ and $\mathscr{G}_k$ in the three frameworks, and in particular, in hazard rate and transition intensity estimation. In these inequalities, 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 $K \subset \mathbb{R}$,

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

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 the interior of K. Moreover, $V(\mathbf{s_0})$ can be bounded from above as follows when $k = 2$ and $\mathbf{s_0} \in \mathcal{F}_2$.

Consider framework 1, $k = 2$. If the support of $\mathbf{s_0}$ is of finite length L_{supp}, we deduce

$$V(\mathbf{s_0}) \leq 2 L_{\mathrm{supp}}^{1/3} (\sup_{x \in \mathbb{R}} \mathbf{s_0}(x))^{1/3}.$$

Consider now framework 2, $k = 2$, and suppose that X has finite expectation. Then, for all interval $K \subset [0, +\infty)$, $M_E(K)$ is not larger than $\mathbb{E}(X)$ and hence

$$V(\mathbf{s_0}) \leq 2 \left(\mathbb{E}(X)\right)^{1/3} \left(\sup_{x \geq 0} \mathbf{s_0}(x)\right)^{1/3}.$$

As to framework 3, define $T_1 = \int_{I_{\mathrm{obs}}} \mathbb{1}_{X_{t-}=1}\, \mathrm{d}t$ and suppose that $\mathbb{E}(T_1) < \infty$. Then, $M_E(K) \leq \mathbb{E}(T_1)$ and thus

$$V(\mathbf{s_0}) \leq 2 \left(\mathbb{E}(T_1)\right)^{1/3} \left(\sup_{x \in I_{\mathrm{obs}}} \mathbf{s_0}(x)\right)^{1/3} \quad \text{when } k = 2.$$

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 ρ-estimators. In particular, we do not know how to construct a ρ-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 ℓ.

4.1. **Piecewise polynomial estimator selection.** The quality of a ρ-estimator is, in the present paper, assessed by means of Theorem 7. The event on which (13) is valid depends on the collection $(\mathcal{I}_d)_{d\geq 1}$ but not on the model S. In particular, the risk bound remains true when the model S vary randomly among the class of models for which Assumption 2 is fulfilled.

An application of this result is the following. Consider $\ell \geq 1$, $r \geq 0$ and an at most countable collection $\{\hat{s}_\lambda,\, \lambda \in \Lambda\} \subset \mathcal{P}_{\ell,r}$ of non-negative piecewise polynomial estimators of degree at most r based on at most ℓ pieces. We deduce from Proposition 5 that the (random) model $S = \{\hat{s}_\lambda, \lambda \in \Lambda\}$ fulfils Assumption 2 with $\bar{S} = S$, $d_S(\hat{s}_\lambda) = (\ell+2)(2r+1)$ and $\mathcal{I}_d$ the class of unions of at most d intervals.

Instead of dealing with $\mathcal{P}_{\ell,r}$, we may thus restrict our procedure to a random $S = \{\hat{s}_\lambda, \lambda \in \Lambda\}$ subset. This is interesting from a numerical point of view as the optimization problem becomes easier. Indeed, minimizing the criterion $\gamma(\cdot)$ requires, in general, to know $T(\hat{s}_\lambda, \hat{s}_{\lambda'})$, for every pair $(\hat{s}_\lambda, \hat{s}_{\lambda'}) \in S^2$. This is possible in practice when S is finite and not too large.

The resulting estimator is then of the form $\hat{s} = \hat{s}_{\hat{\lambda}}$ and satisfies

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{\lambda}})\right] \leq C \left\{ \inf_{\lambda\in\Lambda} \mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_\lambda)\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n} \right\}, \tag{19}$$

where C is a universal constant. As we see in (19), we should take $S = \{\hat{s}_\lambda, \lambda \in \Lambda\}$ as large as possible to improve on the theoretical performances of the selected estimator. Ideally, $\{\hat{s}_\lambda, \lambda \in \Lambda\} = \mathcal{P}_{\ell,r}$ to recover the risk bound of a ρ-estimator $\hat{s}$ on $\mathcal{P}_{\ell,r}$:

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s})\right] \leq C \left\{ \mathbb{E}\left[h^2(\mathbf{s_0}, \mathcal{P}_{\ell,r})\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n} \right\}. \tag{20}$$

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

4.2. **Selecting among a special collection of piecewise polynomial estimators.** In this section, we propose to deal with a special but possibly very large collection $\{\hat{s}_\lambda, \lambda \in \Lambda\}$ of piecewise polynomial ρ-estimators. This collection being very rich, we hope to recover a theoretical risk bound akin to (20). 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: $Y_{(1)} < Y_{(2)} < \cdots < Y_{(\hat{n})}$. We define the collection $\widehat{\mathcal{M}}$ that gathers all the partitions m of $[Y_{(1)}, Y_{(\hat{n})}]$ of the form

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

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

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

Note that $\widehat{\mathcal{M}}_\ell \subset \mathcal{M}_\ell$, but $\widehat{\mathcal{M}}_\ell \neq \mathcal{M}_\ell$. We consider a random variable $\hat{\ell}$ with values in $\{1, \ldots, \hat{n}-1\}$. For each $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$, we define the piecewise polynomial ρ-estimator $\hat{s}_m$ on $\mathcal{P}_r(m)$ as explained

in Lemma 3. This construction implies that there exist $\hat{\ell}$ intervals on which $\hat{s}_m$ is a polynomial function. Moreover, the set of points on which $\hat{s}_m$ is not smooth is included in $\{Y_{(1)}, Y_{(2)}, \dots, Y_{(\hat{n})}\}$. We now explain how to select 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

$$m' \vee K = \left\{K' \cap K,\, K' \in m',\, K' \cap K \neq \emptyset\right\}. \tag{21}$$

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

$$\gamma_2(\hat{s}_m) = \sum_{K \in m} \sup_{m' \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ T\left(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K\right) - L|m' \vee K| \frac{(r+1)\log_+^2\left(n/(r+1)\right)}{n} \right\}. \tag{22}$$

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 $\gamma_2(\cdot)$:

$$\gamma_2(\hat{s}_{\hat{m}}) = \min_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \gamma_2(\hat{s}_m). \tag{23}$$

Theorem 10. *There exists a universal constant L_0 such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ minimizing (23) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

$$h^2(\mathbf{s_0}, \hat{s}_{\hat{m}}) \leq C \left\{ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + L \frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}. \tag{24}$$

In particular,

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C' \mathbb{E}\left[\inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + L \frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n} \right]. \tag{25}$$

In the above inequalities, C and C' are universal constants. Moreover, the event on which (24) 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 last inequality also says

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C' \mathbb{E}\left[\inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(\mathbf{s_0}, \hat{s}_m) + L \frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n} \right].$$

This risk bound is very similar to the one (19) obtained when the first criterion $\gamma(\cdot)$ is minimized on $S = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ (we only slightly loose on the variance term). The main interest of this procedure with respect to the first one lies in its numerical complexity. Its form is due to numerical considerations. In the literature, a criterion with a similar form is studied in [Sar14] to select a partition among the collection of dyadic partitions defined by the recursive algorithm of [DY90]. To avoid a digression, we defer the discussion on numerical aspects to Section 4.3.

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 $r = 0$ (and when the parameter L is chosen as in the next section). 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 (24) 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 ρ-estimator on $\mathcal{P}_{\hat{\ell},r}$. As the Y_i are allowed to be random, we may choose them accordingly to the data, and in such a way that the bias term becomes comparable to the one we would obtain for the ρ-estimator, at least when $r = 0$.

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

$$N(A) = N(\{Y_i,\, i \in \hat{I}\} \cap A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}). \tag{26}$$

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 $\hat{I} \supset \{i \in \{1, \dots, n\},\, T_{1,0}^{(i)} \in I_{\mathrm{obs}}\}$, and define for all $i \in \hat{I}$, $Y_i = T_{1,0}^{(i)}$.

We may show:

Lemma 4. *Suppose that condition (26) 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 (10) 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 ρ-estimator $\hat{s}_{m'}$ satisfies*

$$h^2(\mathbf{s_0}, \hat{s}_{m'}) \leq C \left\{ h^2(\mathbf{s_0}, \mathcal{P}_0(m)) + \frac{\ell \log_+^2 (n/\ell)}{n} + \xi \log_+(1/\xi) \right\}. \tag{27}$$

We refer to Lemma 5 in Section 4.4 for another result when $r \geq 1$.

Suppose now that $\mathbf{s_0}$ vanishes outside a (possibly unknown) interval, says $[0, L_{\mathrm{supp}}]$. Let $\{Y_i,\, i \in \hat{I}\}$ be the collection described below (26). Then, (26) holds and $\{Y_i,\, i \in \hat{I}\} \subset [0, L_{\mathrm{supp}}]$. We deduce from (25) and (27) when $r = 0$ and L large enough,

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C' \mathbb{E}\left[\inf_{\substack{m \in \mathcal{M}_{\hat{\ell}} \\ m \text{ partition of } [0, L_{\mathrm{supp}}]}} h^2(\mathbf{s_0}, \mathcal{P}_0(m)) + L \frac{\hat{\ell} \log^2 n}{n} \right],$$

where C' is a universal constant. If now $\mathcal{P}_{\hat{\ell},0,[0,L_{\mathrm{supp}}]}$ denotes the collection of step functions based on partitions m of $[0, L_{\mathrm{supp}}]$ belonging to $\mathcal{M}_{\hat{\ell}}$,

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C' \mathbb{E}\left[h^2\big(\mathbf{s_0}, \mathcal{P}_{\hat{\ell},0,[0,L_{\mathrm{supp}}]}\big) + L \frac{\hat{\ell} \log^2 n}{n} \right].$$

The only difference between $h^2(\mathbf{s_0}, \mathcal{P}_{\hat{\ell},0,[0,L_{\mathrm{supp}}]})$ 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 of $[0, L_{\mathrm{supp}}]$. In particular,

$$h^2(\mathbf{s_0}, \mathcal{P}_{\hat{\ell}+2,0,[0,L_{\mathrm{supp}}]}) \leq h^2(\mathbf{s_0}, \mathcal{P}_{\hat{\ell},0}) \leq h^2(\mathbf{s_0}, \mathcal{P}_{\hat{\ell},0,[0,L_{\mathrm{supp}}]}).$$

We recover, up to slight modifications, the risk bound we would obtain for a ρ-estimator on $\mathcal{P}_{\hat{\ell},0}$.

4.3. **About the practical implementation of the procedure.**

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

$$\mathrm{crit}(m) = \sum_{K \in m} F(K). \tag{28}$$

Let $\mathcal{K}$ be the collection of intervals with endpoints in $\{Y_{(i)}, i \in \{1, \dots, \hat{n}\}\}$. Let κ be the number of computations needed to compute all the $F(K)$ when K varies among $\mathcal{K}$. Then, we may minimize or maximize $\mathrm{crit}(\cdot)$ on $\widehat{\mathcal{M}}_{\hat{\ell}}$ by an algorithm of dynamic programming 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, $\gamma_2(\hat{s}_m)$ writes $\gamma_2(\hat{s}_m) = \sum_{K \in m} F(K)$ with

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

and where $\hat{s}_K$, $\hat{s}_{K'}$ are defined in Lemma 3. 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

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

when K' runs among $\mathcal{K}$ in at most $\mathcal{O}(\hat{n}^2)$ operations. Computing all the $F(K)$ with $\mathcal{O}(\hat{n}^2)$ algorithms of dynamic programming 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 $\mathcal{O}(\hat{n}^2(cn + \hat{n}^2\hat{\ell}))$.

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 $\int_{K \cap K'} \hat{s}_K \, \mathrm{d}M$). We begin by storing in memory all the $\hat{s}_K$, all the atoms of N on K, and all the $\int_{K \cap K'} \hat{s}_K \, \mathrm{d}M$. 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 $T\left(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'}\right)$ in at most $\mathcal{O}(nrN(K \cap K') + 1)$ additional operations. We store all the $T\left(\hat{s}_K \mathbb{1}_{K \cap K'}, \hat{s}_{K'} \mathbb{1}_{K \cap K'}\right)$, and use $\mathcal{O}(\hat{n}^2)$ algorithms of dynamic programming 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 favourable 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 (28) 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 $\mathcal{O}(\hat{\ell})$. This collection will appear in Section 4.4.

4.3.2. *About L.* In our procedure, the parameter L is involved in the construction of the estimator $\hat{s}_{\hat{m}}$ and must be chosen by the statistician. Theorem 10 applies when $L \geq L_0$ where L_0 is 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 avoid the choice of a 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 4.1.

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

$$h^2(\mathbf{s_0}, \hat{s}) \leq C \left[\inf_{L \in \mathbb{L}} \left\{ h^2(\mathbf{s_0}, \hat{s}_{\hat{m}_L}) \right\} + \frac{(r+1)\hat{\ell} \log_+^2 \left(n/(r+1)\right)}{n} + \xi \log_+ \left(1/\xi\right) \right],$$

where C is a universal constant. If $\mathbb{L}$ contains at least one number L larger than L_0, we derive from (24),

$$h^2(\mathbf{s_0}, \hat{s}) \leq C' \left[\inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ h^2(\mathbf{s_0}, \mathcal{P}_r(m)) \right\} + \min \left(1, \inf_{\substack{L \in \mathbb{L}, \\ L \geq L_0}} L \right) \frac{\hat{\ell}(r+1) \log_+^2 \left(n/(r+1)\right)}{n} + \xi \log_+(1/\xi) \right],$$

where C' is a universal constant.

Thereby, the estimator $\hat{s}$ no longer depends on the particular choice of a calibration parameter L but rather on a 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 (24).

4.4. **Adaptive piecewise polynomial estimation.** The two preceding sections explain 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 $m \in \widehat{\mathcal{M}}_{\leq \hat{\ell}_{\max}}$,

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

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

$$\gamma_3(\hat{s}_{\hat{m}}) + 2L\frac{(r+1)|\hat{m}|\log_+^2(n/(r+1))}{n} = \inf_{m\in\widehat{\mathcal{M}}_{\leq\hat{\ell}_{\max}}}\left\{\gamma_3(\hat{s}_m) + 2L(r+1)\frac{|m|\log_+^2(n/(r+1))}{n}\right\}. \tag{29}$$

We prove:

Theorem 11. *There exists a universal constant L_0 such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ satisfying (29) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

$$h^2(\mathbf{s_0}, \hat{s}_{\hat{m}}) \leq C \inf_{m\in\widehat{\mathcal{M}}_{\leq\hat{\ell}_{max}}}\left\{h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + L\frac{(r+1)|m|\log_+^2(n/(r+1))}{n} + \xi\log_+(1/\xi)\right\}. \tag{30}$$

In particular,

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C'\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}_{\leq\hat{\ell}_{max}}}\left\{h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + L\frac{(r+1)|m|\log_+^2(n/(r+1))}{n}\right\}\right]. \tag{31}$$

In the above inequalities, C and C' are universal constants. Moreover, the event on which (30) 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 up. Moreover, (31) implies

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C'\mathbb{E}\left[\inf_{1\leq\hat{\ell}\leq\hat{\ell}_{\max}}\left\{\inf_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\left\{h^2(\mathbf{s_0}, \mathcal{P}_r(m))\right\} + L\frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n}\right\}\right].$$

The right-hand side of this inequality corresponds to the bound (25) achieved by the estimator of Section 4.2 when the choice of $\hat{\ell}$ is the best possible among $\{1, \ldots, \hat{\ell}_{\max}\}$.

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 (31) 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 5. *Suppose that $\{Y_i,\, i \in \hat{I}\}$ is chosen in such a way that N satisfies (26). There exists a universal constant C such that for all $\xi > 0$ and probability larger than $1 - e^{-n\xi}$: for all $\ell \geq 1$, $m \in \mathcal{M}_\ell$, there exists $m' \in \widehat{\mathcal{M}}$ such that*

$$h^2(\mathbf{s_0}, \hat{s}_{m'}) \leq C\left\{h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + \frac{\ell(r+1)\log_+^2\ (n/(\ell(r+1)))}{n} + \xi\log_+(1/\xi)\right\}.$$

Moreover, $|m'|\ \leq 2\ell + 3$.

Suppose now that $\hat{\ell}_{\max} = \hat{n} - 1$, that (26) is satisfied, and that $L \geq \max\{1, L_0\}$. We then deduce from (31) that $\hat{s}_{\hat{m}}$ satisfies

$$\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s}_{\hat{m}})\right] \leq C'\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}}\left\{h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + L\frac{|m|(r+1)\log_+^2(n/(r+1))}{n}\right\}\right],$$

where C' is a universal constant. We deduce from Lemma 5,

$$\begin{aligned}\mathbb{E}\left[h^2(\mathbf{s_0},\hat{s}_{\hat{m}})\right] &\leq C''\mathbb{E}\left[\inf_{\substack{m\in\mathcal{M}_\ell\\ \ell\geq 1}}\left\{h^2(\mathbf{s_0},\mathcal{P}_r(m))+L\frac{\ell(r+1)\log_+^2(n/(r+1))}{n}\right\}\right],\\ &\leq C''\mathbb{E}\left[\inf_{\ell\geq 1}\left\{h^2(\mathbf{s_0},\mathcal{P}_{\ell,r})+L\frac{\ell(r+1)\log_+^2(n/(r+1))}{n}\right\}\right],\\ &\leq C''\inf_{\ell\geq 1}\mathcal{R}(\ell),\end{aligned}$$

where

$$\mathcal{R}(\ell)=\mathbb{E}\left[h^2(\mathbf{s_0},\mathcal{P}_{\ell,r})\right]+L\frac{\ell(r+1)\log_+^2(n/(r+1))}{n}.$$

This term $\mathcal{R}(\ell)$ can be interpreted as an upper-bound of the risk of a ρ-estimator on $\mathcal{P}_{\ell,r}$ (up to constants), barely worse than the one given by Theorem 7 and that is written in (20).

5. Numerical simulations

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 ℓ defined in Section 4.2. For each $m\in\widehat{\mathcal{M}}_\ell$, we consider the ρ- 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 4.2 and 4.3.2.

- 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 ρ-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 4.2 specified in (22) and (23) to get a partition $\hat{m}_L\in\widehat{\mathcal{M}}_\ell$. We then use the procedure of Section 4.1 to pick out an estimator among $\{\hat{s}_{\hat{m}_L},\, L\in\mathbb{L}\}$ as explained in Section 4.3.2. 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*

$$\mathbf{s_0}(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\quad \textit{for all } x\in\mathbb{R}.$$

Example 2. $\mathbf{s_0}$ *is the density of a log Normal distribution*

$$\mathbf{s_0}(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2}\log^2 x} \mathbb{1}_{(0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

Example 3. $\mathbf{s_0}$ *is the density of an exponential distribution*

$$\mathbf{s_0}(x) = e^{-x} \mathbb{1}_{[0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

Example 4. $\mathbf{s_0}$ *is the density of a mixture of uniform distributions*

$$\mathbf{s_0}(x) = \frac{1}{2} \times 3\mathbb{1}_{[0,1/3]}(x) + \frac{1}{8} \times 3\mathbb{1}_{[1/3,2/3]}(x) + \frac{3}{8} \times 3\mathbb{1}_{[2/3,1]}(x) \quad \textit{for all } x \in \mathbb{R}.$$

We simulate N_{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^{th} procedure and the i^{th} sample. We evaluate the quality of the estimators by

$$\widehat{R}(k, \ell) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} h^2\big(\mathbf{s_0}, \hat{s}_{\hat{m}^{(k,\ell,i)}}\big).$$

We estimate the probability that the two estimators coincide by

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

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

	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 $n = 50$, $N_{\text{rep}} = 10000$.

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

	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 $n = 100$, $N_{\text{rep}} = 1000$.

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 ℓ of the partitions is well adapted to the underlying density, that is when ℓ 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.

6. Proofs

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

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

Then,

$$\frac{1}{2}\int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} \left(\sqrt{\mathbf{s_0}} - \sqrt{q}\right)^2 \mathrm{d}M = \frac{1}{2}\int_K \frac{\sqrt{s'} - \sqrt{s}}{\sqrt{q}} \mathbf{s_0}\, \mathrm{d}M + \frac{1}{2}\int_K \left(\sqrt{s'} - \sqrt{s}\right)\sqrt{q}\, \mathrm{d}M - \int_K \left(\sqrt{s'} - \sqrt{s}\right)\sqrt{\mathbf{s_0}}\, \mathrm{d}M.$$

Note that

$$h^2(\mathbf{s_0}, s') - h^2(\mathbf{s_0}, s) = \frac{1}{2}\int_K (s' - s)\, \mathrm{d}M + \int_K \sqrt{\mathbf{s_0}} \left(\sqrt{s} - \sqrt{s'}\right) \mathrm{d}M.$$

Therefore,

$$\begin{aligned}\frac{1}{2}\int_K \frac{\sqrt{s'}-\sqrt{s}}{\sqrt{q}}\left(\sqrt{\mathbf{s_0}}-\sqrt{q}\right)^2 \mathrm{d}M &= \frac{1}{2}\int_K \frac{\sqrt{s'}-\sqrt{s}}{\sqrt{q}}\mathbf{s_0}\,\mathrm{d}M + \frac{1}{2}\int_K \left(\sqrt{s'}-\sqrt{s}\right)\sqrt{q}\,\mathrm{d}M \\ &\quad -\frac{1}{2}\int_K (s'-s)\,\mathrm{d}M + h^2(\mathbf{s_0},s') - h^2(\mathbf{s_0},s) \\ &= T_E(s,s') + h^2(\mathbf{s_0},s') - h^2(\mathbf{s_0},s). \qquad (32)\end{aligned}$$

Now,

$$\begin{aligned}\frac{1}{2}\int_K \frac{\sqrt{s'}-\sqrt{s}}{\sqrt{q}}\left(\sqrt{\mathbf{s_0}}-\sqrt{q}\right)^2 \mathrm{d}M &= \int_K \frac{\sqrt{s'}-\sqrt{s}}{\sqrt{s}+\sqrt{s'}}\left(\sqrt{\mathbf{s_0}}-\frac{\sqrt{s}+\sqrt{s'}}{2}\right)^2 \mathrm{d}M \\ &\le \int_K \left(\sqrt{\mathbf{s_0}}-\frac{\sqrt{s}+\sqrt{s'}}{2}\right)^2 \mathrm{d}M \\ &\le \frac{1}{4}\int_K \left(\left(\sqrt{\mathbf{s_0}}-\sqrt{s}\right)+\left(\sqrt{\mathbf{s_0}}-\sqrt{s'}\right)\right)^2 \mathrm{d}M.\end{aligned}$$

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

$$\begin{aligned}\frac{1}{2}\int_K \frac{\sqrt{s'}-\sqrt{s}}{\sqrt{q}}\left(\sqrt{\mathbf{s_0}}-\sqrt{q}\right)^2 \mathrm{d}M &\le \frac{1+\alpha}{4}\int_K \left(\sqrt{\mathbf{s_0}}-\sqrt{s}\right)^2 \mathrm{d}M + \frac{1+\alpha^{-1}}{4}\int_K \left(\sqrt{\mathbf{s_0}}-\sqrt{s'}\right)^2 \mathrm{d}M \\ &\le \frac{1+\alpha}{2}h^2(\mathbf{s_0},s) + \frac{1+\alpha^{-1}}{2}h^2(\mathbf{s_0},s').\end{aligned}$$

We now put this inequality into (32) to get

$$T_E(s,s') \le \frac{3+\alpha}{2}h^2(\mathbf{s_0},s) - \frac{1-\alpha^{-1}}{2}h^2(\mathbf{s_0},s').$$

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'),

$$T_E(s',s) \le 3h^2(\mathbf{s_0},s') - \frac{1}{3}h^2(\mathbf{s_0},s).$$

Yet, $T_E(s,s') = -T_E(s',s)$ and hence $T_E(s,s') \ge \frac{1}{3}h^2(\mathbf{s_0},s) - 3h^2(\mathbf{s_0},s')$ as wished. □.

6.2. **Proof of Theorem 1.** In each framework, the measure N can be put of 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, $\hat{I} = \{i \in \{1,\dots,n\}, T_{1,0}^{(i)} \in I_{\mathrm{obs}}\}$, $Y_i = T_{1,0}^{(i)}$.

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

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

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

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 1. *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 $s \in S'_1$,

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

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 $u \in [0,1]$ by

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

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

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

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

$$\begin{aligned} &\wp_1(0) = \wp_2(0) = 0 \\ &\wp'_1(0) = \wp'_2(0) = \frac{1}{n}\sum_{j \in \hat{J}} \frac{\zeta(Y_j)}{s(Y_j)} - \int_K \zeta \,\mathrm{d}M \\ &\wp''_1(0) = \wp''_2(0) = -\frac{1}{n}\sum_{j \in \hat{J}} \left(\frac{\zeta(Y_j)}{s(Y_j)}\right)^2. \end{aligned}$$

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 ρ-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 $u \in [0,1]$, $s_u = (1-u)\hat{s} + us \in S$,

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

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

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

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. □

6.3. **Sketch of the proof of Theorem 2.** We define the elements Y_i, $\hat{I}$, $\hat{J}$ as in the proof of Theorem 1. Let for $x \in [0, +\infty]$, $\psi_2(x) = \psi(x^2)$ and for $f, f' \in \mathcal{F}$,

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

The proof is very similar to the one of Theorem 1. 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 1 remain valid under these modifications.

As in the proof of Theorem 1, we may suppose that $\hat{J} \neq \emptyset$.

Claim 2. *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 1, $T_2(f, f') = (\mathcal{L}_{K,2}(f') - \mathcal{L}_{K,2}(f))/4$ 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 $u \in [0,1]$ by

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

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

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

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 ψ_2 is concave, $\wp_2$ is also concave, and $\wp_2$ is non-positive on $[0,1]$. In particular, $\wp_2(1) = T_2(f,f') \leq 0$.

□

6.4. **Proof of Theorem 3.** Let $M_{\mathbf{s_0}}$ be the (possibly random) measure defined by

$$M_{\mathbf{s_0}}(A) = \int_A \mathbf{s_0}\,\mathrm{d}M \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

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

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

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 6. *Let $\mathcal{F}$ be a collection of functions of $\boldsymbol{\mathcal{S}}$ 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 α, β, and an event on which: for all $A \in \mathcal{A}$,

$$|Q(A) - \mathbb{E}[Q(A)]| \leq \sqrt{\alpha}\left(\sqrt{Q(A)} + \sqrt{\mathbb{E}[Q(A)]}\right) + \beta. \tag{33}$$

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

$$|Z_Q(f)| \leq C\left\{\sqrt{\alpha v_Q(f)\log_+(1/v_Q(f))} + \alpha + \beta\right\}, \tag{34}$$

where

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

Moreover, for all $\varepsilon > 0$,

$$|Z_Q(f)| \leq \varepsilon \upsilon_Q(f) + C_\varepsilon \alpha \log_+(1/\alpha) \ + C\beta. \tag{35}$$

The constant C is universal whereas C_ε only depends on ε.

The proof of this result is delayed to Section 6.5 below. As N is an empirical measure, we can prove that (33) 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 7. *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 $\lim_{m\to+\infty} \mathbb{1}_{A_m}(t) = \mathbb{1}_A(t)$ 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 (33) 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 6.7.

Lemma 8. *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 $t \in \mathbb{R}$,*

$$\left|Z_{M_{\mathbf{s_0}}}(f)\right| \leq \varepsilon \upsilon_{M_{\mathbf{s_0}}}(f) + C'_\varepsilon \inf_{\alpha\geq 0} \left\{ [\xi + 1/n] \int_0^\alpha \mathbf{s_0} \, \mathrm{d}\mu + \int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu + \xi + 1/n \right\}. \tag{36}$$

Moreover, in framework 2,

$$\left|Z_{M_{\mathbf{s_0}}}(f)\right| \leq \varepsilon \upsilon_{M_{\mathbf{s_0}}}(f) + C''_\varepsilon(\xi + 1/n) \log_+ [1/(\xi + 1/n)] . \tag{37}$$

In these inequalities, C'_ε and C''_ε only depend on ε.

The computation of the infimum in (36) 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 6.8 and from Lemma 7.

Lemma 9. *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 (33) 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 (6). In framework 1, we use Lemmas 6 and 7 to deduce on $\Omega_{\xi,1}$:

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

Sauer's lemma shows that there exists c'_1 such that $\log_+ |S_{\mathcal{A}}(2n)| \leq c'_1 d \log_+(n/d)$. Inequality (6) then follows from elementary computations.

We now turn to frameworks 2 and 3. As in framework 1, Lemmas 6 and 7 imply on $\Omega_{\xi+(\log 2)/n,1}$:

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

where C_ε only depends on ε. It follows from Lemmas 6, 8 and 9 that on $\Omega_{\xi+(\log 2)/n,2}$:

$$\left|Z_{M_{\mathbf{s_0}}}(f)\right| \leq \varepsilon \upsilon_{M_{\mathbf{s_0}}}(f) + C'_\varepsilon \left[\frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right],$$

where C'_ε only depends on ε. Since $|Z(f)| \leq |Z_N(f)| + |Z_{M_{\mathbf{s_0}}}(f)|$, we get on $\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2}$:

$$|Z(f)| \leq \varepsilon(\upsilon_N(f) + \upsilon_{M_{\mathbf{s_0}}}(f)) + (C_\varepsilon + C'_\varepsilon) \left[\frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right]. \tag{38}$$

Now,

$$\upsilon_N(f) \leq \mathbb{E}\left[\int_{\mathbb{R}} f^2 \, \mathrm{d}N \right] = \mathbb{E}\left[\int_{\mathbb{R}} f^2 \, \mathrm{d}M_{\mathbf{s_0}} \right]. \tag{39}$$

In framework 2, we deduce from Lemma 8 with $\varepsilon = 1/2$ that on $\Omega_{\xi+(\log 2)/n,2}$:

$$\begin{aligned} \left| \int_{\mathbb{R}} f^2 \, \mathrm{d}M_{\mathbf{s_0}} - \mathbb{E}\left[\int_{\mathbb{R}} f^2 \, \mathrm{d}M_{\mathbf{s_0}} \right] \right| &\leq \frac{1}{2}\mathbb{E}\left[\int_{\mathbb{R}} f^4 \, \mathrm{d}M_{\mathbf{s_0}} \right] + C(\xi + 1/n) \log_+ \left[1/(\xi + 1/n)\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\int_{\mathbb{R}} f^4 \, \mathrm{d}M_{\mathbf{s_0}} \right] + C \left[\frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right], \end{aligned} \tag{40}$$

where C is universal. In framework 3,

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

is a union of at most $2d$ intervals. Lemma 9 with Lemma 7 show that (40) remains valid on $\Omega_{\xi+(\log 2)/n,2}$. Moreover, as $|f| \leq 1$,

$$\mathbb{E}\left[\int_{\mathbb{R}} f^2 \, \mathrm{d}M_{\mathbf{s_0}} \right] \leq 2 \int_{\mathbb{R}} f^2 \, \mathrm{d}M_{\mathbf{s_0}} + 2C \left[\frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right].$$

By using (39),

$$\upsilon_N(f) \leq 2\upsilon_{M_{\mathbf{s_0}}}(f) + 2C \left[\frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right].$$

We put this inequality in (38) and use $\upsilon_{M_{\mathbf{s_0}}}(f) \leq \upsilon(f)$ to show (6) on $\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2}$. Moreover,

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

□

6.5. **Proof of Lemma 6.** For convenience, and to make the proof more readable, we introduce a new notation. Given $x, y \in \mathbb{R}$, 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 3. *When (33) holds true,*

$$|Q(A) - \mathbb{E}[Q(A)]| \preccurlyeq \sqrt{\alpha \min\{Q(A), \mathbb{E}[Q(A)]\}} + \alpha + \beta. \tag{41}$$

Proof of Claim 3. For reasons of symmetry, we may suppose that $Q(A) \geq \mathbb{E}[Q(A)]$ to prove (41). We derive from (33),

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

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 (41). □

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 (33) 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 $t \in \mathbb{R}$,

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

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

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

we get

$$\begin{aligned} |Z(f)| - \varepsilon v_Q(f) &= \left| \int_0^1 \left(Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]\right)\, \mathrm{d}u \right| - 2\varepsilon \min\left\{ \int_0^1 u Q(A_{f,u})\, \mathrm{d}u, \int_0^1 u \mathbb{E}[Q(A_{f,u})]\, \mathrm{d}u \right\} \\ &\leq \int_0^1 \left\{ |Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| - 2\varepsilon u \min\left\{Q(A_{f,u}), \mathbb{E}[Q(A_{f,u})]\right\} \right\}\, \mathrm{d}u \\ &\leq \int_0^\eta |Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]|\, \mathrm{d}u \\ &\qquad + \int_\eta^1 \left\{ |Q(A_{f,u}) - \mathbb{E}[Q(A_{f,u})]| - 2\varepsilon u \min\left\{Q(A_{f,u}), \mathbb{E}[Q(A_{f,u})]\right\} \right\}\, \mathrm{d}u. \end{aligned} \tag{42}$$

It follows from (41) and the inequality $2\sqrt{xy} \leq \varepsilon x + \varepsilon^{-1} y$,

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

We deduce,

$$|Z(f)| - \varepsilon v_Q(f) \preccurlyeq \sqrt{\alpha} \int_0^\eta \sqrt{\mathbb{E}[Q(A_{f,u})]}\, \mathrm{d}u + (\alpha/\varepsilon) \log(1/\eta) + \alpha + \beta. \tag{43}$$

We now optimize this result with respect to ε and η:

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

It remains to use $\mathbb{E}[Q(A_{f,u})] \leq 1$ to prove (34).

Elementary computations then show (35). 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 $x = \upsilon_Q(f)$ and $y = \alpha$ to get

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

This proves (35). □

6.6. **Proof of Lemma 7.** 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 $\mathbb{E}\left[\int_{\mathbb{R}} |\mathbb{1}_{A_m} - \mathbb{1}_A|\,\mathrm{d}N\right]$ 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 12 (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 $\nu_n(B) = n^{-1}\sum_{i=1}^n \mathbb{1}_B(Z_i)$, $\nu(B) = \mathbb{E}[\mu_n(B)]$ 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 $x > 0$,

$$\mathbb{P}\left(\sup_{B\in\mathcal{B}} \frac{\nu(B) - \nu_n(B)}{\sqrt{\nu(B)}} \geq x \right) \leq 4S_{\mathcal{B}}(2n)e^{-nx^2/4}$$

$$\mathbb{P}\left(\sup_{B\in\mathcal{B}} \frac{\nu_n(B) - \nu(B)}{\sqrt{\nu_n(B)}} \geq x \right) \leq 4S_{\mathcal{B}}(2n)e^{-nx^2/4}.$$

In particular,

$$\mathbb{P}\left(\sup_{B\in\mathcal{B}}\left|\sqrt{\nu_n(B)}-\sqrt{\nu(B)}\right|\geq x\right)\leq 8S_{\mathcal{B}}(2n)e^{-nx^2/4}. \tag{44}$$

Assume that we are within framework 1. Then, the random measure N is the empirical measure of $X_1,\ldots,X_n$. Now (44) with $\mathcal{B}=\mathcal{A}$,

$$x^2=\frac{4}{n}\left(\log 8+\log_+|S_{\mathcal{A}}(2n)|+n\xi\right)$$

shows that (33) 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 $Z_i=(T_{1,0}^{(i)}\mathbb{1}_{T_{1,0}^{(i)}\in I_{\rm obs}},\mathbb{1}_{T_{1,0}^{(i)}\in I_{\rm obs}})$ in framework 3. We apply (44) with $\mathcal{B}=\{A\times\{1\},\,A\in\mathcal{A}\}$. Moreover,

$$\begin{aligned}|S_{\mathcal{B}}(2n)|&\leq\max_{x_1,\ldots,x_{2n}\in\mathbb{R}}|\{\{x_1,\ldots,x_{2n}\}\cap A,A\in\mathcal{A}\}|\\&\leq|S_{\mathcal{A}}(2n)|\,.\end{aligned}$$

We end the proof as in framework 1. □

6.7. **Proof of Lemma 8.**

Claim 4. *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 all Borel set $A\subset[0,+\infty)$, and $k\geq 1$,*

$$\mathbb{E}\left[\left(\int_A\mathbf{s_0}(t)V(t)\,\mathrm{d}t\right)^k\right]\leq k!\int_A\mathbf{s_0}(t)\mathbb{E}[V(t)]\,\mathrm{d}t.$$

The proof of this lemma is deferred to Sections 6.7.1 and 6.7.2 below. We define $\alpha>0$,

$$R_\alpha=\int_0^\alpha\mathbf{s_0}\left(\sqrt{G_n}-\sqrt{G}\right)^2\,\mathrm{d}\mu,$$

and we prove:

Claim 5. *For all $\xi>0$ and probability larger than $1-2e^{-n\xi}$,*

$$R_\alpha\leq 3\xi\int_0^\alpha\mathbf{s_0}\,\mathrm{d}\mu.$$

Proof. Let $u=n/3$ and $\lambda=u/\int_0^\alpha\mathbf{s_0}\,\mathrm{d}\mu$. It follows from Jensen's inequality that

$$\exp\left(u\frac{R_\alpha}{\int_0^\alpha\mathbf{s_0}\,\mathrm{d}\mu}\right)\leq\frac{1}{\int_0^\alpha\mathbf{s_0}\,\mathrm{d}\mu}\int_0^\alpha\mathbf{s_0}\exp\left(u(\sqrt{G_n}-\sqrt{G})^2\right)\,\mathrm{d}\mu.$$

We deduce,

$$\mathbb{E}\left[\exp\left(\lambda R_\alpha\right)\right]\leq\frac{1}{\int_0^\alpha\mathbf{s_0}\,\mathrm{d}\mu}\int_0^\alpha\mathbf{s_0}(t)\mathbb{E}\left[\exp\left(u(\sqrt{G_n(t)}-\sqrt{G(t)})^2\right)\right]\,\mathrm{d}t. \tag{45}$$

Now,

$$\mathbb{E}\left[\exp\left(u(\sqrt{G_n(t)}-\sqrt{G(t)})^2\right)\right] = 1+\int_1^\infty \mathbb{P}\left[\exp\left(u(\sqrt{G_n(t)}-\sqrt{G(t)})^2\right)\geq x\right]\,\mathrm{d}x$$
$$= 1+\int_1^\infty \mathbb{P}\left[\left|\sqrt{G_n(t)}-\sqrt{G(t)}\right|\geq\sqrt{\frac{\log x}{u}}\right]\,\mathrm{d}x.$$

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,

$$\mathbb{E}\left[\exp\left(u(\sqrt{G_n(t)}-\sqrt{G(t)})^2\right)\right]\leq 1+\int_1^\infty\frac{2}{x^{n/u}}\,\mathrm{d}x$$
$$\leq 1+\int_1^\infty\frac{2}{x^3}\,\mathrm{d}x$$
$$\leq 2.$$

By (45), $\mathbb{E}\left[\exp\left(\lambda R_\alpha\right)\right]\leq 2$ and by Markov's inequality,

$$\mathbb{P}\left[R_\alpha\geq n\xi/\lambda\right]=\mathbb{P}\left[e^{\lambda R_\alpha}\geq e^{n\xi}\right]\leq 2e^{-n\xi}.$$

□

We now prove Lemma 8. We have,

$$|Z_{M_{\mathbf{s_0}}}(f)|\leq\int_0^\infty \mathbf{s_0}\left|f\left(G_n-G\right)\right|\,\mathrm{d}\mu$$
$$\leq\int_{G_n\leq G}\mathbf{s_0}\left|f\left(\sqrt{G_n}-\sqrt{G}\right)\left(\sqrt{G}-\sqrt{G_n}+2\sqrt{G_n}\right)\right|\,\mathrm{d}\mu$$
$$+\int_{G_n>G}\mathbf{s_0}\left|f\left(\sqrt{G_n}-\sqrt{G}\right)\left(\sqrt{G_n}-\sqrt{G}+2\sqrt{G}\right)\right|\,\mathrm{d}\mu$$
$$\leq\int_0^\infty\mathbf{s_0}|f|\left(\sqrt{G_n}-\sqrt{G}\right)^2\,\mathrm{d}\mu+2\int_0^\infty\mathbf{s_0}|f|\left|\sqrt{G_n}-\sqrt{G}\right|\sqrt{\min\{G_n,G\}}\,\mathrm{d}\mu.$$

Define

$$R=R_\infty=\int_0^\infty\mathbf{s_0}\left(\sqrt{G_n}-\sqrt{G}\right)^2\,\mathrm{d}\mu.$$

By using $|f|\leq 1$ and Cauchy-Schwarz inequality,

$$|Z_{M_{\mathbf{s_0}}}(f)|\leq R+2\sqrt{R}\sqrt{\int_0^\infty\mathbf{s_0}f^2\min\{G_n,G\}\,\mathrm{d}\mu}$$
$$\leq R+2\sqrt{R}\sqrt{\upsilon_{M_{\mathbf{s_0}}}(f)}.$$

Therefore, for all $\varepsilon>0$, using that $2\sqrt{xy}\leq\varepsilon^{-1}x+\varepsilon y$,

$$|Z_{M_{\mathbf{s_0}}}(f)|\leq\varepsilon\upsilon_{M_{\mathbf{s_0}}}(f)+(1+\varepsilon^{-1})R. \tag{46}$$

It remains to bound R from above. We consider $\alpha \geq 0$ that minimizes

$$[\xi + 1/n] \int_0^\alpha \mathbf{s_0} \, \mathrm{d}\mu + \int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu + \xi + 1/n.$$

Now,

$$R \leq R_\alpha + \int_\alpha^\infty \mathbf{s_0} \left(G_n + G\right) \, \mathrm{d}\mu. \tag{47}$$

With probability larger than $1 - (1/2)e^{-n\xi}$, Claim 5 ensures that

$$R_\alpha \leq 3 \left(\xi + \frac{\log 4}{n} \right) \int_0^\alpha \mathbf{s_0} \, \mathrm{d}\mu. \tag{48}$$

Now,

$$\int_\alpha^\infty \mathbf{s_0} G_n \, \mathrm{d}\mu = \frac{1}{n} \sum_{i=1}^n \int_\alpha^\infty \mathbf{s_0}(t) V_i(t) \, \mathrm{d}t,$$

where $V_i(t) = \mathbb{1}_{X_i \geq t}$ in framework 2 and $V_i(t) = \mathbb{1}_{X_{t-}^{(i)}=1} \mathbb{1}_{I_{\mathrm{obs}}}(t)$ in framework 3. Claim 4 shows that the assumptions of Bernstein's deviation inequality are satisfied (Proposition 2.9 of [Mas07]). Therefore, with probability larger than $1 - (1/2)e^{-n\xi}$,

$$\int_\alpha^\infty \mathbf{s_0} \left[G_n - G\right] \, \mathrm{d}\mu \leq 2 \sqrt{\left(\xi + \frac{\log 2}{n} \right) \int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu} + \xi + \frac{\log 2}{n}. \tag{49}$$

Since $2\sqrt{xy} \ \leq x + y$,

$$\int_\alpha^\infty \mathbf{s_0} G_n \, \mathrm{d}\mu \leq 2 \int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu + 2 \left(\xi + \frac{\log 2}{n} \right). \tag{50}$$

By putting (48) and (50) into (47), we get with probability larger than $1 - e^{-n\xi}$,

$$R \leq 3 \left[\frac{\log 4}{n} + \xi \right] \int_0^\alpha \mathbf{s_0} \, \mathrm{d}\mu + 3 \int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu + 2\xi + \frac{2 \log 2}{n}.$$

It then remains to use the definition of α and (46) to prove (36). Moreover, in framework 2,

$$\int_\alpha^\infty \mathbf{s_0} G \, \mathrm{d}\mu = \int_\alpha^\infty \frac{f_0(t)}{\mathbb{P}\left(T \ \geq t\right)} \mathbb{P}\left(T \geq t\right) \mathbb{P}\left(C \geq t\right) \mathrm{d}t \leq \int_\alpha^\infty f_0 \, \mathrm{d}\mu,$$

and

$$\int_0^\alpha \mathbf{s_0} \, \mathrm{d}\mu = \int_0^\alpha \frac{f_0(t)}{\mathbb{P}\left(T \ \geq t\right)} \, \mathrm{d}t = -\log \left(\int_\alpha^\infty f_0 \, \mathrm{d}\mu \right),$$

which gives (37). □

6.7.1. *Proof of Claim 4 in framework 2.* We define for $k \geq 1$,

$$J_k = \int_{\substack{t_1,\ldots,t_k \in A \\ t_1 < t_2 < \cdots < t_k}} \left(\prod_{j=1}^k \mathbf{s_0}(t_j) \right) \mathbb{P}\left(X \geq t_k\right) \, \mathrm{d}t_1 \, \mathrm{d}t_2 \ldots \, \mathrm{d}t_k.$$

We have,

$$\begin{aligned}\mathbb{E}\left[\left(\int_A \mathbf{s_0}(t)\mathbb{1}_{X\geq t}\,\mathrm{d}t\right)^k\right] &= \mathbb{E}\left[\int_{A^k}\prod_{j=1}^k \mathbf{s_0}(t_j)\mathbb{1}_{X\geq t_j}\,\mathrm{d}t_1\,\mathrm{d}t_2\ldots\,\mathrm{d}t_k\right]\\ &= \int_{A^k}\left(\prod_{j=1}^k \mathbf{s_0}(t_j)\right)\mathbb{P}\left(X\geq \max\{t_1,\ldots,t_k\}\right)\,\mathrm{d}t_1\,\mathrm{d}t_2\ldots\,\mathrm{d}t_k\\ &= k!J_k.\end{aligned}$$

Now,

$$J_k \leq \int_{\substack{t_1,\ldots,t_{k-1}\in A\\ t_1<t_2<\cdots<t_{k-1}}}\left(\prod_{j=1}^{k-1}\mathbf{s_0}(t_j)\right)\left(\int_{t_{k-1}}^\infty \mathbf{s_0}(t_k)\mathbb{P}\left(X\geq t_k\right)\,\mathrm{d}t_k\right)\,\mathrm{d}t_1\,\mathrm{d}t_2\ldots\,\mathrm{d}t_{k-1},$$

and

$$\begin{aligned}\int_{t_{k-1}}^\infty \mathbf{s_0}(t_k)\mathbb{P}\left(X\geq t_k\right)\mathrm{d}t_k &= \int_{t_{k-1}}^\infty f_0(t_k)\mathbb{P}\left(C\geq t_k\right)\mathrm{d}t_k\\ &\leq \left(\int_{t_{k-1}}^\infty f_0(t_k)\,\mathrm{d}t_k\right)\mathbb{P}\left(C\geq t_{k-1}\right)\\ &\leq \mathbb{P}\left(T\geq t_{k-1}\right)\mathbb{P}\left(C\geq t_{k-1}\right)\\ &\leq \mathbb{P}\left(X\geq t_{k-1}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}J_k &\leq \int_{\substack{t_1,\ldots,t_{k-1}\in A\\ t_1<t_2<\cdots<t_{k-1}}}\left(\prod_{j=1}^{k-1}\mathbf{s_0}(t_j)\right)\mathbb{P}\left(X\geq t_{k-1}\right)\,\mathrm{d}t_1\,\mathrm{d}t_2\ldots\,\mathrm{d}t_{k-1}\\ &\leq J_{k-1}.\end{aligned}$$

By induction, $J_k \leq J_1$. □

6.7.2. *Proof of Claim 4 in framework 3.*

Claim 6. *Let $t>0, \mathcal{F}_t=\sigma(X_v,\, v\leq t)$ be the σ-algebra generated by the family of random variables X_v, $v\in[0,t]$. Let B be an event $\mathcal{F}_t$-measurable. Let μ_B be the measure defined for all $A\in\mathcal{B}(\mathbb{R})$ by*

$$\mu_B(A)=\mathbb{P}\left(B \text{ and } T_{1,0}\in A\right).$$

Then, for μ-almost all $u>t$,

$$\frac{\mathrm{d}\mu_B}{du}(u)=\mathbb{P}\left(B \text{ and } X_{u-}=1\right)\mathbf{s_0}(u). \tag{51}$$

Proof. First of all, μ_B is absolutely continuous with respect to the Lebesgue measure μ and admits therefore a Radon-Nikodym derivative. We now aim to show that this derivative is given by (51) 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, $\mathbb{P}\left(Z_h(u)\geq 2\right)=o(h)$ when $h\to 0$. We deduce,

$$\mu_B([u,u+h])=\mathbb{P}\left(B, Z_h(u)=1,\, T_{1,0}\in[u,u+h]\right)+o(h).$$

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

$$\begin{aligned}\mu_B([u,u+h]) &= \mathbb{P}\left(B, Z_h(u) = 1, X_{u-} = 1,\, X_{u+h} = 0\right) + o(h)\\ &= \mathbb{P}\left(B, X_{u-} = 1,\, X_{u+h} = 0\right) + o(h)\\ &= \mathbb{P}\left(B, X_{u-} = 1\right)\mathbb{P}\left(X_{u+h} = 0 \mid B, X_{u-} = 1\right) + o(h).\end{aligned}$$

As B is $\mathcal{F}_t$-measurable and $u > t$,

$$\begin{aligned}\mu_B([u,u+h]) &= \mathbb{P}\left(B, X_{u-} = 1\right)\mathbb{P}\left(X_{u+h} = 0 \mid X_{u-} = 1\right) + o(h)\\ &= \mathbb{P}\left(B, X_{u-} = 1\right)\frac{\mathbb{P}\left(X_{u-} = 1, X_{u+h} = 0\right)}{\mathbb{P}\left(X_{u-} = 1\right)} + o(h). \qquad (52)\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{P}\left(X_{u-} = 1, X_{u+h} = 0\right) &= \mathbb{P}\left(X_{u-} = 1, X_{u+h} = 0,\, Z_h(u) = 1\right) + o(h)\\ &= \mathbb{P}\left(T_{1,0} \in [u, u+h],\, Z_h(u) = 1\right) + o(h)\\ &= \mathbb{P}\left(T_{1,0} \in [u, u+h]\right) + o(h).\end{aligned}$$

Finally, by putting this inequality into (52),

$$\begin{aligned}\mu_B([u,u+h]) &= \mathbb{P}\left(B, X_{u-} = 1\right)\frac{\mathbb{P}\left(T_{1,0} \in [u, u+h]\right)}{\mathbb{P}\left(X_{u-} = 1\right)} + o(h)\\ &= \mathbb{P}\left(B, X_{u-} = 1\right)\frac{hf_0(u)}{\mathbb{P}\left(X_{u-} = 1\right)} + o(h)\\ &= h\mathbb{P}\left(B, X_{u-} = 1\right)\mathbf{s_0}(u) + o(h),\end{aligned}$$

which proves (51). □

We now return to the proof of Claim 4. Without loss of generality, we suppose that $A \subset I_{\text{obs}}$. Define for $k \geq 1$,

$$J_k = \int_{\substack{t_1,\ldots,t_k \in A \\ t_1 < t_2 < \cdots < t_k}} \left(\prod_{j=1}^{k} \mathbf{s_0}(t_j)\right) \mathbb{P}\left(X_{t_1-} = 1, \ldots, X_{t_k-} = 1\right)\, \mathrm{d}t_1\, \mathrm{d}t_2 \ldots\, \mathrm{d}t_k.$$

We have,

$$\begin{aligned}\mathbb{E}\left[\left(\int_A \mathbf{s_0}(t)\mathbb{1}_{X_{t-}=1}\,\mathrm{d}t\right)^k\right] &= \mathbb{E}\left[\int_{A^k} \prod_{j=1}^{k} \mathbf{s_0}(t_j)\mathbb{1}_{X_{t_j-}=1}\,\mathrm{d}t_1\,\mathrm{d}t_2 \ldots\,\mathrm{d}t_k\right]\\ &= \int_{A^k} \left(\prod_{j=1}^{k} \mathbf{s_0}(t_j)\right) \mathbb{P}\left(X_{t_1-} = 1, \ldots, X_{t_k-} = 1\right)\,\mathrm{d}t_1\,\mathrm{d}t_2 \ldots\,\mathrm{d}t_k\\ &= k!J_k.\end{aligned}$$

Yet,

$$J_k \leq \int_{\substack{t_1,\ldots,t_{k-1} \in A \\ t_1 < t_2 < \cdots < t_{k-1}}} \left(\prod_{j=1}^{k-1} \mathbf{s_0}(t_j)\right) \left(\int_{t_{k-1}}^{\infty} \mathbf{s_0}(t_k)\mathbb{P}\left(X_{t_1-} = 1, \ldots, X_{t_k-} = 1\right)\,\mathrm{d}t_k\right) \mathrm{d}t_1\,\mathrm{d}t_2 \ldots\,\mathrm{d}t_{k-1}.$$

Let $B = \left[X_{t_1-} = 1, \dots, X_{t_{k-1}-} = 1\right] \in \mathcal{F}_{t_{k-1}}$. Then,

$$\begin{aligned}\int_{t_{k-1}}^{\infty} \mathbf{s_0}(t_k)\mathbb{P}\left(X_{t_1-} = 1, \dots, X_{t_k-} = 1\right) \mathrm{d}t_k &= \int_{t_{k-1}}^{\infty} \frac{\mathrm{d}\mu_B}{\mathrm{d}t_k}(t_k)\,\mathrm{d}t_k \\ &= \mu_B([t_{k-1}, +\infty)) \\ &= \mathbb{P}\left(X_{t_1-} = 1, \dots, X_{t_{k-1}-} = 1 \text{ and } T_{1,0} \geq t_{k-1}\right) \\ &\leq \mathbb{P}\left(X_{t_1-} = 1, \dots, X_{t_{k-1}-} = 1\right).\end{aligned}$$

Therefore, $J_k \leq J_{k-1}$ and by induction $J_k \leq J_1$. □

6.8. **Proof of Lemma 9.** 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 (33) when $A \in \overline{\mathcal{I}}_{2d}$.

The random measure $M_{\mathbf{s_0}}$ is of the form

$$M_{\mathbf{s_0}}(A) = \frac{1}{n}\sum_{i=1}^{n} \int_A \mathbf{s_0}(t)\mathbb{1}_{X^{(i)}_{t-}=1}\mathbb{1}_{I_{\mathrm{obs}}}(t)\,\mathrm{d}t \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

There exist independent random variables $Z_1, \dots, Z_n$ such that

$$M_{\mathbf{s_0}}(A) = \frac{1}{n}\sum_{i=1}^{n} f_A(Z_i) \quad \text{with } f_A(Z_i) = \int_A \mathbf{s_0}(t)\mathbb{1}_{X^{(i)}_{t-}=1}\mathbb{1}_{I_{\mathrm{obs}}}(t)\,\mathrm{d}t.$$

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 7. *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 $k \geq 1$,*

$$\mathbb{E}\left[\left(f_{A_2}(Z_1) - f_{A_1}(Z_1)\right)^k\right] \leq \frac{k!}{2}\delta^2.$$

Proof. First of all, we only need to prove the claim when δ 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

$$dist(A_1, A_2) = \mathbb{E}\left[M_{\mathbf{s_0}}(A_1 \Delta A_2)\right] \quad \text{where } A_1\Delta A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1).$$

We may write $dist(A_1, A_2)$ as

$$dist(A_1, A_2) = \int_{I_{\mathrm{obs}}} \left|\mathbb{1}_{A_1}(t) - \mathbb{1}_{A_2}(t)\right| f_0(t)\,\mathrm{d}t.$$

We introduce the real valued function F defined for $x \geq 0$ by

$$F(x) = \int_0^x f_0(t)\,\mathrm{d}t.$$

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 ℓ non-negative numbers $(x_1, x_2, \ldots, x_{\ell-1}, x_\ell) \in \{0\} \times \mathbb{Q}^{\ell-2} \times \{+\infty\}$ such that

$$\max_{1 \le i \le \ell-1} \{F(x_{i+1}) - F(x_i)\} \le \delta^2/(8d).$$

We may suppose that $\ell \ge 2d$. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_\ell\}$, and $\overline{\mathcal{I}}_{dis}$ be the collection of unions of at most $2d$ closed intervals whose endpoints belong to $\mathcal{X}$.

When $k \le \ell/2$, 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

$$|\overline{\mathcal{I}}_{dis}| \le \sum_{k=0}^{2d} C_\ell^{2k}.$$

Standard combinatorial arguments (see, for instance, Lemma 6 of [BBM99]) show that $|\overline{\mathcal{I}}_{dis}| \le (\ell e/(2d))^{2d}$. Using now that $\ell \le 8d/\delta^2 + 2$, we derive that

$$\log |\overline{\mathcal{I}}_{dis}| \le cd \log_+(1/\delta^2)$$

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 $f_{A_1} \le f_A \le f_{A_2}$ and $dist(A_1, A_2) \le \delta^2/2$. 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 $a_k \le b_k$. For each $k \in \{1, \ldots, 2d\}$, there exist $a_k^{(1)} \le a_k^{(2)} \le b_k^{(1)} \le b_k^{(2)} \in \mathcal{X}$ such that

$$a_k^{(1)} \le a_k \le a_k^{(2)}, \quad b_k^{(1)} \le b_k \le b_k^{(2)},$$

and

$$F\big(a_k^{(2)}\big) - F\big(a_k^{(1)}\big) \le \delta^2/(8d), \quad F\big(b_k^{(2)}\big) - F\big(b_k^{(1)}\big) \le \delta^2/(8d).$$

Define the closed intervals

$$A_k^{(1)} = \left\{x \in \mathbb{R},\, a_k^{(2)} \le x \le b_k^{(1)}\right\}, \quad A_k^{(2)} = \left\{x \in \mathbb{R},\, a_k^{(1)} \le x \le b_k^{(2)}\right\}.$$

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 $f_{A_1} \le f_A \le f_{A_2}$. 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,

$$\begin{aligned}
dist(A_1, A_2) &\le \sum_{k=1}^{2d} \int_{[a_k^{(1)}, a_k^{(2)}) \cup (b_k^{(1)}, b_k^{(2)}]} f_0(t)\, \mathrm{d}t \\
&\le \sum_{k=1}^{2d} \Big(F\big(a_k^{(2)}\big) - F\big(a_k^{(1)}\big) + F\big(b_k^{(2)}\big) - F\big(b_k^{(1)}\big)\Big) \\
&\le \sum_{k=1}^{2d} \big(\delta^2/(8d) + \delta^2/(8d)\big) \\
&\le \delta^2/2.
\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}\left[(f_{A_2}(Z_1)-f_{A_1}(Z_1))^k\right] &= \mathbb{E}\left[\left(f_{A_2\setminus A_1}(Z_1)\right)^k\right]\\ &\leq k!\mathbb{E}\left[M_{\mathbf{s_0}}(A_2\setminus A_1)\right] \quad \text{thanks to Claim 4}\\ &\leq k! dist(A_1,A_2)\\ &\leq k!\delta^2/2,\end{aligned}$$

which completes the proof with $\mathcal{C}_\delta = \{f_A,\, A\in\overline{\mathcal{I}}_{dis}\}$. □

We will use several times an exponential inequality of [Mas07] to prove Lemma 9. We keep the notation $\preccurlyeq$ introduced at the beginning of Section 6.5.

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 $\sigma\in(0,1]$,

$$E=\sqrt{n}\int_0^\sigma \sqrt{H(u)\wedge n}\,\mathrm{d}u + 2(1+\sigma)H(\sigma).$$

Simple arguments allow to bound E from above, see for instance page 190 of [GN15]: the fundamental theorem of calculus shows

$$\begin{aligned}\sigma\sqrt{\log(e/\sigma)} &= \int_0^\sigma\left(\sqrt{\log(e/u)}\ -\frac{1}{2\sqrt{\log(e/u)}}\right)\mathrm{d}u\\ &\geq \int_0^\sigma \sqrt{\log_+(1/u)}\ \mathrm{d}u - \sigma/2.\end{aligned}$$

Consequently,

$$E\preccurlyeq \sigma\sqrt{nd\log_+(1/\sigma^2)} + d\log_+(1/\sigma^2). \tag{53}$$

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 $j\in J$, $x_j = 2\log(j+1)+1+n\xi$, $\overline{\mathcal{A}}_j = \left\{A\in\overline{\mathcal{I}}_{2d},\, 2^{-j-1}\leq\mathbb{E}[M_{\mathbf{s_0}}(A)]\leq 2^{-j}\right\}$. Claims 4 and 7 show that assumptions of Corollary 6.9 of [Mas07] are satisfied with $\mathcal{F}=\left\{f_A,-f_A,\, A\in\overline{\mathcal{A}}_j\right\}$, $\sigma^2=2^{-j+1}$, $b=1$, and $H(\delta)=c_1 d\log_+(1/\delta^2)$. There exists therefore an event Ω_j such that $\mathbb{P}(\Omega_j)\geq 1-e^{-x_j}$ and on which: for all $A\in\overline{\mathcal{A}}_j$,

$$n\left|M_{\mathbf{s_0}}(A)-\mathbb{E}[M_{\mathbf{s_0}}(A)]\right| \preccurlyeq E+\sigma\sqrt{nx_j}+x_j.$$

Hence,

$$|M_{\mathbf{s_0}}(A)-\mathbb{E}[M_{\mathbf{s_0}}(A)]| \preccurlyeq \sigma\sqrt{\frac{d\log_+(1/\sigma^2)+x_j}{n}} + \frac{d\log_+(1/\sigma^2)+x_j}{n}.$$

As $\sigma^2\leq 4\mathbb{E}[M_{\mathbf{s_0}}(A)]$, $\sigma^2\geq d/n$, and $x_j\preccurlyeq\log_+(n/d)+n\xi$,

$$|M_{\mathbf{s_0}}(A)-\mathbb{E}[M_{\mathbf{s_0}}(A)]| \preccurlyeq \sqrt{\mathbb{E}[M_{\mathbf{s_0}}(A)]}\sqrt{\frac{d\log_+(n/d)+n\xi}{n}} + \frac{d\log_+(n/d)+n\xi}{n}.$$

Let now $\overline{\mathcal{A}} = \left\{A \in \overline{\mathcal{I}}_{2d}, \mathbb{E}[M_{\mathbf{s_0}}(A)] \leq d/(2n)\right\}$. We apply Corollary 6.9 of [Mas07] with $\mathcal{F} = \left\{f_A, -f_A,\, A \in \overline{\mathcal{A}}\right\}$, $b = 1$, $\sigma^2 = \min\{d/n, 2\}$. We deduce that there exists an event Ω' such that $\mathbb{P}(\Omega') \geq 1 - (1/2)e^{-n\xi}$ and on which: for all $A \in \overline{\mathcal{A}}$,

$$|M_{\mathbf{s_0}}(A) - \mathbb{E}[M_{\mathbf{s_0}}(A)]| \preccurlyeq \sigma\sqrt{\frac{d\log_+(1/\sigma^2) + n\xi + \log 2}{n}} + \frac{d\log_+(1/\sigma^2) + n\xi + \log 2}{n}.$$

Since $\sigma \leq \sqrt{d/n} \leq \sqrt{(d\log_+(n/d) + n\xi)/n}$,

$$|M_{\mathbf{s_0}}(A) - \mathbb{E}[M_{\mathbf{s_0}}(A)]| \preccurlyeq \frac{d\log_+(n/d) + n\xi}{n}.$$

We deduce from these computations that (33) holds true with $Q = M_{\mathbf{s_0}}$ on the event $\Omega' \bigcap \left(\cap_{j\in J}\Omega_j\right)$ for all $A \in \bigcup_{j\in J}\overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$ with α, β 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

$$\begin{aligned}\mathbb{P}\left[\left(\Omega' \bigcap \left(\bigcap_{j\in J}\Omega_j\right)\right)^c\right] &\leq \mathbb{P}\left[\Omega'^c\right] + \sum_{j\in J}\mathbb{P}\left[\Omega_j^c\right] \\ &\leq \frac{e^{-n\xi}}{2} + \sum_{j=1}^{\infty}\frac{e^{-n\xi}}{j^2 e} \\ &\leq e^{-n\xi}.\end{aligned}$$

□

6.9. **Proof of Proposition 4.** 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 3 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

$$\begin{aligned}|Z(f)| \leq \varepsilon\sigma^2 &+ \int_0^\eta |N(A_u) - \mathbb{E}[N(A_u)]| \ \mathrm{d}u \\ &+ \int_\eta^1 \{|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u\mathbb{E}[N(A_u)]\} \ \mathrm{d}u.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}|Z(f)|\right] \leq \varepsilon\sigma^2 &+ \int_0^\eta \mathbb{E}\left[\sup_{A_u\in\mathcal{A}_u}|N(A_u) - \mathbb{E}[N(A_u)]|\right] \mathrm{d}u \\ &+ \int_\eta^1 \mathbb{E}\left[\sup_{A_u\in\mathcal{A}_u}\{|N(A_u) - \mathbb{E}[N(A_u)]| - 2\varepsilon u\mathbb{E}[N(A_u)]\}\right] \mathrm{d}u.\end{aligned} \tag{54}$$

Let now $\xi > 0$. As (33) 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 3 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 ξ 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 (54),

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} |Z(f)| \right] \leq \varepsilon \sigma^2 &+ \frac{C''}{n\varepsilon} \int_\eta^1 \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{u} \, \mathrm{d}u \\ &+ \frac{C''}{\sqrt{n}} \int_0^\eta \sqrt{\left(\sup_{A_u \in \mathcal{A}_u} \mathbb{E}[N(A_u)] \right) \left(\log_+ |S_{\mathcal{A}_u}(2n)| \right)} \, \mathrm{d}u + \frac{C''}{n} \int_0^1 \log_+ |S_{\mathcal{A}_u}(2n)| \, \mathrm{d}u. \end{aligned}$$

As $\varepsilon > 0$ and $\eta \in (0,1)$ are arbitrary,

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} |Z(f)| \right] \leq \frac{C'''}{\sqrt{n}} \inf_{\eta \in (0,1)} &\left\{ \sigma \sqrt{\int_\eta^1 \frac{\log_+ |S_{\mathcal{A}_u}(2n)|}{u} \, \mathrm{d}u} + \int_0^\eta \sqrt{\left(\sup_{A_u \in \mathcal{A}_u} \mathbb{E}[N(A_u)] \right) \left(\log_+ |S_{\mathcal{A}_u}(2n)| \right)} \, \mathrm{d}u \right\} \\ &+ \frac{C'''}{n} \int_0^1 \log_+ |S_{\mathcal{A}_u}(2n)| \, \mathrm{d}u, \end{aligned}$$

where C''' is a universal constant. It then remains to bound above these integrals. □

6.10. **Proof of Lemma 2.** 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

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

Then,

$$\begin{aligned}
\int_{\mathbb{R}} \psi^2\left(\frac{s'}{s}\right) \mathbf{s_0}\,\mathrm{d}M &= \int_K \psi^2\left(\frac{s'}{s}\right) \mathbf{s_0}\,\mathrm{d}M \\
&= \frac{1}{4}\int_K \left(\sqrt{s'}-\sqrt{s}\right)^2 \frac{\mathbf{s_0}}{q}\,\mathrm{d}M \\
&= \frac{1}{4}\int_K \left(\sqrt{s'}-\sqrt{s}\right)^2 \left(\sqrt{\frac{\mathbf{s_0}}{q}}-1+1\right)^2 \mathrm{d}M \\
&\le \frac{1}{2}\int_K \left(\sqrt{s'}-\sqrt{s}\right)^2 \left(\sqrt{\frac{\mathbf{s_0}}{q}}-1\right)^2 \mathrm{d}M + \frac{1}{2}\int_K \left(\sqrt{s'}-\sqrt{s}\right)^2 \mathrm{d}M \\
&\le \frac{1}{2}\int_K \frac{\left(\sqrt{s'}-\sqrt{s}\right)^2}{q} \left(\sqrt{\mathbf{s_0}}-\sqrt{q}\right)^2 \mathrm{d}M + h^2(s,s') \\
&\le 2\int_K \left(\sqrt{\mathbf{s_0}}-\sqrt{q}\right)^2 \mathrm{d}M + h^2(s,s') \\
&\le \frac{1}{2}\int_K \left((\sqrt{\mathbf{s_0}}-\sqrt{s})+(\sqrt{\mathbf{s_0}}-\sqrt{s'})\right)^2 \mathrm{d}M + h^2(s,s') \\
&\le \int_K \left(\sqrt{\mathbf{s_0}}-\sqrt{s}\right)^2 \mathrm{d}M + \int_K \left(\sqrt{\mathbf{s_0}}-\sqrt{s'}\right)^2 \mathrm{d}M + h^2(s,s') \\
&\le 2h^2(\mathbf{s_0},s) + 2h^2(\mathbf{s_0},s') + h^2(s,s').
\end{aligned}$$

We complete the proof by using $h^2(s,s') \le 2h^2(\mathbf{s_0},s) + 2h^2(\mathbf{s_0},s')$. □

6.11. **Proof of Proposition 5 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 $|m| \le |m_1| + |m_2| \le 2\ell + 4$. 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 $b_1 < b_2 < \cdots < b_k$ and $k \le r$. 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. □

6.12. **Proof of Theorem 7.** Let for $d \geq 1$,

$$\vartheta(d) = \frac{d}{n} \log_+^2 \left(\frac{n}{d} \right). \tag{55}$$

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

$$h^2(\mathbf{s_0}, \hat{s}) \leq C \inf_{\bar{s} \in \bar{S}} \left\{ h^2(\mathbf{s_0}, \bar{s}) + \vartheta(d_S(\bar{s})) + \xi \log_+(1/\xi) \right\}. \tag{56}$$

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 $s, s' \in \boldsymbol{\mathcal{S}}$. 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

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

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 integer $d \geq 3$,

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

We will apply Theorem 3 to the class $\mathcal{F} = \mathscr{G}_d$. We begin with the following elementary claim:

Claim 8. *The functions $f \in \mathscr{G}_d$ satisfy $|f| \leq 1$. Moreover, the collection*

$$\mathcal{A} = \{ \{t \in \mathbb{R},\ f_+(t) > u\},\ f \in \mathscr{G}_d,\ u \in (0,1) \} \cup \{ \{t \in \mathbb{R},\ f_-(t) > u\},\ f \in \mathscr{G}_d,\ u \in (0,1) \} \tag{57}$$

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

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

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

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

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

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

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

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

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

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

Now, for all $r \geq 1$,

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

Therefore,

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

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 $\bar{\mathcal{J}}_d$. □

Claim 9. *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 $t_1, \ldots, t_{2n} \in \mathbb{R}$ and a non-increasing sequence $(A_r)_{r\geq 1}$. Then,

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

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

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

This implies that

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

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.

□

The lemma below is at the core of the proof of Theorem 7.

Lemma 10. *For all $\xi > 0$, there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which: for all $\varepsilon \in (0, 1/12)$, $s, s' \in \boldsymbol{\mathcal{S}}$,*

$$T(s,s') \leq (3+4\varepsilon)h^2(\mathbf{s_0}, s) - \frac{4-3\varepsilon}{12}h^2(\mathbf{s_0}, s') + c_1 \min\left\{\vartheta(d_{s'}(s)), \vartheta(d_s(s'))\right\} + c_2\xi\log_+(1/\xi). \quad (58)$$

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

Proof. Let $d \geq 3$ be an odd integer. Claims 8 and 9 show that the assumptions of Theorem 3 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 $\varepsilon > 0$, $f \in \mathscr{G}_d$ of the form $f = \psi(s'/s)$, with $s, s' \in \boldsymbol{\mathcal{S}}$,

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

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

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

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

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

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

$$|Z(f)| \leq \varepsilon \upsilon(f) + c' \min\left\{\vartheta(d_s(s')), \vartheta(d_{s'}(s))\right\} + c\xi\log_+(1/\xi). \tag{59}$$

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

$$T(s,s') \leq 3h^2(\mathbf{s_0}, s) - \frac{1}{3}h^2(\mathbf{s_0}, s') + Z(\psi(s'/s)). \tag{60}$$

Therefore, we deduce from Lemma 2 and from (59) that on Ω_ξ: for all $s, s' \in \boldsymbol{\mathcal{S}}$,

$$T(s,s') \leq (3+4\varepsilon)h^2(\mathbf{s_0}, s) - \frac{4-3\varepsilon}{12}h^2(\mathbf{s_0}, s') + c' \min\left\{\vartheta(d_s(s')), \vartheta(d_{s'}(s))\right\} + c\xi\log_+(1/\xi),$$

which proves (58) with $c_1 = c'$ and $c_2 = c$. □

We now finish the proof of Theorem 7. Lemma 10 implies that on Ω_ξ: for all $s, s' \in S$,

$$T(s,s') \leq (3+4\varepsilon)h^2(\mathbf{s_0}, s) - \frac{4-3\varepsilon}{12}h^2(\mathbf{s_0}, s') + c_1\vartheta(d_{s'}(s)) + c_2\xi\log_+(1/\xi). \tag{61}$$

Thus, for all $s \in S$,

$$\gamma(s) \leq (3+4\varepsilon)h^2(\mathbf{s_0}, s) - \frac{4-3\varepsilon}{12}h^2(\mathbf{s_0}, S) + c_1 \sup_{s'\in S}\vartheta(d_{s'}(s)) + c_2\xi\log_+(1/\xi). \tag{62}$$

By using $T(s,s') = -T(s',s)$, we deduce from (61) 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 ρ-estimator $\hat{s}$ satisfies on Ω_ξ: for all $s \in S$,

$$\begin{aligned}
\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0}, \hat{s}) - (3+4\varepsilon)h^2(\mathbf{s_0}, s') - c_1\vartheta(d_{\hat{s}}(s)) - c_2\xi\log_+(1/\xi) &\leq T(\hat{s}, s) \\
&\leq \gamma(\hat{s}) \\
&\leq \gamma(s) + 1/n.
\end{aligned} \tag{63}$$

Using now (62) and $1/n \le \sup_{s' \in S} \vartheta(d_{s'}(s))$, we deduce when $\varepsilon \in (0, 1/12)$,

$$h^2(\mathbf{s_0}, \hat{s}) \le \inf_{s \in S} \left\{ c_{1,\varepsilon} h^2(\mathbf{s_0}, s) - h^2(\mathbf{s_0}, S) + c_{2,\varepsilon} \sup_{s' \in S} \vartheta(d_{s'}(s)) + c_{2,\varepsilon} \xi \log_+(1/\xi) \right\}, \tag{64}$$

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

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, (64) becomes

$$h^2(\mathbf{s_0}, \hat{s}) \le \inf_{s \in \bar{S}} \left\{ c_{1,\varepsilon} h^2(\mathbf{s_0}, s) - h^2(\mathbf{s_0}, S) + c_{2,\varepsilon} \vartheta(d_S(s)) + c_{2,\varepsilon} \xi \log_+(1/\xi) \right\},$$

and it remains to choose ε arbitrarily in $(0, 1/12)$ to prove the theorem. □

6.13. **Proof of Lemma 3.** As in the proof of Theorem 1, the measure N can be put of 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,

$$\mathcal{L}_K(s) = -\int_K s(t)\, \mathrm{d}M(t),$$

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 μ. Then, $G_n = 1$ in framework 1, $G_n(t) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \ge t} \mathbb{1}_{[0,+\infty)}(t)$ in framework 2 and

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

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 α small enough to get $G_n(t) \ge 1/n$ 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

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

and is bounded above by

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

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

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

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 $\|s\|_1 \to +\infty$. 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 1 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 ρ-estimator on $\mathcal{P}_r(m)$.

Finally, $\hat{s}_m$ maximizes

$$\mathcal{L}_{\cup_{K\in m}}(s) = \int_{\bigcup_{K\in m}} \log s \,\mathrm{d}N - \int_{\bigcup_{K\in m}} s \,\mathrm{d}M$$

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 $N(\cup_{K\in m}) = N(\mathbb{R})$. □

6.14. **Proof of Proposition 9.** 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 2η. 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

$$\mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n) \geq 1 - (\mathbb{P}(\mathcal{A}_n^c) + \mathbb{P}(\mathcal{B}_n^c)) \geq 0.13. \tag{65}$$

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,

$$h^2(\mathbf{s_0}, \tilde{s}) \geq \frac{1}{4}\int_{2\eta}^1 \left(\sqrt{\tilde{s}(t)} - \sqrt{\mathbf{s_0}(t)}\right)^2 \mathrm{d}t.$$

Now, on $\mathcal{A}_n$, $\tilde{s}(t) = 1/t$ when $t \geq 2\eta$ and hence using the definition of $\mathbf{s_0}$:

$$h^2(\mathbf{s_0}, \tilde{s}) \geq \frac{1}{4}\int_{2\eta}^1 \frac{1}{t}\left(1 - \sqrt{\frac{\eta}{(n-1)t+\eta}}\right)^2 \mathrm{d}t.$$

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

$$\mathbb{E}[h^2(\mathbf{s_0},\tilde{s})] > 0.04\log(1/(2\eta))\mathbb{P}\left(\mathcal{A}_n\cap\mathcal{B}_n\right).$$

It remains to use (65) to conclude.

We now turn to ρ-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 ρ-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 $[\eta,1)$,

$$\mathbb{E}[h^2(\mathbf{s_0},S)] \leq \frac{1}{2}\int_\eta^1 \left(\sqrt{\mathbf{s_0}(t)}-0\right)^2 G(t)\,\mathrm{d}t = \frac{\mathbb{P}\left(T\in[\eta,1]\right)}{2} = \frac{G(\eta)-G(1)}{2} = \frac{1-\eta}{n} \leq \frac{1}{n}.$$

This concludes the proof. □

6.15. **Proof of Proposition 8.** We begin by proving the following lemma.

Lemma 11. *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*

$$(70+16\xi)\,\frac{\log n}{n} \leq G_n(\hat{\alpha}_\xi) \leq (71+16\xi)\,\frac{\log n}{n}. \tag{66}$$

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 $\hat{s}\in\boldsymbol{\mathcal{S}}$,

$$h_E^2(\mathbf{s_0},\hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) \leq (9/4)h^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]},\hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + (80+18\xi)\frac{\log n}{n} \tag{67}$$

$$h^2(\mathbf{s_0},\hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) \leq 4h_E^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]},\hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + (321+73\xi)\frac{\log n}{n}. \tag{68}$$

Proof of Lemma 11. We need the celebrated Vapnick-Chervonenkis inequalities for relative deviations. We apply Theorem 12 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 $\mathbb{P}\left(\Omega_{1,\xi}\right)\geq 1-(1/2)n^{-\xi}$ 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,

$$h_E^2(\mathbf{s_0}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) \leq h_E^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + h_E^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0)$$
$$\leq (9/4)h^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + h_E^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0).$$

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

$$h_E^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0) = \frac{1}{2}\int_{\hat{\alpha}_\xi}^{\infty} \mathbf{s_0} G \,\mathrm{d}\mu = \frac{1}{2}\int_{\hat{\alpha}_\xi}^{\infty} \mathbf{s_0} G_C G_T \,\mathrm{d}\mu = \frac{1}{2}\int_{\hat{\alpha}_\xi}^{\infty} f_0 G_C \,\mathrm{d}\mu.$$

Therefore,

$$h_E^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0) \leq \left(\frac{1}{2}\int_{\hat{\alpha}_\xi}^{\infty} f_0 \,\mathrm{d}\mu\right) \times G_C(\hat{\alpha}_\xi) \leq \frac{1}{2}\, G_T(\hat{\alpha}_\xi) G_C(\hat{\alpha}_\xi) \leq \frac{1}{2} G(\hat{\alpha}_\xi).$$

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 (67) on $\Omega_{1,\xi}$.

As to (68), note that

$$h^2(\mathbf{s_0}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) \leq h^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + h^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0)$$
$$\leq 4h_E^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + h^2(\mathbf{s_0}\mathbb{1}_{[\hat{\alpha}_\xi,+\infty)}, 0). \tag{69}$$

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 (69) leads to

$$h^2(\mathbf{s_0}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) \leq 4h_E^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}, \hat{s}\mathbb{1}_{[0,\hat{\alpha}_\xi]}) + h^2(\mathbf{s_0}\mathbb{1}_{[\alpha,+\infty)}, 0). \tag{70}$$

It follows from Bernstein's deviation inequality, and more precisely from (50), that there exists an event $\Omega_{\xi,2}$ such that $\mathbb{P}\left[\Omega_{\xi,2}\right] \ \geq 1 - (1/2)n^{-\xi}$ and on which

$$\int_\alpha^\infty \mathbf{s_0} G_n \leq 2\int_\alpha^\infty \mathbf{s_0} G \,\mathrm{d}\mu + 2\frac{\xi \log n + \log 2}{n}.$$

Since $\int_\alpha^\infty \mathbf{s_0} G \,\mathrm{d}\mu \leq G(\alpha)$ (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 (70), 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 (68). □

Lemma 12. *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 (66). There exists an event which holds true with probability larger than $1 - n^{-\xi}$ and on which: for all $\kappa_n > 0$, all estimator $\hat{s} \in \boldsymbol{\mathcal{S}}$, and $\tilde{s} = \min\{\hat{s}, \kappa_n\}\mathbb{1}_{[0,\hat{\alpha}_\xi]}$,*

$$h_E^2(\mathbf{s_0}, \tilde{s}) \leq (9/2)h^2(\mathbf{s_0}, \hat{s}) + 9\mathbb{P}\left[f_0(T) > 70\kappa_n \frac{\log n}{n}\right] + (803 + 183\xi)\frac{\log n}{n}. \tag{71}$$

Proof of Lemma 12. We derive from (67) 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,

$$h_E^2(\mathbf{s_0}, \tilde{s}) \leq (9/2)h^2(\mathbf{s_0}, \hat{s}) + (9/4)h^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}, \kappa_n\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}) + (80 + 18\xi)\frac{\log n}{n}. \tag{72}$$

We now derive from (68),

$$\begin{aligned} h^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}, \kappa_n\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}) &\leq 4h_E^2(\mathbf{s_0}\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}, \kappa_n\mathbb{1}_{[0,\hat{\alpha}_\xi]}\mathbb{1}_{\mathbf{s_0}>\kappa_n}) + (321 + 73\xi)\frac{\log n}{n} \\ &\leq 2\int_0^{\hat{\alpha}_\xi} (\mathbf{s_0} + \kappa_n)\,\mathbb{1}_{\mathbf{s_0}>\kappa_n} G\,\mathrm{d}\mu + (321 + 73\xi)\frac{\log n}{n} \\ &\leq 4\int_0^{\hat{\alpha}_\xi} \mathbf{s_0}\mathbb{1}_{\mathbf{s_0}>\kappa_n} G\,\mathrm{d}\mu + (321 + 73\xi)\frac{\log n}{n} \\ &\leq 4\int_0^{\hat{\alpha}_\xi} f_0\mathbb{1}_{\mathbf{s_0}>\kappa_n}\,\mathrm{d}\mu + (321 + 73\xi)\frac{\log n}{n}. \end{aligned}$$

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,

$$\int_0^{\hat{\alpha}_\xi} f_0 \mathbb{1}_{\mathbf{s_0}>\kappa_n} \,\mathrm{d}\mu \leq \mathbb{P}\left[f_0(T) > 70\kappa_n \frac{\log n}{n}\right].$$

□

We are now in position to state:

Proposition 13. *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 (66). Then, for all estimator $\hat{s} \in \boldsymbol{\mathcal{S}}$, and $\tilde{s} = \min\{\hat{s}, \kappa_n\}\mathbb{1}_{[0,\hat{\alpha}_\xi]}$,*

$$\mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\right] \leq (9/2)\mathbb{E}\left[h^2(\mathbf{s_0}, \hat{s})\right] + (803 + 183\xi)\frac{\log n}{n} + 9\mathbb{P}\left[f_0(T) > 70\kappa_n \frac{\log n}{n}\right] + \frac{1 + \kappa_n \mathbb{E}[X]}{2n^\xi}.$$

Proof of Proposition 13. Let $\mathcal{A}_\xi$ be the event given by Lemma 12. Then $\mathbb{P}(\mathcal{A}_\xi) \geq 1 - n^{-\xi}$. Moreover,

$$\mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\right] \leq \mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\mathbb{1}_{\mathcal{A}_\xi}\right] + \mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\mathbb{1}_{\mathcal{A}_\xi^c}\right].$$

We have,

$$h_E^2(\mathbf{s_0}, \tilde{s}) \leq \frac{1}{2}\int_0^\infty (\mathbf{s_0} + \tilde{s})\, G \,\mathrm{d}\mu \leq \frac{1}{2}\left(1 + \int_0^\infty \kappa_n G \,\mathrm{d}\mu\right) \leq \frac{1 + \kappa_n \mathbb{E}[X]}{2},$$

and hence,

$$\mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\right] \leq \mathbb{E}\left[h_E^2(\mathbf{s_0}, \tilde{s})\mathbb{1}_{\mathcal{A}_\xi}\right] + \frac{1 + \kappa_n \mathbb{E}[X]}{2n^\xi}.$$

It remains to use (71) to finish the proof. □

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

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

to get Proposition 8. Note that there exists a random variable $\hat{\alpha}$ satisfying (16) if $150(\log n)/n < 1$ that is $n \geq 1043$. □

6.16. **Proofs of Theorems 10 and 11.** It is convenient for ease of demonstration to encompass the two procedures into a more general selection rule we now describe. Theorems 10 and 11 follow from Theorem 14 below. Their proofs are given in Sections 6.16.2 and 6.16.3.

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 $\boldsymbol{\mathcal{S}}$. 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 $\hat{S}_\lambda \subset \boldsymbol{\mathcal{S}}$. We associate to each $\lambda \in \hat{\Lambda}$, $s \in \hat{S}_\lambda$, two penalty terms $\mathrm{pen}_{1,\lambda}(s)$ and $\mathrm{pen}_2(\lambda)$. We finally define the criterion γ_4 by

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

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

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

The risk of this estimator is bounded above as follows.

Theorem 14. *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 6.12. Let for $\xi > 0$, Ω_ξ be the event given by Lemma 10. 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 $\boldsymbol{\mathcal{S}}$ such that*

$$d_{\hat{s}_\lambda}(s) \leq d(s) + \Delta(\lambda) \quad \text{for all } \lambda \in \hat{\Lambda},\ s \in \hat{S}_\lambda. \tag{73}$$

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

$$\mathrm{pen}_{1,\lambda}(s) \leq \mathrm{pen}_1(s) + \mathrm{pen}_2(\lambda) \quad \text{for all } s\ \in \hat{S},\ \lambda\ \in \hat{\Lambda}. \tag{74}$$

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

$$\begin{aligned} \mathrm{pen}_{1,\lambda}(s) &\geq L_1 \frac{d(s)}{n} \log_+^2\left(\frac{n}{d(s)}\right) \\ \mathrm{pen}_2(\lambda) &\geq L_1 \frac{\Delta(\lambda)}{n} \log_+^2\left(\frac{n}{\Delta(\lambda)}\right), \end{aligned} \tag{75}$$

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

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

then, on Ω_ξ:

$$h^2(\mathbf{s_0}, \hat{s}_{\hat{\lambda}}) \leq c\left(\inf_{\lambda\in\hat{\Lambda}} \left\{h^2(\mathbf{s_0}, \hat{s}_\lambda) + \mathrm{pen}_2(\lambda)\right\} + \inf_{s\in\hat{S}} \left\{h^2(\mathbf{s_0}, s) + \mathrm{pen}_1(s)\right\} + \xi \log_+(1/\xi)\right).$$

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

6.16.1. *Proof of Theorem 14.* Let $\varepsilon \in (0, 1/12)$. Lemma 10 asserts the following on Ω_ξ: for all $s, s' \in \boldsymbol{\mathcal{S}}$

$$T(s,s') \leq (3+4\varepsilon)h^2(\mathbf{s_0}, s) - \frac{4-3\varepsilon}{12} h^2(\mathbf{s_0}, s') + c_1 \min\left\{\vartheta(d_{s'}(s)), \vartheta(d_s(s'))\right\} + c_2 \xi \log_+(1/\xi) \tag{76}$$

where c_1, c_2 only depend on ε and where $\vartheta(\cdot)$ is given by (55).

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

$$T(\hat{s}_\lambda, s) \leq (3+4\varepsilon)h^2(\mathbf{s_0}, \hat{s}_\lambda) - \frac{4-3\varepsilon}{12} h^2(\mathbf{s_0}, s) + c_1 \vartheta\left(d_{\hat{s}_\lambda}(s)\right) + c_2 \xi \log_+(1/\xi).$$

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

$$\begin{aligned} T(\hat{s}_\lambda, s) &\leq (3+4\varepsilon)h^2(\mathbf{s_0}, \hat{s}_\lambda) + 1.48 c_1 \vartheta(d(s) + \Delta(\lambda)) + c_2 \xi \log_+(1/\xi) \\ &\leq (3+4\varepsilon)h^2(\mathbf{s_0}, \hat{s}_\lambda) + 1.48 c_1 \vartheta(d(s)) + 1.48 c_1 \vartheta(\Delta(\lambda)) + c_2 \xi \log_+(1/\xi). \end{aligned}$$

If L_1 is large enough,

$$T(\hat{s}_\lambda, s) \leq (3+4\varepsilon)h^2(\mathbf{s_0}, \hat{s}_\lambda) + \mathrm{pen}_{1,\lambda}(s) + \mathrm{pen}_2(\lambda) + c_2 \xi \log_+(1/\xi),$$

and hence

$$\gamma_4(\hat{s}_\lambda) \leq (3+4\varepsilon)h^2(\mathbf{s_0}, \hat{s}_\lambda) + \mathrm{pen}_2(\lambda) + c_2 \xi \log_+(1/\xi). \tag{77}$$

We now derive from (76) that for all $s \in \hat{S}_{\hat{\lambda}}$,

$$T(s,\hat{s}_{\hat{\lambda}}) \leq (3+4\varepsilon)h^2(\mathbf{s_0},s) - \frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) + c_1\vartheta(d_{\hat{s}_{\hat{\lambda}}}(s)) + c_2\xi\log_+(1/\xi).$$

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

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) &\leq T(\hat{s}_{\hat{\lambda}},s) + (3+4\varepsilon)h^2(\mathbf{s_0},s) + c_1\vartheta(d_{\hat{s}_{\hat{\lambda}}}(s)) + c_2\xi\log_+(1/\xi)\\ &\leq T(\hat{s}_{\hat{\lambda}},s) + (3+4\varepsilon)h^2(\mathbf{s_0},s) + 1.48c_1\vartheta(d(s)+\Delta(\hat{\lambda})) + c_2\xi\log_+(1/\xi)\\ &\leq T(\hat{s}_{\hat{\lambda}},s) + (3+4\varepsilon)h^2(\mathbf{s_0},s) + 1.48c_1\vartheta(d(s)) + 1.48c_1\vartheta(\Delta(\hat{\lambda})) + c_2\xi\log_+(1/\xi).\end{aligned}$$

If L_1 is large enough,

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) &\leq T(\hat{s}_{\hat{\lambda}},s) + (3+4\varepsilon)h^2(\mathbf{s_0},s) + \frac{1}{2}\mathrm{pen}_{1,\hat{\lambda}}(s) + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda})\\ &\quad + c_2\xi\log_+(1/\xi) - 1/n\\ &\leq \left[T(\hat{s}_{\hat{\lambda}},s) - \mathrm{pen}_{1,\hat{\lambda}}(s)\right] + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda}) + \left[(3+4\varepsilon)h^2(\mathbf{s_0},s) + \frac{3}{2}\mathrm{pen}_{1,\hat{\lambda}}(s)\right]\\ &\quad + c_2\xi\log_+(1/\xi) - 1/n.\end{aligned}$$

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

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) &\leq \gamma_4(\hat{s}_{\hat{\lambda}}) + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda}) + \inf_{s\in\hat{S}}\left\{3(1+\varepsilon)h^2(\mathbf{s_0},s) + \frac{3}{2}\mathrm{pen}_{1,\hat{\lambda}}(s)\right\}\\ &\quad + c_2\xi\log_+(1/\xi) - 1/n.\end{aligned}$$

We deduce from (74),

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) &\leq \gamma_4(\hat{s}_{\hat{\lambda}}) + 2\mathrm{pen}_2(\hat{\lambda}) + \inf_{s\in\hat{S}}\left\{3(1+\varepsilon)h^2(\mathbf{s_0},s) + \frac{3}{2}\mathrm{pen}_1(s)\right\}\\ &\quad + c_2\xi\log_+(1/\xi) - 1/n.\end{aligned}$$

By using the definition of $\hat{\lambda}$ and (77), we get for all $\lambda \in \hat{\Lambda}$,

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(\mathbf{s_0},\hat{s}_{\hat{\lambda}}) &\leq \gamma_4(\hat{s}_{\lambda}) + 2\mathrm{pen}_2(\lambda) + \inf_{s\in\hat{S}}\left\{3(1+\varepsilon)h^2(\mathbf{s_0},s) + \frac{3}{2}\mathrm{pen}_1(s)\right\} + 2c_2\xi\log_+(1/\xi)\\ &\leq (3+4\varepsilon)h^2(\mathbf{s_0},\hat{s}_{\lambda}) + 3\mathrm{pen}_2(\lambda) + \inf_{s\in\hat{S}}\left\{3(1+\varepsilon)h^2(\mathbf{s_0},s) + \frac{3}{2}\mathrm{pen}_1(s)\right\}\\ &\quad + 2c_2\xi\log_+(1/\xi).\end{aligned}$$

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

6.16.2. *Proof of Theorem 10.* We will apply the selection rule developed in Section 6.16 to pick out an estimator among $\{\hat{s}_\lambda,\, \lambda \in \widehat{\Lambda}\} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. 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 14 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 (21). 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 5 that Assumption 2 is satisfied with $S = \mathcal{P}_{\hat{\ell} \vee k, r}$ and $d_{\mathcal{P}_{\hat{\ell} \vee k, r}}(s) = ((\hat{\ell} \vee k) + 2)(2r + 1)$. In particular, for all $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ and $s \in \bar{S}$, we may set since $\hat{s}_m \in \mathcal{P}_{\hat{\ell}, r}$,

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

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

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

We define $d(\cdot)$ arbitrarily when $s \notin \bar{S}$. Note that (73) 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

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

The first penalty satisfies the lower bound (75) since

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

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

Claim 10. *For all $m, m' \in \widehat{\mathcal{M}}$, $|m(\hat{s}_{m'})| \leq |m| + |m'|$.*

Proof of Claim 10. We have,

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

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

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

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

The penalty defined by

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

satisfies therefore (74).

Note now that the selection rules described in Sections 6.16 and 4.2 coincide. Theorem 14 controls the risk of the selected estimator: for all $\xi > 0$, there exists an event Ω_ξ of probability larger than $1 - e^{-n\xi}$, and on which:

$$h^2(\mathbf{s_0}, \hat{s}_{\hat{m}}) \leq C\left(\inf_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\left\{h^2(\mathbf{s_0}, \hat{s}_m) + \mathrm{pen}_2(m)\right\} + \inf_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\left\{h^2(\mathbf{s_0}, \hat{s}_m) + \mathrm{pen}_1(\hat{s}_m)\right\} + \xi\log_+(1/\xi)\right),$$

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

$$h^2(\mathbf{s_0}, \hat{s}_{\hat{m}}) \leq C'\left\{\inf_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\left\{h^2(\mathbf{s_0}, \hat{s}_m)\right\} + L\frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n} + \xi\log_+(1/\xi)\right\},$$

where C' is a universal constant. It then remains to use the fact that $\hat{s}_m$ is a ρ-estimator on $\mathcal{P}_r(m)$ to get a bound on $h^2(\mathbf{s_0}, \hat{s}_m)$ on the same event Ω_ξ (the event that appears in Theorem 7 to control the risk of a ρ-estimator is the same that the one that appears in Theorem 14. It is, in each case, defined by Lemma 10). □

6.16.3. *Proof of Theorem 11.* The proof is almost the same than the one of Theorem 10. 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 6.16:

$$\hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\leq\hat{\ell}_{\max}}\}$$

$$\hat{S}_m = \left\{\sum_{K\in m}\hat{s}_{m_K}\mathbb{1}_K,\, m_K \in \widehat{\mathcal{M}}_{\leq\hat{\ell}_{\max}}\right\} \quad \text{for all } m \in \widehat{\mathcal{M}}_{\leq\hat{\ell}_{\max}}$$

$$\mathrm{pen}_1(\hat{s}_m) = \mathrm{pen}_2(m) = L\frac{(r+1)|m|\log_+^2(n/(r+1))}{n} \quad \text{for all } m \in \widehat{\mathcal{M}}_{\leq\hat{\ell}_{\max}}.$$

□

6.17. **Proofs of Lemmas 4 and 5.** We set $\mathcal{M} = \bigcup_{\ell\geq 1}\mathcal{M}_\ell$. The following claim will be useful in the sequel.

Claim 11. *Let $\xi > 0$ and Ω_ξ be the event given by Lemma 10 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 Ω_ξ: 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 ρ-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 Ω_ξ:

$$T(\hat{s}_m, \hat{s}_{m'}) \leq (3+4\varepsilon) h^2(\mathbf{s_0}, \hat{s}_m) - \frac{4-3\varepsilon}{12} h^2(\mathbf{s_0}, \hat{s}_{m'}) + c_1 \vartheta(d_{\hat{s}_{m'}}(\hat{s}_m)) + c_2 \xi \log_+(1/\xi), \tag{78}$$

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

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

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 5 that we may set

$$d_{\hat{s}'_m}(\hat{s}_m) = (|m| + |m'| + 2)(2r+1).$$

We now bound above $\vartheta(d_{\hat{s}'_m}(\hat{s}_m))$ in (78), 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 7 and Proposition 5. □

Proof of Lemma 4. Let $m \in \mathcal{M}'_\ell$ be a collection written as

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

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

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

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 ρ-estimators on $\mathcal{P}_0(m)$ and $\mathcal{P}_0(\bar{m})$ respectively. We order the intervals of $\bar{m}$ as follows. We define ℓ intervals $I_1, \ldots, I_\ell$ such that $\bar{m} = \{I_1, \ldots, I_\ell\}$ 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 $\{1, \ldots, \hat{n}\}$ 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 $j \in \{2, \ldots, \ell\}$, $J_j = K_j \setminus \bigcup_{i=1}^{j-1} K_i$. 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 $\bar{m}' = \{J_j,\, j \in \{1, \ldots, \ell\}\}$ 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 $j \in \{1, \ldots, \ell\}$ 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 $\left\{Y_{(i)},\, Y_{(i)} \in I_j,\, i \in \{1,\dots,\hat{n}\}\right\}$ 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 $\cup_{\substack{j' \in \{1,\dots,\hat{n}\} \\ j' \neq j}} K_{j'}$. Such $Y_{(i)}$ belong therefore to J_j and satisfy $s(Y_{(i)}) = \hat{s}_{\bar{m}}(Y_{(i)})$. Therefore,

$$\begin{aligned}\left|\left\{Y_{(i)},\, s(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}),\, i \in \{1,\dots,\hat{n}\}\right\}\right| &= \sum_{j=1}^{\ell} \left|\left\{Y_{(i)},\, s(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}),\, Y_{(i)} \in I_j,\, i \in \{1,\dots,\hat{n}\}\right\}\right| \\ &\leq 2\ell. \end{aligned} \tag{79}$$

It follows from $s \leq \hat{s}_{\bar{m}}$, (79) and (26) that $T(s, \hat{s}_{\bar{m}}) \leq 2\ell/n$. We now use Claim 11 to get on Ω_ξ

$$h^2(\mathbf{s_0}, s) \leq C\left\{h^2(\mathbf{s_0}, \hat{s}_{\bar{m}}) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}, \tag{80}$$

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 ρ-estimators on $\mathcal{P}_0(m')$ and $\mathcal{P}_0(\bar{m}')$ respectively. There exists a universal constant C' such that on Ω_ξ:

$$h^2(\mathbf{s_0}, \hat{s}_{m'}) \leq C'\left\{h^2(\mathbf{s_0}, \mathcal{P}_0(m')) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}.$$

By using that $s \in \mathcal{P}_0(\bar{m}') \subset \mathcal{P}_0(m')$ and (80),

$$\begin{aligned} h^2(\mathbf{s_0}, \hat{s}_{m'}) &\leq C'\left\{h^2(\mathbf{s_0}, s) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}, \\ &\leq C''\left\{h^2(\mathbf{s_0}, \hat{s}_{\bar{m}}) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}, \end{aligned} \tag{81}$$

where C'' is universal. Note now that $\hat{s}_m \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]} \in \mathcal{P}_0(\bar{m})$ and thus $T(\hat{s}_{\bar{m}}, \hat{s}_m \mathbb{1}_{[Y_{(1)}, Y_{(\hat{n})}]}) \leq 0$ as $\hat{s}_{\bar{m}}$ is a ρ-estimator on the convex model $\mathcal{P}_0(\bar{m})$ (see Theorem 1 and Lemma 3). Now,

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

Therefore, Claim 11 asserts that

$$h^2(\mathbf{s_0}, \hat{s}_{\bar{m}}) \leq C'''\left\{h^2(\mathbf{s_0}, \hat{s}_m) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}, \tag{82}$$

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

$$h^2(\mathbf{s_0}, \hat{s}_m) \leq C''''\left\{h^2(\mathbf{s_0}, \mathcal{P}_0(m)) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi)\right\}. \tag{83}$$

It remains to put inequalities (81), (82) and (83) together to finish the proof. □

Proof of Lemma 5. Note that we may always suppose that

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

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

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

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 ρ-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 $\alpha_j = \min\left\{Y_{(i)},\, Y_{(i)} \in K_j\right\}$, $\beta_j = \max\left\{Y_{(i)},\, Y_{(i)} \in K_j\right\}$. 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 ρ-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 11 and from $|m'| \leq 2k-1 \leq 2\ell - 1$, that on Ω_ξ

$$h^2(\mathbf{s_0}, \hat{s}_{m'}) \leq C\left\{h^2(\mathbf{s_0}, \tilde{s}_{m'}) + \frac{(r+1)\ell}{n} \log_+^2\left(\frac{n}{(r+1)\ell}\right) + \xi \log_+(1/\xi)\right\}, \tag{84}$$

where C is universal.

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

$$T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}\right) \leq 0, \tag{85}$$

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 $J_{2j-1} = \emptyset$, $T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}\right) = 0$, and thus (85) also holds.

We define

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

We deduce from (85) that $T(\tilde{s}_{m'} \mathbb{1}_A, \hat{s}_m \mathbb{1}_A) \leq 0$. Therefore,

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

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 $|\psi| \leq 1$,

$$T(\tilde{s}_{m'}, \hat{s}_m) \leq \int_{B \cap A^c} \psi\left(\hat{s}_m/0\right) \mathrm{d}N \leq N(B \cap A^c). \tag{86}$$

Now,

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

Since $\alpha_j, \beta_j \in \{Y_i, i \in \hat{I}\}$, we deduce from (26) 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 (86), we get $T(\tilde{s}_{m'}, \hat{s}_m) \leq k/n$. Claim 11 with $\eta = k/n \leq \ell/n$ ensures that on Ω_ξ:

$$h^2(\mathbf{s_0}, \tilde{s}_{m'}) \leq C' \left\{ h^2(\mathbf{s_0}, \hat{s}_m) + \frac{(r+1)\ell}{n} \log_+^2 \left(\frac{n}{(r+1)\ell} \right) + \xi \log_+(1/\xi) \right\},$$

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

$$h^2(\mathbf{s_0}, \hat{s}_m) \leq C'' \left\{ h^2(\mathbf{s_0}, \mathcal{P}_r(m)) + \frac{(r+1)\ \ell}{n} \log_+^2 \left(\frac{n}{(r+1)\ell} \right) + \xi \log_+(1/\xi) \right\},$$

where C'' is universal. It then remains to combine the two last inequalities with (84) to finish the proof. □

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. ρ-estimators for shape restricted density estimation. *Stochastic Processes and their Applications*, 126(12):3888–3912, 2016.

[BB17] Yannick Baraud and Lucien Birgé. ρ-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: ρ-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.

[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.

[Cas99] Gwénaëlle Castellan. Modified akaike's criterion for histogram density estimation. *Technical report*, 1999.

[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.

[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.

[Pla09] Sandra Plancade. Non parametric estimation of hazard rate in presence of censoring. *hal preprint*, 2009.

[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.

Univ Lyon, UJM-Saint-Étienne, CNRS UMR 5208, Institut Camille Jordan, 10 rue Trélilerie, CS 82301, F-42023 Saint-Etienne Cedex 2, France

E-mail address: mathieu.sart@univ-st-etienne.fr